ISSN 2713-3192
DOI 10.15622/ia.2025.24.1
http://ia.spcras.ru

ТОМ 24 № 1

ИНФОРМАТИКА И АВТОМАТИЗАЦИЯ

INFORMATICS AND AUTOMATION

СПб ФИЦ РАН

Санкт-Петербург
2025

INFORMATICS AND AUTOMATION

Volume 24 № 1, 2025

Scientific and educational journal primarily specialized in computer science, automation, robotics, applied mathematics, interdisciplinary research
Founded in 2002

Founder and Publisher
St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS)

Editor: A. S. Viktorova **Interpreter:** Ya. N. Berezina
Art editor: N. A. Dormidontova

Editorial office address
SPC RAS, 39 litera A , 14-th line V.O., St. Petersburg, 199178, Russia
e-mail: ia@spcras.ru, web: http://ia.spcras.ru
The journal is indexed in Scopus
The journal is published under the scientific-methodological supervision of Department for Nanotechnologies and Information Technologies of the Russian Academy of Sciences

ИНФОРМАТИКА И АВТОМАТИЗАЦИЯ

Том 24 № 1, 2025

Научный, научно-образовательный журнал с базовой специализацией
в области информатики, автоматизации, робототехники, прикладной математики
и междисциплинарных исследований.
Журнал основан в 2002 году

Учредитель и издатель

Федеральное государственное бюджетное учреждение науки
«Санкт-Петербургский Федеральный исследовательский центр Российской академии наук»
(СПб ФИЦ РАН)

Выпускающий редактор: А. С. Викторова **Переводчик:** Я. Н. Березина
Художественный редактор: Н. А. Дормидонтова

Адрес редакции

14-я линия В.О., д. 39, лит. А, г. Санкт-Петербург, 199178, Россия

e-mail: ia@spcras.ru, сайт: http://ia.spcras.ru

Журнал индексируется в международной базе данных Scopus

Журнал входит в «Перечень ведущих рецензируемых научных журналов и изданий,
в которых должны быть опубликованы основные научные результаты диссертации
на соискание ученой степени доктора и кандидата наук»

Журнал выпускается при научно-методическом руководстве Отделения нанотехнологий
и информационных технологий Российской академии наук

CONTENTS

СОДЕРЖАНИЕ

УДК 681.5 **DOI** 10.15622/ia.24.1.1

А.Л. Фрадков, Н.А. Бабич

ТРЕХПОЗИЦИОННОЕ УПРАВЛЕНИЕ ТРАНСПОРТНЫМ СРЕДСТВОМ НА ОСНОВЕ НЕЙРОИНТЕРФЕЙСА С ПРИМЕНЕНИЕМ МАШИННОГО ОБУЧЕНИЯ

Фрадков А.Л., Бабич Н.А. **Трехпозиционное управление транспортным средством на основе нейроинтерфейса с применением машинного обучения.**

Аннотация. Интерфейс мозг-компьютер представляет собой сложную систему, позволяющую управлять внешними электронными устройствами с помощью активности головного мозга. Эта система включает в себя несколько элементов – устройство считывания сигналов активности головного мозга, аппаратно-программный комплекс, выполняющий обработку и анализ этих сигналов, а также объект управления. Основную сложность представляет разработка методов и алгоритмов, способных правильно распознавать и предсказывать намерения человека, который использует этот интерфейс, чтобы обеспечить решение задач управления. В данной работе описывается математическая постановка задачи управления транспортным средством и предложенная алгоритмическая структура разработанной системы управления. Описываются методы предобработки сигналов ЭЭГ, их анализа, принятия решений о выдаче сигнала управления, описывается структура программной реализации этих методов, а также результаты экспериментальной проверки работоспособности системы. Для классификации сигналов ЭЭГ используются методы машинного обучения. Предложен новый метод классификации в машинном обучении – метод нечетких почти ближайших соседей, являющийся модификацией классического метода k-ближайших соседей и снижающий зависимость решения от выбора параметра k. Алгоритмы обработки сигналов ЭЭГ и управления реализованы на языке программирования Python. В качестве объекта управления рассматривается инвалидное кресло с дистанционным управлением, а в качестве задачи управления – поворот кресла вправо или влево. Для экспериментальной проверки работоспособности разработанной модели и алгоритмов всего было проведено более 15 испытаний с 5 испытуемыми в общей сложности. Разработанный и описанный в данной статье подход и его программная реализация в ходе испытаний продемонстрировали эффективность в задачах управления поворотом инвалидного кресла. Также было уделено особое внимание ресурсоёмкости программной реализации. Методы и алгоритмы были реализованы с учётом требований, возникающих при выполнении вычислений на малопроизводительных устройствах с ограниченным количеством памяти.

Ключевые слова: нейроинтерфейс, управление, машинное обучение, ЭЭГ, активность мозга, KNN.

1. Введение. В последние годы в технике, медицине, биологии все чаще употребляется термин "интерфейс мозг-компьютер (ИМК)" или "нейроинтерфейс". Под нейроинтерфейсами понимаются технологии и устройства, позволяющие передавать информацию из мозга непосредственно на внешние устройства, в качестве которых могут выступать смартфон, компьютер, протез или транспортное средство, оснащенное электронной системой управления и любые другие электронные устройства [1 – 9].

Работоспособный нейроинтерфейс позволяет создавать различные приборы и устройства, управляемые "силой мысли", обучать компьютер диагностике состояния пациента или домашнего животного, облегчать связь с окружающим миром людей с ограниченными возможностями и т.д. Ниже будут рассматриваться только неинвазивные нейроинтерфейсы, не требующие хирургического вмешательства, которые предпочтительны по критериям удобства и безопасности. Обычно они основаны на измерении сигналов электроэнцефалографа (ЭЭГ).

Одной из интереснейших и важнейших задач для решения которых применяются нейроинтерфейсы, является управление оборудованием «силой мысли». Ее решение позволит расширить возможности реабилитации пациентов с нарушениями двигательного аппарата, улучшить качество робототехнических протезов, откроет новые возможности управления объектами, находящимися в труднодоступных и опасных областях и т.д. Современный подход к решению подобных задач состоит в создании программно-аппаратных комплексов, позволяющих измерять и обрабатывать сигналы электроэнцефалограммы (ЭЭГ) в режиме реального времени [1, 2]. Такие комплексы должны позволять производить предобработку сигналов ЭЭГ с целью снижений уровня шума и удаления артефактов, выделять информативные признаки, фиксировать и классифицировать намерения человека совершать то или иное движение. Результаты классификации передаются на исполнительные устройства, подключенные к оборудованию или транспортным средствам, что при правильной работе всего комплекса позволяет совершить задуманное движение.

Ниже описывается постановка задачи управления оборудованием на основе нейроинтерфейса – системы, осуществляющей взаимодействие мозга человека с компьютером.

Характерной отличительной особенностью рассматриваемой задачи является трехпозиционность управления: на каждом шаге регулятор принимает одно из трех возможных решений (влево-стоп-вправо, вперед-стоп-назад и т.п.). Подобный выбор множества управленческих решений позволяет существенно упростить реализацию системы в условиях неопределенности. Построение алгоритмической структуры системы является одним из основных результатов работы. Другим научным результатом является новый алгоритм классификации в машинном обучении, являющийся модификацией классического метода ближайших соседей.

Опишем структуру данной работы. Обзору основных существующих публикаций по управлению транспортными средствами на основе нейроинтерфейса в России и за рубежом посвящен раздел 2. В разделе 3 описываются технические и программные средства предлагаемого нейроинтерфейса и общий алгоритм взаимодействия узлов установки при работе, описывается структура одного из разработанных комплексов, применимых к управлению роботизированными устройствами, в том числе инвалидными робоколясками. Постановка задачи сбора информации с нейроинтерфейса дается в разделе 4. В разделах 5 и 6 описываются алгоритмические средства предобработки сигналов ЭЭГ, алгоритмы формирования признаков и классификации намерений испытуемого совершить движение на основе машинного обучения, алгоритм выдачи управляющего воздействия. Основной научный результат работы – модифицированный метод ближайших соседей – представлен в разделе 7. В разделах 8 и 9 описываются методика и результаты проведения экспериментов по проверке работоспособности подхода в задаче управления поворотом инвалидной коляски.

2. Работы по управлению мехатронными и робототехническими системами на основе неинвазивных нейроинтерфейсов. В литературе описан целый ряд разработок неинвазивных нейроинтерфейсов для управления мехатронными и робототехническими системами. В течение нескольких лет ведутся разработки в СПБ ФИЦ РАН (СПИИРАН) совместно с СПбПУ Петра Великого [4 – 6]. В работе [4] предложена классификация электроэнцефалографических паттернов воображаемых движений пальцами руки. В работе [5] описан новый подход к классификации пространственно-временных паттернов активности мозга на основе нейроморфных нейронных сетей. В [6] рассматриваются вопросы применения интерфейсов мозг-компьютер в ассистивных технологиях. Описываются и сравниваются классификаторы, основанные на методе опорных векторов, искусственных нейронных сетях и римановой геометрии.

Известны результаты по управлению мобильным роботом ScEdBo (school educational robot) и мехатронной рукой, разработанными в STEM-центре ТУСУРа [7]. В [7] приведен также обзор существующих на потребительском рынке нейроинтерфейсов.

В работе [8] предложено устройство, реализованное на платформе аналого-цифрового регистратора типа Arduino Mega 2560. Устройство позволяет распознавать ЭЭГ-сигналы мозговой активности и вырабатывать сигналы для управления

роботизированными механизмами типа бионические протезы, роботизированные инвалидные коляски, экзоскелеты и др. Принципы построения нейроуправляемого автомобиля для мобилизации людей с двигательным дефицитом – нейромобиля описаны в [9].

Множество публикаций по нейроинтерфейсам появилось за рубежом, начиная с начала 2000-х годов. Опубликован целый ряд обзоров таких публикаций, в том числе посвященных нейроинтерфейсам на основе ЭЭГ [10 – 12].

Многие работы посвящены применению нейроинтерфейсов в медицине. Рассматривались возможности применения нейроинтерфейсов к реабилитации пациентов с поражением головного мозга после инсульта [13], при параличе нижних конечностей [14], в психиатрии [15]. Целый ряд исследований, в том числе экспериментальных исследований, посвящено использованию нейроинтерфейсов для управления инвалидными колясками, обзоры [16, 17], где упомянуто более 40 работ, опубликованных, начиная с 2005 года. В ряде работ, например, [18, 19], проводится сравнение решений, основанных на использовании для классификации различных известных алгоритмов машинного обучения. Делается вывод, что для классификации сигналов ЭЭГ наилучшие результаты дают метод k ближайших соседей (kNN) и метод опорных векторов (SVM). Однако публикаций в российских источниках, где были бы описаны экспериментальные результаты по управлению инвалидными колясками на основе нейроинтерфейсов авторам найти не удалось. Исключение составляют разработки, в которых используются дополнительно, измерения движений глаз человека, т.е. использующих, фактически, интерфейс «мозг-глаз-компьютер» [20].

Разработка подходов к созданию программно-аппаратных комплексов на основе неинвазивных нейроинтерфейсов уже несколько лет ведется в ИПМаш РАН [21 – 24]. Совместно с Институтом мозга человека РАН (ИМЧ РАН) разработаны подходы к использованию сигналов ЭЭГ для ранней диагностики шизофрении в психиатрии [22]. Совместно с кафедрой высшей нервной деятельности и психофизиологии СПбГУ разработаны подходы к исследованию самоинициированных движений на основе вызванных потенциалов [23].

Значительный интерес представляет использование для управления с нейроинтерфейсами сигналов так называемой нейрообратной связи (НОС или ЭЭГ-БОС), позволяющей головному мозгу регулировать параметры своей электрической активности

на основе информации об успехе или неудаче текущих решений, поступающей из внешнего мира непосредственно через органы чувств. В контуре НОС используются данные ЭЭГ, отражающие изменение потенциала электрического поля на поверхности головы (скальпе) испытуемого. После предобработки данных ЭЭГ производится их классификация методами машинного обучения, на основе которой вырабатывается управляющее воздействие для реализации соответствующего движения робота или мехатронного устройства. Информация об успехе или неудаче реализации намерений (решений) позволяет испытуемому скорректировать свою мозговую активность и адаптироваться к условиям эксперимента. Альтернативная реализация принципа НОС состоит в том, что результаты мозговой активности предъявляются испытуемому в виде, например, визуального стимула (высоты столбика на экране, яркости экрана) с заданием изменять эти параметры в желательном направлении. В такой парадигме испытуемый, сосредотачиваясь на сигнале НОС, старается запомнить связь между параметром и своим состоянием. Параметры ЭЭГ и локализация электродов, образующие протокол НОС, выбираются в зависимости от задачи. Подход к реализации НОС на основе адаптивной модели активности головного мозга предложен в [24]. Возможности адаптивных моделей нейрональной активности на основе сетей нейронов ФитцХью-Нагумо исследованы в [25]. Проблема формирования сигнала нейрообратной связи является достаточно сложной, т.к. на данный момент не существует чётких правил предъявления стимула, которых надо придерживаться, чтобы помочь испытуемому наиболее эффективно (например, с точки зрения затраченного времени) справиться с задачей.

3. Задачи управления оборудованием на основе нейроинтерфейса. Технические средства используемого нами нейроинтерфейса включают стандартный медицинский электроэнцефалограф, доступный из имеющихся в продаже. Кроме того, используются стандартные средства связи роботизированного оборудования с компьютером (ноутбуком) по каналу Wi-Fi или Bluetooth (например, реализованные в контроллере конструктора ТРИК) и специально разработанный интерфейсный блок для инвалидного кресла. На рисунке 1 схематически показан процесс взаимодействия между компонентами системы управления и другими блоками.

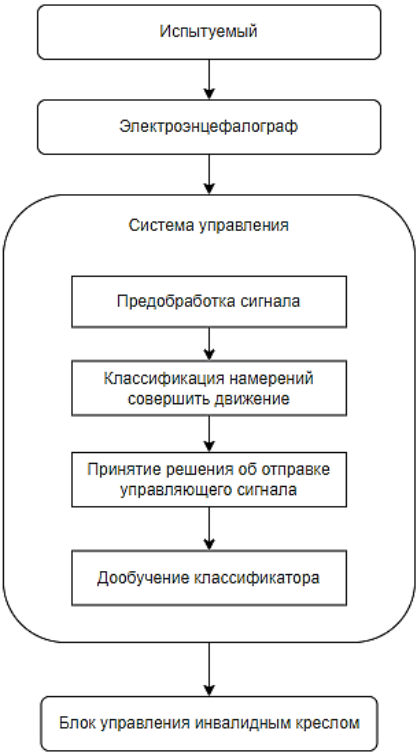


Рис. 1. Схематическое изображение взаимодействия составляющих частей системы

Электроэнцефалограф регистрирует активность мозга испытуемого и отправляет сигналы ЭЭГ на систему управления, там сигналы проходят через предобработку, классификацию, затем принимается решение о подаче управляющего сигнала на блок управления инвалидным креслом. При необходимости происходит дообучение классификатора.

Далее описаны математические постановки основных задач, решаемых системой: задач сбора информации, предобработки измерений, классификации намерений совершить движение и принятия решения об отправке управляющего сигнала. Описаны применяемые алгоритмы и программы, приводится типовой план экспериментов. В качестве примера приводится описание и результаты одной серии экспериментов.

4. Постановка задачи сбора информации с нейроинтерфейса. Опишем математическую постановку задачи управления оборудованием на основе нейроинтерфейса. Пусть $X(t)$, $t>0$ – векторный сигнал ЭЭГ, а $\boldsymbol{x_{i1}, x_{i2},\ldots, x_{iN}}$, где $\boldsymbol{i=1,\ldots,M}$ – последовательные результаты измерений сигнала с i-го канала (отведения) электроэнцефалографа (измеряются на самом деле разности потенциалов между текущим и некоторым референтным электродом). Величины $\boldsymbol{x_{i1}, x_{i2},\ldots, x_{iN}}$ относятся к моментам времени $\boldsymbol{t_{01}, t_{02},\ldots, t_{0N}}$, где $\boldsymbol{t_{0j}=t_0+j\Delta t}$, через $\boldsymbol{t_0}$ обозначено время начала сбора данных, $\boldsymbol{\Delta t}$ – шаг сбора данных (sampling interval) $\boldsymbol{j=1,\ldots,N}$. Обычно $\boldsymbol{t_0}=0$, $\boldsymbol{\Delta t = T/N}$, где $\boldsymbol{T}$ – величина интервала сбора данных (кадра). Результаты измерений сигнала в следующем кадре обозначаются через $\boldsymbol{x_{i,N+1}, x_{i,N+2},\ldots, x_{i,2N}}$ и т.д. Время начала k-го кадра обозначается через $\boldsymbol{t_k}$, так, что $\boldsymbol{t_{kj}=t_k+j\Delta t}$ – моменты измерений в k-м кадре.

Размер кадра T определяется, исходя из того, что, с одной стороны, число отсчетов в кадре должно быть не менее 500-1000, чтобы уменьшить погрешность статистических оценок. С другой стороны, размер кадра T не должен быть большим, чтобы не вносить больших запаздываний в контур управления. В нашей работе было выбрано $T=2c$, $N=500$, $\Delta t=4мс=0.004\ c$. Близкая модель, основанная на анализе вызванных потенциалов P300, описана в [26].

***5.* Предобработка измерений.** Расчет и выдача управляющих воздействий происходит для каждого кадра в моменты $\boldsymbol{t_k=kT}$. Предобработка измерений внутри кадра выполняется по следующему алгоритму.

Для каждого кадра к сигналу $X(t)$ применяется некоторое преобразование $\boldsymbol{G}$, выполняющее фильтрацию сигнала в заданном диапазоне частот: $g(t)=\boldsymbol{G}(t_k,x(t),\ \omega_1,\ \omega_2)$, где $\boldsymbol{g(t)}$ – сигнал после выполнения предобработки сигнала $\boldsymbol{x(t)}$, $\boldsymbol{t_{k-1}<t<t_k}$, ω_1, ω_2 – нижняя и верхняя границы диапазона фильтрации, соответственно. В нашей работе были выбраны значения $\omega_1= 8\ Гц$ и $\omega_2= 30\ Гц$. Такой диапазон фильтрации выбран, исходя из результатов работ на схожую тематику [24, 25], а также вследствие собственных экспериментов – такой диапазон позволял получить лучшие результаты.

К фильтру предъявляется условие гладкости амплитудно-частотной характеристики (АЧХ). В качестве фильтра был выбран фильтр Баттерворта, т.к. его АЧХ максимально гладкая на частотах полосы пропускания и снижается практически до нуля на частотах полосы подавления. В общем случае фильтр 2-го порядка описывается следующим уравнением [27, 28]:

$$Y(t_{kj}, j) = b_0X(t_{kj}) + b_1X(t_{kj-1}) + b_2X(t_{kj-2}) + a_1Y(t_{kj-1}) + a_2Y(t_{kj-2}), \quad (1)$$

где $X(t_{kj})$ – входные данные, $Y(t_{kj})$ – выходные данные, a и b – весовые коэффициенты фильтра, $j=1,2,\ldots N$ – номер отсчета в кадре, $t_{kj}=t_k+j\Delta t$. В промежутках между отсчетами считаем выходной сигнал фильтра постоянным: $g(t)=g(t_{kj})$ при $t_{kj-1}<t<t_{kj}$. Начальные условия фильтра задаются нулевые.

6. Классификация намерений совершить движение. За этапом предобработки следует этап анализа (классификации) измеряемых сигналов и выдачи управляющих сигналов. Он начинается с подачи звукового или светового сигнала, сообщающего испытуемому, в какую сторону он должен повернуть транспортное средство: влево или вправо. В наших экспериментах использовались звуковые сигналы. Повороту влево соответствовал короткий звуковой сигнал (0.5 с), а повороту вправо – длинный (2 с). Производится анализ полученного предобработанного сигнала ЭЭГ $g(t)$ и определение в каждый момент времени значения вектора признаков, используемых далее для обучения системы и распознавания намерения испытуемого.

Следующий этап – обучение системы с целью определения принадлежности текущего значения вектора признаков к одному из двух классов, соответствующих намерениям совершить движение влево или вправо. Этап обучения включает обработку сигналов нескольких кадров (эпохи) и длится до подачи следующего звукового сигнала. Для классификации может использоваться один из известных алгоритмов машинного обучения (SVM, KNN, RF и др.) [29 – 32]. Если эпоха содержит S кадров, то в ней содержится SN измерений.

Часто классифицирующая функция строится как линейная, что соответствует разделению классов гиперплоскостью. Однако во многих практических задачах классы не являются линейно разделимыми и решение задачи усложняется. В нашей работе в качестве классификатора использовался стандартный метод k-ближайших соседей, а также его модификация, так называемый «метод нечетких почти ближайших соседей», который будет описан в следующем разделе.

В конце каждой эпохи, завершающейся подачей нового звукового сигнала, проводится анализ обученной системы. А именно, по значению $H(t_k, g)$, вычисляется матрица ошибок $E(t_k)=||e_{ab}(t_k)||$, $a,b =$ 1,2 предсказания того, к какому классу относится сигнал. Вычисления проводятся следующим образом: элементы матрицы $E(t_k)$, расположенные на главной диагонали, равны числу правильно классифицированных значений сигнала $g(t)$ за предыдущую эпоху, т.е.

при $t_{k-S}<t<t_k$. Иначе говоря, $e_{11}(t_k)$ – число верно распознанных испытуемым поворотов налево, а $e_{22}(t_k)$ – число верно распознанных испытуемым поворотов направо. Соответственно, $e_{12}(t_k)$ и $e_{21}(t_k)$ – количества неверно распознанных поворотов.

Полученная матрица ошибок E используется для принятия решения о том, какой управляющий сигнал следует отправить на объект управления на следующем этапе. Для принятия решения необходимо вычислить доли верно распознанных сигналов (кадров) в общем числе сигналов: $P_L(t_k) = e_{11}(t_k) / S$, и $P_R(t_k) = e_{22}(t_k) / S$, где $0 \le P_L \le 1$ – частота наступления события, состоящего в том, что система верно распознала сигнал поворота налево, $0 \le P_R \le 1$ – частота наступления события, состоящего в том, что система верно распознала сигнал поворота направо, S–число кадров в предыдущей эпохе.

Тогда $\Delta P = P_L - P_R$. Если $\Delta P > P_1$, то сигнал управления y_i принимает значение 1, если $\Delta P < P_2$то y_i принимает значение -1, что соответствует повороту влево и вправо соответственно, P_1иP_2 – некоторые пороговые значения. Если ни одно из этих условий не выполняется (состояние неопределённости), то $y_i = 0$. В нашей работе $P_1 = 0.2$ и $P_2 = -0.2$.

7. Метод ближайших соседей и его модификация. Напомним, что в стандартном методе ближайших соседей (kNN) из выборки S кадров формируется массив (выборка) из SN M-мерных векторов $\boldsymbol{x_i \epsilon R^M}$**,** $\boldsymbol{i=1, \ldots, SN,}$ где $\boldsymbol{N}$ **–** число измерений в кадре (в нашем случае $N=500$)**.** Среди них могут быть векторы из класса А (поворот налево) и из класса В (поворот направо). При появлении нового (тестового) вектора y вычисляются расстояния от y до всех остальных векторов выборки и выбираются k ближайших к y векторов. Для удобства распознавания k берется нечетным. Считается, что вектор y принадлежит классу А, если число векторов $\boldsymbol{x_i}$ из класса А среди ближайших к нему больше, чем число векторов из класса В. В противном случае считается, что вектор y принадлежит классу В. Обычно рекомендуется брать $3<k<11$.

Достоинством метода kNN является его простота, а недостатками – большой объем вычислений, выполняемых нерекуррентно и зависимость результатов от выбора числа k. Метод может быть очень чувствителен к выбору k. Например, при выборе k=9, если среди ближайших соседей есть пять векторов из класса А и четыре вектора из класса В, то тестовый вектор будет отнесен к классу А. Однако, если увеличить число k до k=11 и оба из несколько более далеких дополнительных векторов окажутся принадлежащими классу В, то и тестовый вектор будет отнесен к классу В.

Также достоинством метода kNN является и то, что он не требует разделимости классов A и B гиперплоскостью или какой-то другой простой поверхностью. Важно, что в задачах распознавания сигналов ЭЭГ предположение о разделимости классов, по-видимому, несправедливо и поэтому в экспериментах по анализу ЭЭГ метод kNN зачастую дает более высокий процент верно распознанных тестовых векторов. Эта закономерность подтверждается и в наших экспериментах, что обуславливает выбор метода kNN для дальнейших исследований.

Предлагаемая модификация метода kNN основана на введении так называемых *радиально нечетких векторов*. Радиально нечетким (R-нечетким) вектором будем называть нечеткое множество $X \subset R^n$, имеющее функцию принадлежности $d_X(x)=R(||x-y||/f)$, где $R(a)$ – невозрастающая скалярная функция на $[0,\infty)$, такая, что $R(0)=1$, $R(1)=0$, $R(a)=0$ при $a>1$. Например, можно взять $R(a)=1-a$ при $0 \le a \le 1$, $R(a)=0$ при $a>1$. Вектор $y \in R^n$ называется центром радиально нечеткого вектора X, а число $f>0$ – коэффициентом нечеткости. Очевидно, степень принадлежности радиально нечеткого вектора определяется расстоянием от него до заданного центра. Мы будем рассматривать формально более общий случай, когда ключевую роль играет расстояние не до центра, а до некоторого шара, имеющего тот же центр, а именно, до шара, описанного вокруг множества ближайших соседей.

Модификация метода kNN состоит в том, что все векторы выборки x_i считаются R-нечеткими векторами с функциями принадлежности

$$d_i(x) = \mathrm{R}((\left|\left|\mathrm{x} - \mathrm{x_i}\right|\right| - \mathrm{R}_k)_+ / f)), \qquad (2)$$

где R_k, f – параметры алгоритма (положительные числа), $(.)_+$ – положительная часть числа, равная самому числу, если оно положительно и нулю в противном случае.

Алгоритм метода начинает работать так же, как и алгоритм обычного kNN. Задается целое число k, при поступлении нового тестового вектора y находится $N_k(y)$ – множество k ближайших к y векторов выборки x_i. Затем определяется величина R_k, равная максимуму $||y-x_i||$ по всем $x_i \in N_k(y)$. Очевидно, R_k – радиус шара, описанного вокруг множества ближайших соседей тестового вектора y. Далее вычисляются веса w_i векторов выборки, равные значениям их функций принадлежности $d_i(y)$ (2). Очевидно, веса ближайших соседей равны 1, а веса векторов, расположенных от тестового вектора на

расстоянии больше, чем R_k+f равны нулю. Т.е. достаточно вычислять веса векторов, расположенных от тестового вектора на расстоянии от R_k до R_k+f. Наконец, сравниваются суммы весов векторов из класса A и из класса B. Тестовый вектор относится к тому классу, сумма весов которого больше.

Таким образом, по сравнении с методом kNN в предлагаемом методе кроме соседей учитываются «почти соседи» – векторы обучающей выборки, находящиеся от тестового вектора на расстоянии, не более, чем на величину f превышающем радиус шара, описанного вокруг множества ближайших соседей. При этом вес, с которым учитываются «почти соседи» тем меньше, чем дальше от этого шара они находятся. Если параметр нечеткости f настолько мал, что множество «почти соседей» пусто, то модифицированный метод совпадает с методом kNN. Метод можно назвать методом «нечетких почти ближайших соседей» (Fuzzy Almost Nearest Neighbors – FANN). Описанная модификация легко обобщается на случай распознавания нескольких (более, чем двух) классов.

Следует отметить, что имеется достаточно много публикаций, описывающих «нечеткие» расширения метода kNN [33, 34, 35]. Однако большинство из них, вслед за пионерской работой [33] (имеющей более 3000 цитирований в Google Scholar) основаны на введении нечеткого отношения принадлежности точки к классу, т.е. на нечеткой классификации. Функция принадлежности строится эвристически как функция расстояний векторов выборки до тестового вектора. В предлагаемом алгоритме классификация остается четкой и принимаемое решение совпадает с методом kNN при малой степени вводимой нечеткости. Таким образом можно реализовать плавный переход от метода kNN к новому методу, что позволяет обеспечить преемственность подходов.

Одним из достоинств метода FANN по сравнению с классическим kNN является то, что FANN менее требователен к выбору значения числа k и практически во всех случаях обеспечивает максимальную точность классификации. Сравнение точности классификации для методов kNN и FANN для разных значений k представлено на рисунке 2.

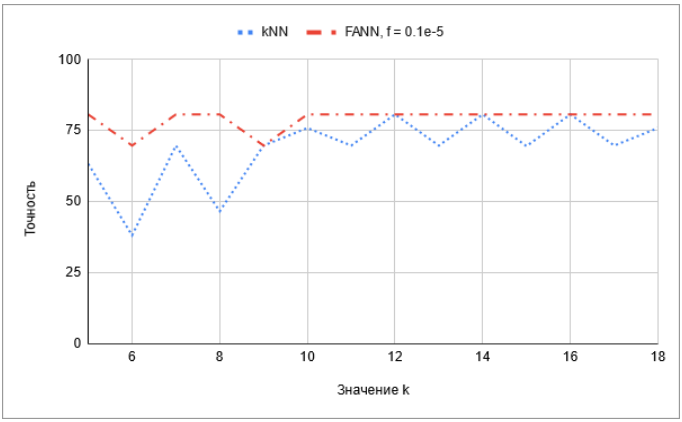


Рис. 2. Сравнение точности классификации для разных значений *k* для методов kNN и FANN

Совокупность алгоритмов сбора и предобработки сигналов ЭЭГ, отбора признаков и классификации сигналов была реализована в виде программной платформы [36]. Программная платформа содержит набор средств для автоматизированной обработки сигналов ЭЭГ и их анализа, в том числе методами машинного обучения. Платформа имеет гибкую архитектуру и состоит из модулей, что позволяет применять её при исследовании данных ЭЭГ для различных целей. Данные могут быть получены как из файлов, так и напрямую от электроэнцефалографа в режиме реального времени. Графический интерфейс предоставляет удобный способ конфигурирования модулей платформы. Рассматриваемая платформа предоставляет открытый для пользователя программный интерфейс взаимодействия. Это позволяет использовать платформу как библиотеку для построения собственных приложений для работы с данными ЭЭГ. Это обстоятельство особенно полезно, например, на этапе, прототипирования нового устройства, выполняющего роль управляющего блока в некоторой системе, использующей данные с электроэнцефалографа. Разработчику достаточно импортировать настройки окружения, реализовать связь с источником данных и объектом управления.

8. Постановка эксперимента. Целью проведения экспериментов была оценка эффективности выбранного подхода. В эксперименте испытуемый сидит в инвалидном кресле, снабженным устройством дистанционного управления, на него надета электродная шапочка беспроводного электроэнцефалографа.

В экспериментах использовалось инвалидное кресло MET COMPACT 15. Кресло оборудовано пультом управления и блоком дистанционного управления, поддерживающим отправку команд по Wi-Fi согласно протокола управления двигателем.

В экспериментах также использовался беспроводной электроэнцефалограф NeoRecCAP [37], который имеет 21 отведение, т.е. M =21. NeoRecCAP может записывать ЭЭГ, метки событий от ИК-триггеров и акселерометра в файлы различных форматов (EDF+ 16 bit, BDF+ 24 bit, GDF 32 bit и т.д.) или передавать эти данные в LSL-поток (Lab Streaming Layer) для анализа сторонним программным обеспечением. NeoRecCAP может применяться для обучения, научных исследований и разработок в области ЭЭГ и нейромашинных интерфейсов. На рисунке 3 схематично показано расположение электродов электроэнцефалографа на поверхности головы (стандартная система 10/20).

И инвалидное кресло, и электроэнцефалограф подключены к устройству управления – ноутбуку.

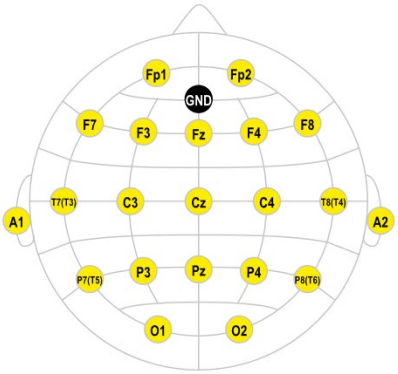


Рис. 3. Схематичное изображение положения электродов на поверхности головы испытуемого (система 10/20)

Все время эксперимента делится на несколько интервалов – эпох, длящихся по 6-10 секунд. Каждая эпоха начинается с подачи звукового сигнала, который говорит испытуемому, в какую сторону нужно повернуть кресло (влево или вправо). Задача испытуемого за каждую эпоху – попытаться представить себе движение левой или правой рукой (в зависимости от звукового сигнала) таким образом, чтобы инвалидное кресло повернулось влево или вправо соответственно.

Таким образом, эксперимент состоит из нескольких этапов: подача звукового сигнала, который говорит испытуемому, в какую сторону нужно повернуть; реакция испытуемого; обработка и анализ сигнала на устройстве управления; принятие решения и выдача управляющего сигнала на объект управления (инвалидное кресло). Схематически этот процесс показан на рисунке 4.

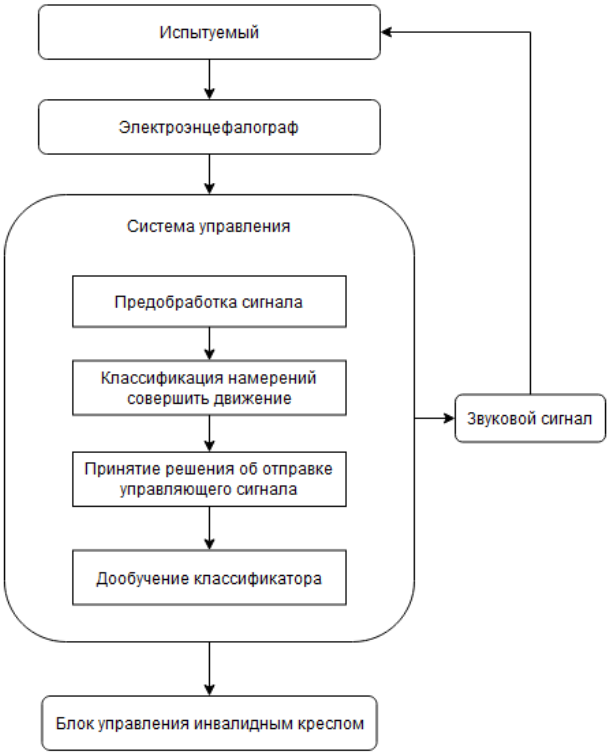


Рис. 4. Схематическое изображение взаимодействия составляющих частей системы в экспериментах

Алгоритм управления был реализован на языке программирования Python. На этапе принятия решения о повороте происходит сохранение ЭЭГ сигнала при удачных попытках управления (когда тип поданного звукового сигнала совпадает с классом распознанного воображаемого движения). Эти сохранённые

отрезки сигнала ЭЭГ дополняют собой датасет для обучения классификатора. Таким образом, возникает эффект адаптации алгоритма под конкретного испытуемого, позволяющий лучше распознавать его намерения.

Дообучение системы управления происходит через определённые промежутки времени, называемые сессиями. Один эксперимент длительностью 20-30 минут может состоять из 3-5 сессий и контрольного тестирования (валидации). Контрольное тестирование выполняется после окончания всех сессий сбора данных и дообучения системы.

Для оценки корректности работы системы в целом также было введено понятие общей точности распознавания $P = D / K$, где D – количество успешных эпох, K – общее количество эпох в сессии. Успешной эпоха считается в том случае, если инвалидное кресло двигалось в рамках этой эпохи в нужном направлении более чем 50% времени.

9. Результаты. Для экспериментальной проверки работоспособности разработанной модели и алгоритмов всего было проведено более 15 экспериментов с пятью испытуемыми в общей сложности. От всех испытуемых было получено информированное согласие на участие в экспериментах. На рисунке 5 показано инвалидное кресло, в котором сидит испытуемый.

Рис. 5. Фотография испытуемого, сидящего в инвалидном кресле во время проведения испытаний

Каждое испытание проводилось по плану, описанному ранее. Одно испытание продолжительностью 27.5 минуты состояло из 4 сессий: первая и вторая сессии длились по 2.5 минуты, остальные две по 10 минут каждая. После этого выполнялось контрольное тестирование, состоящее в том, что испытуемый должен за 2.5 минуты выполнить поворот инвалидного кресла не менее 10 раз.

Испытания показали, что разработанный подход позволяет существенно повысить точность распознавания воображаемых движений. За время испытания происходит адаптация как системы управления (классификатора), так и испытуемого. На рисунке 6 показан график изменения точности распознавания во время первой сессии, а на рисунке 7 – график изменения точности распознавания во время контрольного тестирования.

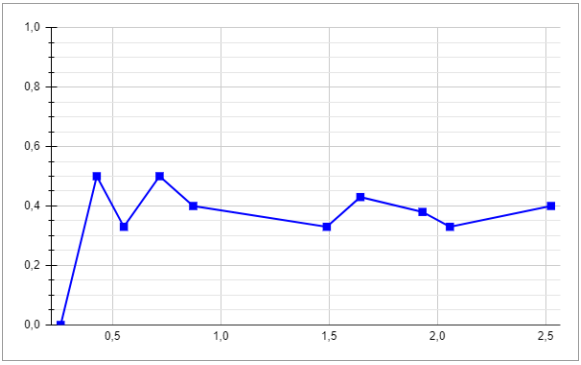


Рис. 6. График изменения точности распознавания (вертикальная ось) во времени (горизонтальная ось, минуты) на первой сессии

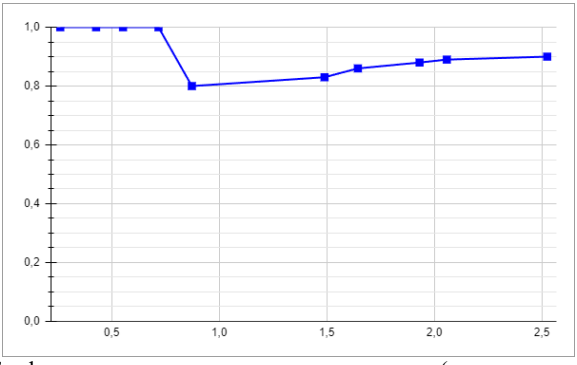


Рис. 7. График изменения точности распознавания (вертикальная ось) во времени (горизонтальная ось, минуты) в процессе контрольного тестирования

Таким образом, сравнивая рисунки 6 и 7 можно отметить, что точность распознавания в ходе испытания повысилась более, чем вдвое. На первой сессии точность распознавания $P \approx 0.4$ (40%), а во время контрольного тестирования – $P \approx 0.9$ (90%).

Проведенные эксперименты показали, что помимо адаптации (обучения) системы управления инвалидным креслом в процессе эксперимента также происходит и адаптация (тренировка) испытуемого. Это означает, что чем дольше испытуемый пытается выполнять задачи управления с помощью активности своего головного мозга, тем чаще у него должно это получаться.

И действительно, было замечено, что в течение одной сессии (т.е. в течение времени, когда параметры обучения неизменны) способность испытуемого совершать воображаемые движения так, чтобы получить правильную реакцию системы управления (и, соответственно, поворот инвалидного кресла в нужную сторону), растёт. Это выражается в том, что значение P возрастает в течение сессии.

Заключение. Разработанный и описанный в данной статье подход и его программная реализация в ходе испытаний продемонстрировали эффективность в задачах управления поворотом инвалидного кресла. В процессе контрольного тестирования люди, участвовавшие в экспериментах, демонстрировали способность выполнять поворот инвалидного кресла вправо и влево с точностью ~90%, что не уступает результатам других известных работ и представляется неплохим результатом для первой обкатки подхода. Предложенная в данной работе модификация классического метода k ближайших соседей, как показано, снижает требования к производительности бортового компьютера кресла и упрощает подбор параметров алгоритма классификации.

Реализованный программный комплекс позволяет автоматизировать процесс обучения системы управления, что упрощает практическое применение подхода в системах управления движением транспортного средства. Этот процесс не требует привлечения специалистов и может быть выполнен самим испытуемым. При этом ноутбук, на котором реализована система, для удобства испытуемого может быть расположен на самом транспортном средстве.

Также было уделено особое внимание ресурсоёмкости программной реализации. Методы и алгоритмы были реализованы с учётом требований, возникающих при их выполнении на низкопроизводительных устройствах с ограниченным объемом

памяти. Программная реализация может быть адаптирована для выполнения на, например, микропроцессорных устройствах Orange Pi 3B [38] без существенной потери скорости обработки сигналов ЭЭГ. Этот микрокомпьютер может иметь до 8 гигабайт оперативной памяти, позволяет выполнять многопоточную обработку данных на 4 процессорных ядрах, при этом имеет умеренное энергопотребление (порядка 6 Вт при максимальной нагрузке), а также небольшие габариты.

Таким образом, предложенный подход к управлению поворотом транспортного средства на основе нейроинтерфейса показал свою перспективность.

Дальнейшие исследования будут направлены на повышение точности распознавания и реализацию задач управления дополнительными типами движения (например, движением вперед и торможением). Представляет интерес также математическое моделирование и исследование процессов обучения и адаптации как системы управления движением, так и самих испытуемых. Перспективным подходом к совершенствованию подобных процессов представляется использование метода рекуррентных целевых неравенств и адаптивных математических моделей нейронов (модели ФитцХью-Нагумо, Хиндмарша-Роуза и т.п) и моделей биологических нейронных сетей, построенных на их основе. Некоторые результаты в этих направлениях были получены недавно в ИПМаш РАН [22, 39].

Литература

1. Левицкая О.С., Лебедев М.А. Интерфейс мозг–компьютер: будущее в настоящем // Вестник РГМУ. 2016. № 2. С. 4–16.
2. Каплан А.Я., Кочетова А.Г., Шишкин С.Л., Басюл И.А., Ганин И.П., Васильев А.Н., Либуркина С.П. Экспериментально-теоретические основания и практические реализации технологии «интерфейс мозг-компьютер» // Бюллетень сибирской медицины. 2013. № 12(2). С. 21–29.
3. Лунев Д.В., Полетыкин С.К., Кудрявцев Д.О. Нейроинтерфейсы: обзор технологий и современные решения // Современные инновации, системы и технологии. 2022. № 2(3). С. 117–126.
4. Станкевич Л.А., Сонькин К.М., Нагорнова Ж.В., Хоменко Ю.Г., Шемякина Н.В. Классификация электроэнцефалографических паттернов воображаемых движений пальцами руки для разработки интерфейса мозг-компьютер // Труды СПИИРАН. 2015. Т. 3(40). С. 163–182.
5. Гунделах Ф.В., Станкевич Л.А. Классификация пространственно-временных паттернов на основе нейроморфных сетей // Информатика и автоматизация. 2024. Т. 23. № 3. С. 886–908.
6. Гунделах Ф.В., Станкевич Л.А., Сонькин К.М., Нагорнова Ж.В., Шемякина Н.В. Применение интерфейсов «мозг-компьютер» в ассистивных технологиях // Труды СПИИРАН. 2020. Т. 19. № 2. С. 277–301.
7. Лобода Ю.О., Функ А.В., Гасымов З.А., Рачкован О.А. Управление мехатронными системами нейроинтерфейсом // XIII Международная научно-

практическая конференция «Электронные средства и системы управления», посвященная 55-летию ТУСУРа (г. Томск, 29 ноября – 1 декабря 2017 г.). 2017. С. 143–146.

8. Бодин О.Н., Солодимова Г.А., Спиркин А.Н. Нейроинтерфейс для управления роботизированными устройствами // Измерение. Мониторинг. Управление. Контроль. 2019. № 4(30). С. 70–76.
9. Миронов В.И., Лобов С.А., Крылова Н.П., Гордлеева С.Ю., Каплан А.Я., Буйлова Т.В., Бахшиев А.В., Щуровский Д.В., Вагнер В.О., Кастальский И.А., Ли А.Н., Казанцев В.Б. Разработка нейроуправляемого автомобиля для мобилизации людей с двигательным дефицитом – нейромобиля // Современные технологии в медицине (СТМ). 2018. Т. 10. № 4. С. 49–56.
10. Rashid Mamunur, Sulaiman Norizam, P.P. Abdul Majeed Anwar, Musa Rabiu Muazu, Ab. Nasir Ahmad Fakhri, Bari Bifta Sama, Khatun Sabira. Current Status, Challenges, and Possible Solutions of EEG-Based Brain-Computer Interface: A Comprehensive Review. Frontiers in Neurorobotics. 2020. vol. 14. DOI: 10.3389/fnbot.2020.00025.
11. Varbu K., Muhammad N., Muhammad Y. Past, Present, and Future of EEG-Based BCI Applications. Sensors. 2022. vol. 22. no. 9. DOI: 10.3390/s22093331.
12. Yadav H., Maini S. Electroencephalogram based brain-computer interface: Applications, challenges, and opportunities. Multimedia Tools and Applications. 2023. vol. 82. pp. 47003–47047.
13. Сазонова Н.Н., Дегтярев С.В., Сазонова Е.С. Аппаратно-программный комплекс на основе нейроинтерфейса и vr-технологии в системе реабилитации пациентов с поражением головного мозга после инсульта // Информационные технологии в управлении, автоматизации и мехатронике: Сборник научных статей 4-й Международной научно-технической конференции (г. Курск, 7 апреля 2022 г.). 2022. С. 180–183.
14. Боброва Е.В., Решетникова В.В., Вершинина Е.А., Гришин А.А., Исаев М.Р., Бобров П.Д., Герасименко Ю.П. Оценка эффективности управления мозг-компьютерным интерфейсом при обучении воображению движений верхних и нижних конечностей // Журнал высшей нервной деятельности им. И.П. Павлова. 2023. Т. 73. № 1. С. 52–61.
15. Каплан А.Я. Некоторые теоретические и практические основания к реализации нейроинтерфейсных технологий в психиатрии // Психическое здоровье человека и общества. Актуальные междисциплинарные проблемы. Научно-практическая конференция (г. Москва, 30 октября 2017 г.). Москва: КДУ, 2018. С. 366–372.
16. Wang H., Yan F., Xu T., Yin H., Chen P., Yue H., Chen C., Zhang H., Xu L., He Y., Bezerianos A. Brain-Controlled Wheelchair Review: From Wet Electrode to Dry Electrode, From Single Modal to Hybrid Modal, From Synchronous to Asynchronous. IEEE Access. 2021. vol. 9. pp. 55920–55938.
17. Fernandez-Rodriguez A., Velasco-Alvarez F., Ron-Angevin R. Review of real brain-controlled wheelchairs. Journal of neural engineering. 2016. vol. 13. no. 6. DOI: 10.1088/1741-2560/13/6/061001.
18. Sha'abani M.N.A.H., Fuad N., Jamal N., Ismail M.F. kNN and SVM Classification for EEG: A Review. Lecture Notes in Electrical Engineering. Springer, Singapore. 2020. vol. 632. pp. 555–565.
19. Palumbo A., Gramigna V., Calabrese B., Ielpo N. Motor-Imagery EEG-Based BCIs in Wheelchair Movement and Control: A Systematic Literature Review. Sensors 2021. vol. 21. no. 18. DOI: 10.3390/s21186285.
20. Яшин А.С., Васильев А.Н., Шишкин С.Л. Чем квазидвижения полезны для изучения произвольных движений? Взгляд со стороны нейронаук, психологии и философии. Гены и Клетки. 2023. Т. 18. № 4. С. 649–652.

21. Babbysh N. Computing brain rhythm indicators of EEG signal. 5th Scientific School Dynamics of Complex Networks and their Applications (DCNA), Kaliningrad, Russian Federation. IEEE, 2021. pp. 32–35.
22. Shanarova N., Pronina M., Lipkovich M., Ponomarev V., Müller A., Kropotov J. Application of Machine Learning to Diagnostics of Schizophrenia Patients Based on Event-Related Potentials. Diagnostics. 2023. vol. 13. no. 3. DOI: 10.3390/diagnostics13030509.
23. Lipkovich M., Knyazeva V., Aleksandrov A., Shanarova N., Sagatdinov A., Fradkov A. Evoked Potentials Detection During Self-Initiated Movements Using Machine Learning Approach. 2023 Fifth International Conference on Neurotechnologies and Neurointerfaces (CNN). Kaliningrad, Russian Federation, 2023. pp. 47–50.
24. Овод И.В., Осадчий А.Е., Пупышев А.А., Фрадков А.Л. Формирование нейрообратной связина основе адаптивной модели активности головного мозга // Нейрокомпьютеры: разработка, применение. 2012. № 2. С. 36–41.
25. Rybalko A., Fradkov A. Identification of Two-Neuron FitzHugh-Nagumo Model Based on the Speed-Gradient and Filtering. Chaos: An Interdisciplinary Journal of Nonlinear Science. 2023. vol. 33. no. 8. DOI: 10.1063/5.0159132.
26. Обухов С.А., Степанов В.П., Рудаков И.В. Математическая модель нейрокомпьютерного интерфейса на основе анализа вызванных потенциалов P300 // Вестник РГГУ. Серия: Информатика. Информационная безопасность. Математика. 2021. № 2. С. 48–67.
27. Chen Z, Wang Y, Song Z. Classification of Motor Imagery Electroencephalography Signals Based on Image Processing Method. Sensors (Basel). 2021. vol. 21(14). DOI: 10.3390/s21144646.
28. Lun X., Liu J., Zhang Y., Hao Z., Hou Y., A Motor Imagery Signals Classification Method via the Difference of EEG Signals Between Left and Right Hemispheric Electrodes. Frontiers in Neuroscience. 2022. vol. 16. DOI: 10.3389/fnins.2022.865594.
29. Васильев В.П., Муро Э.Л., Смольский С.М. Основы теории и расчета цифровых фильтров: учебное пособие, 2-еизд., стер. // М.: НИЦ ИНФРА-М. 2024. 272 с.
30. Hussin S.F., Birasamy G., Hamid Z. Design of Butterworth Band-Pass Filter // Politeknik and Kolej Komuniti Journal of Engineering and Technology. 2016. vol. 1. no. 1. pp. 32–46.
31. Haykin S. Neural Networks: A Comprehensive Foundation Second Edition. 2019. 1104 p.
32. Mohri M., Rostamizadeh A., Talwalkar A. Foundations of Machine Learning. The MIT Press. 2012. 504 p.
33. Капралов Н.В., Нагорнова Ж.В., Шемякина Н.В. Методы классификации ЭЭГ-паттернов воображаемых движений // Информатика и автоматизация. 2021. Т. 20. № 1. С. 94–132.
34. Keller J.M., Gray M.R., Givens J.A. A fuzzy K-nearest neighbor algorithm // IEEE Transactions on Systems, Man, and Cybernetics. 1985. vol. 15. no. 4. pp. 580–585.
35. Kumbure M.M., Luukka P. A generalized fuzzy k-nearest neighbor regression model based on Minkowski distance // Granular Computing. 2022. vol. 7. pp. 657–671.
36. Бабич Н.А. Программная платформа для чтения, обработки и анализа данных ЭЭГ // Программная инженерия. 2023. № 5. С. 254–260.
37. Официальная страница электроэнцефалографа NeoRecCAP. URL: https://mks.ru/product/neoreccap/ (дата обращения: 10.04.2024).
38. Официальная страница процессорного модуля Orange-Pi-3B. URL: http://www.orangepi.org/html/hardWare/computerAndMicrocontrollers/details/Orange-Pi-3B.html (дата обращения: 03.05.2024).

39. Kovalchukov A. Approximate Hindmarsh-Rose model identification: application to EEG data // 7th Scientific School on Dynamics of Complex Networks and their Applications (DCNA). 2023. pp. 151–154.

Фрадков Александр Львович — д-р техн. наук, профессор, главный научный сотрудник, Институт проблем машиноведения РАН. Область научных интересов: математическое моделирование, адаптивное и интеллектуальное управление нелинейными системами, машинное обучение, кибернетическая физика. Число научных публикаций — 800+. fradkov@mail.ru; Большой проспект В.О., 61, 199178, Санкт-Петербург, Россия.

Бабич Николай Александрович — аспирант, Институт проблем машиноведения РАН. Область научных интересов: математическое моделирование, обработка сигналов, машинное обучение. Число научных публикаций — 15. nickware@mail.ru; Большой проспект В.О., 61, 199178, Санкт-Петербург, Россия.

Поддержка исследований. Работа поддержана Министерством науки и высшего образования Российской Федерации (проект № 124041500008-1).

DOI 10.15622/ia.24.1.1

A. FRADKOV, N. BABICH

THREE-POSITION VEHICLE CONTROL BASED ON NEURAL INTERFACE USING MACHINE LEARNING

Fradkov A., Babich N. **Three-Position Vehicle Control Based on Neural Interface Using Machine Learning.**

Abstract. The brain-computer interface is a complex system that allows you to control external electronic devices using brain activity. This system includes several elements – a device for reading brain activity signals, a hardware and software complex that processes and analyzes these signals, and a control object. The main challenge here is the development of methods and algorithms that can correctly recognize and predict the intentions of the person who uses this interface to provide solutions to control problems. This paper describes the mathematical formulation of the equipment control problem. Methods for preprocessing EEG signals, analyzing them, and making decisions about issuing a control signal are described; the structure of the software implementation of these methods is described, as well as a plan for experimental testing of the performance of the entire system that forms the brain-computer interface. For classification of EEG signals the methods of machine learning are used. A modification of the k-nearest neighbors method is proposed – the so-called fuzzy almost nearest neighbors method. An algorithm for the adaptive classification of EEG taking into account the drift of the parameters of the subject's model based on the method of recurrent objective inequalities (ROI) has also been developed. The control algorithm was implemented in the Python programming language. A remote-controlled wheelchair is considered as a control object, and turning the chair to the right or left is considered as a control task. To experimentally test the performance of the developed model and algorithms, more than 15 tests were carried out with five subjects in total. The approach developed and described in this article and its software implementation during testing demonstrated its effectiveness in the tasks of controlling the rotation of a wheelchair. Special attention was also paid to the resource intensity of the software implementation. Methods and algorithms were implemented taking into account the requirements that arise when performing calculations on low-performance devices with a limited amount of memory.

Keywords: neural interface, BCI, control, machine learning, EEG, brain activity, KNN.

References

1. Levickaja O.S., Lebedev M.A. [Brain-computer interface: the future in the present]. Vestnik RGMU – Bulletin of the Russian State Medical University. 2016. no. 2. pp. 4–16. (In Russ.).
2. Kaplan A.Ja., Kochetova A.G., Shishkin S.L., Basjul I.A., Ganin I.P., Vasil'ev A.N., Liburkina S.P. [Experimental and theoretical foundations and practical implementations of the brain-computer interface technology]. Bjulleten' sibirskoj mediciny – Bulletin of Siberian Medicine. 2013. no. 12(2). pp. 21–29. (In Russ.).
3. Lunev D.V., Poletykin S.K., Kudrjavcev D.O. [Neural interfaces: review of technologies and modern solutions]. Sovremennye innovacii, sistemy i tehnologii – Modern innovations, systems and technologies. 2022. no. 2(3). pp. 117–126. (In Russ.).
4. Stankevich L.A., Son'kin K.M., Nagornova Zh.V., Homenko Ju.G., Shemjakina N.V. [Classification of electroencephalographic patterns of imaginary movements of the fingers for the development of a brain-computer interface]. Trudy SPIIRAN – SPIIRAN Proceedings. 2015. vol. 3(40). pp. 163–182. (In Russ.).

5. Gundelah F.V., Stankevich L.A. [Classification of spatiotemporal patterns based on neuromorphic networks]. Informatika i avtomatizaciya – Informatics and Automation. 2024. vol. 23. no. 3. pp. 886–908. (In Russ.).
6. Gundelah F.V., Stankevich L.A., Son'kin K.M., Nagornova Zh.V., Shemjakina N.V. [Application of brain-computer interfaces in assistive technologies]. Trudy SPIIRAN – SPIIRAN Proceedings. 2020. vol. 19. no. 2. pp. 277–301. (In Russ.).
7. Loboda Ju.O., Funk A.V., Gasymov Z.A., Rachkovan O.A. Upravlenie mehatronnymi sistemami nejrointerfejsom [Control of mechatronic systems using a neural interface]. XIII Mezhdunarodnaja nauchno-prakticheskaja konferencija, posvjashhennaja 55-letiju TUSURa [XIII International scientific and practical conference dedicated to the 55th anniversary of TUSUR]. 2017. pp. 143–146. (In Russ.).
8. Bodin O.N., Solodimova G.A., Spirkin A.N. [Neural interface for controlling robotic devices]. Izmerenie. Monitoring. Upravlenie. Kontrol' – Measurement. Monitoring. Management. Control. 2019. no. 4(30). pp. 70–76. (In Russ.).
9. Mironov V.I., Lobov S.A., Krylova N.P., Gordleeva C.Ju., Kaplan A.Ja., Bujlova T.V., Kazancev V.B. [Development of a neuro-controlled car to mobilize people with motor deficits – a neuromobile]. Sovremennye tehnologii v medicine – Modern technologies in medicine. 2018. vol. 10. no. 4. pp. 49–56. (In Russ.).
10. Rashid Mamunur, Sulaiman Norizam, P.P. Abdul Majeed Anwar, Musa Rabiu Muazu, Ab. Nasir Ahmad Fakhri, Bari Bifta Sama, Khatun Sabira. Current Status, Challenges, and Possible Solutions of EEG-Based Brain-Computer Interface: A Comprehensive Review. Frontiers in Neurorobotics. 2020. vol. 14. DOI: 10.3389/fnbot.2020.00025.
11. Varbu K., Muhammad N., Muhammad Y. Past, Present, and Future of EEG-Based BCI Applications. Sensors. 2022. vol. 22. no. 9. DOI: 10.3390/s22093331.
12. Yadav H., Maini S. Electroencephalogram based brain-computer interface: Applications, challenges, and opportunities. Multimedia Tools and Applications. 2023. vol. 82. pp. 47003–47047.
13. Sazonova N.N., Degtjarev S.V., Sazonova E.S. Apparatno-programmnyj kompleks na osnove nejrointerfejsa i vr-tehnologii v sisteme reabilitacii pacientov s porazheniem golovnogo mozga posle insul'ta [Hardware and software complex based on a neural interface and VR technology in the rehabilitation system for patients with brain damage after a stroke]. Informacionnye tehnologii v upravlenii, avtomatizacii i mehatronike: Sbornik nauchnyh statej 4-j Mezhdunarodnoj nauchno-tehnicheskoj konferencii [Information technologies in control, automation and mechatronics: Collection of scientific articles of the 4th International Scientific and Technical Conference]. 2022. pp. 180–183.
14. Bobrova E.V., Reshetnikova V.V., Vershinina E.A., Grishin A.A., Isaev M.R., Bobrov P.D., Gerasimenko Ju.P. [Evaluation of the effectiveness of brain-computer interface control in teaching the imagination of movements of the upper and lower extremities]. Zhurnal vysshej nervnoj dejatel'nosti im. I.P. Pavlova – Journal of Higher Nervous Activity named after I.P. Pavlov. 2023. vol. 73. no. 1. pp. 52–61.
15. Kaplan A.Ja. ekotorye teoreticheskie i prakticheskie osnovanija k realizacii nejrointerfejsnyh tehnologij v psihiatrii [Some theoretical and practical grounds for the implementation of neural interface technologies in psychiatry] Psihicheskoe zdorov'e cheloveka i obshhestva. Aktual'nye mezhdisciplinarnye problemy. Nauchno-prakticheskaja konferencija [Mental health of individuals and society. Current interdisciplinary problems. Scientific and practical conference]. Moscow: CDU, 2018. pp. 366–372. (In Russ.).
16. Wang H., Yan F., Xu T., Yin H., Chen P., Yue H., Chen C., Zhang H., Xu L., He Y., Bezerianos A. Brain-Controlled Wheelchair Review: From Wet Electrode to Dry Electrode, From Single Modal to Hybrid Modal, From Synchronous to Asynchronous. IEEE Access. 2021. vol. 9. pp. 55920–55938.

17. Fernandez-Rodriguez A., Velasco-Alvarez F., Ron-Angevin R. Review of real brain-controlled wheelchairs. Journal of neural engineering. 2016. vol. 13. no. 6. DOI: 10.1088/1741-2560/13/6/061001.
18. Sha'abani M.N.A.H., Fuad N., Jamal N., Ismail M.F. kNN and SVM Classification for EEG: A Review. Lecture Notes in Electrical Engineering. Springer, Singapore. 2020. vol. 632. pp. 555–565.
19. Palumbo A., Gramigna V., Calabrese B., Ielpo N. Motor-Imagery EEG-Based BCIs in Wheelchair Movement and Control: A Systematic Literature Review. Sensors 2021. vol. 21. no. 18. DOI: 10.3390/s21186285.
20. Yashin A.S., Vasil'ev A.N., SHishkin S.L. [How are quasi-movements useful for learning voluntary movements? A view from neuroscience, psychology and philosophy]. Geny i Kletki – Genes and Cells. 2023. vol. 18. no. 4. pp. 649–652. (In Russ.).
21. Babbysh N. Computing brain rhythm indicators of EEG signal. 5th Scientific School Dynamics of Complex Networks and their Applications (DCNA), Kaliningrad, Russian Federation. IEEE, 2021. pp. 32–35.
22. Shanarova N., Pronina M., Lipkovich M., Ponomarev V., Müller A., Kropotov J. Application of Machine Learning to Diagnostics of Schizophrenia Patients Based on Event-Related Potentials. Diagnostics. 2023. vol. 13. no. 3. DOI: 10.3390/diagnostics13030509.
23. Lipkovich M., Knyazeva V., Aleksandrov A., Shanarova N., Sagatdinov A., Fradkov A. Evoked Potentials Detection During Self-Initiated Movements Using Machine Learning Approach. 2023 Fifth International Conference on Neurotechnologies and Neurointerfaces (CNN). Kaliningrad, Russian Federation, 2023. pp. 47–50.
24. Ovod I.V., Osadchij A.E., Pupyshev A.A., Fradkov A.L. [Formation of neurofeedback based on an adaptive model of brain activity]. Nejrokomp'jutery – Neurocomputers. 2012. no. 2. pp. 36–41. (In Russ.).
25. Rybalko A., Fradkov A. Identification of Two-Neuron FitzHugh-Nagumo Model Based on the Speed-Gradient and Filtering. Chaos: An Interdisciplinary Journal of Nonlinear Science. 2023. vol. 33. no. 8. DOI: 10.1063/5.0159132.
26. Obuhov S.A., Stepanov V.P., Rudakov I.V. [Mathematical model of a brain-computer interface based on the analysis of evoked potentials P300]. Vestnik RGGU. Serija: Informatika. Informacionnaja bezopasnost'. Matematika – Bulletin of the Russian State University for the Humanities. Series: Computer Science. Information Security. Mathematics. 2021. no. 2. pp. 48–67. (In Russ.).
27. Chen Z, Wang Y, Song Z. Classification of Motor Imagery Electroencephalography Signals Based on Image Processing Method. Sensors (Basel). 2021. vol. 21(14). DOI: 10.3390/s21144646.
28. Lun X., Liu J., Zhang Y., Hao Z., Hou Y., A Motor Imagery Signals Classification Method via the Difference of EEG Signals Between Left and Right Hemispheric Electrodes. Frontiers in Neuroscience. 2022. vol. 16. DOI: 10.3389/fnins.2022.865594.
29. Vasil'ev V.P., Muro Je.L., Smol'skij S.M. Osnovy teorii i rascheta cifrovyh fil'trov: uchebnoe posobie, 2-e izd., ster. [Fundamentals of the theory and calculation of digital filters: textbook, 2nd ed.]. Moscow: NIC INFRA-M. 2024. 272 p. (In Russ.).
30. Hussin S.F., Birasamy G., Hamid Z. Design of Butterworth Band-Pass Filter. Politeknik and Kolej Komuniti Journal of Engineering and Technology. 2016. vol. 1. no. 1. pp. 32–46.
31. Haykin S. Neural Networks: A Comprehensive Foundation Second Edition. 2019. 1104 p.
32. Mohri M., Rostamizadeh A., Talwalkar A. Foundations of Machine Learning. The MIT Press. 2012. 504 p.

33. Kapralov N.V., Nagornova Zh.V., Shemjakina N.V. [Methods for classifying EEG patterns of imaginary movements]. Informatika i avtomatizacija – Informatics and Automation. 2021. vol. 20. no. 1. pp. 94–132.
34. Keller J.M., Gray M.R., Givens J.A. A fuzzy K-nearest neighbor algorithm. IEEE Transactions on Systems, Man, and Cybernetics. 1985. vol. 15. no. 4. pp. 580–585.
35. Kumbure M.M., Luukka P. A generalized fuzzy k-nearest neighbor regression model based on Minkowski distance. Granular Computing. 2022. vol. 7. pp. 657–671.
36. Babich N.A. [Software platform for reading, processing and analyzing EEG data].Programmnaja inzhenerija – Software Engineering. 2023. no. 5. pp. 254–260. (In Russ.).
37. Oficial'naja stranica jelektrojencefalografa NeoRecCAP [Official page of the electroencephalograph NeoRecCAP]. Available at: https://mks.ru/product/neoreccap/ (accessed 10.04.2024). (In Russ.).
38. Oficial'naja stranica processornogo modulja Orange-Pi-3B [Official page of the Orange-Pi-3B processor module]. Available at: http://www.orangepi.org/html/hardWare/computerAndMicrocontrollers/details/Orange-Pi-3B.html (accessed 03.05.2024).
39. Kovalchukov A. Approximate Hindmarsh-Rose model identification: application to EEG data. 7th Scientific School on Dynamics of Complex Networks and their Applications (DCNA). 2023. pp. 151–154.

Fradkov Alexander — Ph.D., Dr.Sci., Professor, Chief researcher, Institute for Problems in Mechanical Engineering of the Russian Academy of Sciences. Research interests: mathematical modeling, adaptive and intelligent control, nonlinear systems, machine learning, cybernetic physics. The number of publications — 800+. fradkov@mail.ru; 61, Bolshoi Av. of V.O., 199178, St. Petersburg, Russia; office phone.

Babich Nickolay — Postgraduate student, Institute for Problems in Mechanical Engineering of the Russian Academy of Sciences. Research interests: mathematical modeling, signal processing, machine learning. The number of publications — 15. nickware@mail.ru; 61, Bolshoi Av. of V.O., 199178, St. Petersburg, Russia; office phone.

Acknowledgements. The research was supported by the Ministry of Science and Higher Education of the Russian Federation (project no. 124041500008-1).

DOI 10.15622/ia.24.1.2

A. GHARBI, M. AYARI, Y. EL TOUATI

INTELLIGENT AGENT-CONTROLLED ELEVATOR SYSTEM: ALGORITHM AND EFFICIENCY OPTIMIZATION

Gharbi A., Ayari M., El Touati Y. **Intelligent Agent-Controlled Elevator System: Algorithm and Efficiency Optimization.**

Abstract. The study introduces an innovative intelligent agent-controlled elevator system specially designed to improve passenger wait times and enhance the efficiency of high-rise buildings. By utilizing the classic single-agent planning model, we developed a unique strategy for handling calls from halls and cars, and combined with this strategy we significantly improved the overall performance of the elevator system. Our intelligent control methods are in-depth compared with conventional elevator systems, assessing three important performance indicators: response time, system capacity to handle multiple active elevator cars simultaneously, and average passenger waiting time. The results of the full simulation show that an intelligent agent-based model consistently exceeds conventional elevator systems in all measured criteria. Intelligent control systems have significantly reduced response times, and improved simultaneous elevator management and passenger wait times, especially during high traffic. These advances not only improved traffic flow efficiency, but also greatly contributed to passenger satisfaction and brought smoother and more reliable transport experiences within the building. Furthermore, the increased efficiency of our systems is in line with the goals of building energy management, as it minimizes unnecessary movements and idle time. The results demonstrate the system's ability to meet dynamic, high-occupation environment requirements and mark a significant step forward in intelligent infrastructure management.

Keywords: intelligent agent control, elevator system optimization, passenger waiting time.

1. Introduction.

Elevator systems are integral to vertical transportation in modern buildings, but traditional control methods often result in suboptimal performance, particularly under fluctuating traffic patterns. Recent advancements in artificial intelligence offer new opportunities to enhance system efficiency. This paper introduces an intelligent agent-controlled elevator system designed to minimize passenger wait times through dynamic real-time optimization.

This paper addresses the complexities inherent in elevator system management within the context of intelligent agent control, focusing on optimizing operations to enhance performance. It introduces a classical single-agent project planning model specifically designed for elevator systems, incorporating advanced strategies for both hall call and car call processing. The primary objective is to improve elevator system performance by minimizing passenger waiting times and managing traffic flow more effectively. To achieve this, the paper outlines detailed procedures and algorithms tailored to these tasks.

Additionally, the proposed algorithms are validated through simulations, providing empirical evidence of their superior efficiency compared to conventional methods. The key contribution of this work is its

comprehensive approach to elevator system optimization, which offers a pathway to developing smarter, more adaptive elevator solutions for modern buildings.

The state of the art in intelligent agent-controlled elevator systems reflects a convergence of technologies, including machine learning, real-time data analytics, IoT integration, and advanced algorithms. These advancements have paved the way for more efficient, adaptive, and user-centric elevator systems capable of optimizing transportation operations and minimizing passenger wait times.

Machine learning techniques, including reinforcement and deep learning, are applied to elevator systems for intelligent decision-making. These techniques enable agents to learn the best strategies through observation and interaction, improving efficiency and adaptability [1 – 4]. However, integration of machine learning into elevators brings challenges such as data collection burdens, model interpretation in safety-critical systems, adaptation to dynamic environments, avoidance of over-adaptation, addressing algorithmic biases, ensuring robustness and efficiently managing computational resources.

Real-time data analysis uses data such as elevator positions and passenger demand to optimize operations and reduce waiting times [5, 6]. However, in complex elevator environments, the need for reliable data flows, robust infrastructure, and effective data processing poses challenges, which can hinder computer resources and affect decision-making speed. To ensure data accuracy, reliability, and security, rigorous monitoring and validation processes are also required.

IoT integration in elevator systems includes sensors and devices to optimize operations using real-time data [7 – 9]. However, cybersecurity risks increase with IoT connections, which requires robust security measures such as encryption and regular audits. The management of a wide range of IoT devices requires scalable connections, efficient data processing, and solutions for interoperability challenges between different devices and platforms.

Advanced planning and optimization algorithms dynamically allocate elevator assignments, optimize traffic flow, and reduce waiting times taking into account passenger demand, elevator capacity and building traffic patterns in real time [10, 11]. However, this integration is subject to challenges due to the complexity of the computation, especially in large systems. The development and maintenance of sophisticated adaptive algorithms requires significant resources and expertise. System uncertainty and instability require robust validation, testing and reactive mechanisms for reliability and safety.

Advanced planning and optimization algorithms dynamically allocate the assignment of elevators, optimize traffic flow, and reduce wait times, considering passenger demand, elevator capacity and real-time traffic patterns [10, 11]. However, this integration is faced with challenges due to the complexity of the calculation, especially in large systems. Developing and maintaining sophisticated adaptive algorithms require considerable resources and expertise. Uncertainty and instability of the system require robust validation, testing and reactive mechanisms for reliability and safety.

Elevator systems are often complex and involve multiple elevators serving different floors simultaneously. Multi-agent systems employ intelligent agents to coordinate and synchronize the actions of multiple elevators, resulting in improved traffic flow, reduced congestion, and enhanced transportation efficiency.

The state of the art in intelligent agent-controlled elevator systems reflects a convergence of technologies, including machine learning, real-time data analytics, IoT integration, and advanced algorithms. These advancements have paved the way for more efficient, adaptive, and user-centric elevator systems capable of optimizing transportation operations and minimizing passenger wait times.

This paper makes significant contributions by proposing new controls and strategies for optimizing the handling of hall calls, moving car procedures and handling of car calls in elevator systems. The paper further validates these contributions by simulation-based assessments, highlighting improvements in key performance metrics such as reduced passenger wait times, optimized traffic flow, increased energy efficiency and increased system response.

The remainder of this paper is organized as follows: Section 2 elaborates on the elevator problem and the challenges associated with traditional control approaches. Section 3 discusses the proposed control actions and strategies for optimizing elevator system efficiency. Section 4 presents the single-agent classical planning task model and its components, including hall call handling, move car procedures, and car call handling. Section 5 details the simulation setup and results, showcasing the efficacy of the proposed algorithms. Finally, Section 6 concludes the paper with a summary of contributions, insights into future research directions, and implications for the field of elevator system management.

2. An Elevator Problem. The miconic planning domain [15 – 19] involves transporting several passengers between different floors of a building using an elevator that can move up or down between floors, and passengers can enter or leave the elevator on each floor. Each passenger has an origin and a destination floor, and the initial floor f0 of the elevator is

given. There are two types of actions in this domain: for each floor f, the "up(f)" and "down(f)" actions are considered, signifying the change in the current floor resulting from the action taken. This conceptualization yields a total of 2np + 2(nf – 1) distinct actions, where np represents the count of passengers and nf signifies the number of floors in the system. It is imperative to note that the system architecture excludes the provision of an "up" action from the topmost floor or a "down" action from the bottommost floor, reflecting the physical constraints inherent in such transport systems.

Figure 1 shows the resulting automaton where:

– The diagram has two states, "i" and "j", which represent the current floor of the elevator.

– Both states are initially in an "Idle" state, indicating that the elevator is not moving.

– The transitions between states are represented by "MoveDown" and "MoveUp" actions. These actions are triggered by conditions R_i, which represent requests for the elevator to move to a different floor.

Figure 1 illustrates how the elevator system responds to these requests by changing its state (i.e., moving to a different floor).

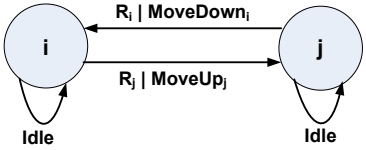


Fig. 1. Elevator automaton

The input set, represented as $I = \{R_i\}$, delineates the requested floors within the system, where R_i signifies a request for the i^{th} floor. Correspondingly, the output set $O = \{d_j, n, u_j\}$ encapsulates directives regarding the elevator's movement, including the intended direction and destination floor. For instance, d_i denotes a downward movement to the i^{th} floor, u_j denotes an upward movement to the j^{th} floor, and n indicates the elevator should remain in an idle state.

The set of states, denoted as $S = \{S_i\}$, characterizes the elevator's current floor position within the system. Illustrated in Figure 1, when the elevator is situated at state S_i and receives a request for the j^{th} floor (R_j), it will actuate a downward movement (d_j) to floor j, thereby transitioning from state S_i to state S_j. This delineation establishes a clear framework for modeling the elevator's dynamic behavior based on input requests and current state conditions.

3. Control Actions. Elevator operations are constrained by a predefined set of permissible actions dictated by the system's operational rules. When an elevator is positioned on the floor, it faces the decision to either ascend or descend. While in transit between floors, it must choose between halting at the upcoming floor or bypassing it. These actions, however, are subject to constraints influenced by passenger expectations and operational guidelines. Notably, the elevator is obliged to accommodate passenger requests for disembarkation or change of direction before proceeding past a floor. Moreover, certain principles are integrated into the system to reflect foundational knowledge. These principles dictate that the elevator must halt only if there are passengers intending to enter or exit, avoid picking up passengers if another elevator is already stationed on that floor, and prioritize upward movement over downward movement. Consequently, the available choices for each elevator are limited to halting or continuing its trajectory. As the time taken to execute these actions varies among elevators, they perform their actions asynchronously, leading to staggered completion times.

To evaluate the cost associated with a plan π, the following formula can be employed:

$$C(\pi) = \sum_{ai \in \pi} c(a_i),$$

where $\sum_{ai \in \pi}$ means to sum up the cost of each action in the plan π, and c(ai) is the cost of each individual action.

The flowchart delineated in Figure 2 elucidates the operational sequence of the proposed model, which initiates the generation of a hall call within the elevator system. Initially, the model assesses the availability of elevators, discerning if multiple elevators are unoccupied. In such instances, the model employs a selective approach to designate the elevator with the shortest waiting time to respond to the hall call and dispatch it to the requested floor. Conversely, when only a single elevator is available, it is directed to the requested floor using the same selective approach. However, in scenarios where no elevators are vacant, the model undertakes a more intricate decision-making process. This involves calculating the movement direction and assessing the capacity of each elevator. By aligning these parameters with the requirements of the hall call, the model employs a collective approach to determine the most suitable elevator for the task and assigns it to respond to the request.

The model utilizes two distinct categories of variables: state variables and action command variables. State variables encapsulate the current operational status and conditions of the elevators within the system,

aiding in decision-making processes. Conversely, action command variables are triggered upon the generation of hall calls or car calls, initiating decision pathways and directing elevator movements accordingly. This structured approach optimizes elevator response times, resource utilization, and overall system efficiency, enhancing passenger experience and operational performance within the elevator system.

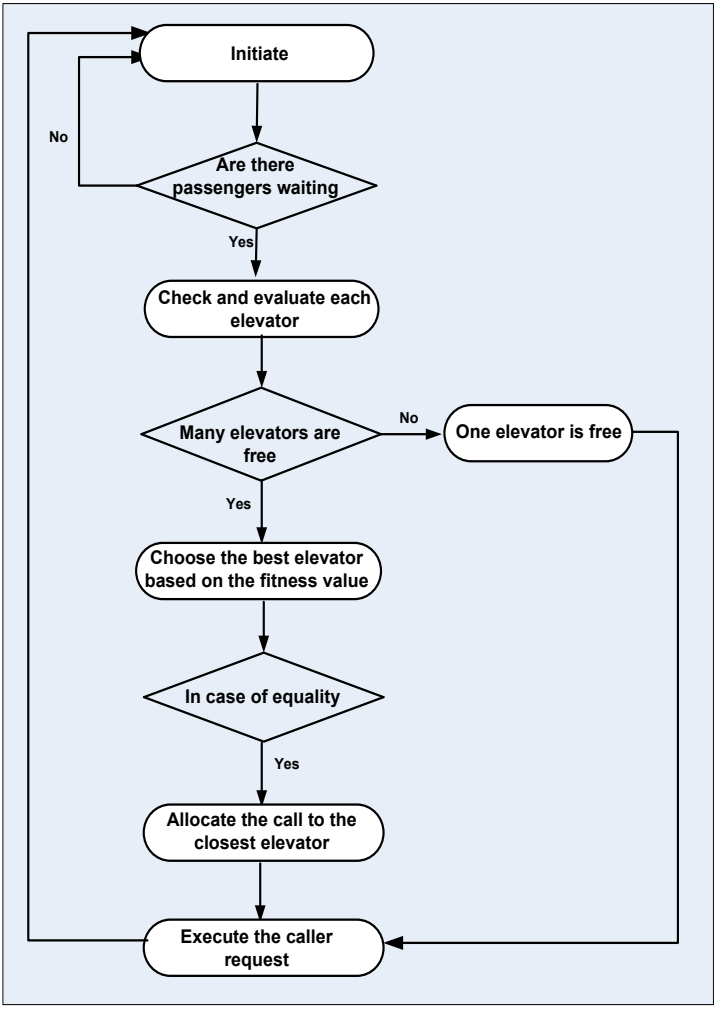


Fig. 2. Flowchart of conventional elevator control system

4. Single-agent classical planning task for an elevator system. The elevator system operates within a building encompassing multiple floors, denoted by numbers ranging from 1 to n (where n represents the total floors in the building). The elevator's movement capability is limited to traversing one floor at a time, either ascending or descending, with the flexibility to halt at any floor along its route. Additionally, the elevator has a predefined passenger capacity of c, and each passenger is associated with a specific destination floor within the building. Passengers can only embark and disembark at floors where the elevator makes stops, adhering to operational protocols.

4.1. Intelligent Agent-Controlled Elevator System. The intelligent control agent for managing multiple elevators and calls operates by implementing a closed-loop system that continuously uses feedback to reassess key input parameters for the decision-making process. The agent's primary goal is to minimize passenger waiting times by dynamically assigning elevators to calls based on real-time conditions. Various contextual parameters, including the current floor, movement direction, current status, and waiting time, are considered. The waiting time (T waiting) for arrival at the departure floor is dynamically calculated in real time, considering the current direction of movement and the stop-request queue.

The input data for the proposed system consists of fundamental operational information, including:

1) The current movement direction of each elevator (upward, downward, or stationary).

2) The precise position of each elevator and the corresponding landing floors.

3) The log of active car calls within each cabin.

During each iteration, the control agent evaluates potential assignments, selecting the one with the optimal fitness (i.e., the best match between elevator and call that minimizes waiting time). Once the best assignment is determined, it is finalized, and the system recalculates the remaining options, factoring in the newly fixed allocation and any changes in the system's state, such as updated elevator positions or new calls.

When the control loop is first executed in a building with (k) elevators and (n) landing calls, the agent must evaluate a maximum number of decision-making processes. With each subsequent iteration, one less landing call needs to be evaluated, as the best allocation from the previous iteration is fixed and removed from the pool. This process continues until all landing calls have been assigned to elevators.

By performing this iterative process, the intelligent control agent effectively reduces the complexity of the problem, despite the large number

of possible solutions. The algorithm is designed to find an optimal or near-optimal solution to this NP-hard problem without imposing significant computational demands. The agent not only optimizes individual elevator-call assignments but also dynamically adjusts the order in which calls are dispatched, ensuring efficient and responsive elevator operation in real time.

In practical implementation, the comprehensive up-peak energy-saving scheduling strategy is outlined as follows, with the corresponding flowchart depicted in Figure 3. During each scheduling cycle, the intelligent controller agent, using the proposed algorithm, independently retrieves the current state of each elevator, all active call signals, and the average waiting time from the previous scheduling cycle. The intelligent controller agent then determines the number of elevators required to serve the waiting passengers, guided by an energy-time piecewise function.

Subsequently, the intelligent controller agent selects the specific elevators to deploy, based on an optimization scheduling solution aimed at minimizing waiting time for passengers. After iterative optimization, an optimal scheduling plan is derived, and all elevators are dispatched accordingly.

Once the elevators are scheduled based on the optimal plan, the scheduling algorithm evaluates whether each elevator has been dispatched in the current cycle. Each elevator's operation is governed by specific logic based on its active or inactive state, as illustrated by the four operational logics in Figure 3:

a) If an elevator is dispatched and currently ascending with passengers, it will complete the ongoing transport task and then return to the lobby floor to pick up new waiting passengers.

b) If an elevator is dispatched but currently stationary, it will switch from inactive to active and proceed directly to the lobby to serve passengers.

c) If an elevator is not dispatched but currently transporting passengers, it will complete its current task, remain stationary, and await a new task in the next scheduling cycle.

d) If an elevator was not dispatched in both the previous and current scheduling cycles, it will remain inactive.

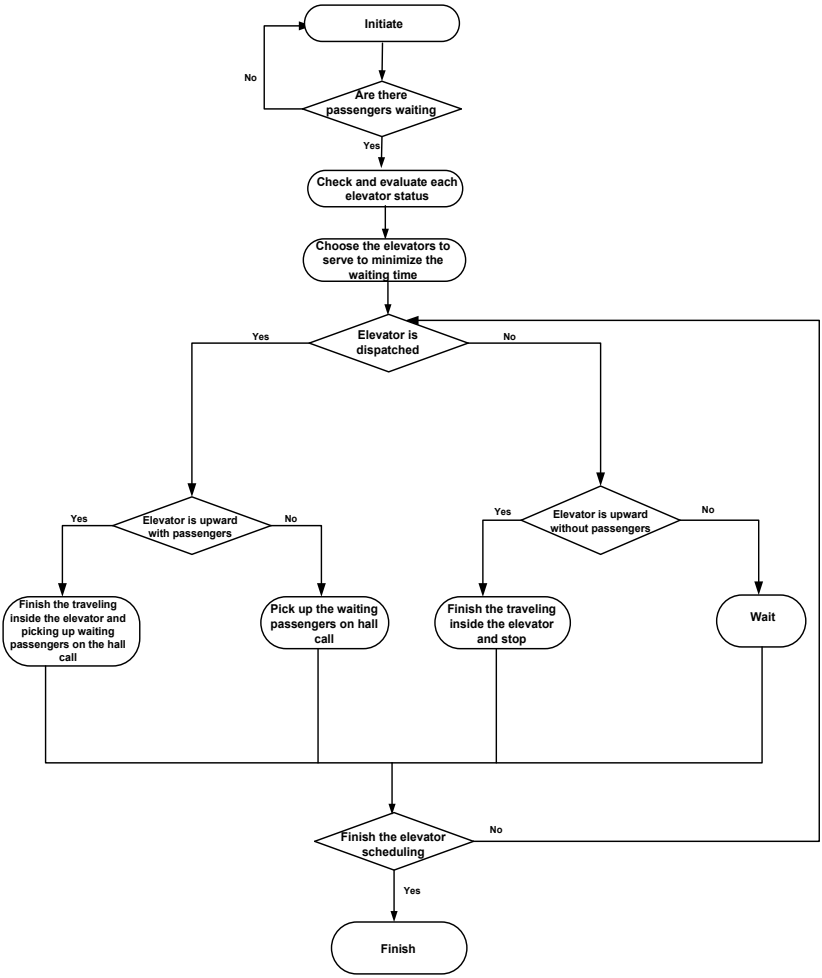


Fig. 3. Flowchart of intelligent elevator control system

Running example.

Suppose we have an intelligent control system with three elevators and eleven floors. Elevator 1 is on floor 4, Elevator 2 is on floor 3, and Elevator 3 is on floor 7. Elevator 1 wants to move up to floor 8, Elevator 2 wants to move down to floor 2, and Elevator 3 to move up to floor 10 (Figure 4).

To minimize waiting times, the intelligent control system allocates Hall Call 6 to Elevator 1, Hall Call 2 to Elevator 2, and Hall Call 9 to Elevator 3.

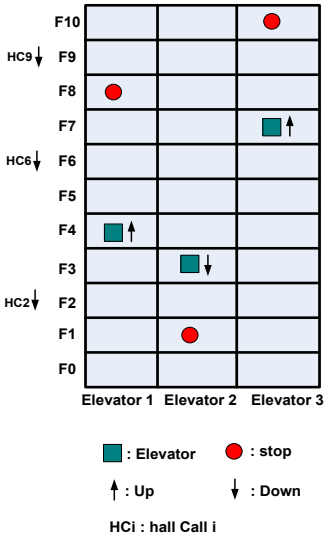


Fig. 4. A simplified example of an intelligent elevator control system composed of 11 floors and 3 elevators

4.2. Intelligent Agent Control in Action. To illustrate how the intelligent agent controls elevator efficiency based on the outlined algorithm, let's consider a scenario in a busy office building. The building has several elevators serving multiple floors, and the intelligent agent is tasked with optimizing elevator operations to minimize wait times and improve overall efficiency.

– Hall Call Handling: When a passenger on a floor requests an elevator by pressing the up or down button, the HALL_CALL procedure is activated. The agent checks the status of each elevator, identifying the one closest to the requested floor and moving in the same direction. If all elevators are idle, the first available one is assigned. The selected elevator is then directed to the requested floor, minimizing travel time and improving efficiency.

– Car Call Handling: When a passenger inside an elevator presses a button to select a destination floor, the CAR_CALL procedure is invoked. The destination floor is added to the elevator's destination queue. If the elevator is idle, it is activated, and its direction is set based on the selected floor relative to its current position. The elevator then moves towards the destination floor, optimizing its path to efficiently serve passengers.

– Move Car Procedure: The MOVE_CAR procedure controls the elevator's movement based on its current state and direction. If the elevator

is moving up, it visits the lowest floor in the up queue or destination queue. If there are no requests in either queue, all requests in the destination queue are cleared, and the elevator's state is set to idle. Similarly, when moving down, the elevator visits the highest floor in the down queue or destination queue, adjusting its direction and state accordingly.

By employing these procedures, the intelligent agent optimizes elevator operations, reducing wait times for passengers, and enhancing overall efficiency in the building's transportation system.

4.2.1. Hall Call Handling. The first procedure, HALL_CALL, handles hall calls from a floor with a given direction. If the direction is up, the floor is inserted into the Up_wQ queue; otherwise, it is inserted into the Down_wQ queue. If the elevator is idle, its state is set to active, and the elevator is moved in the direction of the hall call. The MOVE_CAR procedure is called to move the elevator.

We added a loop to check the distance of each elevator to the requested floor and selected the one with the closest position and in the same direction. If all elevators are idle, we will use the first one found in the loop. Then we assigned the request to the selected elevator and added it to its destination queue. If the selected elevator is idle, we set its moving direction to the direction of the destination queue and start moving the car using the MOVE_CAR procedure (Listing 1) to assign the new request to the elevator with the closest current position and in the same direction.

```
min_distance ← infinity
selected_elevator ← null
for each elevator e in the system
            if (e.state = IDLE)
                selected_elevator ← e
                break
            else if ((d = UP) and (e.current_position ≤ f) and (e.moving_direction = UP)
    and ((f - e.current_position) < min_distance))
                selected_elevator ← e
                min_distance ← f - e.current_position
            else if ((d = DOWN) and (e.current_position ≥ f) and (e.moving_direction =
    DOWN) and ((e.current_position - f) < min_distance))
                 selected_elevator ← e
                 min_distance ← e.current_position - f
        end if
end for
if (selected_elevator = null)
            if (d = UP)
                insert(f, Up_wQ)
            else
                insert(f, Down_wQ)
```

```
            end if
else
      selected_elevator.dest_q.add(f)
      if (selected_elevator.state = IDLE)
                  selected_elevator.state ← ACTIVE
                  selected_elevator.moving_direction ← d
      end if
      selected_elevator.MOVE_CAR()
end if
END
```

Listing 1. Procedure HALL CALL(source floor f, Direction d)

4.2.2. Move Car Procedure. The second procedure, MOVE_CAR, moves the elevator based on the direction it is traveling. If the elevator is moving up, it visits the lowest floor in the Up_wQ or the dest_q (destination queue) in the up direction. If there are no requests in either queue, all requests in the dest_q are removed. If all queues are empty, the elevator state is set to idle. If the elevator is moving down, it visits the highest floor in the Down_wQ or the dest_q in the down direction. If there are no requests in either queue, all requests in the dest_q are removed. If all queues are empty, the elevator state is set to idle. The procedure then sets the moving direction to the opposite direction and calls the VISIT procedure (Listing 2).

```
While (ElevState = ACTIVE) do
      If (MovingDirection = UP) then
                  VISIT (Lowest floor in Up_wQ or dest_q in this direction)
                  If no request in Up_wQ or dest_q in this direction then
                              Remove all requests in dest_q
                              If all queues are empty then
                                          ElevState  ← IDLE
                              else
                                          MovingDirection← DOWN
                                          VISIT (Highest floor in Down_wQ or dest_q)
                              End if
                   End if
      Else // moving direction is down
                  VISIT (Highest floor in Down_wQ or dest_q in this direction)
                  If no request in Down_wQ or dest_q in this direction then
                              Remove all requests in dest_q
                              If all queues are empty then
                                          ElevState  ← IDLE
                              Else
                                          MovingDirection← UP
```

```
                                    VISIT (Lowest floor in Up_wQ or dest_q in
                            destination)
                            End if
                    End if
        End if
    End while
    End
```

Listing 2. Procedure MOVE_CAR()

4.2.3. Car Call Handling. The third procedure, CAR_CALL, handles car calls to a destination floor. The floor is inserted into the dest_q queue. If the elevator is idle, its state is set to active, and the elevator is moved in the direction of the car call. The MOVE_CAR procedure is called to move the elevator (Listing 3).

```
Insert (f, dest_q)
If (ElevState = IDLE) then
    ElevState ← ACTIVE
    //If f is higher than current car position then
    If (f > currentElvPos) then
            MovingDirection ← UP
    else
            MovingDirection ← DOWN
    End if
    MOVE_CAR()
End if
  End
```

Listing 3. CAR_CALL(destination floor f)

5. Simulation. In this study, we conduct a comparative analysis between two distinct scenarios: our prior research involving a conventional elevator system [20] and a novel intelligent control elevator system designed to optimize passenger service based on proximity. The conventional elevator system represents a baseline model characterized by traditional operational algorithms and rules governing elevator movements and passenger interactions. In contrast, the intelligent control elevator system integrates advanced algorithms and decision-making processes to dynamically adjust elevator operations based on passenger location and demand, prioritizing minimizing average waiting time. The comparison between these two systems aims to assess the efficacy, performance improvements, and potential benefits of adopting intelligent control strategies in elevator systems, particularly in scenarios where passenger proximity influences service allocation and response times.

Figure 5 presents the response times of an intelligent control system compared to a conventional system over a 3600-second period. From 0 to 600 seconds, the intelligent system's response time increases from 0.01 to 0.12 seconds, while the conventional system rises from 0.015 to 0.20 seconds, with the intelligent system maintaining a consistent advantage. Between 600 and 1500 seconds, the intelligent system peaks at 0.20 seconds, whereas the conventional system reaches 0.40 seconds, highlighting a significant performance gap. In the later stages (1600 to 3600 seconds), the intelligent control system stabilizes, fluctuating between 0.05 and 0.29 seconds, while the conventional system plateaus between 0.10 and 0.49 seconds, indicating less adaptability. Overall, the intelligent control system consistently shows lower response times across all intervals, with a maximum of 0.29 seconds compared to 0.49 seconds for the conventional system, suggesting greater efficiency in managing increasing passenger demand.

Overall, the intelligent control system is significantly more effective than the conventional system in handling varying passenger arrival rates. Its ability to maintain lower response times, even as demand increases, highlights its potential for enhancing operational efficiency in vertical transportation systems.

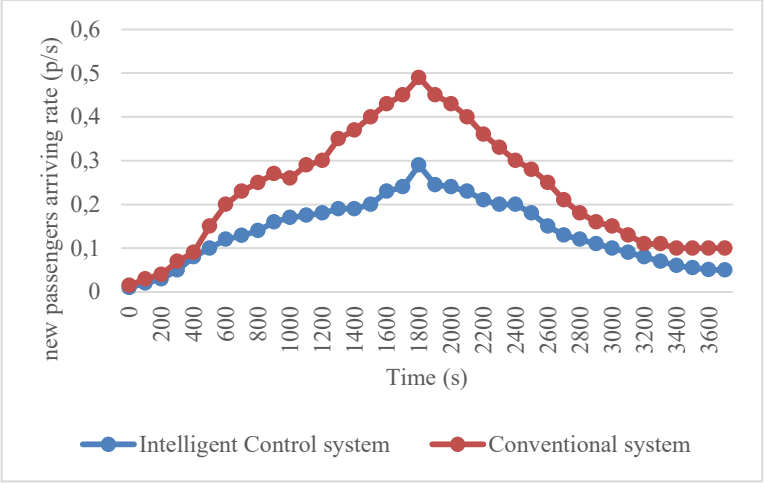


Fig. 5. Curves of up-peak new passengers arriving rate

Figure 6 presents a comparison of the performance of an Intelligent Control System and a Conventional System over time, focusing on their ability to manage active cars simultaneously. The data indicates that the

Intelligent Control System consistently manages fewer active cars compared to the Conventional System, suggesting that it is more proficient in optimizing the utilization of operational cars. Additionally, the Intelligent Control System appears to sustain its performance over a longer duration, indicating greater resilience to fluctuations in demand and the ability to maintain a high level of service for extended periods, contrasting with the performance of the Conventional System.

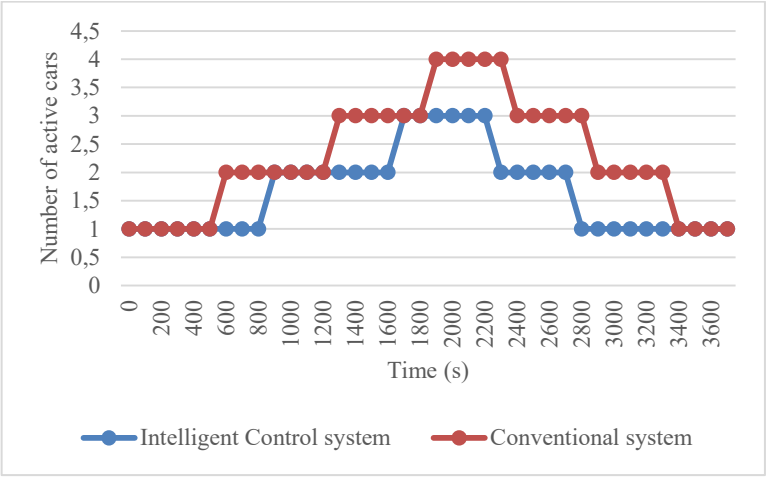


Fig. 6. Number of Elevators Dispatched During Up-Peak Periods

The simulation is set to run for a fixed period of one hour, allowing for the collection of sufficient data. During this time, the passenger arrival rate varies across different floors to simulate both peak and off-peak periods. For each control method, calls are assigned to elevators, and the elevators' responses are simulated. The key metric for evaluating performance is the average waiting time, which measures the time passengers wait for an elevator under both control scenarios.

Figure 7 presents the average waiting time in 21 simulations revealing that the intelligent control system significantly exceeds the conventional system, with an average waiting time of approximately 17.14 seconds compared to 29.14 seconds for the conventional system. The intelligent system shows constant low variability, with waiting times of 14 to 20 seconds, while the conventional system shows higher fluctuations, reaching a maximum of 33 seconds. Significant differences are noted in simulations 1, 3 and 11, where the efficiency of the intelligent system is particularly pronounced. Overall, the implementation of intelligent control

mechanisms can improve the efficiency of operations and improve passenger experiences in vertical transportation systems, making them a more suitable choice for managing dynamic passenger requirements.

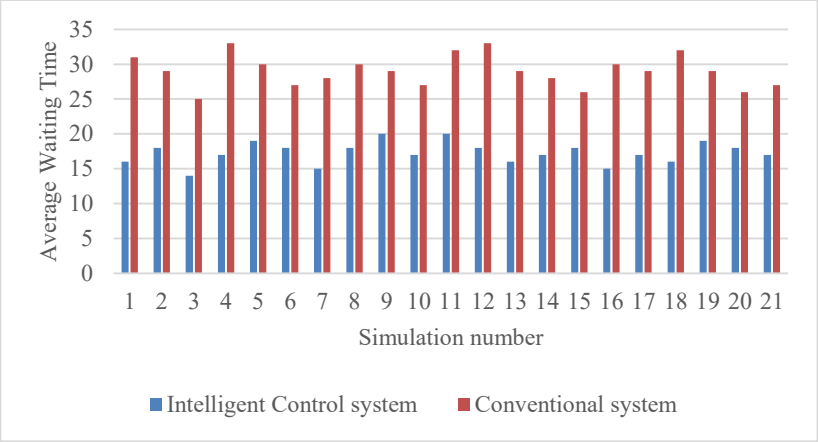


Fig. 7. Comparison of Average Waiting Times: Intelligent Control System vs. Conventional System

6. Conclusion. In conclusion, the proposed intelligent agent-controlled elevator system effectively reduces passenger wait times and optimizes traffic flow in real time. The study contributes a novel approach to elevator control, demonstrating significant improvements over traditional systems. Simulation-based evaluations validate the effectiveness of the proposed algorithms and demonstrate significant improvements in system performance metrics. However, the proposed intelligent control system faces limitations in integrating with existing elevator infrastructures, particularly in older buildings, where retrofitting could be costly and complex. Scalability is also a concern, as the system's effectiveness may decrease in larger buildings, leading to potential delays and data management challenges. Additionally, the system's reliance on accurate, real-time data could be problematic in environments with outdated sensors and communication networks.

Future research may focus on the assessment of implementations and scalability of intelligent agent-controlled elevator systems in the real world, the development of machine learning for elevator operations, the integration of IoT for real-time data optimization, and the collaboration with industrial partners for practical implementation, the improvement of urban mobility and construction management.

References

1. Shen L.J., Lukose J., Young L.C. Predictive maintenance on an elevator system using machine learning. Journal of Applied Technology and Innovation. 2021. vol. 5(1). pp. 75–82.
2. Hsu C.Y., Qiao Y., Wang C., Chen S.T. Machine learning modeling for failure detection of elevator doors by three-dimensional video monitoring. IEEE Access, 2020. vol. 8. pp. 211595–211609.
3. Arrieta A., Ayerdi J., Illarramendi M., Agirre A., Sagardui G., Arratibel M. Using machine learning to build test oracles: an industrial case study on elevators dispatching algorithms. IEEE/ACM International Conference on Automation of Software Test (AST). 2021. pp. 30–39.
4. Syed F.I., Alshamsi M., Dahaghi A.K., Neghabhan S. Artificial lift system optimization using machine learning applications. Petroleum. 2022. vol. 8(2). pp. 219–226.
5. Gichane M.M., Byiringiro J.B., Chesang A.K., Nyaga P.M., Langat R.K., Smajic H., Kiiru C.W. Digital triplet approach for real-time monitoring and control of an elevator security system. Designs. 2020. vol. 4(2). DOI: 10.3390/designs4020009.
6. Zubair M.U., Zhang X. Explicit data-driven prediction model of annual energy consumed by elevators in residential buildings. Journal of Building Engineering. 2020. vol. 31. DOI: 10.1016/j.jobe.2020.101278.
7. Yao W., Jagota V., Kumar R., Ather D., Jain V., Quraishi S.J., Osei-Owusu J. Study and application of an elevator failure monitoring system based on the internet of things technology. Scientific Programming, 2022. DOI: 10.1155/2022/2517077.
8. Mao J., Chen L., Cheng H., Wang C. Elevator fault diagnosis and maintenance method based on Internet of Things. In International Conference on Mechatronics and Intelligent Control (ICMIC 2023). SPIE. 2023. vol. 12793. pp. 18–23.
9. Sharma A., Chatterjee A., Thakur P.K., Jha S., Sriharipriya K. IoT Based Automated Elevator Emergency Alert System Using Android Mobile Application. IEEE International Conference on Data Science and Information System (ICDSIS). 2022. pp. 1–6.
10. Wu Y., Yang J. Directional optimization of elevator scheduling algorithms in complex traffic patterns. Applied Soft Computing. 2024. vol. 158. DOI: 10.1016/j.asoc.2024.111567.
11. Wang H., Zhang M., Zhang R., Liu L. Research on predictive sliding mode control strategy for horizontal vibration of ultra-high-speed elevator car system based on adaptive fuzzy. Measurement and Control. 2021. vol. 54(3-4). pp. 360–373.
12. Le S., Lei Q., Wei X., Zhong J., Wang Y., Zhou J., Wang W. Smart Elevator Control System Based on Human Hand Gesture Recognition. IEEE 6th International Conference on Computer and Communications (ICCC). 2020. pp. 1378–1385.
13. Lu L., Jiang C., Hu G., Liu J., Yang B. Flexible noncontact sensing for human–machine interaction. Advanced Materials. 2021. vol. 33(16). DOI: 10.1002/adma.202100218.
14. Huang J., Wang H., Li J.A., Zhang S., Li H., Ma Z., Xin M., Yan K., Cheng W., He D., Wang X., Shi Y., Pan L. High-performance flexible capacitive proximity and pressure sensors with spiral electrodes for continuous human–machine interaction. ACS Materials Letters. 2022. vol. 4(11). pp. 2261–2272. DOI: 10.1021/acsmaterialslett.2c00860.
15. Koehler J., Schuster K. Elevator Control as a Planning Problem. In AIPS. 2000. pp. 331–338.
16. Nivasanon C., Srikun I., Aungkulanon P. Aggregate production planning: A case study of installation elevator company. Proceedings of the International Conference on Industrial Engineering and Operations Management. 2021. pp. 5357–5365.

17. Bagenda D.N., Basjaruddin N.C., Darwati E., Rakhman E. Development of an elevator simulator to support problem-based electric motor control practicum for vocational high school student. INVOTEK: Jurnal Inovasi Vokasional dan Teknologi. 2021. vol. 21(2). pp. 139–148.
18. Cortes P., Munuzuri J., Vazquez-Ledesma A., Onieva L. Double deck elevator group control systems using evolutionary algorithms: Interfloor and lunchpeak traffic analysis. Computers & Industrial Engineering. 2021. vol. 155. DOI: 10.1016/j.cie.2021.107190.
19. Tartan E.O., Ciflikli C. Esra (elevator simulation, research & analysis): an open-source software tool for elevator traffic simulation, research, and analysis. Journal of Simulation. 2024. pp. 1–18.
20. Gharbi A., Exploring Heuristic and Optimization Approaches for Elevator Group Control Systems. Applied Sciences. 2024. vol. 14(3). DOI: 10.3390/app14030995.

Gharbi Atef — Ph.D., Dr.Sci., Associate Professor, Department member, Information systems department, faculty of computing and information technology, Northern Border University; laboratory member, Lisi laboratory, National Institute of Applied Sciences and Technology (INSAT), University of Carthage. Research interests: specification of model, verification of properties related to functional safety, implementation of software solutions to ensure functional safety, multi-agent system. The number of publications — 45. Atef.gharbi@nbu.edu.sa; 37, Al-Qadisyiah, Rafha, Northern Borders, Saudi Arabia; office phone: +(916)545635491.

Ayari Mohamed — Ph.D., Dr.Sci., Associate Professor, Department member, Information systems department, faculty of computing and information technology, Northern Border University; Member in syscom laboratory, Syscom laboratory, National Engineering School of Tunis, University of Tunis El-Manar. Research interests: electromagnetic (EM) fields, numerical EM methods, and computer-aided design of microwave circuits and antennas. The number of publications — 56. mohamed.ayari@nbu.edu.sa; 37, Al-Qadisyiah, Rafha, Northern Borders, Saudi Arabia; office phone: +(916)545633598.

El Touati Yamen — Ph.D., Dr.Sci., Associate Professor, Department member, Computer science department, faculty of computing and information technology, Northern Border University. Research interests: petri Net, formal specification and verification, artificial intelligence, machine learning. The number of publications — 37. yamen.touati@nbu.edu.sa; 37, Al-Qadisyiah, Rafha, Northern Borders, Saudi Arabia; office phone: +(916)542476840.

УДК 006.72 **DOI** 10.15622/ia.24.1.2

А. ГАРБИ, М. АЙЯРИ, Я. ЭЛЬ ТУАТИ

ИНТЕЛЛЕКТУАЛЬНАЯ СИСТЕМА ЛИФТОВ, УПРАВЛЯЕМАЯ АГЕНТАМИ: АЛГОРИТМ И ОПТИМИЗАЦИЯ ЭФФЕКТИВНОСТИ

Гарби А., Айяри М., Эль Туати Я. **Интеллектуальная система лифтов, управляемая агентами: алгоритм и оптимизация эффективности.**

Аннотация. В исследовании представлена инновационная интеллектуальная система лифтов, управляемая агентами, специально разработанная для сокращения времени ожидания пассажиров и повышения эффективности высотных зданий. Используя классическую модель планирования с одним агентом, мы разработали уникальную стратегию обработки вызовов из коридоров и автомобилей, и в сочетании с этой стратегией мы значительно улучшили общую производительность лифтовой системы. Наши интеллектуальные методы управления подробно сравниваются с традиционными системами лифтов, при этом оцениваются три важных показателя эффективности: время отклика, способность системы одновременно обрабатывать несколько активных кабин лифта и среднее время ожидания пассажира. Результаты полного моделирования показывают, что интеллектуальная модель на основе агентов неизменно превосходит обычные системы лифтов по всем измеряемым критериям. Интеллектуальные системы управления значительно сократили время отклика и улучшили одновременное управление лифтами и время ожидания пассажиров, особенно во время большой загруженности. Эти усовершенствования не только повысили эффективность потока трафика, но и в значительной степени способствовали удовлетворенности пассажиров и обеспечили более плавное и надежное перемещение внутри здания. Кроме того, повышенная эффективность наших систем соответствует целям управления энергопотреблением зданий, поскольку она сводит к минимуму ненужные движения и время простоя. Результаты демонстрируют способность системы соответствовать требованиям динамичной среды с высокой загруженностью и знаменуют собой значительный шаг вперед в интеллектуальном управлении инфраструктурой.

Ключевые слова: интеллектуальное управление агентом, оптимизация системы лифта, время ожидания пассажира.

Литература

1. Shen L.J., Lukose J., Young L.C. Predictive maintenance on an elevator system using machine learning. Journal of Applied Technology and Innovation. 2021. vol. 5(1). pp. 75–82.
2. Hsu C.Y., Qiao Y., Wang C., Chen S.T. Machine learning modeling for failure detection of elevator doors by three-dimensional video monitoring. IEEE Access, 2020. vol. 8. pp. 211595–211609.
3. Arrieta A., Ayerdi J., Illarramendi M., Agirre A., Sagardui G., Arratibel M. Using machine learning to build test oracles: an industrial case study on elevators dispatching algorithms. IEEE/ACM International Conference on Automation of Software Test (AST). 2021. pp. 30–39.
4. Syed F.I., Alshamsi M., Dahaghi A.K., Neghabhan S. Artificial lift system optimization using machine learning applications. Petroleum. 2022. vol. 8(2). pp. 219–226.

5. Gichane M.M., Byiringiro J.B., Chesang A.K., Nyaga P.M., Langat R.K., Smajic H., Kiiru C.W. Digital triplet approach for real-time monitoring and control of an elevator security system. Designs. 2020. vol. 4(2). DOI: 10.3390/designs4020009.
6. Zubair M.U., Zhang X. Explicit data-driven prediction model of annual energy consumed by elevators in residential buildings. Journal of Building Engineering. 2020. vol. 31. DOI: 10.1016/j.jobe.2020.101278.
7. Yao W., Jagota V., Kumar R., Ather D., Jain V., Quraishi S.J., Osei-Owusu J. Study and application of an elevator failure monitoring system based on the internet of things technology. Scientific Programming, 2022. DOI: 10.1155/2022/2517077.
8. Mao J., Chen L., Cheng H., Wang C. Elevator fault diagnosis and maintenance method based on Internet of Things. In International Conference on Mechatronics and Intelligent Control (ICMIC 2023). SPIE. 2023. vol. 12793. pp. 18–23.
9. Sharma A., Chatterjee A., Thakur P.K., Jha S., Sriharipriya K. IoT Based Automated Elevator Emergency Alert System Using Android Mobile Application. IEEE International Conference on Data Science and Information System (ICDSIS). 2022. pp. 1–6.
10. Wu Y., Yang J. Directional optimization of elevator scheduling algorithms in complex traffic patterns. Applied Soft Computing. 2024. vol. 158. DOI: 10.1016/j.asoc.2024.111567.
11. Wang H., Zhang M., Zhang R., Liu L. Research on predictive sliding mode control strategy for horizontal vibration of ultra-high-speed elevator car system based on adaptive fuzzy. Measurement and Control. 2021. vol. 54(3-4). pp. 360–373.
12. Le S., Lei Q., Wei X., Zhong J., Wang Y., Zhou J., Wang W. Smart Elevator Control System Based on Human Hand Gesture Recognition. IEEE 6th International Conference on Computer and Communications (ICCC). 2020. pp. 1378–1385.
13. Lu L., Jiang C., Hu G., Liu J., Yang B. Flexible noncontact sensing for human–machine interaction. Advanced Materials. 2021. vol. 33(16). DOI: 10.1002/adma.202100218.
14. Huang J., Wang H., Li J.A., Zhang S., Li H., Ma Z., Xin M., Yan K., Cheng W., He D., Wang X., Shi Y., Pan L. High-performance flexible capacitive proximity and pressure sensors with spiral electrodes for continuous human–machine interaction. ACS Materials Letters. 2022. vol. 4(11). pp. 2261–2272. DOI: 10.1021/acsmaterialslett.2c00860.
15. Koehler J., Schuster K. Elevator Control as a Planning Problem. In AIPS. 2000. pp. 331–338.
16. Nivasanon C., Srikun I., Aungkulanon P. Aggregate production planning: A case study of installation elevator company. Proceedings of the International Conference on Industrial Engineering and Operations Management. 2021. pp. 5357–5365.
17. Bagenda D.N., Basjaruddin N.C., Darwati E., Rakhman E. Development of an elevator simulator to support problem-based electric motor control practicum for vocational high school student. INVOTEK: Jurnal Inovasi Vokasional dan Teknologi. 2021. vol. 21(2). pp. 139–148.
18. Cortes P., Munuzuri J., Vazquez-Ledesma A., Onieva L. Double deck elevator group control systems using evolutionary algorithms: Interfloor and lunchpeak traffic analysis. Computers & Industrial Engineering. 2021. vol. 155. DOI: 10.1016/j.cie.2021.107190.
19. Tartan E.O., Ciflikli C. Esra (elevator simulation, research & analysis): an open-source software tool for elevator traffic simulation, research, and analysis. Journal of Simulation. 2024. pp. 1–18.
20. Gharbi A., Exploring Heuristic and Optimization Approaches for Elevator Group Control Systems. Applied Sciences. 2024. vol. 14(3). DOI: 10.3390/app14030995.

Гарби Атеф — Ph.D., Dr.Sci., доцент, сотрудник отдела, отдел информационных систем факультета вычислительной техники и информационных технологий, Университет Северной границы; сотрудник лаборатории, лаборатория lisi, Национальный институт прикладных наук и технологий (INSAT) Карфагенского университета. Область научных интересов: спецификация модели, проверка свойств, связанных с функциональной безопасностью, внедрение программных решений для обеспечения функциональной безопасности, многоагентная система. Число научных публикаций — 45. Atef.gharbi@nbu.edu.sa; Аль-Кадисия, 37, Рафха, Северные границы, Саудовская Аравия; р.т.: +(916)545635491.

Айяри Мохаммед — Ph.D., Dr.Sci., доцент, сотрудник отдела, отдел информационных систем факультета вычислительной техники и информационных технологий, Университет Северной границы; сотрудник лаборатории, лаборатория syscom, Национальная инженерная школа Туниса, Тунисский университет Эль-Манар. Область научных интересов: электромагнитные поля, численные электромагнитные методы и автоматизированное проектирование микроволновых схем и антенн. Число научных публикаций — 56. mohamed.ayari@nbu.edu.sa; Аль-Кадисия, 37, Рафха, Северные границы, Саудовская Аравия; р.т.: +(916)545633598.

Эль Туати Ямен — Ph.D., Dr.Sci., доцент, сотрудник отдела, кафедра компьютерных наук факультета вычислительной техники и информационных технологий, Университет Северной границы. Область научных интересов: сеть Петри, формальная спецификация и верификация, искусственный интеллект, машинное обучение. Число научных публикаций — 37. yamen.touati@nbu.edu.sa; Аль-Кадисия, 37, Рафха, Северные границы, Саудовская Аравия; р.т.: +(916)542476840.

DOI 10.15622/ia.24.1.3

A. HAMMOUD, A. ISKANDAR, B. KOVÁCS

DYNAMIC FORAGING IN SWARM ROBOTICS: A HYBRID APPROACH WITH MODULAR DESIGN AND DEEP REINFORCEMENT LEARNING INTELLIGENCE

Hammoud A., Iskandar A., Kovács B. **Dynamic Foraging in Swarm Robotics: A Hybrid Approach with Modular Design and Deep Reinforcement Learning Intelligence.**

Abstract. This paper proposes a hybrid approach that combines intelligent algorithms and modular design to solve a foraging problem within the context of swarm robotics. Deep reinforcement learning (RL) and particle swarm optimization (PSO) are deployed in the proposed modular architecture. They are utilized to search for many resources that vary in size and exhibit a dynamic nature with unpredictable movements. Additionally, they transport the collected resources to the nest. The swarm comprises 8 E-Puck mobile robots, each equipped with light sensors. The proposed system is built on a 3D environment using the Webots simulator. Through a modular approach, we address complex foraging challenges characterized by a non-static environment and objectives. This architecture enhances manageability, reduces computational demands, and facilitates debugging processes. Our simulations reveal that the RL-based model outperforms PSO in terms of task completion time, efficiency in collecting resources, and adaptability to dynamic environments, including moving targets. Notably, robots equipped with RL demonstrate enhanced individual learning and decision-making abilities, enabling a level of autonomy that fosters collective swarm intelligence. In PSO, the individual behavior of the robots is more heavily influenced by the collective knowledge of the swarm. The findings highlight the effectiveness of a modular design and deep RL for advancing autonomous robotic systems in complex and unpredictable environments.

Keywords: swarm robotics, foraging task, modular design, reinforcement learning, particle swarm optimization.

1. Introduction. Swarm robotics realizes the principle of decentralized control among multiple robots, drawing inspiration from natural phenomena where collective behaviors emerge from simple individual actions, such as in ant colonies or bird flocks. This field leverages the scalability, robustness, and flexibility inherent to swarms, which represent the main features of swarm robots' systems. They allow them to solve complex tasks more efficiently than individuals do [1]. To achieve the goal required of robots in the swarm concept, the robots should act with a high level of autonomy and have local knowledge about their environment. Namely, a microscopic view. While coordination and communication among robots reflect the concept of collective behavior, which arises from the aggregation of simple rules followed by individual robots, leading to emergent behaviors that require no centralized oversight, this represents the macroscopic level [2]. For example, aggregation behavior occurs when robots come together to form clusters or groups while foraging behavior emerges during the coordinated effort of the robots to search, identify, collect, and transport resources to a destination called the nest. Other

examples include collective searching and exploration, pattern formation, and others [3]. In general, collective behavior is obtained by various algorithms that enable individual robots to make decisions based on local perception and the cumulative contribution of all robots in achieving the desired outcomes. This involves communication and collaboration among the robots as well. Most of the methods used in swarm engineering can be divided into two basic categories: Behavioral design methods like the probabilistic finite state machine (PFSM) method, where the robots' behaviors are broken down into a finite number of states and transitions between these states. These transitions are triggered by events or conditions [4]. On the other hand, automatic design methods such as RL and PSO stand out. RL fosters the robot to learn the required behavior at a microscopic level through interaction with the environment. By deploying RL, the robots learn and adapt their actions through a continuous process, gradually refining their policies. Conversely, PSO is a mathematical model that reflects social behavioral patterns and guides individuals within the swarm toward optimal solutions [5]. Each method has its own set of advantages. RL is renowned for its adaptability to the environment's changes, enhancing autonomy and decision-making abilities, while PSO is lauded for its simplicity and straightforward implementation. Nonetheless, RL's computational cost and PSO's potential ineffectiveness in dynamic settings pose significant challenges. Many RL algorithms can potentially be used for generating collective behavior, like Deep-Q networks [6], SARSA, and other value-based methods [7]. In addition to policy-based methods including DDPG [8], PPO, and others. Proximal Policy Optimization (PPO) is particularly favored due to its efficient balance between sample efficiency and implementation simplicity, making it well-suited for the dynamic and uncertain scenarios often encountered in swarm robotics [9].

The swarm architecture and design process, including the collective behavior, depend on several factors, such as the nature of the task required by a swarm, the environment, whether it's dynamic or static, and others. This process emphasizes the importance of understanding both the microscopic (individual agents and interactions) and macroscopic (collective behavior and task achievement) in creating effective swarm systems. For example, one of the challenges in creating an effective foraging swarm is adapting methods for searching and navigation in a continuous environment. Another challenge is ensuring the scalability of learned behaviors for different swarm sizes and configurations. Addressing the challenges requires sophisticated and adaptive algorithms that can effectively deal with the dynamic nature of swarm challenges.

2. Related works. The integration of the RL approach in swarm robotics offers significant benefits such as adaptability. This allows robots to adjust their behavior to dynamic environments through trial and error. It enhances autonomy by allowing robots to make self-decisions until they learn the optimal strategies. Moreover, RL adapts with scalability and flexibility, making the swarm applicable across various tasks [10, 11]. Particularly, RL in the foraging task optimizes resource collection and exploration, contributing to more efficient environmental mapping and more robust mission execution. However, some challenges persist, including managing the complexity of dynamic environments, ensuring the scalability of learned behaviors, overcoming communication limitations, balancing exploration with exploitation, and addressing energy and computational constraints [12]. Most recent studies have been conducted on strategies for deploying RL on generating foraging collective behavior. These challenges are addressed in the Table 1 with proposed solutions encompassing a spectrum of strategies.

Table 1. Challenges of deploying RL in foraging swarms.

Challenge	**Description**
Complexity of Dynamic Environments	Adapting to unpredictable changes, including moving targets and obstacles.
Scalability of Learning	Applying learned behaviors to swarms of varying sizes and compositions.
Communication Limitations	Coordinating actions without overwhelming the network or centralized control.
Balancing Exploration and Exploitation	Optimizing foraging efficiency by finding the right balance.
Energy and Computational Constraints	Managing energy consumption and computational demands for efficient operation.

Many researchers address the "complexity of dynamic environments" challenge in swarm robotics through various approaches. One study developed a macroscopic foraging behavior using deep RL, combined with microscopic behaviors controlled by fuzzy logic for obstacle avoidance and low-level navigation. This hybrid approach simplifies the RL search space, and achieves robust and scalable foraging behavior in swarms, even in scenarios that are not encountered during the training phase [13]. Another paper optimized the foraging strategy for active particles using Multi-Agent Reinforcement Learning (MARL). It addressed the problem by enabling the active particles to locate and efficiently forage from randomly occurring food sources. The paper demonstrated that individual optimization could lead to the emergence of collective behaviors that are beneficial to the swarm's overall foraging efficiency [14]. Additionally, a distinct approach utilizes deep RL with

curriculum learning to sequentially tackle navigation tasks, enhancing learning efficiency and adaptability to the environment. These methods collectively demonstrate the potential of advanced computational strategies to navigate and adapt to the uncertainty of dynamic environments [15].

Studies have addressed the scalability challenge by proposing various strategies. In [16], the authors focused on improving the system performance without increasing the complexity of individual robots or the need for heavy communication among them. They proposed leveraging simple, decentralized interactions among robots to achieve complex tasks. Also, [17] proposed a self-organizing task allocation model that enables swarm robots to dynamically distribute complex foraging tasks. This approach utilizes a response threshold model, which ensures efficient task allocation without the need for centralized control or extensive communication. It guarantees robust performance under varying conditions by effectively managing task distribution through local interactions. This makes it a scalable solution for complex swarm robotics applications. In addressing the challenges of communication limitations within swarm robotics, researchers have highlighted the development of innovative solutions to enhance the robustness and efficiency of systems in constrained communication environments. These strategies include the use of federated learning and deep RL to improve generalization and performance [8], alongside the adoption of biologically inspired communication mechanisms that enable decentralized operations [18]. These approaches significantly enhance the adaptability of swarm systems to dynamic conditions without the need for complex individual robot capabilities. For balancing exploration and exploitation, researchers have explored methods like Mutual-Information Upper Confidence Bound (MI-UCB) [19] and virtual pheromone mechanisms [20]. MI-UCB enhances drone coordination through decentralized Monte Carlo Tree Search. It improves surveillance performance by balancing information gain with reward maximization. Meanwhile, the use of virtual pheromones allows minimalist agents to effectively switch between exploring new resources and exploiting known ones. Finally, [21] conducted a study on the challenge of energy and computational constraints. It proposed a mobile edge computing solution integrated with a mobility-aware deep RL model for computation considerations. This approach reduces the computational cost of the robots, allowing for energy-efficient operation while meeting computation latency requirements. Compared to approaches that individually address challenges using deep RL with fuzzy logic, MARL, MI-UCB, curriculum learning, and others, our research emphasizes the advantages of a modular approach combined with the adaptability and efficiency of the PPO in dynamic environments. It enhances manageability,

adaptability, and efficiency in non-static environments. Specifically, it overcomes the observed limitations by providing an adaptive system that simplifies the debugging and computational demands, showcasing superior performance in terms of task completion and resource collection. Our approach aims to achieve individual learning and decision-making capabilities within a collective swarm intelligence framework.

3. System Description

3.1. Environment setup and system structure. The swarm system was implemented in a 3D robot simulator called Webots [22] where the E-Puck mobile robot was selected to form the swarm. A foraging collective behavior was produced to search for small and big boxes through the environment and then transport them to the nest. This behavior was evaluated through the environment as shown in Figure 1. The dimensions of the workspace were defined as $3 \times 3m^2$, forming a square area surrounded by four walls. The parameters of E-Puck robots were set as follows: linear velocity $V = [0, 0.25]m/s$, angular velocity $W = [-3.14, 3.14]rad/s$, and light sensors' readings LS_0, LS_7 corresponded to the light intensity $[0, 4095]$.

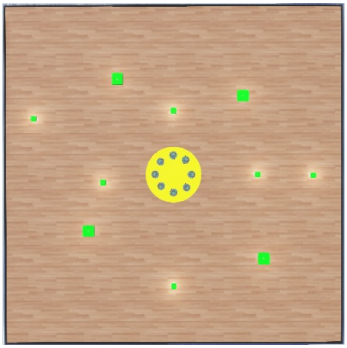


Fig. 1. Foraging environment

3.2. Modular Design. The system's modular design segregates the foraging task into discrete, manageable components, as shown in Figure 2:

– **Searching**: robots use light sensors to locate boxes, which emit light at varying intensities using PPO or PSO.

– **Gripping**: once a box is located, robots catch the box.

– **Waiting**: this state is for big boxes which require cooperative effort.

– **Transporting**: robots navigate back to the nest using PPO.

– **Release**: upon reaching the nest, robots release the box.

– **Return**: robots return to the searching phase, creating a continuous operational loop.

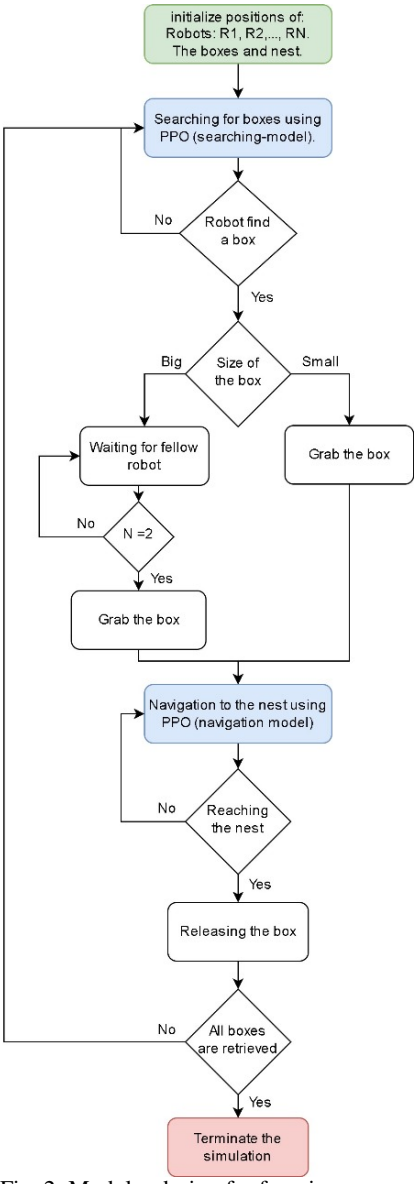


Fig. 2. Modular design for foraging swarm

In Figure 2, at the beginning, the system initializes the positions of all the entities involved: the robots, the foraging boxes (targets), and the nest. After initialization, the robots start the search phase. This first phase uses a PPO model or PSO algorithm specifically tuned for the search task. The algorithm guides the robots towards the targets, based on inputs from light sensors. These sensors assume that the boxes emit light, making them detectable. After finding a box, a decision determines the next steps based on the size of the box. If the box is small, a single robot can retrieve it. If the box is big, the robot waits until another robot arrives to help, ensuring cooperative transportation. This reflects a real-world situation where tasks may require different levels of effort and collaboration. After catching the box, the robots use another PPO model for navigation to return to the nest. This model processes the distance and angle between the robot and the nest to optimize the path. After successful delivery, the robot checks if any boxes are still uncollected. If so, it reverts to the search phase, creating a cycle of the foraging process. Once all targets have been retrieved, the simulation ends.

3.3. PPO architecture. It is used in the searching and transporting phases, chosen for its stability and robust performance in environments with high uncertainty. PPO operates via a policy gradient method that maximizes an objective function by using a clipped surrogate objective to keep updates stable. The main architecture of the PPO neural networks, consisting of an actor and a critic with fully connected layers, are shown in Figure 3.

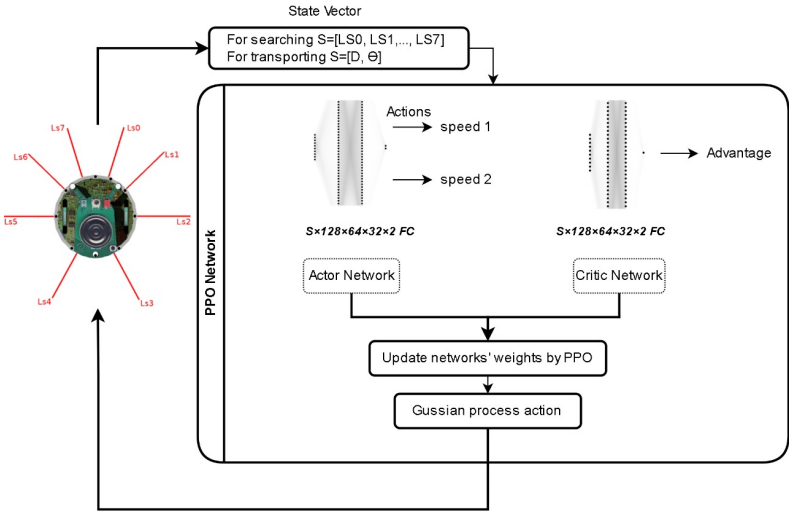


Fig. 3. PPO architecture

The problem is formulated as a MDP represented by the tuple (S, A, T, R, γ). The state space S comprised of light sensor readings during searching phase in addition to the distance between each robot and the nest D, and the angle between the robot and the nest θ in the transport phase. The action space A included the two velocities of the left and right motors. The transition function T describes the dynamics of the system, while the reward function R provides feedback based on the achieved objectives.

3.3.1. PPO Searching. The PPO's actor-critic networks receive input from the light sensors and outputs wheel velocities, adjusting the robot's trajectory towards the light source. A shaping reward is provided based on the change in light intensity, encouraging the robot to move towards brighter areas, i.e., closer to the boxes. The reward function $R(t)$ at time t is given as in Equation 1, 2:

$$R(t) = \frac{(LS_0^{(t)} - LS_0^{(t-1)}) + (LS_7^{(t)} - LS_7^{(t-1)})}{2} + r_{\text{box}}(t), \tag{1}$$

$$r_{\text{box}}(t) = \begin{cases} 1.1 & \text{if } LS_0^{(t)} > \text{FindThreshold}_{\text{searching}} \\ 1.1 & \text{if } LS_7^{(t)} > \text{FindThreshold}_{\text{searching}} \\ 0 & \text{otherwise} \end{cases}, \tag{2}$$

where $LS_0^{(t)}$, $LS_7^{(t)}$ – the normalized current readings of light sensors 0 and 7, respectively, at time t,
$LS_0^{(t-1)}$, $LS_7^{(t-1)}$ – the previous normalized readings of light sensors 0 and 7, respectively, at time $t-1$,
$\text{FindThreshold}_{\text{searching}}$ – the threshold value of the light sensor where the box is found. The normalized readings sensors are more than 0.85 that means the robot reaches the boundray of the box. It is defined experimentally,
$r_{\text{box}}(t)$ – the additional reward when the robot finds the box. The common approach to choose values of rewards like1.1 are defined experimentally to fit the environment.

3.3.2. PPO in Transporting. For transporting phase, the inputs of the PPO network are the robot's current distance and angle relative to the nest, with outputs modifying the wheel velocities to navigate the nest effectively. Rewards are sparse for successful delivery as in Equation 3, and the shaping method for leveraging the experience each time step to speed up the learning

process by considering the angle and the distance to know if the robot is in the direction of the nest or not as in Equation 4.

– **Reward in case of collaboration is not required (one robot is enough to transport a small box to the nest)**:

$$r_{\text{nest}} = \begin{cases} 0.1 & \text{if } d_{\text{current}} < \text{FindThreshold} \\ 0 & \text{otherwise} \end{cases}, \tag{3}$$

$$\text{reward} = (d_{\text{prev}} - d_{\text{current}}) + r_{\text{nest}} + \frac{\cos(\alpha_{\text{current}})}{1000}, \tag{4}$$

where r_{nest} – the obtained reward when robot reaches the nest.
FindThreshold – the Threshold to consider that the robot inside the nest. When the value of Threshold is less than 0.2. The nest circle has a radius 0.2 so when the distance between the robot and the center of the nest less than 0.2, the robot is in the nest,
d_{current} – the normalized distance between the robot and the nest at time t with $[0,3]$,
d_{prev} – the normalized distance between the robot and the nest at time t-1 with $[0,3]$,
α_{current} – the angle between the robot and the nest with $[0,2\pi]$.

This overall reward is designed to incentivize the robot to decrease its distance to the target (higher reward for reduced distance) and to orient itself towards the target (using the cosine of the angle).

– **Reward in case of collaboration is required (two robots have to transport a big box to the nest together**: Distance reward (dis_{reward}): When another robot (a "friend") is present within a certain distance range, a positive or negative reward is given based on the distance between the two robots (d_{robots}), as in Equation 5.

$$dis_{reward} = \begin{cases} 0.01 & \text{if } 0.035 \leqslant d_{\text{robots}} \leqslant 0.1 \\ -0.001 & \text{if } d_{\text{robots}} < 0.035 \\ -\frac{d_{\text{robots}}}{100} & \text{otherwise} \end{cases}, \tag{5}$$

when the calculated Euclidian distance between two robots is between 0.035 and 0.1, the two robots are rewarded for learning to stay together. They are punished if they get closer to less than 0.035 or go farther from each other. All numerical values are chosen by the trial-error approach. Nest Reward (r_{nest}): a

reward is given when the current distance to the nest ($d_{current}$) is less than a defined threshold (Threshold), as in Equation 6.

$$r_{\text{nest}} = \begin{cases} 0.15 & \text{if } d_{\text{current}} < \text{Threshold and friend is present} \\ 0 & \text{otherwise} \end{cases}. \quad (6)$$

The overall reward includes the difference in distance to the nest between the previous and current time step, the nest reward, the cosine of the current angle to the nest, and the distance reward,as in Equation 7. Let d_{prev} be the previous distance to the nest, and $d_{current}$ be the current distance to the nest, and $\alpha_{current}$ be the current angle to the nest.

$$\text{reward} = (d_{\text{prev}} - d_{\text{current}}) + r_{\text{nest}} + \frac{\cos(\alpha_{\text{current}})}{1000} + dis_{reward}. \quad (7)$$

3.4. PSO for comparative analysis. We implemented PSO to compare the performance. PSO is a biologically inspired computational algorithm that simulates the social behavior of organisms. Robots adjust their trajectory based on the collective movement of the swarm, aiming to find the optimal path by sharing information about individual successes. The PSO pseudo-code is shown in Algorithm 1, where $r1, r2$ – random values generally used to maintain diversity in the swarm's search behavior:

$$v_{rightmotor} = V + W + V_r, \quad (8)$$

$$v_{leftmotor} = V + W + V_l, \quad (9)$$

V_r – Avoiding speed for right wheel.
V_l – Avoiding speed for left wheel.
W – Turning velocity.
V – Linear velocity.

Algorithm 1. PSO algorithm

Initialize the environment and robots with positions and velocities
Define fitness function ← The average of the values of the sensors $[LS_0, LS_1, LS_6, LS_7]$.
repeat
  for each robot **do**
    Evaluate robot fitness using fitness function
    if current robot fitness is better than previous best robot fitness **then**
      Update robot best position to current position
    end if
    GET all robots fitness and positions
    Update global best position according to best robot fitness
  end for
  for each robot **do**
    Define $w = 0.7$, $c1 = 2$, $c2 = 2$
    Update robot velocity using equation:
    $v = w \times v + c1 \times r1 \times (\text{robot best} - \text{current position}) + c2 \times r2 \times (\text{global best} - \text{current position})$
    Update robot position using equation:
    position = position + velocity
    Calculate the linear velocity $V \leftarrow$ Distance to the new position
    Calculate the turning velocity $W \leftarrow$ Angle to the new position
    Calculate the avoiding speed for each wheel Vl, Vr
    Apply left motor speed $\leftarrow V + W + Vl$
    Apply right motor speed $\leftarrow V + W + Vr$
  end for
until robot reaches the box

4. Results and discussion. The proposed modular design, as outlined in the flowchart in Figure 2, leverages the application of PPO and PSO to enhance decision-making during the searching and navigation. Simultaneously, it maintains simplicity for less computation-intensive tasks, like gripping or waiting. This approach not only increases computational efficiency but also allows for specialized optimization when necessary. The modular approach offers multiple advantages:

– Specialized Optimization: PPO is deployed in modules that significantly influence task performance, such as search and transportation. This approach ensures that PPO's strengths are effectively utilized.

– Computational Efficiency: As a computationally intensive algorithm, the selective implementation of PPO optimizes the computational load, which is crucial for managing a large swarm of robots with limited processing power.

– Simplicity in Less Complex Modules: Certain tasks that do not require complex decision-making, such as gripping or releasing objects, benefit from simpler control mechanisms, facilitating ease of programming and system maintenance.

– Reduced Overfitting Risk: Limiting the usage of PPO to more complex tasks mitigates the risk of overfitting, ensuring that the model remains generalizable and applicable to diverse scenarios.

– Faster Training Time: By focusing on specific modules, PPO reduces the overall training time, speeding up the system deployment and adaptation.

– Maintaining Predictability in Certain Behaviors: Some modules prioritize predictability and reliability over adaptability, where rule-based behaviors are more appropriate.

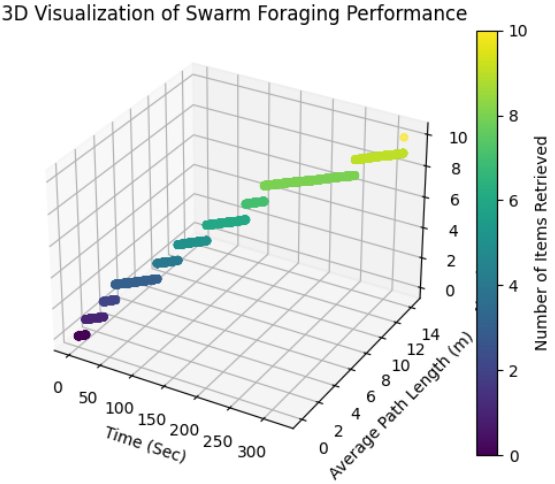
– Reward Design: The reward structure is carefully crafted to align with the objectives of each module. Designing the reward function just for two phases ensures that the main objective of the system will be achieved. It will prevent the system from engaging in unintended behaviors.

– End-to-end System Autonomy: An end-to-end system that includes one deep RL architecture to learn all behaviors like navigation, searching, gripping, and others for all robots' swarm. Relying solely on autonomous decision-making may not yield optimal efficiency. Therefore, combining autonomous and rule-based modules can create a more resilient system.

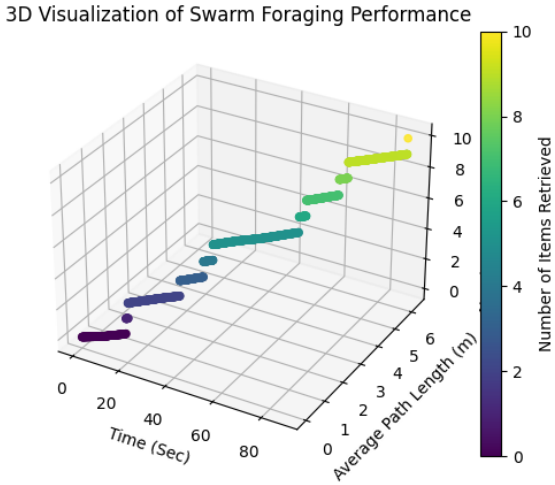
4.1. Foraging performance (RL vs PSO). The provided 3D visualizations in Figure 4 demonstrate the foraging behavior's characteristics of the swarm driven by PSO-PPO and PPO-PPO in addition to neumircal samples given in the Table 2.

Table 2. Foraging performance metrics

Retrieved items	PPO-PPO		PSO-PPO	
	Time (sec)	Average Path (m)	Time (sec)	Average Path (m)
1	13.024	0.726	7.552	0.377
2	13.696	0.774	25.024	1.173
3	27.712	1.767	36.928	1.719
4	33.92	2.243	78.08	3.513
5	36.512	2.434	98.112	4.451
6	59.488	4.129	125.92	5.725
7	61.408	4.268	163.776	7.436
8	69.92	4.942	181.6	8.256
9	72.192	5.126	272.64	12.051
10	88.096	6.305	320.096	14.211



a) PSO-PPO-driven swarm



b) PPO-PPO-driven swarm

Fig. 4. Foraging performance

In both graphs 4-a and 4-b, number of retrieved items N in the proposed problem is in the range $[0, 10]$. ΔT is the time required to collect all items and transport them to the nest. P_N is the average path length of the entire swarm needed to retrieve all items. Efficiency E is conceptualized as the number of items retrieved per unit of time and effort. To calculate efficiency, we can use the change in time ΔT, the average path lenght P_N, and the number of items retrieved N, as in Equation 10.

$$E = \frac{N}{\Delta T \times P_N}. \tag{10}$$

Based on Equation 10:

$$E_{PPO-PPO} = 10/(88.096 \times 6.31) = 18 \times 10^{-3},$$

$$E_{PSO-PPO} = 10/(320.096 \times 14.21) = 2.19 \times 10^{-3}.$$

The PPO-PPO-driven swarm has outperformed the PSO-PPO-driven swarm in terms of efficiency. Based on the data, it seems that the PPO model allows the swarm to retrieve items faster, as indicated by the steeper slope of the number of items over time. Additionally, the average path lengths taken are shorter for the PPO system. On the other hand, the PSO graph demonstrates a less steep slope, indicating a slower completation time in the foraging task. These results highlight the superiority of the PPO in rapidly adapting and efficiently solving the foraging task. The PPO not only learns faster but also appears to sustain its performance. This is due to PPO's policy gradient optimization, which allows fine-tuning adjustments to the robot's actions based on the received rewards, leading to a refined and more effective strategy. PSO, on the other hand, tends to converge to local optima and lacks the ability to fine-tune. Another reason is that PSO relies on collective swarm dynamics, which can also be a limitation if individual robots do not effectively follow the swarm's behavior or share information.

4.2. Dynamic behavior and autonomy. Based on the two sets of figures provided in Figure 5 and Figure 6, each shows the behavior of a swarm in a dynamic foraging task to collect two moving boxes. In the proposed dynamic situation with moving green boxes, the robots follow each box until they grip it. The box turns its color to red as an indicator of gripping and stopping its dynamic nature. Then, it is transported to the nest (yellow area).

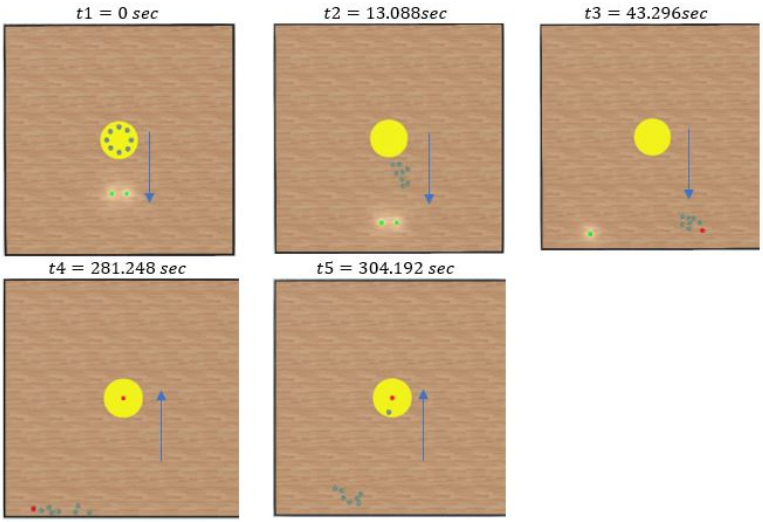


Fig. 5. Dynamic Forgaing performance/ PSO-driven swarm

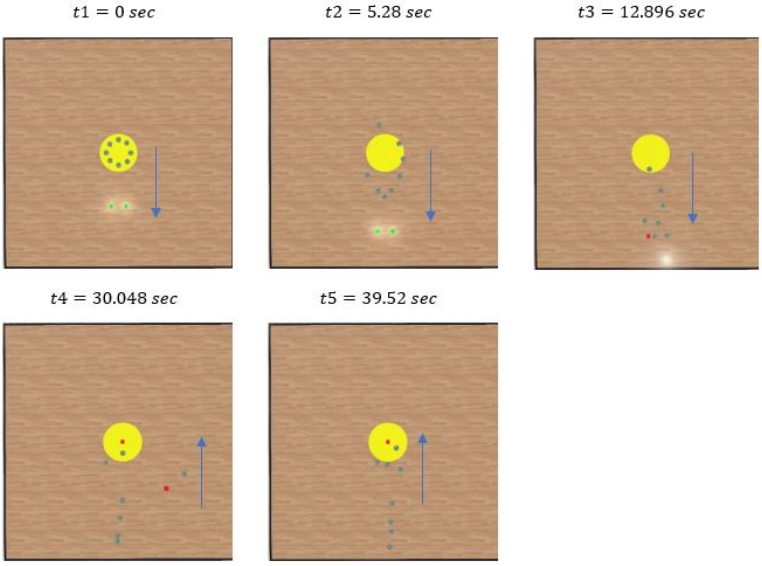


Fig. 6. Dynamic Forgaing performance/RL-driven swarm

Behavior analysis for PSO:

– Initial response (t1-t2): The swarm starts in the nest as an initial position and quickly follows one moving box, indicating a strong initial collective drive and less autonomy.

– Mid-phase (t3): As the swarm finds the green box, it clusters around it. Robots' movements are heavily influenced by their surroundings.

– Gripping action (t4-t5): When the box is gripped (indicated by a change in color to red), the robot that catches the box returns to the nest. The robot releases the box and returns to PSO mode. This demonstrates that PSO can localize and transport the target in dynamic scenarios.

Behavior analysis for RL:

– Initial Response (t1-t2): The RL swarm starts with a distributed, exploratory pattern with no immediate convergence, suggesting an exploratory approach. Which allows the swarm to follow both moving boxes at the same time with a high level of autonomy.

– Mid-phase (t3): The swarm gradually adjusts to the moving target, taking less time to locate the box compared to the PSO, which represents a better response to changes in the target's movement.

– Gripping action (t3-t4-t5): Once the box has been gripped, the robot navigates to the nest with a gripped box. It releases it and returns to the PPO searching box.

5. Conclusion. This study addresses a dynamic foraging task for a swarm of mobile robots. The proposed solution combines a modular design for handling processes like gripping, waiting for aid in carrying the big box, and releasing the box in the nest, with an intelligence algorithm for driving the searching and transportation processes, such as deep RL and PSO. This model maintains simplicity to allow for specialized optimization when necessary, like searching and transporting in a continuous environment while preventing overfitting. It introduces a module for testing various algorithms. Therefore, a comparative analysis was conducted on PPO and PSO. The results revealed that PPO achieved a faster retrieval rate, as well as better efficiency due to its high adaptability and autonomy. In contrast, PSO did not demonstrate the same level of efficiency or autonomy. The findings of this research emphasize the importance of selecting appropriate optimization algorithms depending on the specific task requirements. For tasks that require rapid adaptation and sustainable performance in dynamic environments, RL-PPO stands out as the most effective technique. The study also emphasizes the benefits of a modular approach to swarm robotics, offering a foundation for future developments in this field that require both efficiency and flexibility.

References

1. Cheraghi A., Shahzad S., Graffi K. Past, present, and future of swarm robotics. In Intelligent Systems and Applications: Proceedings of the 2021 Intelligent Systems Conference (IntelliSys). Springer International Publishing. 2022. vol. 3. pp. 190–233.
2. Brambilla M., Ferrante E., Birattari M., Dorigo M. Swarm robotics: a review from the swarm engineering perspective. Swarm Intelligence. 2013. vol. 7. pp. 1–41.
3. Schranz M., Umlauft M., Sende M., Elmenreich W. Swarm robotic behaviors and current applications. Frontiers in Robotics and AI. 2020. vol. 7. DOI: 10.3389/frobt.2020.00036.
4. Li J., Tan Y. A probabilistic finite state machine based strategy for multi-target search using swarm robotics. Applied Soft Computing. 2019. vol. 77. pp. 467–483.
5. Iskandar A., Kovacs B. A Survey on Automatic Design Methods for Swarm Robotics Systems. Carpathian Journal of Electronic and Computer Engineering. 2021. vol. 14. no. 2. pp. 1–5.
6. Jin B., Liang Y., Han Z., Ohkura K. Generating collective foraging behavior for robotic swarm using deep reinforcement learning. Artificial Life and Robotics. 2020. vol. 25. pp. 588–595.
7. Kakish Z., Elamvazhuthi K., Berman S. Using reinforcement learning to herd a robotic swarm to a target distribution. In Distributed Autonomous Robotic Systems: 15th International Symposium. Springer International Publishing. 2022. pp. 401–414.
8. Na S., Roucek T., Ulrich J., Pikman J., Krajnik T., Lennox B., Arvin F. Federated reinforcement learning for collective navigation of robotic swarms. IEEE Transactions on Cognitive and Developmental Systems. 2023. vol. 15. no. 4. pp. 2122–2131.
9. Wang Y., Damani M., Wang P., Cao, Y., Sartoretti G. Distributed reinforcement learning for robot teams: A review. Current Robotics Reports. 2022. vol. 3. no. 4. pp. 239–257.
10. Blais M., Akhloufi M. Reinforcement learning for swarm robotics: An overview of applications, algorithms and simulators. Cognitive Robotics. 2023. vol. 3. pp. 226–256. DOI: 10.1016/j.cogr.2023.07.004.
11. Dias P., Silva M., Rocha Filho G., Vargas P., Cota L., Pessin G. Swarm robotics: A perspective on the latest reviewed concepts and applications. Sensors. 2021. vol. 21. no. 6. DOI: 10.3390/s21062062.
12. Orr J., Dutta A. Multi-agent deep reinforcement learning for multi-robot applications: a survey. Sensors. 2023. vol. 23. no. 7. DOI: 10.3390/s23073625.
13. Aznar F., Pujol M., Rizo, R. Learning a swarm foraging behavior with microscopic fuzzy controllers using deep reinforcement learning. Applied Sciences. 2021. vol. 11. no. 6. DOI: 10.3390/app11062856.
14. Loffler R., Panizon E., Bechinger C. Collective foraging of active particles trained by reinforcement learning. Scientific Reports. 2023. vol. 13. no. 1. DOI: 10.1038/s41598-023-44268-3.
15. Alaa I., Bela K. Curriculum learning for deep reinforcement learning in swarm robotic navigation task. Multidiszciplinaris Tudomanyok. 2023. vol. 13. no. 3. pp. 175–187.
16. Altshuler Y. Recent Developments in the Theory and Applicability of Swarm Search. Entropy. 2023. vol. 25. no. 5. DOI: 10.3390/e25050710.
17. Lee W., Vaughan N., Kim D. Task allocation into a foraging task with a series of subtasks in swarm robotic system. IEEE Access. 2020. vol. 8. pp. 107549–107561.
18. Adams S., Jarne O, Mazo J. A self-guided approach for navigation in a minimalistic foraging robotic swarm. Autonomous Robots. 2023. vol. 47. no. 7. pp. 905–920.
19. Lee K., Kong F., Cannizzaro R., Palmer J., Johnson D., Yoo C., Fitch R. An upper confidence bound for simultaneous exploration and exploitation in heterogeneous multi-robot systems. In 2021 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2021. pp. 8685–8691.

20. Talamali M., Bose T., Haire M., Xu X., Marshall J., Reina A. Sophisticated collective foraging with minimalist agents: A swarm robotics test. Swarm Intelligence. 2020. vol. 14. no. 1. pp. 25–56.
21. Wang X., Guo H. Mobility-aware computation offloading for swarm robotics using deep reinforcement learning. In 2021 IEEE 18th Annual Consumer Communications and Networking Conference (CCNC). IEEE, 2021. pp. 1–4.
22. Michel O. Cyberbotics ltd. webots™: professional mobile robot simulation. International Journal of Advanced Robotic Systems. 2004. vol. 1. no. 1. DOI: 10.5772/5618.

Hammoud Ali — Ph. D. student, Faculty of applied informatics, Federal State Budgetary Educational Institution of Higher Education "Kuban State Agrarian University named after I.T. Trubilin". Research interests: multi-agent systems and decision-making. The number of publications — 1. ali-hammoud@mail.ru; 13, Kalinina St., 350044, Krasnodar, Russia; office phone: +7(861)221-5942.

Iskandar Alaa — Ph. D. student, Faculty of mechanical engineering and informatics (Istvan Salyi Doctoral School of Mechanical Engineering Sciences – mathematic institute), University of Miskolc. Research interests: reinforcement learning for swarm robotics, Navigation, and foraging behaviors. The number of publications — 3. iskandar.alaa@student.uni-miskolc.hu; Egyetemvaros, 3515, Miskolc city, Hungary; office phone: +(36)46-565-111.

Kovács Béla — Ph.D., Dr.Sci., Associate professor, Faculty of mechanical engineering and information, Institute of mathematics department of analysis, University of Miskolc. Research interests: mechanical engineering, differential equations. The number of publications — 99. matmn@uni-miskolc.hu; Egyetemvaros, 3515, Miskolc city, Hungary; office phone: +(36)46-565-111.

УДК 006.72 **DOI** 10.15622/ia.24.1.3

А. Хаммуд, А. Искандар, Б. Ковач

ДИНАМИЧЕСКОЕ ФУРАЖИРОВАНИЕ В РОЕВОЙ РОБОТОТЕХНИКЕ: ГИБРИДНЫЙ ПОДХОД С МОДУЛЬНОЙ КОНСТРУКЦИЕЙ И ГЛУБОКИМ ОБУЧЕНИЕМ С ПОДКРЕПЛЕНИЕМ

Хаммуд А., Искандар А., Ковач Б. **Динамическое фуражирование в роевой робототехнике: гибридный подход с модульной конструкцией и глубоким обучением с подкреплением.**

Аннотация. В этой статье предлагается гибридный подход, который объединяет интеллектуальные алгоритмы и модульную конструкцию для решения проблемы фуражирования в контексте роевой робототехники. Глубокое обучение с подкреплением (RL) и оптимизация роя частиц (PSO) используются в предлагаемой модульной архитектуре. Они используются для поиска множества ресурсов, которые различаются по размеру и демонстрируют динамическую природу с непредсказуемыми движениями. Кроме того, они транспортируют собранные ресурсы в гнездо. Рой состоит из 8 мобильных роботов E-Puck, каждый из которых оснащен датчиками света. Предлагаемая система построена на трехмерной среде с использованием симулятора Webots. С помощью модульного подхода мы решаем сложные проблемы фуражирования, характеризующиеся нестатичной средой и целями. Эта архитектура повышает управляемость, снижает вычислительные требования и упрощает процессы отладки. Наше моделирование показывает, что модель на основе RL превосходит PSO по времени выполнения задач, эффективности сбора ресурсов и адаптивности к динамическим средам, включая движущиеся цели. В частности, роботы, оснащенные RL, демонстрируют улучшенные способности к индивидуальному обучению и принятию решений, обеспечивая уровень автономии, который способствует коллективному интеллекту роя. В PSO коллективные знания роя в большей степени влияют на индивидуальное поведение роботов. Полученные результаты подчеркивают эффективность модульной конструкции и глубокого RL для продвижения автономных роботизированных систем в сложных и непредсказуемых условиях.

Ключевые слова: роевая робототехника, задача поиска пищи, модульное проектирование, обучение с подкреплением, оптимизация роя частиц.

Литература

1. Cheraghi A., Shahzad S., Graffi K. Past, present, and future of swarm robotics. In Intelligent Systems and Applications: Proceedings of the 2021 Intelligent Systems Conference (IntelliSys). Springer International Publishing. 2022. vol. 3. pp. 190–233.
2. Brambilla M., Ferrante E., Birattari M., Dorigo M. Swarm robotics: a review from the swarm engineering perspective. Swarm Intelligence. 2013. vol. 7. pp. 1–41.
3. Schranz M., Umlauft M., Sende M., Elmenreich W. Swarm robotic behaviors and current applications. Frontiers in Robotics and AI. 2020. vol. 7. DOI: 10.3389/frobt.2020.00036.
4. Li J., Tan Y. A probabilistic finite state machine based strategy for multi-target search using swarm robotics. Applied Soft Computing. 2019. vol. 77. pp. 467–483.
5. Iskandar A., Kovacs B. A Survey on Automatic Design Methods for Swarm Robotics Systems. Carpathian Journal of Electronic and Computer Engineering. 2021. vol. 14. no. 2. pp. 1–5.

6. Jin B., Liang Y., Han Z., Ohkura K. Generating collective foraging behavior for robotic swarm using deep reinforcement learning. Artificial Life and Robotics. 2020. vol. 25. pp. 588–595.
7. Kakish Z., Elamvazhuthi K., Berman S. Using reinforcement learning to herd a robotic swarm to a target distribution. In Distributed Autonomous Robotic Systems: 15th International Symposium. Springer International Publishing. 2022. pp. 401–414.
8. Na S., Roucek T., Ulrich J., Pikman J., Krajnik T., Lennox B., Arvin F. Federated reinforcement learning for collective navigation of robotic swarms. IEEE Transactions on Cognitive and Developmental Systems. 2023. vol. 15. no. 4. pp. 2122–2131.
9. Wang Y., Damani M., Wang P., Cao, Y., Sartoretti G. Distributed reinforcement learning for robot teams: A review. Current Robotics Reports. 2022. vol. 3. no. 4. pp. 239–257.
10. Blais M., Akhloufi M. Reinforcement learning for swarm robotics: An overview of applications, algorithms and simulators. Cognitive Robotics. 2023. vol. 3. pp. 226–256. DOI: 10.1016/j.cogr.2023.07.004.
11. Dias P., Silva M., Rocha Filho G., Vargas P., Cota L., Pessin G. Swarm robotics: A perspective on the latest reviewed concepts and applications. Sensors. 2021. vol. 21. no. 6. DOI: 10.3390/s21062062.
12. Orr J., Dutta A. Multi-agent deep reinforcement learning for multi-robot applications: a survey. Sensors. 2023. vol. 23. no. 7. DOI: 10.3390/s23073625.
13. Aznar F., Pujol M., Rizo, R. Learning a swarm foraging behavior with microscopic fuzzy controllers using deep reinforcement learning. Applied Sciences. 2021. vol. 11. no. 6. DOI: 10.3390/app11062856.
14. Loffler R., Panizon E., Bechinger C. Collective foraging of active particles trained by reinforcement learning. Scientific Reports. 2023. vol. 13. no. 1. DOI: 10.1038/s41598-023-44268-3.
15. Alaa I., Bela K. Curriculum learning for deep reinforcement learning in swarm robotic navigation task. Multidiszciplinaris Tudomanyok. 2023. vol. 13. no. 3. pp. 175–187.
16. Altshuler Y. Recent Developments in the Theory and Applicability of Swarm Search. Entropy. 2023. vol. 25. no. 5. DOI: 10.3390/e25050710.
17. Lee W., Vaughan N., Kim D. Task allocation into a foraging task with a series of subtasks in swarm robotic system. IEEE Access. 2020. vol. 8. pp. 107549–107561.
18. Adams S., Jarne O, Mazo J. A self-guided approach for navigation in a minimalistic foraging robotic swarm. Autonomous Robots. 2023. vol. 47. no. 7. pp. 905–920.
19. Lee K., Kong F., Cannizzaro R., Palmer J., Johnson D., Yoo C., Fitch R. An upper confidence bound for simultaneous exploration and exploitation in heterogeneous multi-robot systems. In 2021 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2021. pp. 8685–8691.
20. Talamali M., Bose T., Haire M., Xu X., Marshall J., Reina A. Sophisticated collective foraging with minimalist agents: A swarm robotics test. Swarm Intelligence. 2020. vol. 14. no. 1. pp. 25–56.
21. Wang X., Guo H. Mobility-aware computation offloading for swarm robotics using deep reinforcement learning. In 2021 IEEE 18th Annual Consumer Communications and Networking Conference (CCNC). IEEE, 2021. pp. 1–4.
22. Michel O. Cyberbotics ltd. webots™: professional mobile robot simulation. International Journal of Advanced Robotic Systems. 2004. vol. 1. no. 1. DOI: 10.5772/5618.

Хаммуд Али — аспирант, факультет прикладной информатики, Федеральное государственное бюджетное образовательное учреждение высшего образования «Кубанский государственный аграрный университет имени И.Т. Трубилина». Область научных

интересов: мультиагентные системы и принятие решений. Число научных публикаций — 1. ali-hammoud@mail.ru; улица Калинина, 13, 350044, Краснодар, Россия; р.т.: +7(861)221-5942.

Искандар Алаа — аспирант, факультет машиностроения и информатики (докторантура машиностроительных наук иштвана сали – математический институт), Университет Мишкольца. Область научных интересов: обучение с подкреплением для роевой робототехники, навигации и поиска пищи. Число научных публикаций — 3. iskandar.alaa@student.uni-miskolc.hu; Эгьетемварош, 3515, Мишкольц, Венгрия; р.т.: +(36)46-565-111.

Ковач Бела — Ph.D., Dr.Sci., доцент, факультет машиностроения и информатики института математики, кафедра анализа, Университет Мишкольца. Область научных интересов: машиностроение, дифференциальные уравнения. Число научных публикаций — 99. matmn@uni-miskolc.hu; Эгьетемварош, 3515, Мишкольц, Венгрия; р.т.: +(36)46-565-111.

DOI 10.15622/ia.24.1.4

T. MUSLIMOV

COLLISION AVOIDANCE IN CIRCULAR MOTION OF A FIXED-WING DRONE FORMATION BASED ON ROTATIONAL MODIFICATION OF ARTIFICIAL POTENTIAL FIELD

Muslimov T. **Collision Avoidance in Circular Motion of a Fixed-Wing Drone Formation Based on Rotational Modification of Artificial Potential Field.**

Abstract. In coordinated circular motion of a group of autonomous unmanned aerial vehicles (UAVs or drones), it is important to ensure that collisions between them are avoided. A typical situation occurs when one of the drones in circular formation needs to overtake the drone ahead. The reason for such an overtake may be due to a given geometry of the UAV formation, when this configuration of a given relative position of the drones has changed for some reason. In this case, the limited maneuverability of UAVs of exactly fixed-wing type requires taking into account the peculiarities of their dynamics in the synthesis of the collision avoidance algorithm. The impossibility of the airspeed for a fixed-wing type UAV to drop below a certain minimum value also plays a role here. In this paper, we propose to use an approach based on vortex vector fields, which are essentially a rotational modification of the artificial potential field (APF) method. In this case, the path following algorithm developed in our previous works provides the circular motion. As a result, a collision avoidance algorithm has been developed that works efficiently by maintaining coordinated circular motion of the autonomous drone formation without unnecessary turns. The proposed algorithm was named Artificial Potential Field for Circular Motion (abbreviated as APFfCM). Using the direct Lyapunov method, it is shown that the trajectories of the formation system have uniform boundedness (UB) when using the proposed control algorithm. Due to the boundedness of the candidate Lyapunov function, it is guaranteed that no collision event between drones will occur. Thus, the control objective of providing coordinated circular motion for an autonomous fixed-wing type drone formation without collisions is achieved. Fixed-wing ("flying wing") UAV models in MATLAB/Simulink environment demonstrate the effective performance of the proposed algorithm. These models have both full nonlinear dynamics and implementation of tuned autopilots stabilizing angular and trajectory motion.

Keywords: collision avoidance, drone teams, multi-UAV system, artificial potential field method, vortex vector field.

1. Introduction. The artificial potential field (APF) method [1, 2] has been widely developed for various types of robots: both industrial manipulators and mobile types. From its basic version, this approach has received in its development many modifications, for the possibilities of use in specific models of controlled objects. For example, significantly modified APF methods have been used for high-level models in the form of a double integrator [3], as well as for wheeled mobile robots [4-7]. However, the use of the APF method on unmanned aerial vehicles (UAVs or drones) has its limitations [8]. This is especially evident in the case of fixed-wing type drones, which have limited maneuverability due to nonholonomic dynamics.

A rotationally modified vector field-based APF (vortex vector field) method for motion planning and collision avoidance in robot control was

probably first proposed and studied in detail in [9, 10]. This approach was further developed for mobile robots [11]. The application of rotational vector field for mobile robots avoiding collisions on an opposite course was studied in [12]. In [13], a strategy for avoiding collisions between two UAVs while moving on an opposite course is considered. In [14], this concept was applied to a single quadrocopter avoiding a collision with a fixed obstacle. In [15], fuzzy logic theory and genetic algorithm are used to further modify such an improved potential field. In [16, 17], the rotational modification of the potential field is successfully applied to a single quadrocopter and a stationary obstacle. In [18], a new modified algorithm based on the vortex vector field is developed to enable a wheeled mobile robot to effectively avoid collision with a stationary obstacle. The main advantage of the vortex vector field can be described as the efficient escape of the robot from the local minimum state, the simplicity of the algorithm tuning, and the insignificant influence of the evasive maneuver on the final performance of the main mission. In the case of circular formation motion performed while tracking some object, this strategy should be significantly modified. The special feature here is that the UAVs must not only evade the collision, but also successfully return to the circular motion orbit.

In [19], a rotational modification of the artificial potential field is applied to the formation of quadrocopters, but the stability preservation of such a modification has not been analyzed. In [20], the application of rotational modification for the artificial potential field in controlling a group of rotary-wing drones was discussed. The paper [21] also studied the motion of rotary-wing type drone models jointly avoiding collision with a stationary obstacle. However, this study has the following significant differences from [20, 21]: first, the application of the modification on fixed-wing type UAVs has additional complexity due to the limited maneuverability of this type of aircraft; second, it is the circular motion that is studied here, which imposes certain requirements for maintaining the consistency of the group's flight. Some papers propose strategies similar to the vortex vector field, e.g., [22] uses an orthogonal component that shifts the robot from a local minimum state. However, this approach is fundamentally different in the way the control algorithm is designed.

Let us consider in more detail the difference between this paper and our previous one. In [20] a special vector $\mathbf{f}^{escape}(\mathbf{q}_i, \mathbf{q}_{i+1})$ was used:

$$\mathbf{f}^{escape}(\mathbf{q}_i, \mathbf{q}_{i+1}) \triangleq \mu_i (\mathbf{q}_i - \mathbf{q}_{i+1}) / \|\mathbf{q}_i - \mathbf{q}_{i+1}\|_2^2,$$

where μ_i is an adjustable positive coefficient; $\mathbf{q}_i$ and $\mathbf{q}_{i+1}$ are the vectors of UAV positions numbered i and $i+1$ in the global coordinate system. The UAVs

under these numbers must perform collision avoidance. A "danger zone" is introduced as a region of flight space within which the potential repulsion field starts to operate. The radius of this region is denoted as d_o ("safety radius"). The vector $\mathbf{f}^{escape}(\mathbf{q}_i, \mathbf{q}_{i+1})$ allows a pair of UAVs to leave the "danger zone" and continue circular motion according to the original mission.

If we consider $\|\mathbf{q}_i - \mathbf{q}_{i+1}\|_2 = d_o$ as the equilibrium position for the nominal system (not using $\mathbf{f}^{escape}(\mathbf{q}_i, \mathbf{q}_{i+1})$ in the control algorithm), then $\mathbf{f}^{escape}(\mathbf{q}_i, \mathbf{q}_{i+1})$ is a non-vanishing perturbation. Note an important point: it is not shown in [20] that the equilibrium position $\|\mathbf{q}_i - \mathbf{q}_{i+1}\|_2 = d_o$ is Lyapunov stable (in contrast to the equilibrium positions of the formation after leaving the "danger zone"). Moreover, it can be easily shown that $\|\mathbf{q}_i - \mathbf{q}_{i+1}\|_2 = d_o$ is Lyapunov unstable (e.g., using Chetaev's theorem). However, it was found in [20] that in the region $\|\mathbf{q}_i - \mathbf{q}_{i+1}\|_2 \leqslant d_o$ the derivative of the candidate Lyapunov function (denoted as V) is nonpositive. This fact means that the candidate Lyapunov function V is decreasing in the region $\|\mathbf{q}_i - \mathbf{q}_{i+1}\|_2 \leqslant d_o$, that is, $V(t) \leqslant V(0)$. In practical terms, it follows that two quadrocopters placed inside the region $\|\mathbf{q}_i - \mathbf{q}_{i+1}\|_2 \leqslant d_o$, but not at the collision point $\|\mathbf{q}_i - \mathbf{q}_{i+1}\|_2 = 0$, will not hit this collision point at subsequent times. This is justified by the fact that at the collision point $\|\mathbf{q}_i - \mathbf{q}_{i+1}\|_2 = 0$ candidate Lyapunov function tends to plus infinity, which contradicts the condition $V(t) \leqslant V(0)$. Thus, we explained that the component $\mathbf{f}^{escape}(\mathbf{q}_i, \mathbf{q}_{i+1})$ was introduced specifically to destabilize the nominal system that uses only the artificial potential field-based algorithm. This destabilization allows drones performing collision avoidance to successfully leave the "danger zone" in a finite amount of time.

In our current study, the original mission is the coordinated circular motion of the team. Therefore, instead of the vector $\mathbf{f}^{escape}(\mathbf{q}_i, \mathbf{q}_{i+1})$, an additional vector field component is applied to follow the circular path in combination with a rotational modification of the vector field. This additional component can also be considered as a non-vanishing perturbation for a nominal system that uses only the vector field component derived from the artificial potential field.

Thus, the main contribution of this paper is as follows:

– a rotational modification of the artificial potential field (vortex vector field) is considered specifically for autonomous fixed-wing drones, and in the problem of circular coordinated motion;

– using the direct Lyapunov method, the stability of the equilibrium is analyzed; the uniform boundedness (UB) of the trajectories (so-called Lagrange stability) guarantees that no collision event will occur;

– the proposed algorithm is simulated on full nonlinear models of fixed-wing type UAVs; the obtained results clearly illustrate the efficiency of the proposed algorithm.

2. Collision avoidance algorithm for a group of fixed-wing type drones. Consider the motion of two fixed-wing type UAVs in a formation making a circular motion. In this case, UAV No.$(i+1)$ is ahead of UAV No.i in a circular orbit. However, according to the formation control algorithm, the UAV No.i should be ahead of the UAV No.$(i+1)$ at a predetermined distance in steady-state. If the collision avoidance algorithm is not applied, a collision will occur between the two UAVs as UAV No.i will try to overtake UAV No.$(i+1)$ while traveling in a circular orbit.

Objectives: when automating the flight of a circular formation of small UAVs on the basis of decentralized interaction, it is necessary to prevent collisions between drones in a group while maintaining the specified flight altitude by each of the vehicles. In this case, circular motion is based on the path following algorithm in autonomous mode, that is, the drones autonomously execute the embedded control algorithm. A specific technical realization of autonomous circular formation can be based on ZigBee [23] or a mesh network using 900 MHz RF modems [24]. The collision avoidance algorithm should be developed specifically for the above-described flight mode of the drone formation.

Next, we consider the main modifications of the standard artificial potential field (APF) algorithm that lead to the collision avoidance algorithm proposed in this paper. The first modification of the standard APF algorithm: the Attractive Potential Field is not used in the algorithm. Instead of this field, a Path Following Vector Field is used, which is generated by the formation control algorithm for circular motion.

The second modification compared to the work of [20]: no rotationally modified vector field (zero-rotor in this case) is applied for the UAV ahead (in this case, it is UAV No.$(i+1)$). The reason for this is that, due to the limited maneuverability of the fixed-wing drone, the application of a rotationally modified vector field causes the drone to undesirably turn completely around. At the same time, UAV No.$(i+1)$ will also deviate significantly from the basic circular motion trajectory following UAV No.i, since the rotationally modified vector field assumes a circular trajectory around the obstacle. This problem is illustrated in Figure 1, which shows the motion trajectories when using the vector field with rotational modification on both UAV No.i and UAV No.$(i+1)$. This illustration clearly demonstrates the inefficiency of using the rotationally modified vector field for UAV No.$(i+1)$.

The third modification: for a UAV using a vortex vector field, an additional condition is imposed to turn off the evasive maneuver if it overtakes a UAV flying ahead. The overtaking event can be defined by computing the triple product of the vectors: $T_{prod} = \mathbf{n} \cdot (\mathbf{d}_i \times \mathbf{d}_{i+1})$, where $\mathbf{n} = \begin{bmatrix} 0, & 0, & 1 \end{bmatrix}^{\mathrm{T}}$, and $\mathbf{d}_*$ is the distance vector from the center of the circular path to the UAV with the lower index corresponding to the ordinal number of the UAV in the formation. If $T_{prod} < 0$, this means that UAV No.i is behind UAV No.$(i+1)$ and will try to overtake it. The value $T_{prod} > 0$ will mean that the overtaking has already occurred. Note that switching off the evasive maneuver should not be done immediately after overtaking, but after reaching a certain distance (this is shown later in the control algorithm).

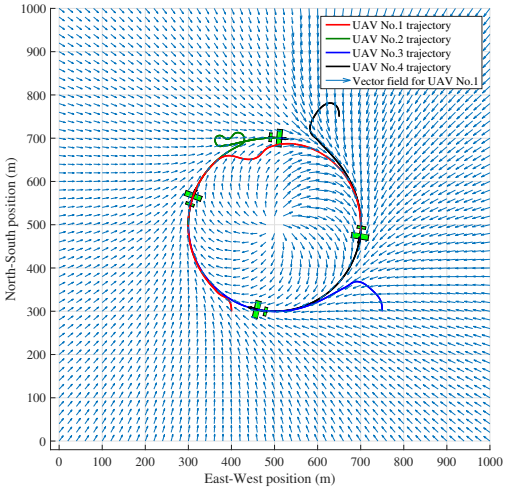


Fig. 1. Trajectories of fixed-wing type UAVs using a vector field with rotational modification

The complete potential field-based collision avoidance algorithm designed for circular motion of UAV formation (we named "Artificial Potential Field for Circular Motion" – APFfCM) is as follows. First, the ordinal numbers of the “overtaking” (No.i) and “overtaken” (No.$(i+1)$) UAVs on the circular path are determined. For this purpose, the triple product $T_{prod} = \mathbf{n} \cdot (\mathbf{d}_i \times \mathbf{d}_{i+1})$ sign should be used. In this case, it is important how exactly this value is calculated initially. The order of the product in brackets plays a role in the expression $\mathbf{n} \cdot (\mathbf{d}_i \times \mathbf{d}_{i+1})$, i.e. which UAV is selected as the i-th and which as

the $(i+1)$-th. In the formation control algorithm, it is assumed that the current angle between the UAVs is given by the center angle α when computing T_{prod}, so the value of this angle is assumed to be less than or equal to π radians. However, also initially the angle can be given through the value of $2\pi - \alpha$, and this should be provided in the structure of the formation control algorithm itself.

Thus, in order to determine the correct order or disorder in the formation, it is necessary to know the inherent rule for determining the angles. For example, for a formation of three UAVs, the inherent rule that the center angle should not exceed π radians makes sense. Therefore, it will be considered that the UAV number i is ahead of the UAV number $i+1$ if $T_{prod} > 0$. Otherwise, the order will be broken and the UAV number i will try to overtake the UAV number $i+1$. As a result, without a collision avoidance algorithm, a collision between UAVs may occur.

Figure 2 on the left shows that if the algorithm is applied to a formation of two UAVs, the originally laid down method of angle calculation can be changed during flight so that overtaking will not occur, since the UAV number $i+1$ can move to a given position by making a turn in an arc of greater length. However, as can be seen from Figure 2 on the right, such a change in the embedded method is no longer possible in the case of a formation of three or more UAVs, because in this case, the UAV number $i+1$ will meet the UAV number $i+2$ on its way (and the rest when the number of UAVs in the formation increases, due to which the number of required obstacle avoidance maneuvers becomes larger than in case of exchanging places of vehicles with numbers i and $i+1$). Therefore, there is a need to develop a collision avoidance algorithm for a decentralized formation of UAVs making such a circular motion.

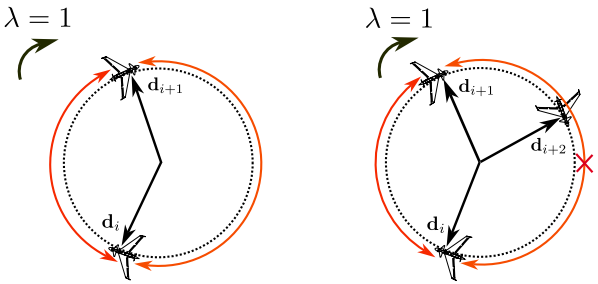


Fig. 2. Disorder in the formation of UAVs

A high-level model of an autopilot-stabilized UAV is considered as

$$\begin{aligned} \mathbf{q}_i &= \begin{bmatrix} p_i^e, & p_i^n \end{bmatrix}^{\mathrm{T}}, \\ \dot{\mathbf{q}}_i &= \begin{bmatrix} v_i \sin\psi_i, & v_i \cos\psi_i \end{bmatrix}^{\mathrm{T}}, \end{aligned} \tag{1}$$

where p_i^e is the position along the coordinate axis pointing East (east axis); p_i^n is the position along the coordinate axis pointing North (north axis); v_i is the airspeed; ψ_i is the course. These states according to the lower index refer to UAV No.i. In this case, the course is measured from the north axis.

Let us define the repulsive artificial potential field (APF) function in the following form of the classical artificial potential, in this case the FIRAS function was chosen [2]:

$$U_r^{\mathrm{APFfCM}}(\mathbf{q}_i, \mathbf{q}_{i+1}) = \begin{cases} \frac{1}{2} k_{r_i} \left(\frac{1}{d(\mathbf{q}_i, \mathbf{q}_{i+1})} - \frac{1}{d_o} \right)^2, & \text{if} \quad d(\mathbf{q}_i, \mathbf{q}_{i+1}) \leqslant d_o; \\ 0, & \text{if} \quad d(\mathbf{q}_i, \mathbf{q}_{i+1}) > d_o, \end{cases} \tag{2}$$

where constant coefficient $k_{r_i} \in \mathbb{R}_{>0}$;
$d(\mathbf{q}_i, \mathbf{q}_{i+1})$ is the distance between UAV No.i and UAV No.$(i+1)$;
d_o is the radius of the so-called "danger zone" within which the action of the potential field of repulsion begins.

Since the UAVs are on the same path line during the circular motion, applying only the standard artificial potential field on both UAVs performing the evasive maneuver may cause a local minimum effect. For this reason, the repulsive artificial potential field for the algorithm is proposed as a function of the artificial potential with rotational modification.

Next, for UAV No.i, we specify the component $\mathbf{f}_r^i(\mathbf{q}_i, \mathbf{q}_{i+1}) \in \mathbb{R}^2$ used in the control law and defined by multiplying the gradient with the opposite sign of the repulsive potential field U_r^{APFfCM} (2) by the rotation matrix $\mathbf{R}(\lambda)$. Thus, we can obtain:

$$\begin{aligned} &\mathbf{f}_r^i(\mathbf{q}_i, \mathbf{q}_{i+1}) = \\ &= -\nabla U_r^{\mathrm{APFfCM}}(\mathbf{q}_i, \mathbf{q}_{i+1}) \mathbf{R}(\lambda) \\ &= \begin{cases} k_{r_i} \left(\frac{1}{d(\mathbf{q}_i, \mathbf{q}_{i+1})} - \frac{1}{d_o} \right) \frac{\mathbf{q}_i - \mathbf{q}_{i+1}}{d^3(\mathbf{q}_i, \mathbf{q}_{i+1})} \mathbf{R}(\lambda), & \text{if} \quad d(\mathbf{q}_i, \mathbf{q}_{i+1}) \leqslant d_o; \\ 0, & \text{if} \quad d(\mathbf{q}_i, \mathbf{q}_{i+1}) > d_o, \end{cases} \end{aligned} \tag{3}$$

where

$$\mathbf{R} = \begin{cases} \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} & \text{if} \quad \lambda = -1 \Rightarrow \text{clockwise}, \\ \\ \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} & \text{if} \quad \lambda = 1 \Rightarrow \text{counterclockwise}. \end{cases} \tag{4}$$

Let us clarify the role of the parameter λ in equation (4). In this case, it specifies the direction of rotation of the formation. The value $\lambda = 1$ indicates clockwise rotation of the formation (when viewed from above), in which case UAV No.i should perform the collision avoidance maneuver according to the vector field with counterclockwise rotational modification. Exactly such a choice of rotation is due to the fact that at some point in time UAV No.i starts to return to the circular motion trajectory. Accordingly, the total path traveled by UAV No.i will be less in the case of maneuvering along the inner region of the circular area bounded by the circumference of the formation flight path. Thus, UAV No.i will be faster to overtake UAV No.$(i+1)$, which in turn, in the final part of the return trajectory will move along an arc of greater length than UAV No.i.

The control algorithm atan2 $\left(\mathbf{f}_i^{\text{APFfCM}}[1]; \mathbf{f}_i^{\text{APFfCM}}[2]\right)$ for the course of UAV No.i is computed using (3) through the vector $\mathbf{f}_i^{\text{APFfCM}} \in \mathbb{R}^2$, defined finally in the following way as shown in Sub-algorithm 1.

if $d(\mathbf{q}_i, \mathbf{q}_{i+1}) \leqslant d_o$ and $T_{prod} < T_{tresh}$ then
$\mathbf{f}_i^{\text{APFfCM}}(\mathbf{q}_i, \mathbf{q}_{i+1}) = \mathbf{f}_r^i(\mathbf{q}_i, \mathbf{q}_{i+1}) + \eta_i \mathbf{f}^{VF}(\mathbf{q}_i)$
$= -\nabla_{\mathbf{q}_i} U_r^{\text{APFfCM}}(\mathbf{q}_i, \mathbf{q}_{i+1}) \mathbf{R}(\lambda)$
$+ \eta_i \mathbf{f}^{VF}(\mathbf{q}_i)$
if $d(\mathbf{q}_i, \mathbf{q}_{i+1}) > d_o$ or $d(\mathbf{q}_i, \mathbf{q}_{i+1}) \leqslant d_o \wedge T_{prod} \geqslant T_{tresh}$ then
$\mathbf{f}_i^{\text{APFfCM}}(\mathbf{q}_i, \mathbf{q}_{i+1}) = \mathbf{f}^{VF}(\mathbf{q}_i)$
end

Listing 1. Sub-algorithm 1

The following notation is used here:
T_{tresh} is a positive parameter chosen in advance;
$\mathbf{f}^{VF}(\mathbf{q}_i)$ is a vector given by circular path following vector field, the essence of which is disclosed in [25, 26];
$\eta_i \in \mathbb{R}_{>0}$ is an adjustable coefficient, which should be chosen sufficiently small, as will be explained later.

For the sake of clarity, we show an expanded form of the course control algorithm for UAV No.i in the case of clockwise motion in Sub-algorithm 2 below.

$$
\begin{aligned}
&\textbf{if } d(\mathbf{q}_i,\mathbf{q}_{i+1}) \leqslant d_o \text{ and } T_{prod} < T_{tresh} \text{ then}\\
&\quad \chi_c^{APF_i} = \operatorname{atan2}\left(v_e^{APF_i}; v_n^{APF_i}\right), \text{ where}\\
&\quad v_e^{APF_i} = \eta_i \sin\chi_i^c - k_{r_i}\left(\frac{1}{d(\mathbf{q}_i,\mathbf{q}_{i+1})} - \frac{1}{d_o}\right)\frac{p_i^n - p_{i+1}^n}{d^3(\mathbf{q}_i,\mathbf{q}_{i+1})}\\
&\quad v_n^{APF_i} = \eta_i \cos\chi_i^c + k_{r_i}\left(\frac{1}{d(\mathbf{q}_i,\mathbf{q}_{i+1})} - \frac{1}{d_o}\right)\frac{p_i^e - p_{i+1}^e}{d^3(\mathbf{q}_i,\mathbf{q}_{i+1})}\\
&\textbf{if } d(\mathbf{q}_i,\mathbf{q}_{i+1}) > d_o \text{ or } d(\mathbf{q}_i,\mathbf{q}_{i+1}) \leqslant d_o \wedge T_{prod} \geqslant T_{tresh} \text{ then}\\
&\quad v_e^{APF_i} = \eta_{v_i}\sin\chi_i^c,\ v_n^{APF_i} = \eta_{v_i}\cos\chi_i^c\\
&\text{end}
\end{aligned}
$$

Listing 2. Sub-algorithm 2

The following notation is used here:
$v_e^{APF_i}$ is the control component along the east axis;
$v_n^{APF_i}$ is the control component along the north axis;
χ_i^c is the course command given by the path following vector field;
$\eta_{v_i} \in \mathbb{R}_{>0}$ is a coefficient that depends on the implementation of the path following algorithm.

The control algorithm for UAV No.$(i+1)$ course will eventually look as follows as in Sub-algorithm 3.

$$
\begin{aligned}
&\textbf{if } d(\mathbf{q}_i,\mathbf{q}_{i+1}) \leqslant d_o \text{ and } T_{prod} < T_{tresh} \text{ then}\\
&\quad \mathbf{f}_{i+1}^{\text{APFfCM}}(\mathbf{q}_{i+1},\mathbf{q}_i) = \mathbf{f}^{APF}(\mathbf{q}_{i+1},\mathbf{q}_i) + \eta_{i+1}\mathbf{f}^{VF}(\mathbf{q}_{i+1})\\
&\qquad\qquad = -\nabla_{\mathbf{q}_{i+1}} U_r^{\text{APFfCM}}(\mathbf{q}_{i+1},\mathbf{q}_i)\\
&\qquad\qquad + \eta_{i+1}\mathbf{f}^{VF}(\mathbf{q}_{i+1})\\
&\textbf{if } d(\mathbf{q}_i,\mathbf{q}_{i+1}) > d_o \text{ or } d(\mathbf{q}_i,\mathbf{q}_{i+1}) \leqslant d_o \wedge T_{prod} \geqslant T_{tresh} \text{ then}\\
&\quad \mathbf{f}_{i+1}^{\text{APFfCM}}(\mathbf{q}_{i+1},\mathbf{q}_i) = \mathbf{f}^{VF}(\mathbf{q}_{i+1})\\
&\text{end}
\end{aligned}
$$

Listing 3. Sub-algorithm 3

The following notation is used here:
$\mathbf{f}^{APF}(\mathbf{q}_{i+1},\mathbf{q}_i)$ is the gradient with the opposite sign of the chosen artificial potential function;
$\mathbf{f}^{VF}(\mathbf{q}_{i+1})$ is a vector similar to $\mathbf{f}^{VF}(\mathbf{q}_i)$, but chosen for the drone number $i+1$;
$\eta_{i+1} \in \mathbb{R}_{>0}$ is a coefficient similar to η_{i+1},but chosen for the drone number $i+1$.

For the sake of clarity, we show an expanded form of the course control algorithm for UAV No.$(i+1)$ in the case of clockwise motion in Sub-algorithm 4 below.

if $d(\mathbf{q}_i, \mathbf{q}_{i+1}) \leqslant d_o$ and $T_{prod} < T_{tresh}$ then

$$v_e^{APF_{i+1}} = \eta_{i+1} \sin \chi_{i+1}^c + k_{r_{i+1}} \left(\frac{1}{d(\mathbf{q}_i, \mathbf{q}_{i+1})} - \frac{1}{d_o} \right) \frac{p_{i+1}^e - p_i^e}{d^3(\mathbf{q}_i, \mathbf{q}_{i+1})}$$

$$v_n^{APF_{i+1}} = \eta_{i+1} \cos \chi_{i+1}^c + k_{r_{i+1}} \left(\frac{1}{d(\mathbf{q}_i, \mathbf{q}_{i+1})} - \frac{1}{d_o} \right) \frac{p_{i+1}^n - p_i^n}{d^3(\mathbf{q}_i, \mathbf{q}_{i+1})}$$

if $d(\mathbf{q}_i, \mathbf{q}_{i+1}) > d_o$ or $d(\mathbf{q}_i, \mathbf{q}_{i+1}) \leqslant d_o \wedge T_{prod} \geqslant T_{tresh}$ then

$$v_e^{APF_{i+1}} = \eta_{v_{i+1}} \sin \chi_{i+1}^c, \; v_n^{APF_{i+1}} = \eta_{v_{i+1}} \cos \chi_{i+1}^c$$

end

Listing 4. Sub-algorithm 4

The following notation is used here:
$v_e^{APF_{i+1}}$ is the control component along the east axis;
$v_n^{APF_{i+1}}$ is the control component along the north axis;
$\chi_c^{APF_{i+1}}$ is the course command given by the path following vector field;
$k_{r_{i+1}} \in \mathbb{R}_{>0}$ is an adjustable constant coefficient;
$\eta_{v_{i+1}} \in \mathbb{R}_{>0}$ is a coefficient that depends on the implementation of the path following algorithm.

Let us denote "$d(\mathbf{q}_i, \mathbf{q}_{i+1}) \leqslant d_o$ and $T_{prod} < T_{tresh}$" as Condition 1 and "$d(\mathbf{q}_i, \mathbf{q}_{i+1}) > d_o$ or $d(\mathbf{q}_i, \mathbf{q}_{i+1}) \leqslant d_o \wedge T_{prod} \geqslant T_{tresh}$" as Condition 2.

The control sub-algorithm for absolute values of UAV velocities (i.e., speeds) is proposed as follows (v_i^c for No.i and v_{i+1}^c for No.$(i+1)$):

$$\begin{aligned} v_i^c &= \begin{cases} \left(\left[v_e^{APF_i} \right]^2 + \left[v_e^{APF_i} \right]^2 \right)^{1/2}, & \text{if Condition 1;} \\ v_{VF_i}^c, & \text{if Condition 2,} \end{cases} \\ v_{i+1}^c &= \begin{cases} \left(\left[v_e^{APF_{i+1}} \right]^2 + \left[v_e^{APF_{i+1}} \right]^2 \right)^{1/2}, & \text{if Condition 1;} \\ v_{VF_{i+1}}^c, & \text{if Condition 2.} \end{cases} \end{aligned} \tag{5}$$

In this case, it is necessary to set speed limits in the control algorithm itself, taking into account the peculiarities of fixed-wing UAV dynamics:

$$v_i^c \in [v_{\min}; v_{\max}] \wedge v_{i+1}^c \in [v_{\min}; v_{\max}].$$

The following notation is used here:
$v_{\max}$ is the selectable maximum speed value;
$v_{\min}$ is the selectable minimum speed value;

$v^c_{VF_i}$ is the speed command obtained through the path following vector field for UAV No.i;
$v^c_{VF_{i+1}}$ is the speed command obtained through the path following vector field for UAV No.$(i+1)$.

It is also possible to implement a simplified version of Sub-algorithm (5) in the form:

$$\begin{aligned} v^c_i &= \begin{cases} v_{\max}, & \text{if} \quad \text{Condition 1;} \\ v^c_{VF_i}, & \text{if} \quad \text{Condition 2,} \end{cases} \\ v^c_{i+1} &= \begin{cases} v_{\min}, & \text{if} \quad \text{Condition 1;} \\ v^c_{VF_{i+1}}, & \text{if} \quad \text{Condition 2.} \end{cases} \end{aligned} \tag{6}$$

In this case, the artificial potential field is used only for course control.

The choice of values $v_{\max}$ and $v_{\min}$ should be made taking into account the dynamics of the UAV, since a significant decrease in speed also causes a strong loss of flight altitude. The operation of Sub-algorithm (6) implies the acceleration of the overtaking maneuver by increasing the speed of the overtaking UAV and reducing the speed of the overtaken UAV by higher values than it is provided by the path following vector field algorithm. These conditions (5)-(6) themselves imply that the velocity components in Sub-algorithms 2 and 4 are multiplied by the same positive coefficient if the velocity modulus is outside the limit values. Therefore, for clarity of presentation, this multiplication by a coefficient is not written out in the stability analysis below. This can be done since there is no influence on the process of the corresponding reasoning.

Together, Sub-algorithms 1-4, (5)-(6) constitute an overall Algorithm for collision avoidance between UAVs in the circular formation motion problem, which we denoted earlier as APFfCM.

Consider a system of two fixed-wing UAVs numbered i and $i+1$ that need to perform collision avoidance. For such a system, the following theorem is satisfied when using the proposed algorithm. A proof of stability in the case of applying an artificial potential field in the object tracking problem was given, for example, in [27]. However, the algorithm in that paper is fundamentally different from the one we consider, so the proof itself is also disparate.

The considered system of two fixed-wing UAV No.i and No.$(i+1)$, taking into account (1) can be represented in a generalized form:

$$\dot{d}\left(\mathbf{q}_i, \mathbf{q}_{i+1}\right) = \left(\nabla_{\mathbf{q}_i - \mathbf{q}_{i+1}} d\left(\mathbf{q}_i, \mathbf{q}_{i+1}\right)\right)^{\mathrm{T}} \left(\dot{\mathbf{q}}_i - \dot{\mathbf{q}}_{i+1}\right), \tag{7}$$

$$\begin{aligned}\dot{\hat{\mathbf{d}}} &= \left(\dot{\hat{d}}_k\right)_{k=\overline{i,\,i+1}} \\ &\triangleq \left(\begin{bmatrix} \dot{p}_i^n & \dot{p}_i^e \end{bmatrix}\begin{bmatrix} \cos\varphi_i & \sin\varphi_i \end{bmatrix}^{\mathrm{T}}\right)\otimes\begin{bmatrix} 1 \\ 0 \end{bmatrix} \\ &\quad + \left(\begin{bmatrix} \dot{p}_{i+1}^n & \dot{p}_{i+1}^e \end{bmatrix}\begin{bmatrix} \cos\varphi_{i+1} & \sin\varphi_{i+1} \end{bmatrix}^{\mathrm{T}}\right)\otimes\begin{bmatrix} 0 \\ 1 \end{bmatrix}.\end{aligned} \tag{8}$$

The following notation is used here:

$\varphi_{k\in\{i,\,i+1\}}$ is the phase angle of rotation for the k-th UAV;
$d_{k\in\{i,\,i+1\}}$ is the distance to the center of rotation for the k-th UAV;
ρ is the radius of the rotation orbit for the formation;
block vector $\hat{\mathbf{d}} \in \mathbb{R}^{2\times 1}$ is defined as

$$\hat{\mathbf{d}} = \left(\hat{d}_k\right)_{k=\overline{i,\,i+1}} \triangleq \left(\|\mathbf{q}_i - \mathbf{c}\|_2 - \rho\right)\otimes\begin{bmatrix} 1 \\ 0 \end{bmatrix} + \left(\|\mathbf{q}_{i+1} - \mathbf{c}\|_2 - \rho\right)\otimes\begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

through the $\mathbf{c} \in \mathbb{R}^{2\times 1}$ as the position vector of the rotation center for the UAV formation. That is, the elements of the vector $\hat{\mathbf{d}}$ are the distances to the circular path given by the distances $d_{k\in\{i,\,i+1\}}$ from the considered UAV to the rotation center of the formation:

$$\hat{d}_{k\in\{i,\,i+1\}} \triangleq d_{k\in\{i,\,i+1\}} - \rho.$$

The equilibrium position $d(\mathbf{q}_i, \mathbf{q}_{i+1}) = d_o$ is considered for the nominal system (6), i.e., in the case when the control algorithm fulfills the condition $\eta_i = \eta_{i+1} = 0$. Then the summands $\eta_i \sin\chi_i^c$, $\eta_i \cos\chi_i^c$, $\eta_{i+1}\sin\chi_{i+1}^c$, $\eta_{i+1}\cos\chi_{i+1}^c$ act as non-vanishing perturbations.

Let us introduce the domains $\mathscr{D}_1$ and $\mathscr{D}_2$:

$$\mathscr{D}_1 \triangleq \{d(\mathbf{q}_i, \mathbf{q}_{i+1}) \in \mathbb{R} \mid d(\mathbf{q}_i, \mathbf{q}_{i+1}) \leqslant d_o \wedge d(\mathbf{q}_i, \mathbf{q}_{i+1}) \neq 0\},$$

$$\mathscr{D}_2 \triangleq \{d(\mathbf{q}_i, \mathbf{q}_{i+1}) \in \mathbb{R} \mid d(\mathbf{q}_i, \mathbf{q}_{i+1}) > d_o\}.$$

We consider the behavior of the system (8) only in the domain $\mathscr{D}_2$, since we are interested in the convergence of the drones to a circular orbit of rotation only after they have performed an evasive maneuver and left the "danger zone".

Theorem, 1. Consider domain $\mathscr{D}_1 \cup \mathscr{D}_2$. Under the action of the APFfCM algorithm, the trajectories of system (7) are uniformly bounded (UB), and the equilibrium position of system (8) is locally asymptotically stable in the domain $\mathscr{D}_2$. In addition, there is no collision event between UAV No.i and UAV No.$(i+1)$.

Proof. As noted earlier, the final APFfCM control algorithm uses a rotationally modified artificial potential function for UAV No.i and a FIRAS artificial potential function for UAV No.$(i+1)$. For the distances to the center of rotation $\|\mathbf{d}_i\|_2$ and $\|\mathbf{d}_{i+1}\|_2$, the condition that $\|\mathbf{d}_i\|_2 \neq 0 \wedge \|\mathbf{d}_{i+1}\|_2 \neq 0$ is assumed since the calculation method of some parameters is not defined when this condition is violated.

Let us introduce the following function $\mathbb{V}$ as a candidate Lyapunov function using (2):

$$\mathbb{V} = \begin{cases} \frac{1}{2}\kappa \left\{ \begin{array}{l} U_r^{\text{APFfCM}}(\mathbf{q}_i, \mathbf{q}_{i+1}) + \\ + U_r^{\text{APFfCM}}(\mathbf{q}_{i+1}, \mathbf{q}_i) \end{array} \right\}, & \text{if} \quad d(\mathbf{q}_i, \mathbf{q}_{i+1}) \leqslant d_o; \\ \\ \frac{1}{2}\hat{\mathbf{d}}^{\mathrm{T}}\hat{\mathbf{d}}, & \text{if} \quad d(\mathbf{q}_i, \mathbf{q}_{i+1}) > d_o, \end{cases} \tag{9}$$

where constant coefficient $\kappa \in \mathbb{R}_{>0}$.

The derivative of the candidate Lyapunov function $\mathbb{V}$ (9) can be represented in this form:

$$\dot{\mathbb{V}} = \begin{cases} \frac{1}{2}\kappa \left\{ \begin{array}{l} \dot{U}_r^{\text{APFfCM}}(\mathbf{q}_i, \mathbf{q}_{i+1}) + \\ + \dot{U}_r^{\text{APFfCM}}(\mathbf{q}_{i+1}, \mathbf{q}_i) \end{array} \right\}, & \text{if} \quad d(\mathbf{q}_i, \mathbf{q}_{i+1}) \leqslant d_o; \\ \\ \hat{\mathbf{d}}^{\mathrm{T}}\dot{\hat{\mathbf{d}}}, & \text{if} \quad d(\mathbf{q}_i, \mathbf{q}_{i+1}) > d_o. \end{cases} \tag{10}$$

Consider the case of $d(\mathbf{q}_i, \mathbf{q}_{i+1}) \leqslant d_o$. We can obtain the following representation for the derivative $\dot{U}_r^{\text{APFfCM}}(\mathbf{q}_i, \mathbf{q}_{i+1})$:

$$\dot{U}_r^{\text{APFfCM}}(\mathbf{q}_i, \mathbf{q}_{i+1}) = \left(\nabla_{\mathbf{p}_{i,i+1}} U_r^{\text{APFfCM}}\right)^{\mathrm{T}} \dot{\mathbf{p}}_{i,i+1},$$

where $\mathbf{p}_{i,i+1} \stackrel{\Delta}{=} \mathbf{q}_i - \mathbf{q}_{i+1}$. Similarly

$$\dot{U}_r^{\text{APFfCM}}(\mathbf{q}_{i+1}, \mathbf{q}_i) = \left(\nabla_{\mathbf{p}_{i+1,i}} U_r^{\text{APFfCM}}\right)^{\text{T}} \dot{\mathbf{p}}_{i+1,i}.$$

Note that the following relation holds:

$$\begin{aligned}\nabla_{\mathbf{p}_{i,i+1}} U_r^{\text{APFfCM}}(\mathbf{q}_i, \mathbf{q}_{i+1}) &= \nabla_{\mathbf{q}_i} U_r^{\text{APFfCM}}(\mathbf{q}_i, \mathbf{q}_{i+1}) \\ &= -\nabla_{\mathbf{q}_{i+1}} U_r^{\text{APFfCM}}(\mathbf{q}_i, \mathbf{q}_{i+1}) \\ &= -\nabla_{\mathbf{q}_{i+1}} U_r^{\text{APFfCM}}(\mathbf{q}_{i+1}, \mathbf{q}_i).\end{aligned}$$

Taking into account the above, we can obtain

$$\begin{aligned}\dot{U}_r^{\text{APFfCM}}(\mathbf{q}_i, \mathbf{q}_{i+1}) + \\ +\dot{U}_r^{\text{APFfCM}}(\mathbf{q}_{i+1}, \mathbf{q}_i)\end{aligned} &= \left\{ \begin{array}{l} \left(\nabla_{\mathbf{p}_{i,i+1}} U_r^{\text{APFfCM}}(\mathbf{q}_i, \mathbf{q}_{i+1})\right)^{\text{T}} \dot{\mathbf{p}}_{i,i+1} + \\ +\left(\nabla_{\mathbf{p}_{i+1,i}} U_r^{\text{APFfCM}}(\mathbf{q}_{i+1}, \mathbf{q}_i)\right)^{\text{T}} \dot{\mathbf{p}}_{i+1,i} \end{array} \right\} \\ &= \left\{ \begin{array}{l} \left(\nabla_{\mathbf{q}_i} U_r^{\text{APFfCM}}(\mathbf{q}_i, \mathbf{q}_{i+1})\right)^{\text{T}} (\dot{\mathbf{q}}_i - \dot{\mathbf{q}}_{i+1}) + \\ +\left(\nabla_{\mathbf{q}_{i+1}} U_r^{\text{APFfCM}}(\mathbf{q}_{i+1}, \mathbf{q}_i)\right)^{\text{T}} (\dot{\mathbf{q}}_{i+!} - \dot{\mathbf{q}}_i) \end{array} \right\} \\ &= \left(\nabla_{\mathbf{q}_i} \left\{ \begin{array}{l} U_r^{\text{APFfCM}}(\mathbf{q}_i, \mathbf{q}_{i+1}) + \\ +U_r^{\text{APFfCM}}(\mathbf{q}_{i+1}, \mathbf{q}_i) \end{array} \right\} \right)^{\text{T}} \dot{\mathbf{q}}_i + \\ &\quad + \left(\nabla_{\mathbf{q}_{i+1}} \left\{ \begin{array}{l} U_r^{\text{APFfCM}}(\mathbf{q}_i, \mathbf{q}_{i+1}) + \\ +U_r^{\text{APFfCM}}(\mathbf{q}_{i+1}, \mathbf{q}_i) \end{array} \right\} \right)^{\text{T}} \dot{\mathbf{q}}_{i+1} \\ &= 2\left(\nabla_{\mathbf{q}_i} U_r^{\text{APFfCM}}(\mathbf{q}_i, \mathbf{q}_{i+1})\right)^{\text{T}} \dot{\mathbf{q}}_i + \\ &\quad + 2\left(\nabla_{\mathbf{q}_{i+1}} U_r^{\text{APFfCM}}(\mathbf{q}_{i+1}, \mathbf{q}_i)\right)^{\text{T}} \dot{\mathbf{q}}_{i+1}.\end{aligned}$$

Given the APFfCM algorithm, the derivative of the candidate Lyapunov function (10) takes the form:

$$\frac{1}{2}\kappa\left\{\begin{array}{l}\dot{U}_r^{\mathrm{APFfCM}}(\mathbf{q}_i,\mathbf{q}_{i+1})+\\ +\dot{U}_r^{\mathrm{APFfCM}}(\mathbf{q}_{i+1},\mathbf{q}_i)\end{array}\right\}=$$

$$=\frac{1}{2}\kappa\left\{\begin{array}{c}2\left(\nabla_{\mathbf{q}_i}U_r^{\mathrm{APFfCM}}(\mathbf{q}_i,\mathbf{q}_{i+1})\right)^{\mathrm{T}}\dot{\mathbf{q}}_i+\\ +2\left(\nabla_{\mathbf{q}_{i+1}}U_r^{\mathrm{APFfCM}}(\mathbf{q}_{i+1},\mathbf{q}_i)\right)^{\mathrm{T}}\dot{\mathbf{q}}_{i+1}\end{array}\right\}=$$

$$=\kappa\left\{\begin{array}{c}\left(\nabla_{\mathbf{q}_i}U_r^{\mathrm{APFfCM}}(\mathbf{q}_i,\mathbf{q}_{i+1})\right)^{\mathrm{T}}\times\\ \times\left[-\nabla_{\mathbf{q}_i}U_r^{\mathrm{APFfCM}}(\mathbf{q}_i,\mathbf{q}_{i+1})\,\mathbf{R}(\lambda)+\eta_i\mathbf{f}^{VF}(\mathbf{q}_i)\right]+\\ +\left(\nabla_{\mathbf{q}_{i+1}}U_r^{\mathrm{APFfCM}}(\mathbf{q}_{i+1},\mathbf{q}_i)\right)^{\mathrm{T}}\times\\ \times\left[-\nabla_{\mathbf{q}_{i+1}}U_r^{\mathrm{APFfCM}}(\mathbf{q}_{i+1},\mathbf{q}_i)+\eta_{i+1}\mathbf{f}^{VF}(\mathbf{q}_{i+1})\right]\end{array}\right\}.$$

This equation is transformed to the following form:

$$\frac{1}{2}\kappa\left\{\begin{array}{l}\dot{U}_r^{\mathrm{APFfCM}}(\mathbf{q}_i,\mathbf{q}_{i+1})+\\ +\dot{U}_r^{\mathrm{APFfCM}}(\mathbf{q}_{i+1},\mathbf{q}_i)\end{array}\right\}=$$

$$=\kappa\left\{\begin{array}{l}-\mathrm{K}\left[\left(p_i^n-p_{i+1}^n\right)\eta_i\cos\chi_i^c+\left(p_i^e-p_{i+1}^e\right)\eta_i\sin\chi_i^c\right]-\\ -\mathrm{K}\left[\left(p_{i+1}^n-p_i^n\right)\eta_{i+1}\cos\chi_{i+1}^c+\left(p_{i+1}^e-p_i^e\right)\eta_{i+1}\sin\chi_{i+1}^c\right]\\ \underbrace{+\mathrm{K}^2\left(p_i^n-p_{i+1}^n\right)\left(p_i^e-p_{i+1}^e\right)-\mathrm{K}^2\left(p_i^n-p_{i+1}^n\right)\left(p_i^e-p_{i+1}^e\right)}_{=0}-\\ -\mathrm{K}^2\left[\left(p_i^n-p_{i+1}^n\right)^2+\left(p_i^e-p_{i+1}^e\right)^2\right]\end{array}\right\}\quad(11)$$

$$=\kappa\left\{\begin{array}{l}-\mathrm{K}\left[\begin{array}{l}\left(p_i^n-p_{i+1}^n\right)\left(\eta_i\cos\chi_i^c-\eta_{i+1}\cos\chi_{i+1}^c\right)\\ +\left(p_i^e-p_{i+1}^e\right)\left(\eta_i\sin\chi_i^c-\eta_{i+1}\sin\chi_{i+1}^c\right)\end{array}\right]-\\ -\mathrm{K}^2\left[\left(p_i^n-p_{i+1}^n\right)^2+\left(p_i^e-p_{i+1}^e\right)^2\right]\end{array}\right\}.$$

Here $\mathrm{K}\triangleq k_{r_i}\left(\frac{1}{d(\mathbf{q}_i,\mathbf{q}_{i+1})}-\frac{1}{d_o}\right)\frac{1}{d^3(\mathbf{q}_i,\mathbf{q}_{i+1})}\geqslant 0$.

Let us introduce the notation:

$$x\triangleq p_i^e-p_{i+1}^e,\quad y\triangleq p_i^n-p_{i+1}^n,$$
$$\delta_x\triangleq\eta_i\sin\chi_i^c-\eta_{i+1}\sin\chi_{i+1}^c,\quad \delta_y\triangleq\eta_i\cos\chi_i^c-\eta_{i+1}\cos\chi_{i+1}^c.$$

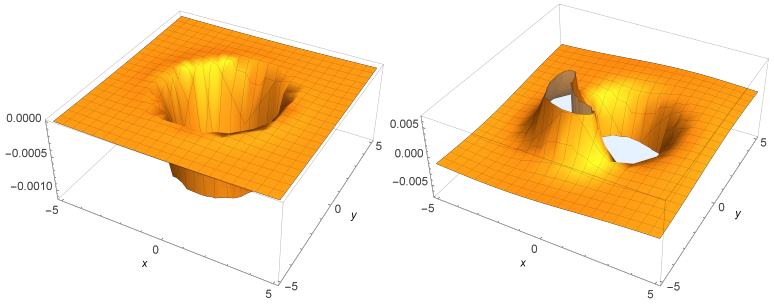
Then we represent the obtained expression (11) in the form

$$\Gamma \triangleq \underbrace{\kappa \mathrm{K}^2 \left(-x^2 - y^2\right)}_{\prec 0} + \kappa \mathrm{K} \left(-\delta_x x - \delta_y y\right), \tag{12}$$

where $\delta_x \in \mathbb{R}$ and $\delta_y \in \mathbb{R}$ are values that should turn out to be sufficiently small, as will be explained later. It is possible to make them so by choosing small parameters η_i and η_{i+1}. The second summand in expression (12) is indefinite. Note that the condition $d\left(\mathbf{q}_i, \mathbf{q}_{i+1}\right) \triangleq \left\|\mathbf{q}_i - \mathbf{q}_{i+1}\right\|_2 \to 0$ corresponds to the divergence of the system trajectories from the equilibrium position. At the same time, the condition $d\left(\mathbf{q}_i, \mathbf{q}_{i+1}\right) \triangleq \left\|\mathbf{q}_i - \mathbf{q}_{i+1}\right\|_2 \to d_o$ corresponds, on the contrary, to the convergence of the trajectories to the equilibrium position. In the case $d\left(\mathbf{q}_i, \mathbf{q}_{i+1}\right) \to 0$, the quadratic summand $\kappa \mathrm{K}^2 \left(-x^2 - y^2\right)$ in (12) will dominate the linear summand $\kappa \mathrm{K} \left(-\delta_x x - \delta_y y\right)$ if it takes positive values. If it takes negative values, then the condition we need is satisfied. Thus, Γ (12) is negative semidefinite except for some regions near the equilibrium position. However, this region can be made as small as desired by the choice of parameters δ_x and δ_y small enough. The trajectories of the system entering this region will not be able to leave it, so uniform boundedness (UB) of trajectories is observed. Note that such a consideration was used in many works, for example, in [28].

To clearly illustrate the above reasoning, we made a graphical representation of the functions. Figure 3(a) shows the three-dimensional view in the case when the summand $\kappa \mathrm{K} \left(-\delta_x x - \delta_y y\right)$ is excluded from the function Γ (12), i.e., the function $\Gamma^{\mathrm{def}} \triangleq \underbrace{\kappa \mathrm{K}^2 \left(-x^2 - y^2\right)}_{\prec 0}$ is considered. Figure 3(b) shows the case where the summand $\kappa \mathrm{K} \left(-\delta_x x - \delta_y y\right)$ is present, that is, the function Γ (12) itself is plotted. The parameters chosen were: $d_o = 3, \quad \delta_x = \delta_y = 0.25, \quad \kappa = 1$. As can be seen, Γ^{def} is always negative or equal to zero, but Γ has, near the equilibrium position, some elevation entering the positive region of values. For an additional illustration, Figure 4 is presented, showing in blue the region near the equilibrium position where the derivative of the Lyapunov function candidate takes positive values. The white color in this case shows the region where conversely the derivative takes negative values. Here the parameters chosen were: $d_o = 3, \quad \delta_x = \delta_y = 0.0015, \quad \kappa = 1$.

Obviously, if the above result holds for the entire region $\mathscr{D}_1$, then it also holds when the additional condition $T_{prod} < T_{tresh}$ is imposed.



a) function Γ^{def} b) function Γ

Fig. 3. Graphical plotting of the candidate Lyapunov function derivative

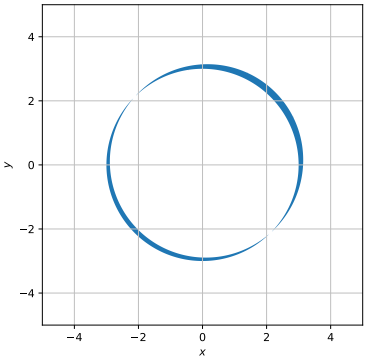


Fig. 4. Unstable region near the equilibrium position

If $d(\mathbf{q}_i, \mathbf{q}_{i+1}) > d_o$, then, given the circular formation control algorithm [25, 26], the derivative of the candidate Lyapunov function (10) takes the form:

$$\begin{aligned}\hat{\mathbf{d}}^T\dot{\hat{\mathbf{d}}} &= \hat{d}_{k\in\{i,i+1\}}\left(\dot{p}^n_{k\in\{i,i+1\}}\cos\varphi_{k\in\{i,i+1\}}\right)+\\ &+\hat{d}_{k\in\{i,i+1\}}\left(\dot{p}^e_{k\in\{i,i+1\}}\sin\varphi_{k\in\{i,i+1\}}\right)\\ &= k^{vf}_{k\in\{i,i+1\}}\hat{d}_{k\in\{i,i+1\}}\cos\varphi_{k\in\{i,i+1\}}\cos\chi^c_{k\in\{i,i+1\}}+\\ &+k^{vf}_{k\in\{i,i+1\}}\hat{d}_{k\in\{i,i+1\}}\sin\varphi_{k\in\{i,i+1\}}\sin\chi^c_{k\in\{i,i+1\}}\\ &= k^{vf}_{k\in\{i,i+1\}}\hat{d}_{k\in\{i,i+1\}}\cos\left(\chi^c_{k\in\{i,i+1\}}-\varphi_{k\in\{i,i+1\}}\right).\end{aligned}$$

Here $\hat{d}_{k\in\{i,i+1\}}$ is the deviation of the distance to the rotation center of the formation, defined earlier in (8); $\varphi_{k\in\{i,\,i+1\}}$ is the angle also defined earlier in (8); $\chi^c_{k\in\{i,i+1\}}$ is the course control command given according to the algorithm for circular path following [25, 26]; $k^{vf}_{\max} > k^{vf}_{k\in\{i,i+1\}} > k^{vf}_{\min} > 0$ is the parameter that is determined by the condition on the UAV's speed in the same control algorithm for circular path following. In view of the way the value $\chi^c_{k\in\{i,i+1\}}$ is calculated in this control algorithm for circular path following [25, 26], we can obtain:

$$\begin{aligned}\hat{\mathbf{d}}^{\mathrm{T}}\dot{\hat{\mathbf{d}}} &= -k^{vf}_{k\in\{i,i+1\}}\hat{d}_{k\in\{i,i+1\}}\sin\left(\arctan\left(k_o\hat{d}_{k\in\{i,i+1\}}\right)\right)\\ &= -\hat{d}^2_{k\in\{i,i+1\}}k^{vf}_{k\in\{i,i+1\}}k_o\left(1+\left(k_o\hat{d}_{k\in\{i,i+1\}}\right)^2\right)^{-1/2},\end{aligned}$$

where $k_o \in \mathbb{R}_{>0}$ is a tunable parameter. Note that in this case the function $\hat{\mathbf{d}}^{\mathrm{T}}\dot{\hat{\mathbf{d}}}$ turns out to be negative definite along the trajectories of the system. Given the positive definiteness of function $\mathbb{V}$ (9), this means the convergence of drones to a circular orbit of rotation.

Note that for the sake of brevity, the case $d(\mathbf{q}_i,\mathbf{q}_{i+1}) \leqslant d_o \wedge T_{prod} \geqslant T_{tresh}$ is not considered separately, since the stability considerations are similar to those given for the case $d(\mathbf{q}_i,\mathbf{q}_{i+1}) > d_o$. At the same time, we assume that the danger of a collision when $T_{prod} \geqslant T_{tresh}$ vanishes due to the occurrence of the overtaking event.

In summary, the derivative of the candidate Lyapunov function (10) along the trajectories of system (7)-(8), in the case of the proposed APFfCM algorithm and the circular path following algorithm [25, 26], takes the form:

$$\dot{\mathbb{V}} = \begin{cases} \kappa\left\{\begin{array}{l} -\mathrm{K}\left[\begin{array}{l}\left(p^n_i-p^n_{i+1}\right)\left(\eta_i\cos\chi^c_i-\eta_{i+1}\cos\chi^c_{i+1}\right)\\ +\left(p^e_i-p^e_{i+1}\right)\left(\eta_i\sin\chi^c_i-\eta_{i+1}\sin\chi^c_{i+1}\right)\end{array}\right]\\ -\mathrm{K}^2\left[\left(p^n_i-p^n_{i+1}\right)^2+\left(p^e_i-p^e_{i+1}\right)^2\right]\end{array}\right\},\\ \qquad\qquad\qquad\qquad\qquad\qquad \text{if } d(\mathbf{q}_i,\mathbf{q}_{i+1}) \leqslant d_o;\\ -\hat{d}^2_{k\in\{i,i+1\}}\times\\ \times k^{vf}_{k\in\{i,i+1\}}k_o\left(1+\left(k_o\hat{d}_{k\in\{i,i+1\}}\right)^2\right)^{-1/2}, \quad \text{if } d(\mathbf{q}_i,\mathbf{q}_{i+1}) > d_o.\end{cases}$$

Hence, we conclude that the derivative of the candidate Lyapunov function along the trajectories of the system (8) is negative definite in the

domain $\mathscr{D}_2$. From this, taking into account the positive definiteness of the function (9), the local asymptotic stability is obtained.

Since uniform boundedness (UB) (so-called Lagrange stability) is satisfied for all trajectories of the system (7), the candidate Lyapunov function itself is bounded for bounded states. The event of a collision between UAVs No.i and No.$(i+1)$ will take the candidate Lyapunov function (9) to $+\infty$. Thus, we can be sure that such a collision will not occur. As a result, by using the direct Lyapunov method, it is possible to guarantee the absence of a collision event. This completes the proof.

3. Simulations on full nonlinear fixed-wing drone models. Further, we present the results of modeling the algorithm in MATLAB/Simulink on full nonlinear models of four small "flying wing" type Zagi UAVs equipped with tuned autopilots. These models are built according to the monograph [29]. In the same monograph, the model parameters and features of its implementation in MATLAB/Simulink environment, as well as the details of autopilot synthesis are described.

Figure 5 shows the UAV flight trajectories in the case of using the APFfCM algorithm proposed in this study.

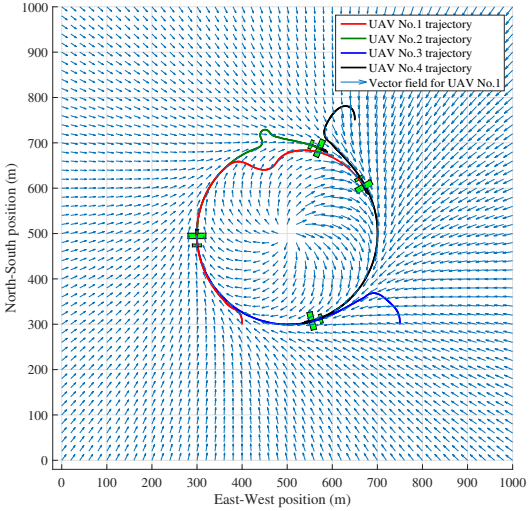


Fig. 5. UAV flight trajectories in case of using APFfCM algorithm

The radius of the "danger zone" ("safety radius") was chosen to be $d_o = 100$ m. We also selected $T_{tresh} = 5 \cdot 10^3$. In the specified figure, the

deviations of the trajectories as a result of performing an evasive maneuver by both UAV No.i and UAV No.$(i+1)$ (in this case, UAV No.1 and UAV No.2) are clearly visible.

Figure 6 shows the distance between UAV No.i and UAV No.$(i+1)$ during the simulation. In this graph, the peak occurring at about 32 seconds is due to the fact that starting at about 32 seconds, the UAVs commence to move closer together on converging courses. This in turn is caused by the fact that after moving away to a reasonably safe distance by about 32 seconds, the UAVs begin to return to collective circular motion, which assumes converging courses while moving.

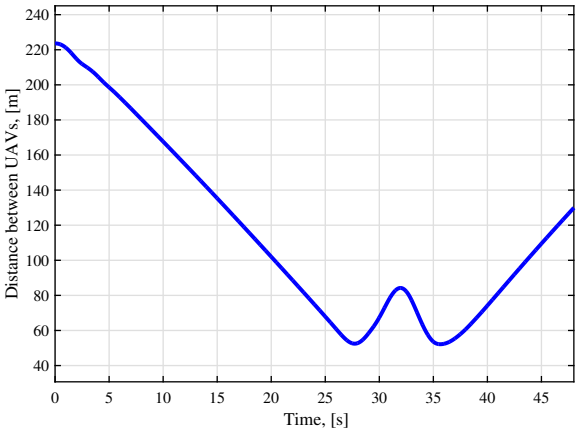


Fig. 6. Distance between UAVs during a flight in case of using APFfCM algorithm

It can also be observed in Figure 6 that the evasive maneuver does not start immediately after exceeding the “safety radius”. This is due to the fact that the path following vector field, rather than the artificial potential field, has a stronger influence at first. As a result, there is a balance between the necessities of maintaining the overall circular strategy and collision avoidance. Figure 7 shows how the value of the triple product T_{prod} changed. In this graph, the moment of sign change corresponds to the moment one UAV overtakes another UAV. From the comparison of Figures 1 and 5, the advantage of the approach proposed in this paper over the algorithm from [20] is clearly visible: UAVs do not make unnecessary meaningless turnarounds and save resources. This advantage arises mainly because the algorithm from [20] was designed for rotary-wing drones and therefore is not directly applicable to fixed-wing UAVs.

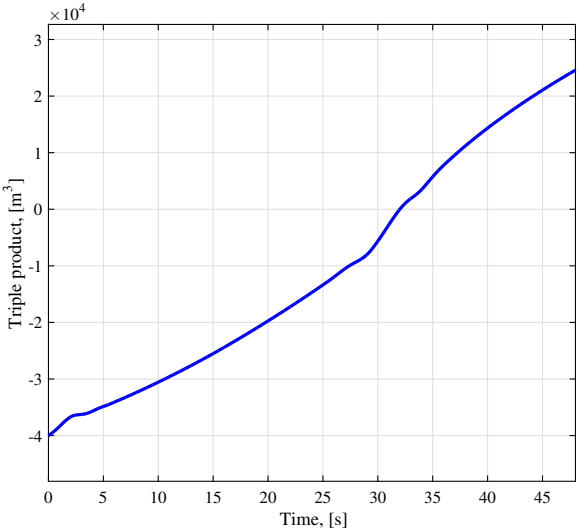


Fig. 7. Changes in the value of the triple product T_{prod}

To numerically compare the proposed algorithm with the standard algorithm, the following metrics are considered. Integral course angle I^{course}:

$$I^{\text{course}} \triangleq \int_{t_0}^{T} \left(\left| \chi^i \right| + \left| \chi^{i+1} \right| \right) dt,$$

where t_0 is the start time of the collision avoidance maneuver; T is the end time of this maneuver; χ^i is the course angle of the UAV No.i; χ^{i+1} is the course angle of the UAV No.$(i+1)$.

Integral path error I^{path}:

$$I^{\text{path}} \triangleq \int_{t_0}^{T} \left(\left| e^i \right| + \left| e^{i+1} \right| \right) dt,$$

where e^i is the path error of the UAV No.i; $e^i + 1$ is the path error of the UAV No.$(i+1)$.

Integral control effort U:

$$U \triangleq \frac{1}{\alpha\chi} \int_{t_0}^{T} \left\{ \left| \chi_c^i - \chi^i \right| + \left| \chi_c^{i+1} - \chi^{i+1} \right| \right\} dt,$$

where α^{χ} is the coefficient determined by the course angle control loop; χ_c^i is the commanded course angle of the UAV No.i; χ_c^{i+1} is the commanded course angle of the UAV No.$(i+1)$.

The maximum total angle of deviation from the initial course χ^{sum}:

$$\chi^{\text{sum}} \triangleq \max_{\chi^i \in \mathbb{R}_{>0}} \chi^i + \max_{\chi^i \in \mathbb{R}_{<0}} \left|\chi^i\right| + \max_{\chi^{i+1} \in \mathbb{R}_{>0}} \chi^{i+1} + \max_{\chi^{i+1} \in \mathbb{R}_{<0}} \left|\chi^{i+1}\right|.$$

Table 1 shows the numerical values obtained from the simulation to compare the APFfCM and standard APF algorithms.

Table 1. Algorithm comparison results

Algorithm variant	I^{course}, rad	I^{path}, m	U, rad/s	χ^{sum}, rad
APFfCM	62.38	798.1	43.76	4.69
Standard APF	107.42	966.1	43.31	7.76

From the comparison of the algorithms, we can conclude that the control effort is almost the same. At the same time, the integral course angle and the maximum total angle of deviation from the original course are significantly larger for the standard APF algorithm. This result is explained by the fact that in the standard APF algorithm, a complete turn of one of the UAVs occurs, which also affects the final trajectory of the other UAV involved in collision avoidance. The integral path error is much larger for the standard APF, which affects the time to return to the final circular path line and the subsequent construction of the given formation geometry by the drones.

4. Conclusion. In this study, a collision avoidance algorithm is proposed for coordinated circular motion of a group of autonomous fixed-wing type UAVs (drones). In this case, one of the drones must overtake the other in order to build a given formation geometry. At the same time, the formation must keep following a circular path line. Compared to the algorithm for rotary-wing UAVs, a modification was required to account for the limited maneuverability of fixed-wing drones due to their nonholonomic dynamics. Also, the uniform boundedness (UB) of the system trajectories allowed to guarantee that collision events between UAVs will not occur. The simulation results on the full nonlinear models of the fixed-wing type UAVs clearly show the advantages of the proposed approach.

References

1. Platonov A.K., Kiril'chenko A.A., Kolganov M.A. [The Potential Field Approach in the Path Finding Problem: History and Perspectives]. Preprinty Instituta im. M.V. Keldysha RAN – Keldysh Institute preprints. 2001. KIAM Preprint № 40. (In Russ.).
2. Khatib O. Real-Time Obstacle Avoidance for Manipulators and Mobile Robots. The international journal of robotics research. 1986. vol. 5. no. 1. pp. 90–98.

3. Bojian L., Jun F., Yunxiao Q., Aijun L. Precise Formation Control for the Multi-agent Systems Based on Tilted Potential Field with Collision Avoidance. International Conference on Guidance, Navigation and Control. Lecture Notes in Electrical Engineering. 2023. vol. 845. pp. 84–93. DOI: 10.1007/978-981-19-6613-2 _10.
4. Alhaddad M., Mironov K., Staroverov A., Panov A. Neural Potential Field for Obstacle-Aware Local Motion Planning. 2023. arxiv preprint arXiv:2310.16362. DOI: 10.48550/arXiv.2310.16362.
5. Filimonov A.B., Filimonov N.B. The Concept of Fairway in Problems of Potential Guidance of Mobile Robots. Optoelectronics, Instrumentation and Data Processing. 2022. vol. 58. no. 4. pp. 366–372. DOI: 10.3103/S8756699022040057.
6. Szczepanski R., Tarczewski T., Erwinski K. Energy Efficient Local Path Planning Algorithm Based on Predictive Artificial Potential Field. IEEE Access. 2022. vol. 10. pp. 39729–39742. DOI: 10.1109/ACCESS.2022.3166632.
7. Mikishanina E.A., Platonov P.S. [Control of a Wheeled Robot on a Plane with Obstacles]. Mekhatronika, Avtomatizatsiya, Upravlenie – Mechatronics, Automation, Control. 2024. vol. 25. no. 2. pp. 93–100. (In Russ.).
8. Xi Z., Han H., Cheng J., Lv M. Reducing Oscillations for Obstacle Avoidance in a Dense Environment Using Deep Reinforcement Learning and Time-Derivative of an Artificial Potential Field. Drones. 2024. vol. 8. no. 3. DOI: 10.3390/drones8030085.
9. De Medio C., Nicolò F., Oriolo G. Robot Motion Planning Using Vortex Fields. New Trends in Systems Theory. Progress in Systems and Control Theory. Boston: Birkhäuser Boston. 1991. vol. 7. pp. 237–244.
10. De Medio C., Oriolo G. Robot Obstacle Avoidance Using Vortex Fields. Advances in Robot Kinematics. Vienna: Springer Vienna. 1991. pp. 227–235. DOI: 10.1007/978-3-7091-4433-6 _26.
11. De Luca A., Oriolo G. Local incremental planning for nonholonomic mobile robots. Proc. IEEE Int. Conf. Robot. Autom. 1994. pp. 104–110. DOI: 10.1109/ROBOT.1994.351003.
12. Martis W.P., Rao S. Cooperative Collision Avoidance in Mobile Robots using Dynamic Vortex Potential Fields. 9th Int. Conf. Autom. Robot. Appl. (ICARA 2023). 2023. pp. 60–64. DOI: 10.1109/ICARA56516.2023.10125851.
13. Choi D., Lee K., Kim D. Enhanced Potential Field-Based Collision Avoidance for Unmanned Aerial Vehicles in a Dynamic Environment. AIAA Scitech 2020 Forum. 2020. DOI: 10.2514/6.2020-0487.
14. Choi D., Kim D., Lee K. Enhanced Potential Field-Based Collision Avoidance in Cluttered Three-Dimensional Urban Environments. Appl. Sci. 2021. vol. 11. no. 22. DOI: 10.3390/app112211003.
15. Choi D., Chhabra A., Kim D. Intelligent cooperative collision avoidance via fuzzy potential fields. Robotica. 2022. vol. 40. no. 6. pp. 1919–1938. DOI: 10.1017/S0263574721001454.
16. Batinovic A., Goricanec J., Markovic L., Bogdan S. Path Planning with Potential Field-Based Obstacle Avoidance in a 3D Environment by an Unmanned Aerial Vehicle. International Conference on Unmanned Aircraft Systems (ICUAS). 2022. pp. 394–401. DOI: 10.1109/ICUAS54217.2022.9836159.
17. Goricanec J., Milas A., Markovic L., Bogdan S. Collision-Free Trajectory Following With Augmented Artificial Potential Field Using UAVs. IEEE Access. 2023. vol. 11. pp. 83492–83506. DOI: 10.1109/ACCESS.2023.3303109.
18. Szczepanski R. Safe Artificial Potential Field – Novel Local Path Planning Algorithm Maintaining Safe Distance From Obstacles. IEEE Robot. Autom. Lett. 2023. vol. 8. no. 8. pp. 4823–4830. DOI: 10.1109/LRA.2023.3290819.

19. Su Y.-H., Bhowmick P., Lanzon A. A Fixed-time Formation-containment Control Scheme for Multi-agent Systems with Motion Planning: Applications to Quadcopter UAVs. IEEE Transactions on Vehicular Technology. 2024. vol. 73. no. 7. pp. 9495–9507. DOI: 10.1109/TVT.2024.3382489.
20. Muslimov T. Curl-Free Vector Field for Collision Avoidance in a Swarm of Autonomous Drones. Lecture Notes in Computer Science. Interactive Collaborative Robotics (ICR 2023). 2023. vol. 14214. pp. 369–379.
21. Muslimov T., Kozlov E., Munasypov R. Drone Swarm Movement without Collisions with Fixed Obstacles Using a Hybrid Algorithm Based on Potential Functions. International Russian Automation Conference (RusAutoCon). 2023. pp. 781–785.
22. Gao Y., Bai C., Fu R., Quan Q. A non-potential orthogonal vector field method for more efficient robot navigation and control. Rob. Auton. Syst. 2023. vol. 159. DOI: 10.1016/j.robot.2022.104291.
23. Park C., Cho N., Lee K., Kim Y. Formation Flight of Multiple UAVs via Onboard Sensor Information Sharing. Sensors. 2015. vol. 15. no. 7. pp. 17397–17419.
24. Kim S., Cho H., Jung D. Circular Formation Guidance of Fixed-Wing UAVs Using Mesh Network. IEEE Access. 2022. vol. 10. pp. 115295–115306. DOI: 10.1109/ACCESS.2022.3218673.
25. Muslimov T. Cooperative Circumnavigation with Robust Vector Field Guidance for Multiple UAVs in Unknown Wind Environments. J. Intell. Robot. Syst. 2023. vol. 109. no. 84. DOI: 10.1007/s10846-023-02000-3.
26. Muslimov T.Z. Metody i algoritmy gruppovogo upravlenija bespilotnymi letatel'nymi apparatami samoletnogo tipa [Methods and algorithms for formation control of fixed wing unmanned aerial vehicles]. Sistemnaya Inzheneriya i Informacionny'e Texnologii (SIIT) [Systems Engineering and Information Technologies]. 2024. vol. 6. no. 1(16). pp. 3–15. (In Russ.).
27. Lafmejani A.S., Farivarnejad H., Sorkhabadi M.R., Zahedi F., Doroudchi A., Berman S. Collision-Free Velocity Tracking of a Moving Ground Target by Multiple Unmanned Aerial Vehicles. The 4th International Symposium on Swarm Behavior and Bio-Inspired Robotics. 2021.
28. Peters S.C., Bobrow J.E., Iagnemma K. Stabilizing a vehicle near rollover: An analogy to cart-pole stabilization. IEEE International Conference on Robotics and Automation. 2010. pp. 5194–5200. DOI: 10.1109/ROBOT.2010.5509367.
29. Beard R.W., McLain T.W. Small unmanned aircraft: Theory and practice. Princeton and Oxford: Princeton University Press. 2012. 320 p.

Muslimov Tagir — Ph.D., Senior researcher, Ufa University of Science and Technology (UUST). Research interests: robotics, control theory, autonomous multi-robot systems. The number of publications — 25. tagir.muslimov@gmail.com; 12, Karl Marx St., 450008, Ufa, Russia; office phone: + 7(347)229-9616.

Acknowledgements. This work was supported by the Ministry of Science and Higher Education of the Russian Federation (Agreement No. 075-15-2021-1016).

УДК 681.5 **DOI** 10.15622/ia.24.1.4

Т.З. Муслимов

ПРЕДОТВРАЩЕНИЕ СТОЛКНОВЕНИЙ ПРИ КРУГОВОМ ДВИЖЕНИИ ГРУППЫ ДРОНОВ САМОЛЕТНОГО ТИПА НА ОСНОВЕ ВРАЩАТЕЛЬНОЙ МОДИФИКАЦИИ ИСКУССТВЕННОГО ПОТЕНЦИАЛЬНОГО ПОЛЯ

Муслимов Т.З. **Предотвращение столкновений при круговом движении группы дронов самолетного типа на основе вращательной модификации искусственного потенциального поля.**

Аннотация. При согласованном круговом движении группы автономных беспилотных летательных аппаратов (БПЛА или дронов) важно обеспечить предотвращение столкновений между ними. Характерная ситуация возникает в том случае, если один из дронов круговой формации должен обогнать впереди летящего. Причина необходимости такого обгона может заключаться в заданной геометрии формации БПЛА, когда эта конфигурация заданного взаимного положения дронов поменялась по какой-либо причине. При этом ограниченная маневренность БПЛА именно самолетного требует учета особенностей их динамики при синтезе алгоритма предотвращения столкновений. Здесь также играет роль невозможность падения воздушной скорости БПЛА самолетного типа ниже определенного минимального значения. В данной статье предлагается использовать подход на основе вихревых векторных полей, которые по сути являются вращательной модификацией метода искусственного потенциального поля (APF). При этом круговое движение обеспечивается разработанным в предыдущих наших работах алгоритмом следования вдоль линии пути. В итоге был предложен алгоритм предотвращения столкновений, который работает эффективно, сохраняя согласованное круговое движение автономной формации дронов без излишних разворотов. Данный алгоритм был назван «Artificial Potential Field for Circular Motion» (сокращенно APFfCM). С помощью прямого метода Ляпунова показано, что траектории системы формации обладают равномерной ограниченностью при использовании предлагаемого алгоритма управления. За счет ограниченности кандидата на функцию Ляпунова при этом гарантировано, что не произойдет события столкновения между дронами. Таким образом цель управления по обеспечению согласованного кругового движения без столкновений для автономной группы дронов самолетного типа достигается. Эффективная работа предлагаемого алгоритма продемонстрирована на моделях БПЛА самолетного типа («летающее крыло») в среде MATLAB/Simulink. Эти модели обладают как полной нелинейной динамикой, так и реализацией настроенных автопилотов, стабилизирующих угловое и траекторное движение.

Ключевые слова: предотвращение столкновений, группы дронов, система из нескольких БПЛА, метод искусственного потенциального поля, вихревое векторное поле.

Литература

1. Платонов А.К., Кирильченко А.А., Колганов М.А. Метод потенциалов в задаче выбора пути: история и перспективы. Препринты Института прикладной математики им. М.В. Келдыша РАН. 2001. Препринт ИПМ № 40.
2. Khatib O. Real-Time Obstacle Avoidance for Manipulators and Mobile Robots. The international journal of robotics research. 1986. vol. 5. no. 1. pp. 90–98.

3. Bojian L., Jun F., Yunxiao Q., Aijun L. Precise Formation Control for the Multi-agent Systems Based on Tilted Potential Field with Collision Avoidance. International Conference on Guidance, Navigation and Control. Lecture Notes in Electrical Engineering. 2023. vol. 845. pp. 84–93. DOI: 10.1007/978-981-19-6613-2 _10.
4. Alhaddad M., Mironov K., Staroverov A., Panov A. Neural Potential Field for Obstacle-Aware Local Motion Planning. 2023. arxiv preprint arXiv:2310.16362. DOI: 10.48550/arXiv.2310.16362.
5. Filimonov A.B., Filimonov N.B. The Concept of Fairway in Problems of Potential Guidance of Mobile Robots. Optoelectronics, Instrumentation and Data Processing. 2022. vol. 58. no. 4. pp. 366–372. DOI: 10.3103/S8756699022040057.
6. Szczepanski R., Tarczewski T., Erwinski K. Energy Efficient Local Path Planning Algorithm Based on Predictive Artificial Potential Field. IEEE Access. 2022. vol. 10. pp. 39729–39742. DOI: 10.1109/ACCESS.2022.3166632.
7. Микишанина Е.А., Платонов П.С. Алгоритмизация управления мобильным колесным роботом в среде с препятствиями методом потенциальных полей. Мехатроника, автоматизация, управление. 2024. Т. 25. № 2. С. 93–100.
8. Xi Z., Han H., Cheng J., Lv M. Reducing Oscillations for Obstacle Avoidance in a Dense Environment Using Deep Reinforcement Learning and Time-Derivative of an Artificial Potential Field. Drones. 2024. vol. 8. no. 3. DOI: 10.3390/drones8030085.
9. De Medio C., Nicolò F., Oriolo G. Robot Motion Planning Using Vortex Fields. New Trends in Systems Theory. Progress in Systems and Control Theory. Boston: Birkhäuser Boston. 1991. vol. 7. pp. 237–244.
10. De Medio C., Oriolo G. Robot Obstacle Avoidance Using Vortex Fields. Advances in Robot Kinematics. Vienna: Springer Vienna. 1991. pp. 227–235. DOI: 10.1007/978-3-7091-4433-6 _26.
11. De Luca A., Oriolo G. Local incremental planning for nonholonomic mobile robots. Proc. IEEE Int. Conf. Robot. Autom. 1994. pp. 104–110. DOI: 10.1109/ROBOT.1994.351003.
12. Martis W.P., Rao S. Cooperative Collision Avoidance in Mobile Robots using Dynamic Vortex Potential Fields. 9th Int. Conf. Autom. Robot. Appl. (ICARA 2023). 2023. pp. 60–64. DOI: 10.1109/ICARA56516.2023.10125851.
13. Choi D., Lee K., Kim D. Enhanced Potential Field-Based Collision Avoidance for Unmanned Aerial Vehicles in a Dynamic Environment. AIAA Scitech 2020 Forum. 2020. DOI: 10.2514/6.2020-0487.
14. Choi D., Kim D., Lee K. Enhanced Potential Field-Based Collision Avoidance in Cluttered Three-Dimensional Urban Environments. Appl. Sci. 2021. vol. 11. no. 22. DOI: 10.3390/app112211003.
15. Choi D., Chhabra A., Kim D. Intelligent cooperative collision avoidance via fuzzy potential fields. Robotica. 2022. vol. 40. no. 6. pp. 1919–1938. DOI: 10.1017/S0263574721001454.
16. Batinovic A., Goricanec J., Markovic L., Bogdan S. Path Planning with Potential Field-Based Obstacle Avoidance in a 3D Environment by an Unmanned Aerial Vehicle. International Conference on Unmanned Aircraft Systems (ICUAS). 2022. pp. 394–401. DOI: 10.1109/ICUAS54217.2022.9836159.
17. Goricanec J., Milas A., Markovic L., Bogdan S. Collision-Free Trajectory Following With Augmented Artificial Potential Field Using UAVs. IEEE Access. 2023. vol. 11. pp. 83492–83506. DOI: 10.1109/ACCESS.2023.3303109.
18. Szczepanski R. Safe Artificial Potential Field – Novel Local Path Planning Algorithm Maintaining Safe Distance From Obstacles. IEEE Robot. Autom. Lett. 2023. vol. 8. no. 8. pp. 4823–4830. DOI: 10.1109/LRA.2023.3290819.

19. Su Y.-H., Bhowmick P., Lanzon A. A Fixed-time Formation-containment Control Scheme for Multi-agent Systems with Motion Planning: Applications to Quadcopter UAVs. IEEE Transactions on Vehicular Technology. 2024. vol. 73. no. 7. pp. 9495–9507. DOI: 10.1109/TVT.2024.3382489.
20. Muslimov T. Curl-Free Vector Field for Collision Avoidance in a Swarm of Autonomous Drones. Lecture Notes in Computer Science. Interactive Collaborative Robotics (ICR 2023). 2023. vol. 14214. pp. 369–379.
21. Muslimov T., Kozlov E., Munasypov R. Drone Swarm Movement without Collisions with Fixed Obstacles Using a Hybrid Algorithm Based on Potential Functions. International Russian Automation Conference (RusAutoCon). 2023. pp. 781–785.
22. Gao Y., Bai C., Fu R., Quan Q. A non-potential orthogonal vector field method for more efficient robot navigation and control. Rob. Auton. Syst. 2023. vol. 159. DOI: 10.1016/j.robot.2022.104291.
23. Park C., Cho N., Lee K., Kim Y. Formation Flight of Multiple UAVs via Onboard Sensor Information Sharing. Sensors. 2015. vol. 15. no. 7. pp. 17397–17419.
24. Kim S., Cho H., Jung D. Circular Formation Guidance of Fixed-Wing UAVs Using Mesh Network. IEEE Access. 2022. vol. 10. pp. 115295–115306. DOI: 10.1109/ACCESS.2022.3218673.
25. Muslimov T. Cooperative Circumnavigation with Robust Vector Field Guidance for Multiple UAVs in Unknown Wind Environments. J. Intell. Robot. Syst. 2023. vol. 109. no. 84. DOI: 10.1007/s10846-023-02000-3.
26. Муслимов Т.З. Методы и алгоритмы группового управления беспилотными летательными аппаратами самолетного типа. Системная инженерия и информационные технологии. 2024. Т. 6. № 1(16). С. 3–15.
27. Lafmejani A.S., Farivarnejad H., Sorkhabadi M.R., Zahedi F., Doroudchi A., Berman S. Collision-Free Velocity Tracking of a Moving Ground Target by Multiple Unmanned Aerial Vehicles. The 4th International Symposium on Swarm Behavior and Bio-Inspired Robotics. 2021.
28. Peters S.C., Bobrow J.E., Iagnemma K. Stabilizing a vehicle near rollover: An analogy to cart-pole stabilization. IEEE International Conference on Robotics and Automation. 2010. p. 5194–5200. DOI: 10.1109/ROBOT.2010.5509367.
29. Beard R.W., McLain T.W. Small unmanned aircraft: Theory and practice. Princeton and Oxford: Princeton University Press. 2012. 320 p.

Муслимов Тагир Забирович — кандидат техн. наук; старший научный сотрудник Уфимского университета науки и технологий (УУНиТ). Область научных интересов: робототехника, теория управления, автономные мультироботные системы. Число научных публикаций — более 25. ORCID: https://orcid.org/0000-0002-9264-529X Email: tagir.muslimov@gmail.com; ФГБОУ ВО «УУНиТ», ул. К.Маркса, д. 12, г. Уфа, 450008, РФ.

Поддержка исследований. Работа выполнена при поддержке Министерства науки и высшего образования Российской Федерации (Соглашение № 075-15-2021-1016).

DOI 10.15622/ia.24.1.5

Y. IMAMVERDIYEV, E. BAGHIROV, I.J. CHUKWU

DETECTING OBFUSCATED MALWARE INFECTIONS ON WINDOWS USING ENSEMBLE LEARNING TECHNIQUES

Imamverdiyev Y., Baghirov E., Chukwu I.J. **Detecting Obfuscated Malware Infections on Windows Using Ensemble Learning Techniques.**

Abstract. In the internet and smart devices era, malware detection has become crucial for system security. Obfuscated malware poses significant risks to various platforms, including computers, mobile devices, and IoT devices, by evading advanced security solutions. Traditional heuristic-based and signature-based methods often fail against these threats. Therefore, a cost-effective detection system was proposed using memory dump analysis and ensemble learning techniques. Utilizing the CIC-MalMem-2022 dataset, the effectiveness of decision trees, gradient-boosted trees, logistic Regression, random forest, and LightGBM in identifying obfuscated malware was evaluated. The study demonstrated the superiority of ensemble learning techniques in enhancing detection accuracy and robustness. Additionally, SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) were employed to elucidate model predictions, improving transparency and trustworthiness. The analysis revealed vital features significantly impacting malware detection, such as process services, active services, file handles, registry keys, and callback functions. These insights are crucial for refining detection strategies and enhancing model performance. The findings contribute to cybersecurity efforts by comprehensively assessing machine learning algorithms for obfuscated malware detection through memory analysis. This paper offers valuable insights for future research and advancements in malware detection, paving the way for more robust and effective cybersecurity solutions in the face of evolving and sophisticated malware threats.

Keywords: malware detection, machine learning, malware analysis, cybersecurity.

1. Introduction. In the rapidly evolving digital era, malware continues to be a severe and prevalent threat to computer systems worldwide. Windows operating systems, in particular, are frequent targets due to their extensive user base and the variety of vulnerabilities that malicious actors can exploit. Malware can lead to severe consequences, including data theft, system damage, and financial loss, necessitating robust detection mechanisms to safeguard systems and users. The ongoing battle between malware distributors and the extensive efforts mobilized for malware detection persists, driven by the destructive potential of malware. This includes significant financial losses, disruption of critical services, and even human casualties in critical infrastructures such as SCADA. Cybercriminals leverage malware as a weapon due to its capacity to inflict widespread harm and chaos [1].

Traditional signature-based methods, which rely on identifying known malware signatures, have proven insufficient in the face of new and sophisticated malware variants. These methods often fail to detect novel threats and zero-day attacks that lack predefined signatures. As a result, the focus has shifted

towards machine learning and behavioral-based detection approaches, which offer the potential to identify previously unseen malware by analyzing patterns and behaviors indicative of malicious intent [2]. This shift is driven by the increasing complexity and obfuscation techniques employed by malware authors, making static analysis methods less effective [3].

On the other hand, dynamic analysis involves executing the software in a controlled environment, such as a sandbox, to observe its behavior and interactions with the system. This method can detect malware that evades static analysis by monitoring runtime behavior, including network activity, file modifications, and registry changes. The strength of dynamic analysis lies in its ability to identify zero-day threats and polymorphic malware. However, it is resource-intensive and time-consuming, requiring a secure environment to execute potentially harmful software. Additionally, sophisticated malware can detect when it is running in a sandbox and alter its behavior to avoid detection, reducing the effectiveness of dynamic analysis [4].

Machine learning-based malware detection leverages the power of statistical analysis and pattern recognition to detect malicious activities in real time. Techniques such as ensemble methods, which combine multiple machine learning models, have shown significant promise in enhancing detection accuracy and robustness. By integrating the strengths of various base classifiers, ensemble techniques can improve detection performance and reduce false positives, making them a valuable tool in the fight against malware [5]. Recent studies have demonstrated the effectiveness of various machine learning and deep learning approaches in malware detection, highlighting the need for continual advancement in this field [6–8].

Moreover, profound and self-supervised learning advancements have opened new avenues for malware detection. Techniques such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have shown potential in identifying complex patterns within large datasets, improving detection rates for previously unseen malware [4]. Self-supervised learning approaches, which do not require large labeled datasets, offer promising solutions for developing efficient and scalable malware detection systems [9].

Despite these advancements, challenges remain in deploying machine learning-based malware detection systems. Issues such as model interpretability, adversarial attacks, and the need for large labeled datasets must be addressed to ensure the effectiveness and reliability of these systems in real-world scenarios. Research efforts continue to explore innovative solutions to these challenges, aiming to develop more robust and adaptable malware detection frameworks [10, 11].

The primary objectives and contributions of this paper can be summarized as follows:

Interpretability of Model Predictions: To enhance the interpretability of our model predictions, we employ SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME). These techniques elucidate the contribution of various system and process-related features to the model predictions, improving transparency and trustworthiness.

Identification of Key Features: Our study identifies vital features that significantly impact malware detection, such as the number of process services, active services, file handles, registry keys, and callback functions. Understanding these features helps refine detection strategies and improve model accuracy.

Future Research Directions: The paper outlines future research directions, including exploring deep learning models, real-time detection systems, dataset expansion, and the integration of behavioral analysis. These directions aim to advance the field of malware detection further and enhance the robustness of detection systems.

The remainder of this paper is structured as follows: Section 3 reviews related works in the field of malware detection, discussing previous studies and their methodologies. Section 4 details the dataset and the machine learning models used in the analysis. In contrast, section 5 presents the results, including the performance metrics of different models and the explainability results using SHAP and LIME to understand the model predictions. Finally, section 6 describes the conclusion and outlines future research directions, summarizing the essential findings and suggesting areas for further investigation.

2. Problem statement. Malware continues to pose a significant threat to Windows operating systems, exploiting the extensive user base and diverse vulnerabilities. Traditional heuristic and signature-based detection methods are increasingly inadequate against sophisticated threats, particularly obfuscated malware designed to evade detection by altering its appearance and behavior. This study aims to develop a robust, cost-effective system for detecting obfuscated malware using memory dump analysis and ensemble learning techniques. By evaluating machine learning algorithms on the CIC-MalMem-2022 dataset and employing SHAP and LIME for model interpretability, this research seeks to enhance detection accuracy, robustness, and transparency in malware detection.

The study utilizes ensemble learning techniques, including methods such as Random Forest, Gradient Boosting, and LightGBM, known for their robustness in handling complex datasets and their ability to generalize well across various conditions. Robustness is ensured through the use of ensemble

methods, which reduce the likelihood of overfitting by aggregating predictions from multiple models, thus enhancing the stability and reliability of the system. Cost-effectiveness is a key consideration in the selection of models and the design of the system. By leveraging memory dump analysis – a method that focuses on analyzing snapshots of system memory – we reduce the computational overhead associated with real-time monitoring and analysis. Additionally, the choice of LightGBM, known for its efficiency in both training time and memory usage, further contributes to the cost-effectiveness of the system. This approach allows for the deployment of the detection system in environments with limited computational resources, making it practical for widespread use.

3. Related work. Several studies, drawing from IEEE Xplore, Web of Science, and Scopus databases,, have focused on applying machine learning and deep learning techniques in the field of malware detection and adversarial attack mitigation [12–32]. In particular, some studies focus on leveraging machine learning models to enhance malware detection capabilities while addressing the vulnerabilities posed by adversarial attacks [13, 15, 17, 18, 20, 22, 23, 26, 27, 29–31]. These efforts include the development of frameworks like EvadeDroid [12], which applies a practical evasion attack on Android malware detectors, and MEME [13], a model-based reinforcement learning algorithm designed to create adversarial malware capable of bypassing detection systems. Additionally, other works have explored the use of deep learning architectures, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), to classify malware based on various features, including API call sequences and syscall subsequences, as seen in studies focusing on these methods [18–20, 22, 31]. These approaches not only demonstrate the effectiveness of machine learning in identifying malware but also highlight the challenges posed by adversarial examples, necessitating the development of robust defense mechanisms, such as adversarial training and randomized smoothing techniques, to safeguard these systems.

A novel opcode-based methodology that leverages multiple behavioral target variables to enhance static malware classification was proposed by the authors in [6]. Their methodology's robustness against random opcode injection attacks was validated on the AMDArgus and MOTIF datasets, achieving superior mean classification accuracy and F1 scores compared to other convolution-based architectures. While the proposed opcode-based malware classification approach shows promise, it also has some limitations. One significant drawback is the assumption that all weak target variables are independent, neglecting the potentially complex relationships between them.

This simplification might limit the model's effectiveness in capturing nuanced patterns in the data.

A novel malware detection scheme for Smart IoT environments called Mal3S, which leverages a multi-spatial pyramid pooling network, was suggested by the authors in [3].Their approach involves static analysis to extract features such as bytes, opcodes, API calls, strings, and dynamic link libraries (DLLs). These are then converted into images of different sizes for training the SPP-net model. Evaluating Mal3S on three malware datasets, they achieved an average detection accuracy of 98.02% and a classification accuracy of 98.43%, outperforming existing techniques. This method also demonstrated effective generalization capabilities across different types of malware. However, the approach's reliance on static analysis means it might struggle to detect malware that extensively obfuscates its code or dynamically modifies its behavior. Future work could explore integrating dynamic analysis techniques to address these limitations and enhance detection accuracy.

The approach for malware classification using self-supervised learning, named MalSSL, was suggested in [4], addressing the challenges of requiring large labeled datasets. MalSSL utilizes image representation, contrastive learning, and data augmentation to classify malware without needing labeled data. The model is first trained on an unlabeled Imagenette dataset as a pretext task and then retrained on an unlabeled malware dataset for downstream tasks, including malware family and benign classification. The results show an accuracy of 98.4% for the malware family classification on the Malimg dataset and 96.2% for the malware and benign classification on the Maldeb dataset, outperforming other self-supervised methods. However, the reliance on pretraining with a dataset like Imagenette might limit its adaptability to more diverse or complex malware datasets. Future work could explore enhancing the model's adaptability and testing its efficacy on varied and evolving malware datasets.

A comparative performance analysis of malware detection algorithms based on various texture features and classifiers was suggested by the authors in [10] to address challenges. Their method includes four stages: converting malware to grayscale, extracting features using segmentation-based fractal texture analysis (SFTA), Local Binary Pattern (LBP), Haralick, Gabor, and Tamura, classifying with Gaussian Discriminant Analysis (GDA), k-Nearest Neighbor (KNN), Logistic, Support Vector Machines (SVM), Random Forest (RF), and Extreme Learning Machine (Ensemble), and evaluating the results. The study used the Malimg imbalanced and MaleVis balanced datasets to assess classifier performance and feature effectiveness. Results indicated that KNN outperformed other classifiers in accuracy, error, F1, and precision, with

SVM and RF as runners-up. Gabor performed better in MaleVis, while SFTA excelled in the Malimg dataset. The SFTA-KNN and Gabor-KNN methods achieved 96.29% and 98.02% accuracy, respectively, surpassing current state-of-the-art approaches. However, the study relied on specific feature extraction methods and comparative analysis using balanced and imbalanced datasets, revealing that balanced datasets significantly improved accuracy and precision while reducing error compared to imbalanced datasets.

The application of several machine-learning algorithms to build a malware detection model for Android systems was suggested by the authors in [11]. Traditional methods of detecting malware using anti-virus software often fall short due to the rapid increase in applications and potentially embedded advertisements or unwanted software. To address this, the authors developed unweighted and weighted models to handle unbalanced data. Their experiments indicated that the weighted random forest model achieved the best performance with an accuracy of 98.94%. However, the study primarily focuses on static analysis and may not account for dynamically changing malware behaviors. Future research could explore incorporating dynamic analysis techniques to enhance detection capabilities further.

A cost-effective obfuscated malware detection system, utilizing diverse machine-learning algorithms through memory dump analysis, was proposed by the authors in [33]. The research focused on the CIC-MaIMem-2022 dataset, simulating real-world scenarios to evaluate the effectiveness of decision trees, ensemble methods, and neural networks in detecting obfuscated malware. Despite the balanced nature of the dataset, with equal malware and benign samples (50%), the authors highlight the application of undersampling and oversampling methods to address potential imbalances within specific malware categories. However, in real-world scenarios, these methods often do not have a positive impact, as they can lead to overfitting or underfitting, reducing the model's generalizability.

Common problems in malware detection research include the over-reliance on static analysis, which struggles against malware with dynamic behavior or advanced obfuscation techniques, and the frequent issue of imbalanced datasets that lead to overfitting or underfitting in models. Simplifying complex relationships between features by assuming their independence can result in models that miss intricate patterns, reducing classification accuracy. Additionally, the use of specific feature extraction methods may limit the generalizability of models to different types of malware. Pretraining with non-malware-specific datasets further hampers adaptability, underscoring the need for integrating dynamic analysis techniques, enhancing

dataset diversity, and capturing interdependencies more effectively in future research.

4. Methodology. Dataset description. The CIC-MalMem-2022 dataset [34], used for this study, comprises memory dumps categorized into four classes: benign, spyware, ransomware, and trojan. Detailed memory features such as process counts, threads, handles, and DLLs are extracted, which help identify malicious patterns. The dataset is balanced, with a total of 58,596 records. Specifically, it includes 29,298 benign records (50%), 10,020 spyware records (17.1%), 9,791 ransomware records (16.7%), and 9,487 trojan records (16.2%).

It is important to note that this is a multiclass classification problem rather than a binary classification task. The distribution of classes reflects a realistic scenario where benign processes are more common, while the various types of malware are less prevalent. This class imbalance (50% benign and the remaining 50% distributed among the three types of malware) adds complexity to the classification task and aligns with real-world situations where malware constitutes a smaller, yet significant, portion of system processes.

Enviromental setup. Our analysis and modeling experiments were conducted using the robust Dataiku platform with advanced data analytics and machine learning capabilities. Dataiku offers a comprehensive suite of tools for data preparation, feature engineering, model development, and evaluation, making it an ideal environment for our research. For this study, we utilized Dataiku version 10.0.5 (licensed), with the notebook server running version 5.4.0-dku10.0-0 and Python 3.6.8.

Model description. In this section, we delve into the machine learning models employed in this study, examining their fundamental principles, loss functions, activation functions, and mathematical formulations. We also discuss each model's strengths, weaknesses, and limitations, providing a comprehensive understanding of their capabilities in the context of obfuscated malware detection.

Decision Tree. Decision Trees are used for classification tasks by recursively splitting the data into subsets based on input feature values. The split criterion, typically Gini impurity or entropy, evaluates the quality of splits. Gini impurity is shown in Equation 1.

$$Gini = \sum_{i=1}^{n} p_i(1 - p_i). \tag{1}$$

In this formula, n represents the total number of classes, and p_i is the probability of an element being classified to class i.

Entropy is shown in Equation 2:

$$Entropy = -\sum_{i=1}^{n} p_i \log(p_i). \tag{2}$$

Strengths include interpretability and ease of implementation, while weaknesses involve susceptibility to overfitting and poor performance on complex datasets [35].

Gradient Boosted Trees. Gradient Boosting builds models sequentially, correcting the errors of previous models. It minimizes a specified loss function using gradient descent. The loss function is shown in Equation 3:

$$L_m(y, F(x)) = \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + v \cdot h_m(x_i)). \tag{3}$$

This formula represents the loss function used in gradient boosting, where $L_m(y, F(x))$ is the loss for each instance i, y_i is the actual value, $F_{m-1}(x_i)$ is the prediction from the previous iteration, v is the learning rate, and $h_m(x_i)$ is the new model to be added.

Strengths include high accuracy and robustness against overfitting, while weaknesses include longer training times and complexity in tuning hyperparameters [36].

LightGBM. LightGBM (Light Gradient Boosting Machine) is designed for speed and performance, using a histogram-based approach to find the best-split points, reducing memory usage and increasing training speed. Histogram-based decision tree learning is shown in Equation 4:

$$G = \sum_{i=1}^{n} \frac{g_i^2}{h_i}, \tag{4}$$

where n is the number of instances, g_i is the gradient of the loss function concerning the prediction, for example, i, and h_i is the Hessian (second derivative) of the loss function concerning the prediction for instance i.

Strengths include high accuracy, scalability, and efficiency, while weaknesses might involve sensitivity to hyperparameter settings [37].

Logistic Regression. Logistic Regression models the probability of a binary classification by applying the logistic function to a linear combination

of input features. The logistic function has the following form:

$$P(y=1|x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n)}}, \tag{5}$$

where $P(y=1|x)$ is the probability of the binary outcome y being one given the input features x. The expression $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$ is a linear combination of the input features $x_1, x_2, \ldots, x_n$ with their respective coefficients $\beta_0, \beta_1, \beta_2, \ldots, \beta_n$. The logistic function (sigmoid function) transforms this linear combination into a probability value between 0 and 1.

Strengths include simplicity and interpretability, while weaknesses involve limitations in handling non-linear relationships [38].

Random Forest. Random Forest constructs multiple decision trees during training and outputs the mode of the classes (classification) or mean prediction (Regression) of the individual trees. The Random Forest algorithm is shown in Equation 6:

$$\dot{f}(x) = \frac{1}{B}\sum_{b=1}^{B} f_b(x). \tag{6}$$

Here, $\dot{f}(x)$ is the final prediction, B is the number of individual models in the ensemble, and $f_b(x)$ is the prediction of the b-th individual model. The ensemble prediction is obtained by averaging the predictions of all individual models.

Strengths include robustness and reduced overfitting, while weaknesses involve complexity and longer training times for large datasets [35].

XGBoost. XGBoost (Extreme Gradient Boosting) is an optimized gradient boosting library for high efficiency, flexibility, and portability. It uses a more regularized model formalization to control overfitting. The regularized objective is shown in Equation 7:

$$L(\theta) = \sum_{i=1}^{n} l(y_i, \dot{y_i}) + \sum_{k=1}^{K} \Omega(f_k). \tag{7}$$

In this formula, $L(\theta)$ represents the total loss, where the first term $\sum_{i=1}^{n} l(y_i, \dot{y_i})$ accounts for the loss for each instance i, with y_i being the true value and $\dot{y_i}$ being the predicted value. The second term $\sum_{k=1}^{K} \Omega(f_k)$ is the regularization term that penalizes the complexity of the model, where $\Omega(f_k)$ applies to each feature k to prevent overfitting.

Strengths include high performance and efficiency, while weaknesses involve complexity in implementation and tuning [39].

Evaluation metrics. To comprehensively understand our models' performance and robustness, we evaluated using these metrics:

Accuracy. Accuracy measures the proportion of correctly predicted instances out of the total cases. While accuracy provides a general performance measure, it may not be suitable for imbalanced datasets.

Precision. Precision, also known as the positive predictive value, indicates the proportion of accurate positive predictions out of all positive predictions. Precision is crucial when the cost of false positives is high.

Recall. Recall, also known as sensitivity or actual positive rate, measures the proportion of accurate positive predictions out of all actual positive instances. Recall is necessary when the cost of false negatives is high.

F1-Score. The F1-score is the harmonic mean of precision and recall, providing a balanced measure of both metrics. It is beneficial when dealing with imbalanced datasets.

ROC-AUC. The ROC-AUC score evaluates the model's ability to discriminate between positive and negative classes. The ROC curve plots the actual positive rate (recall) against the false positive rate. The AUC represents the area under this curve, with a value closer to 1 indicating better model performance.

The evaluation metrics have been shown in Table 1, where the following formulas are used.

Table 1. Metrics used

Metric	**Formula**
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN} \times 100$
Precision	$\frac{TP}{TP+FP}$
Recall	$\frac{TP}{TP+FN}$
F1-Score	$2 \times \frac{Precision \times Recall}{Precision+Recall}$

Explainability. By incorporating explainability into our methodology, we ensured that our machine-learning models for detecting obfuscated malware are transparent and interpretable. While decision trees, regression models, and ensemble methods are generally considered interpretable, the level of interpretability can vary significantly depending on the specific model and its complexity. Simple models like linear regression or single decision trees offer straightforward interpretations; their predictions can be easily understood

by examining coefficients or the structure of the tree, respectively. However, as models become more complex – particularly with the use of ensemble techniques such as boosting and bagging – their interpretability diminishes. Ensemble methods, by their nature, involve the aggregation of predictions from multiple models, often hundreds or thousands of decision trees, each contributing to the final output.

In these complex scenarios, the simple interpretability associated with individual models becomes obscured. It is no longer practical to visualize or directly understand the contribution of each feature across all the constituent models within an ensemble. The final prediction emerges from the collective behavior of many models, making it difficult to deconstruct the prediction into understandable parts.

We used SHapley Additive exPlanations (SHAP) [40] and Local Interpretable Model-agnostic Explanations (LIME) [41] to enhance the explainability of our machine-learning models for detecting obfuscated malware. These techniques assisted us in understanding and interpreting the decisions made by complex models, ensuring transparency and trust.

SHAP (SHapley Additive exPlanations). SHAP values are based on cooperative game theory, providing a unified feature importance measure. They explain how each feature contributes to the prediction by averaging over all possible orderings of features. SHAP ensures three properties: local accuracy, missingness, and consistency. Local accuracy ensures that the sum of feature attributions matches the model output for each instance. Missingness guarantees that features not present in the model have no impact. Consistency ensures that if a model changes such that a feature's contribution increases or stays the same, the attribution should not decrease. SHAP values can be computed using various methods such as Kernel SHAP, which approximates the values for any model type [40].

The formula for the SHAP values is given in Equation 8:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \left[f(S \cup \{i\}) - f(S) \right], \tag{8}$$

where ϕ_i is the SHAP value for feature i, S is a subset of all features N excluding i, $f(S \cup \{i\})$ represents the prediction for the model including feature i in the subset S, and $f(S)$ represents the prediction excluding feature i. The term $\frac{|S|!(|N|-|S|-1)!}{|N|!}$ is a weighting factor based on the size of the subset.

LIME (Local Interpretable Model-agnostic Explanations). LIME explains the predictions of any classifier by approximating it locally with an

interpretable model. It perturbs the data around the instance to be presented and trains a simple, interpretable model (like linear Regression) on these perturbed samples. This local model can provide insights into how each feature influences the prediction in that particular vicinity of the instance. LIME's essence is to balance interpretability and fidelity to the original model [41].

The formula for the LIME explanation model is given in Equation 9:

$$\xi(x) = \arg\min_{g \in G} \sum_{z \in Z} \pi_x(z) \left(f(z) - g(z)\right)^2 + \Omega(g). \tag{9}$$

In this formula, $\xi(x)$ is the explanation model for the instance x, $g \in G$ represents a family of interpretable models, $\pi_x(z)$ is a proximity measure between z and x, $f(z)$ is the prediction of the complex model, and $g(z)$ is the prediction of the interpretable model. The term $\Omega(g)$ is a regularization term to ensure simplicity in the explanation model.

5. Results of the experiments. Our study applied multiple machine-learning models to the CIC-MalMem-2022 dataset to evaluate their performance in detecting obfuscated malware. The models were evaluated based on accuracy, precision, recall, F1-score, and ROC AUC. The results are summarized in Table 2.

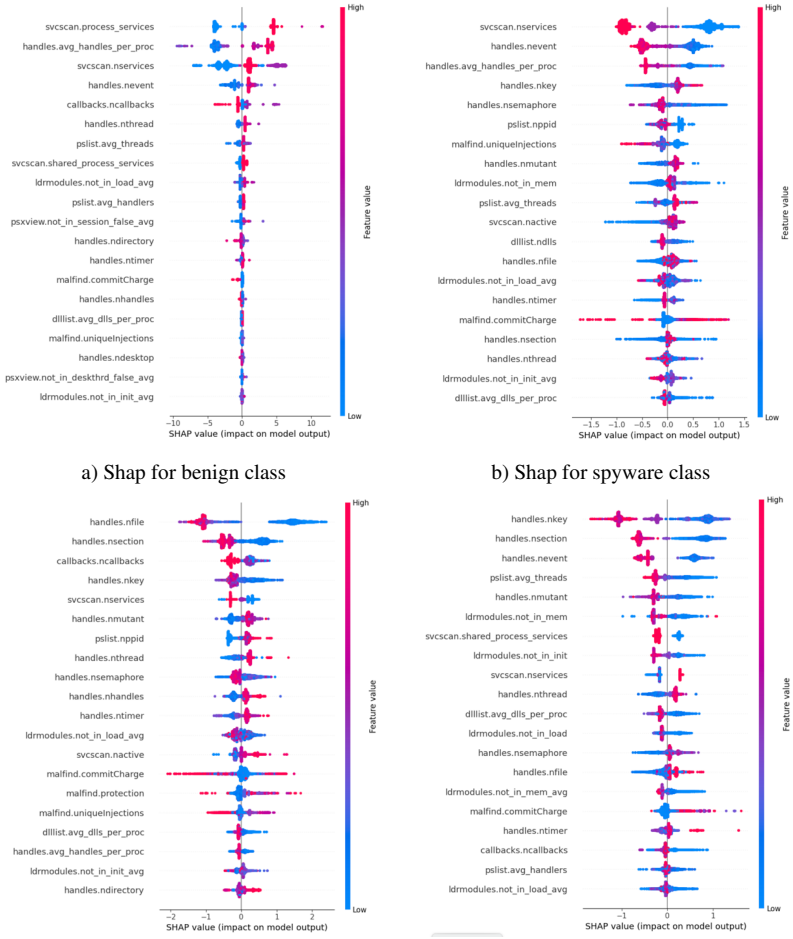
Table 2. Performance comparison of machine learning models for malware detection

Model	Train Time	Accuracy	Precision	Recall	F1-score	ROC AUC
Decision Tree	6s	0.76	0.67	0.64	0.64	0.84
Gradient Boosted Trees	23s	0.81	0.72	0.72	0.72	0.91
LightGBM	28s	0.87	0.81	0.81	0.81	0.95
Logistic Regression	1m 14s	0.74	0.62	0.61	0.61	0.83
Random Forest	1m 9s	0.87	0.80	0.80	0.80	0.95
XGBoost	21s	0.82	0.73	0.73	0.73	0.92

The performance analysis of the machine-learning models on the CIC-MalMem-2022 dataset reveals several key insights. LightGBM achieved the highest F1 score at 0.81, indicating a balanced performance in terms of precision and recall, which is notably higher than other authors' results on the same dataset, with [42] reporting an F1 score of 68% and [43] achieving 70.33%. This score demonstrates that LightGBM is highly effective in identifying true positives while minimizing false positives and false negatives. Similarly, LightGBM also achieved the highest ROC AUC score of 0.95, highlighting

its excellent capability to distinguish between benign and malicious memory dumps. The high ROC AUC score suggests that LightGBM is highly effective in distinguishing between the positive and negative classes.

SHAP explanation. The SHAP (SHapley Additive exPlanations) values for each class of malware and benign processes are presented in Figures 1(a), 1(b), 1(c), and 1(d).



a) Shap for benign class

b) Shap for spyware class

c) Shap for trojan class

d) Shap for ransomware class

Fig. 1. SHAP explanations for different classes: a) benign; b) spyware; c) trojan; d) ransomware

These figures illustrate the impact of various features on the model output for each class. After analyzing the results, we can draw several significant conclusions about the behavior and characteristics of each type of malware and benign processes.

Ransomware is characterized by high values of handles.nkey indicates that these processes heavily utilize registry keys, possibly for configuration and execution. Extensive use of memory sections (handles.nsection) suggests complex memory operations. High event handle usage indicates the reliance on synchronization mechanisms (handles.nevent). Furthermore, ransomware processes tend to have a higher average number of threads (pslist.avg_threads), indicating parallel operations and multitasking. High values of mutant handles (handles.nmutant) point to advanced process control and manipulation.

Trojans exhibit a distinctive pattern where they frequently use many file handles (handles.nfile), likely for file manipulation or monitoring activities. They also make extensive use of memory sections, similar to ransomware. Higher callback counts (callbacks.ncallbacks) suggest that trojans hook into numerous system processes to maintain control and monitor activities. Their extensive interaction with the registry, utilizing numerous registry keys (handles.nkey), is notable. The presence of many running services (svcscan.nservices) indicates that trojans might rely on system services for persistence and functionality.

Spyware processes are marked by their involvement with multiple services (svcscan.nservices), possibly for data collection and transmission. High event handle usage indicates significant synchronization operations within spyware processes. High average handle usage per process (handles.avg_handles_per_proc) suggests intensive interaction with system resources. Frequently engaging with the registry (handles.nkey) and using semaphores (handles.nsemaphore) indicate multiple concurrent processes.

Benign processes, in contrast, show a pattern of routine service operations (svcscan.process_services). Higher average handle usage per process is typical in benign software, reflecting standard interactions with system resources. Everyday system events are frequent in benign processes (handles.nevent), and the presence of callbacks is typical for maintaining standard system functionality. In summary, malicious processes (ransomware, trojan, spyware) generally use handles and registry keys more, indicating manipulation and monitoring activities. Event and memory section handles are frequently used in ransomware and spyware, suggesting complex synchronization and memory usage patterns. Service and callback usage are significant across all classes, differentiating between types of malware and benign processes.

Ransomware is identified by high thread and mutant handle usage, trojans by extensive file and section handle usage, and spyware by intensive service and handle operations. Benign processes exhibit regular service and handle usage patterns, which are typical of standard system operations.

LIME explanation. The provided LIME (Local Interpretable Model-agnostic Explanations) explanation, shown in Figure 2, visualizes the prediction of the LightGBM model for a particular instance. This instance is classified as 'Spyware' with a probability of 0.52. The LIME plot illustrates how different features contribute to the prediction, showing their impact on the model's decision.

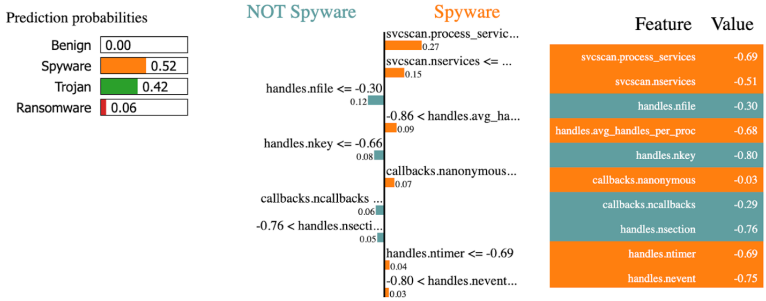


Fig. 2. LIME explanation for spyware instance

The model predicts the instance as 'Spyware' with a probability of 0.52, followed by 'Trojan' with a probability of 0.42. The probabilities for 'Benign' and 'Ransomware' are much lower, at 0.00 and 0.06, respectively. The bar graph on the right-hand side of the LIME plot lists the features and their respective values that contributed to the prediction. Features that increase the likelihood of the instance being classified as 'Spyware' are highlighted in orange, whereas features that decrease the possibility are shown in teal.

We observe that the number of process services (svcscan.process_services) had the highest positive impact on the Spyware classification, with a value of -0.69. This suggests that the number of process services is indicative of spyware activity. Similarly, the number of active services (svcscan.nservices), with a value of -0.51, also positively influenced the classification towards 'Spyware,' implying that the number of active services is a significant factor. The number of file handles (handles.nfile), with a value of -0.30, contributed positively to the classification, suggesting that spyware processes involve numerous file handles.

Further, the average number of handles per process (handles.avg_handles_per_proc), with a value of -0.68, indicates higher activity for spyware. The number of registry keys (handles.nkey), with a value of -0.80, significantly influenced the model toward predicting 'Spyware.' Additionally, anonymous callback functions (callbacks.nanonymous), with a value of -0.03, and the total number of callback functions (callbacks.ncallbacks), with a value of -0.29, also played a role in the classification. Memory-related features such as the number of memory sections (handles.nsection), with a value of -0.76, and the number of timer handles (handles.ntimer), with a value of -0.69, were also influential. The number of event handles (handles.nevent), with a value of -0.75, was another significant factor.

From the LIME explanation, it is clear that the LightGBM model relies heavily on specific system and process-related features to distinguish between different types of malware. For this particular instance, classified as 'Spyware', the number of process services, active services, file handles, and registry keys were vital indicators. In addition, the average number of handles per process, anonymous callback functions, and memory-related features were critical to the prediction.

6. Conclusion and future works. This paper demonstrates a comprehensive approach to detecting obfuscated malware through memory dump analysis using various machine-learning algorithms. Our study leveraged the CIC-MalMem-2022 dataset, which simulates real-world scenarios to evaluate the effectiveness of machine-learning models in identifying obfuscated malware. We implemented and assessed multiple algorithms, including decision trees, gradient-boosted trees, logistic regression, random forest, and LightGBM, to understand their strengths and limitations in malware detection.

The results of the study confirm that the proposed system achieves both robustness and cost-effectiveness, meeting the goals outlined at the outset. The use of ensemble learning techniques, particularly LightGBM, ensures that the system remains robust even when faced with challenging data conditions, such as obfuscated malware samples. Furthermore, the system's efficiency in terms of computational resource usage makes it cost-effective, allowing it to be deployed in environments where resources are limited. This combination of robustness and cost-effectiveness is crucial for practical applications in real-world cybersecurity scenarios, where systems must not only perform accurately but also operate efficiently.

Our findings highlight the superior performance of ensemble learning techniques, particularly LightGBM, in achieving higher detection accuracy and robustness across diverse malware types. We further enhanced the

interpretability of our models using SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME), which provided valuable insights into the contribution of various system and process-related features to the model predictions. Features such as the number of process services, active services, file handles, registry keys, and callback functions were identified as significant indicators in distinguishing between different types of malware and benign processes.

In conclusion, integrating advanced machine learning algorithms and interpretability techniques offers a promising solution to improve malware detection capabilities. This study paves the way for further research in developing robust, interpretable, and practical cybersecurity solutions to combat the ever-evolving landscape of malware threats.

Although this study provides a comprehensive approach to obfuscated malware detection using memory dump analysis and machine learning, several avenues for future research and enhancement remain. Future work could explore the application of deep learning models, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), which have shown promise in various complex classification tasks. Implementing real-time detection systems that can analyze memory dumps and detect malware on the fly is another crucial step. Expanding the dataset to include more diverse and recent malware samples, including those targeting different operating systems and platforms (e.g., macOS, Linux, Android, and IoT devices), would improve the generalizability of the models. Additionally, incorporating benign samples from a broader range of applications and user behaviors could further enhance the model's ability to distinguish between benign and malicious activities. Conducting a more in-depth investigation into feature engineering and selection, studying the impact of adversarial attacks, and exploring other explainable AI techniques would further improve model performance and transparency. Integrating the proposed detection system with existing security frameworks, incorporating behavioral analysis, and developing collaborative defense mechanisms where different systems share threat intelligence could enhance the overall cybersecurity landscape. Addressing regulatory and ethical considerations in the deployment of machine learning-based malware detection systems is also essential. By pursuing these future research directions, we can further advance the field of malware detection, creating more robust, efficient, and interpretable solutions to protect against the ever-evolving landscape of cyber threats.

Declaration of Generative AI and AI-assisted technologies in the writing process. While preparing this work, the authors used ChatGPT for language editing and refinement. After using this tool/service, the author

reviewed and edited the content as needed and took full responsibility for the content of the publication.

References

1. Baghirov E. Evaluating the performance of different machine learning algorithms for Android malware detection. In 2023 5th International Conference on Problems of Cybernetics and Informatics (PCI). IEEE, 2023. pp. 1–4. DOI: 10.1109/PCI60110.2023.10326006.
2. Baghirov E. Comprehensive framework for malware detection: Using ensemble methods, feature selection, and hyperparameter optimization. In 2023 IEEE 17th International Conference on Application of Information and Communication Technologies (AICT). IEEE, 2023. pp. 1–5. DOI: 10.1109/AICT59525.2023.10313179.
3. Jeon J., Jeong B., Baek S., Jeong Y.-S. Static Multi Feature-Based Malware Detection Using Multi SPP-net in Smart IoT Environments. IEEE Transactions on Information Forensics and Security. 2024. vol. 19. pp. 2487–2500. DOI: 10.1109/TIFS.2024.3350379.
4. Ismail S.J.I., Hendrawan Rahardjo B., Juhana T., Musashi Y. MalSSL – Self-Supervised Learning for Accurate and Label-Efficient Malware Classification. IEEE Access. 2024. vol. 12. pp. 58823–58835. DOI: 10.1109/ACCESS.2024.3392251.
5. Baghirov E. Malware detection based on opcode frequency. Journal of Problems of Information Technology, 2023. vol. 14(1). pp. 3–7. DOI: 10.25045/jpit.v14.i1.01.
6. Egitmen A., Yavuz A.G., Yavuz S. TRConv: Multi-Platform Malware Classification via Target Regulated Convolutions. IEEE Access. 2024. vol. 12. pp. 71492–71504. DOI: 10.1109/ACCESS.2024.3401627.
7. Gungor A., Dogru I.A., Barisci N., Toklu S. Malware detection using image-based features and machine learning methods. Journal of the Faculty of Engineering and Architecture of Gazi University. 2023. vol. 38. no. 3. pp. 1781–1792. DOI: 10.17341/gazimmfd.994289.
8. Mesbah A., Baddari I., Riahla M.A. LongCGDroid: Android malware detection through longitudinal study for machine learning and deep learning. Jordanian Journal of Computers and Information Technology. 2023. vol. 9. no. 4. pp. 328–346. DOI: 10.5455/jjcit.71-1693392249.
9. Howard A., Hope B., Saltaformaggio B., Avena E., Ahmadi M., Duncan M., McCann R., Cukierski W. Microsoft Malware Prediction. Kaggle, 2018. Available at: https://kaggle.com/competitions/microsoft-malware-prediction. (accessed 26.10.2024).
10. Ahmed I.T., Hammad B.T., Jamil N.A Comparative Performance Analysis of Malware Detection Algorithms Based on Various Texture Features and Classifiers. IEEE Access. 2024. vol. 12. pp. 11500–11519. DOI: 10.1109/ACCESS.2024.3354959.
11. Xie W., Zhang X. The Application of Machine Learning in Android Malware Detection. 2024 4th International Conference on Neural Networks, Information and Communication Engineering (NNICE). 2024. pp. 1–4. DOI: 10.1109/NNICE61279.2024.10498936.
12. Bostani H.; Moonsamy V. EvadeDroid: A practical evasion attack on machine learning for black-box Android malware detection. Computers and Security. 2024. vol. 139. DOI: 10.1016/j.cose.2023.103676.
13. Rigaki M., Garcia S. The Power of MEME: Adversarial Malware Creation with Model-Based Reinforcement Learning. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2024. pp. 44–64. DOI: 10.1007/978-3-031-51482-1 _3.
14. Rudd E.M., Krisiloff D., Coull S., Olszewski D., Raff E., Holt J. Efficient Malware Analysis Using Metric Embeddings. Digital Threats: Research and Practice. 2024. vol. 5(1). pp. 1–20. DOI: 10.1145/3615669.

15. Zhan D., Zhang Y., Zhu L., Chen J., Xia S., Guo S., Pan Z. Enhancing reinforcement learning based adversarial malware generation to evade static detection. Alexandria Engineering Journal. 2024. vol. 98. pp. 32–43. DOI: 10.1016/j.aej.2024.04.024.
16. Aljabri M., Alhaidari F., Albuainain A., Alrashidi S., Alansari J., Alqahtani W., Alshaya J. Ransomware detection based on machine learning using memory features. Egyptian Informatics Journal. 2024. vol. 25. DOI: 10.1016/j.eij.2024.100445.
17. Ban Y., Kim M., Cho H. An Empirical Study on the Effectiveness of Adversarial Examples in Malware Detection. CMES – Computer Modeling in Engineering and Sciences. 2024. vol. 139(3). pp. 3535–3563. DOI: 10.32604/cmes.2023.046658.
18. Zhang Y., Jiang J., Yi C., Li H., Min S., Zuo R., An Z., Yu Y. A Robust CNN for Malware Classification against Executable Adversarial Attack. Electronics. 2024. vol. 13(5). DOI: 10.3390/electronics13050989.
19. Dam T.Q., Nguyen N.T., Le T.V., Le T.D., Uwizeyemungu S., Le-Dinh T. Visualizing Portable Executable Headers for Ransomware Detection: A Deep Learning-Based Approach. Journal of Universal Computer Science. 2024. vol. 30(2). pp. 262–286. DOI: 10.3897/jucs.104901.
20. Gibert D., Zizzo G., Le Q. Towards a Practical Defense Against Adversarial Attacks on Deep Learning-Based Malware Detectors via Randomized Smoothing. Lecture Notes in Computer Science. 2024. vol. 14399. pp. 683–699. DOI: 10.1007/978-3-031-54129-2_40.
21. Zhang P., Wu C., Wang Z. BINCODEX: A comprehensive and multi-level dataset for evaluating binary code similarity detection techniques. BenchCouncil Transactions on Benchmarks, Standards and Evaluations. 2024. vol. 4(2). DOI: 10.1016/j.tbench.2024.100163.
22. Gibert D., Zizzo G., Le Q., Planes J. Adversarial Robustness of Deep Learning-Based Malware Detectors via (De)Randomized Smoothing. IEEE Access. 2024. vol. 12. pp. 61152–61162. DOI: 10.1109/ACCESS.2024.3392391.
23. Louthanova P., Kozak M., Jurecek M., Stamp M., Di Troia F. A comparison of adversarial malware generators. Journal of Computer Virology and Hacking Techniques. 2024. vol. 20. pp. 623–639. DOI: 10.1007/s11416-024-00519-z.
24. Qian L., Cong L. Channel Features and API Frequency-Based Transformer Model for Malware Identification. Sensors. 2024. vol. 24(2). DOI: 10.3390/s24020580.
25. Surendran R., Uddin M.M., Thomas T., Pradeep G. Android Malware Detection Based on Informative Syscall Subsequences. IEEE Access. 2023. vol. 11. DOI: 10.1109/ACCESS.2024.3387475.
26. Kozak M., Jurecek M., Stamp M., Troia F.D. Creating valid adversarial examples of malware. Journal of Computer Virology and Hacking Techniques. 2024. vol. 20. pp. 607–621. DOI: 10.1007/s11416-024-00516-2.
27. Imran M., Appice A., Malerba D. Evaluating Realistic Adversarial Attacks against Machine Learning Models for Windows PE Malware Detection. Future Internet. 2024. vol. 16(5). DOI: 10.3390/fi16050168.
28. Saha S., Afroz S., Rahman A. H. MALIGN: Explainable static raw-byte based malware family classification using sequence alignment. Computers and Security. 2024. vol. 139. DOI: 10.1016/j.cose.2024.103714.
29. Li D., Cui S., Li Y., Xu J., Xiao F., Xu S. PAD: Towards Principled Adversarial Malware Detection Against Evasion Attacks. IEEE Transactions on Dependable and Secure Computing. 2024. vol. 21. no. 2. pp. 920–936. DOI: 10.1109/TDSC.2023.3265665.
30. Zhang F., Li K., Ren Z. Improving Adversarial Robustness of Ensemble Classifiers by Diversified Feature Selection and Stochastic Aggregation. Mathematics. 2024. vol. 12(6). DOI: 10.3390/math12060834.

31. Alzaidy S., Binsalleeh H. Adversarial Attacks with Defense Mechanisms on Convolutional Neural Networks and Recurrent Neural Networks for Malware Classification. Applied Sciences. 2024. vol. 14(4). DOI: 10.3390/app14041673.
32. Zhou K., Wang P., He B. Comparative Study: Mouth Brooding Fish (MBF) as a Novel Approach for Android Malware Detection. International Journal of Advanced Computer Science and Applications. 2024. vol. 15(5). DOI: 10.14569/IJACSA.2024.0150521.
33. Rakib H., Dhakal S.M. Obfuscated Malware Detection: Investigating Real-World Scenarios Through Memory Analysis. In 5th IEEE International Conference on Telecommunications and Photonics (ICTP 2023). 2023. DOI: 10.1109/ICTP60248.2023.10490701.
34. Carrier T., Victor P., Tekeoglu A., Lashkari A.H. Detecting Obfuscated Malware using Memory Feature Engineering. Proceedings of the 8th International Conference on Information Systems Security and Privacy (ICISSP). 2022. vol. 1. pp. 177–188. DOI: 10.5220/0010908200003120.
35. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and prediction (2nd ed.). Springer, 2009. 745 p.
36. Friedman J.H. Greedy function approximation: A gradient boosting machine. Annals of Statistics. 2001. vol. 29(5). pp. 189–1232. DOI: 10.1214/aos/1013203451.
37. Ke G., Meng Q., Finley T., Wang T., Chen W., Ma W., Ye Q., Liu T. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017. pp. 31496–3157. DOI: 10.5555/3294996.3295074.
38. Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay E. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011. vol. 12. pp. 2825–2830. DOI: 10.5555/1953048.2078195.
39. Chen T., Guestrin C. XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. pp. 785–794. DOI: 10.1145/2939672.2939785.
40. Lundberg S.M., Lee S.-I. A Unified Approach to Interpreting Model Predictions. 2017. arXiv preprint arXiv:1705.07874. DOI: 10.48550/arXiv.1705.07874.
41. Ribeiro M.T., Singh S., Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. pp. 1135–1144. DOI: 10.1145/2939672.293977.
42. Cevallos-Salas D., Grijalva F., Estrada-Jimenez J., Bentez D., Andrade R. Obfuscated Privacy Malware Classifiers Based on Memory Dumping Analysis. IEEE Access. 2024. vol. 12. pp. 17481–17498. DOI: 10.1109/ACCESS.2024.3358840.
43. Roy K.S., Ahmed T., Udas P.B. Karim M.E., Majumdar S. MalHyStack: A hybrid stacked ensemble learning framework with feature engineering schemes for obfuscated malware analysis. Intelligent Systems with Applications. 2023. vol. 20. DOI: 10.1016/j.iswa.2023.200283.

Imamverdiyev Yadigar — Ph.D., Dr.Sci., Head of the department, Cyber security department, Azerbaijan Technical University. Research interests: information security management systems, malware analysis, security of smart systems, security of industrial control systems, web security, cloud security, applied cryptography, biometrics, applications of AI in cyber security. The number of publications — 50. yadigar.imamverdiyev@aztu.edu.az; 25, H. Javid Av., AZ 1073, Baku, Azerbaijan; office phone: +994(50)540-7464.

Baghirov Elshan — Ph.D. candidate, Institute of Information Technology of The Ministry of Science and Education of the Azerbaijan Republic; Senior data scientist, Kapital Bank OJSC. Research interests: malware analysis, data science, information security incident management. The number of publications — 20. elsenbagirov1995@gmail.com; 5/13, A. Kunanbayev St., AZ 1009, Binagadi district, Baku, Azerbaijan; office phone: +994(51)444-1933.

Ikechukwu John Chukwu — Graduate student, Kadir Has University; Ss. Cyril and Methodius University in Skopje (UKIM). Research interests: public-key and lightweight cryptography, quantum optimization problems, malware analysis. The number of publications — 2. cikechukwujohn@stu.khas.edu.tr; Fatih, 34083, Istanbul, Turkey; office phone: +90(212)533-6532.

DOI 10.15622/ia.24.1.5

Я. Имамвердиев, Э. Багиров, Д. Икечукву

ОБНАРУЖЕНИЕ ОБФУСЦИРОВАННЫХ ВРЕДОНОСНЫХ ПРОГРАММ В WINDOWS С ПОМОЩЬЮ МЕТОДОВ АНСАМБЛЕВОГО ОБУЧЕНИЯ

Имамвердиев Я., Багиров Э., Икечукву Д. **Обнаружение обфусцированных вредоносных программ в Windows с помощью методов ансамблевого обучения.**

Аннотация. В эпоху Интернета и смарт-устройств обнаружение вредоносных программ стало важным фактором для безопасности системы. Обфусцированные вредоносные программы создают значительные риски для различных платформ, включая компьютеры, мобильные устройства и устройства IoT, поскольку не позволяют использовать передовые решения для обеспечения безопасности. Традиционные эвристические и сигнатурные методы часто не справляются с этими угрозами. Поэтому была предложена экономически эффективная система обнаружения с использованием анализа дампа памяти и методов ансамблевого обучения. На основе набора данных CIC-MalMem-2022 была оценена эффективность деревьев решений, градиентного бустинга деревьев, логистической регрессии, метода случайного леса и LightGBM при выявлении обфусцированных вредоносных программ. Исследование продемонстрировало превосходство методов ансамблевого обучения в повышении точности и надежности обнаружения. Кроме того, SHAP (аддитивные объяснения Шелли) и LIME (локально интерпретируемые объяснения, не зависящие от устройства модели) использовались для выяснения прогнозов модели, повышения прозрачности и надежности. Анализ выявил важные особенности, существенно влияющие на обнаружение вредоносных программ, такие как службы процессов, активные службы, дескрипторы файлов, ключи реестра и функции обратного вызова. Эти идеи имеют большое значение для совершенствования стратегий обнаружения и повышения производительности модели. Полученные результаты вносят вклад в усилия по обеспечению кибербезопасности путем всесторонней оценки алгоритмов машинного обучения для обнаружения обфусцированных вредоносных программ с помощью анализа памяти. В этой статье представлены ценные идеи для будущих исследований и достижений в области обнаружения вредоносных программ, прокладывая путь для более надежных и эффективных решений в области кибербезопасности перед лицом развивающихся и сложных вредоносных угроз.

Ключевые слова: обнаружение вредоносного ПО, машинное обучение, анализ вредоносного ПО, кибербезопасность.

Литература

1. Baghirov E. Evaluating the performance of different machine learning algorithms for Android malware detection. In 2023 5th International Conference on Problems of Cybernetics and Informatics (PCI). IEEE, 2023. pp. 1–4. DOI: 10.1109/PCI60110.2023.10326006.
2. Baghirov E. Comprehensive framework for malware detection: Using ensemble methods, feature selection, and hyperparameter optimization. In 2023 IEEE 17th International Conference on Application of Information and Communication Technologies (AICT). IEEE, 2023. pp. 1–5. DOI: 10.1109/AICT59525.2023.10313179.

3. Jeon J., Jeong B., Baek S., Jeong Y.-S. Static Multi Feature-Based Malware Detection Using Multi SPP-net in Smart IoT Environments. IEEE Transactions on Information Forensics and Security. 2024. vol. 19. pp. 2487–2500. DOI: 10.1109/TIFS.2024.3350379.
4. Ismail S.J.I., Hendrawan Rahardjo B., Juhana T., Musashi Y. MalSSL – Self-Supervised Learning for Accurate and Label-Efficient Malware Classification. IEEE Access. 2024. vol. 12. pp. 58823–58835. DOI: 10.1109/ACCESS.2024.3392251.
5. Baghirov E. Malware detection based on opcode frequency. Journal of Problems of Information Technology, 2023. vol. 14(1). pp. 3–7. DOI: 10.25045/jpit.v14.i1.01.
6. Egitmen A., Yavuz A.G., Yavuz S. TRConv: Multi-Platform Malware Classification via Target Regulated Convolutions. IEEE Access. 2024. vol. 12. pp. 71492–71504. DOI: 10.1109/ACCESS.2024.3401627.
7. Gungor A., Dogru I.A., Barisci N., Toklu S. Malware detection using image-based features and machine learning methods. Journal of the Faculty of Engineering and Architecture of Gazi University. 2023. vol. 38. no. 3. pp. 1781–1792. DOI: 10.17341/gazimmfd.994289.
8. Mesbah A., Baddari I., Riahla M.A. LongCGDroid: Android malware detection through longitudinal study for machine learning and deep learning. Jordanian Journal of Computers and Information Technology. 2023. vol. 9. no. 4. pp. 328–346. DOI: 10.5455/jjcit.71-1693392249.
9. Howard A., Hope B., Saltaformaggio B., Avena E., Ahmadi M., Duncan M., McCann R., Cukierski W. Microsoft Malware Prediction. Kaggle, 2018. Available at: https://kaggle.com/competitions/microsoft-malware-prediction. (accessed 26.10.2024).
10. Ahmed I.T., Hammad B.T., Jamil N.A Comparative Performance Analysis of Malware Detection Algorithms Based on Various Texture Features and Classifiers. IEEE Access. 2024. vol. 12. pp. 11500–11519. DOI: 10.1109/ACCESS.2024.3354959.
11. Xie W., Zhang X. The Application of Machine Learning in Android Malware Detection. 2024 4th International Conference on Neural Networks, Information and Communication Engineering (NNICE). 2024. pp. 1–4. DOI: 10.1109/NNICE61279.2024.10498936.
12. Bostani H.; Moonsamy V. EvadeDroid: A practical evasion attack on machine learning for black-box Android malware detection. Computers and Security. 2024. vol. 139. DOI: 10.1016/j.cose.2023.103676.
13. Rigaki M., Garcia S. The Power of MEME: Adversarial Malware Creation with Model-Based Reinforcement Learning. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2024. pp. 44–64. DOI: 10.1007/978-3-031-51482-1 _3.
14. Rudd E.M., Krisiloff D., Coull S., Olszewski D., Raff E., Holt J. Efficient Malware Analysis Using Metric Embeddings. Digital Threats: Research and Practice. 2024. vol. 5(1). pp. 1–20. DOI: 10.1145/3615669.
15. Zhan D., Zhang Y., Zhu L., Chen J., Xia S., Guo S., Pan Z. Enhancing reinforcement learning based adversarial malware generation to evade static detection. Alexandria Engineering Journal. 2024. vol. 98. pp. 32–43. DOI: 10.1016/j.aej.2024.04.024.
16. Aljabri M., Alhaidari F., Albuainain A., Alrashidi S., Alansari J., Alqahtani W., Alshaya J. Ransomware detection based on machine learning using memory features. Egyptian Informatics Journal. 2024. vol. 25. DOI: 10.1016/j.eij.2024.100445.
17. Ban Y., Kim M., Cho H. An Empirical Study on the Effectiveness of Adversarial Examples in Malware Detection. CMES – Computer Modeling in Engineering and Sciences. 2024. vol. 139(3). pp. 3535–3563. DOI: 10.32604/cmes.2023.046658.

18. Zhang Y., Jiang J., Yi C., Li H., Min S., Zuo R., An Z., Yu Y. A Robust CNN for Malware Classification against Executable Adversarial Attack. Electronics. 2024. vol. 13(5). DOI: 10.3390/electronics13050989.
19. Dam T.Q., Nguyen N.T., Le T.V., Le T.D., Uwizeyemungu S., Le-Dinh T. Visualizing Portable Executable Headers for Ransomware Detection: A Deep Learning-Based Approach. Journal of Universal Computer Science. 2024. vol. 30(2). pp. 262–286. DOI: 10.3897/jucs.104901.
20. Gibert D., Zizzo G., Le Q. Towards a Practical Defense Against Adversarial Attacks on Deep Learning-Based Malware Detectors via Randomized Smoothing. Lecture Notes in Computer Science. 2024. vol. 14399. pp. 683–699. DOI: 10.1007/978-3-031-54129-2_40.
21. Zhang P., Wu C., Wang Z. BINCODEX: A comprehensive and multi-level dataset for evaluating binary code similarity detection techniques. BenchCouncil Transactions on Benchmarks, Standards and Evaluations. 2024. vol. 4(2). DOI: 10.1016/j.tbench.2024.100163.
22. Gibert D., Zizzo G., Le Q., Planes J. Adversarial Robustness of Deep Learning-Based Malware Detectors via (De)Randomized Smoothing. IEEE Access. 2024. vol. 12. pp. 61152–61162. DOI: 10.1109/ACCESS.2024.3392391.
23. Louthanova P., Kozak M., Jurecek M., Stamp M., Di Troia F. A comparison of adversarial malware generators. Journal of Computer Virology and Hacking Techniques. 2024. vol. 20. pp. 623–639. DOI: 10.1007/s11416-024-00519-z.
24. Qian L., Cong L. Channel Features and API Frequency-Based Transformer Model for Malware Identification. Sensors. 2024. vol. 24(2). DOI: 10.3390/s24020580.
25. Surendran R., Uddin M.M., Thomas T., Pradeep G. Android Malware Detection Based on Informative Syscall Subsequences. IEEE Access. 2023. vol. 11. DOI: 10.1109/ACCESS.2024.3387475.
26. Kozak M., Jurecek M., Stamp M., Troia F.D. Creating valid adversarial examples of malware. Journal of Computer Virology and Hacking Techniques. 2024. vol. 20. pp. 607–621. DOI: 10.1007/s11416-024-00516-2.
27. Imran M., Appice A., Malerba D. Evaluating Realistic Adversarial Attacks against Machine Learning Models for Windows PE Malware Detection. Future Internet. 2024. vol. 16(5). DOI: 10.3390/fi16050168.
28. Saha S., Afroz S., Rahman A. H. MALIGN: Explainable static raw-byte based malware family classification using sequence alignment. Computers and Security. 2024. vol. 139. DOI: 10.1016/j.cose.2024.103714.
29. Li D., Cui S., Li Y., Xu J., Xiao F., Xu S. PAD: Towards Principled Adversarial Malware Detection Against Evasion Attacks. IEEE Transactions on Dependable and Secure Computing. 2024. vol. 21. no. 2. pp. 920–936. DOI: 10.1109/TDSC.2023.3265665.
30. Zhang F., Li K., Ren Z. Improving Adversarial Robustness of Ensemble Classifiers by Diversified Feature Selection and Stochastic Aggregation. Mathematics. 2024. vol. 12(6). DOI: 10.3390/math12060834.
31. Alzaidy S., Binsalleeh H. Adversarial Attacks with Defense Mechanisms on Convolutional Neural Networks and Recurrent Neural Networks for Malware Classification. Applied Sciences. 2024. vol. 14(4). DOI: 10.3390/app14041673.
32. Zhou K., Wang P., He B. Comparative Study: Mouth Brooding Fish (MBF) as a Novel Approach for Android Malware Detection. International Journal of Advanced Computer Science and Applications. 2024. vol. 15(5). DOI: 10.14569/IJACSA.2024.0150521.
33. Rakib H., Dhakal S.M. Obfuscated Malware Detection: Investigating Real-World Scenarios Through Memory Analysis. In 5th IEEE International

Conference on Telecommunications and Photonics (ICTP 2023). 2023. DOI: 10.1109/ICTP60248.2023.10490701.

34. Carrier T., Victor P., Tekeoglu A., Lashkari A.H. Detecting Obfuscated Malware using Memory Feature Engineering. Proceedings of the 8th International Conference on Information Systems Security and Privacy (ICISSP). 2022. vol. 1. pp. 177–188. DOI: 10.5220/0010908200003120.
35. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and prediction (2nd ed.). Springer, 2009. 745 p.
36. Friedman J.H. Greedy function approximation: A gradient boosting machine. Annals of Statistics. 2001. vol. 29(5). pp. 189–1232. DOI: 10.1214/aos/1013203451.
37. Ke G., Meng Q., Finley T., Wang T., Chen W., Ma W., Ye Q., Liu T. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017. pp. 31496–3157. DOI: 10.5555/3294996.3295074.
38. Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay E. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011. vol. 12. pp. 2825–2830. DOI: 10.5555/1953048.2078195.
39. Chen T., Guestrin C. XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. pp. 785–794. DOI: 10.1145/2939672.2939785.
40. Lundberg S.M., Lee S.-I. A Unified Approach to Interpreting Model Predictions. 2017. arXiv preprint arXiv:1705.07874. DOI: 10.48550/arXiv.1705.07874.
41. Ribeiro M.T., Singh S., Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. pp. 1135–1144. DOI: 10.1145/2939672.293977.
42. Cevallos-Salas D., Grijalva F., Estrada-Jimenez J., Bentez D., Andrade R. Obfuscated Privacy Malware Classifiers Based on Memory Dumping Analysis. IEEE Access. 2024. vol. 12. pp. 17481–17498. DOI: 10.1109/ACCESS.2024.3358840.
43. Roy K.S., Ahmed T., Udas P.B. Karim M.E., Majumdar S. MalHyStack: A hybrid stacked ensemble learning framework with feature engineering schemes for obfuscated malware analysis. Intelligent Systems with Applications. 2023. vol. 20. DOI: 10.1016/j.iswa.2023.200283.

Имамвердиев Ядигар — Ph.D., Dr.Sci., заведующий кафедрой, заведующий кафедрой кибербезопасности, Азербайджанский технический университет. Область научных интересов: системы управления информационной безопасностью, анализ вредоносных программ, безопасность интеллектуальных систем, безопасность промышленных систем управления, веб-безопасность, облачная безопасность, прикладная криптографию, биометрия, применение искусственного интеллекта в кибербезопасности. Число научных публикаций — 50. yadigar.imamverdiyev@aztu.edu.az; проспект Х. Джавида, 25, AZ 1073, Баку, Азербайджан; р.т.: +994(50)540-7464.

Багиров Эльшан — аспирант, Институт информационных технологий Министерства науки и образования Азербайджанской Республики; старший специалист по обработке данных, ОАО "Капитал Банк". Область научных интересов: анализ вредоносных программ, обработка данных, управление инцидентами информационной безопасности. Число научных публикаций — 20. elsenbagirov1995@gmail.com; улица А. Кунанбаева, 5/13, AZ 1009, Бинагадинский район, Баку, Азербайджан; р.т.: +994(51)444-1933.

Икечукву Джон Чукву — аспирант, Университет Кадир Хас; Университет святых Кирилла и Мефодия в Скопье (UKIM). Область научных интересов: криптография с открытым ключом и облегченная криптография, проблемы квантовой оптимизации, анализ вредоносных программ. Число научных публикаций — 2. cikechukwujohn@stu.khas.edu.tr; Фатих, 34083, Стамбул, Турция; р.т.: +90(212)533-6532.

УДК 004.056 **DOI** 10.15622/ia.24.1.6

В.Е. Боровков, П.Г. Ключарёв, Д.И. Денисенко

МЕТОДИКА ОЦЕНИВАНИЯ РЕЗУЛЬТАТИВНОСТИ ФУНКЦИОНИРОВАНИЯ СИСТЕМ ОБНАРУЖЕНИЯ ВЕБ-БЭКДОРОВ

Боровков В.Е., Ключарёв П.Г., Денисенко Д.И. **Методика оценивания результативности функционирования систем обнаружения веб-бэкдоров.**

Аннотация. В настоящее время наблюдается значительный рост инцидентов информационной безопасности, связанных с атаками на веб-ресурсы. Получение несанкционированного доступа к веб-ресурсам остается одним из основных методов проникновения в корпоративные сети организаций и расширения возможностей злоумышленников. В связи с этим множество исследований направлено на разработку систем обнаружения веб-бэкдоров (СОВБ), однако существует задача оценивания результативности функционирования данных систем. Цель данного исследования заключается в разработке объективного подхода для оценки результативности функционирования СОВБ. В данной работе было установлено, что объективно результативность СОВБ проявляется в процессе их использования, поэтому тестирование таких систем необходимо проводить в условиях, максимально приближенных к реальным. В связи с этим в статье предложена методика оценивания результативности функционирования СОВБ. Она основана на расчете трех групп частных показателей, характеризующих действенность, ресурсоемкость и оперативность работы системы обнаружения, а также вычислении обобщенного показателя результативности. На основе анализа исследований в данной области была составлена классификация веб-бэкдоров, встраиваемых злоумышленником в исходный код веб-приложений. Эта классификация используется при формировании тестовых наборов данных для вычисления частных показателей действенности. Разработанная методика применима для СОВБ, которые работают на основе анализа исходного кода веб-страниц. Также для ее использования необходим ряд исходных данных, таких как допустимые предельные ошибки частных показателей действенности и вероятность нахождения их в доверительном интервале, а также весовые коэффициенты частных показателей действенности, которые подбираются экспертными методами. Данная работа может быть полезной для специалистов и исследователей в области информационной безопасности, которые хотят проводить объективную оценку своих СОВБ.

Ключевые слова: кибербезопасность, веб-уязвимости, веб-бэкдоры, веб-шеллы, машинное обучение, методы и средства тестирования.

1. Введение. В настоящее время наблюдается увеличение числа компьютерных инцидентов, включая успешные атаки на веб-ресурсы организаций. Как пишут эксперты из компании Positive Technologies [1], захват злоумышленником веб-сайта приводит к расширению его возможностей – от искажения сайта до полной компрометации сети [2]. Так в 2022 году инциденты с веб-ресурсами приводили к нарушениям деятельности организаций в 53% случаях [1], а в 2023 году число атак на веб-ресурсы возросло на 40% по сравнению с аналогичным периодом 2022 года [3].

Тенденция указывает на то, что кибератаки на веб-приложения будут расти как по количеству, так и по уровню сложности. В связи с этим кибербезопасность должна постоянно развиваться и усовершенствоваться, чтобы обеспечить их надежную защиту. Небрежное отношение к обеспечению безопасности веб-приложений может привести к утечке конфиденциальных данных, а также к финансовым и репутационным потерям для организации.

Веб-сайт зачастую не является конечной целью злоумышленника, он может использоваться для проведения атак во внутренней сети организации, а для «закрепления» на веб-ресурсе часто используются веб-бэкдоры. Также в работе [4] было выявлено, что многие исследователи при разработке интеллектуальных методов защиты веб-приложений от уязвимостей, а также веб-бэкдоров, не проводили проверку своих результатов в реальных условиях. Отсюда следует, что результаты их проверки на собственных тестовых наборах могут быть необъективными. Тем самым у специалистов по информационной безопасности (ИБ) может возникнуть вопрос о выборе наиболее результативной системы обнаружения вэб-бэкдоров. Согласно [5] под «результативностью» будем понимать степень реализации запланированной деятельности и достижения запланированных результатов. В контексте систем обнаружения веб-бэкдоров (СОВБ) результативность проявляется в их способности обнаруживать веб-бэкдоры.

Настоящая статья построена следующим образом: в разделе 2 рассмотрены существующие методики тестирования систем обнаружения и защиты от вредоносного программного обеспечения (ВПО); в разделе 3 представлен пример влияния различных наборов тестовых данных на результаты оценивания СОВБ; в разделах 4-6 предлагается методика оценивания результативности функционирования СОВБ, апробация которой представлена в разделе 7.

2. Существующие методики тестирования систем обнаружения и защиты от вредоносного программного обеспечения. *Веб-бэкдор* является скрытым механизмом, который позволяет злоумышленнику получать удаленный несанкционированный доступ к веб-серверу для выполнения различных операций. Следует отметить, что существует множество разновидностей веб-бэкдоров, которые могут быть встроены в веб-приложения, а также в веб-сервера (например, *Apache-бэкдоры*, *Nginx*-бэкдоры [6]) и т.д. Хотя некоторые уязвимости веб-приложений или веб-серверов могут представлять собой веб-бэкдоры, в данной статье под последними понимается вредоносные сценарии (такие как веб-шеллы, веб-загрузчики и прочее [4]),

встраиваемые злоумышленником в исходный код веб-приложения на интерпретируемом языке программирования. Но для встраивания таких бэкдоров в веб-приложение злоумышленник должен иметь возможность изменения файлов на сервере. Это может быть достигнуто различными способами, в том числе через уязвимости веб-приложения [7]. В любом случае веб-бэкдор будет являться ВПО.

Существует множество факторов, влияющих на результаты тестирования систем обнаружения и защиты от ВПО. Различные методы тестирования могут давать разные результаты, так как они могут не учитывать все возможные варианты атак. Кроме того, качество обновлений и скорость реакции на новые угрозы также могут влиять на результаты тестирования. Наконец, каждый пользователь имеет свои собственные потребности и предпочтения, которые могут повлиять на выбор лучшей системы обнаружения и защиты от ВПО для его конкретных нужд. Однако в отношении СОВБ не было обнаружено четкой методики оценивания их результативности, исходя из анализа документов, представленных в [4, 8].

В настоящее время существуют следующие методики тестирования систем обнаружения и защиты от ВПО:

1) *On-demand scan (ODS).* Представляет собой статический анализ ВПО с помощью систем защиты [9]. Статический анализ подразумевает проверку ВПО без его фактического выполнения; например, система обнаружения может применять сигнатурный метод для выявления ВПО. В этой методике предполагается, что результативность системы обнаружения определяется количеством обнаруженных ВПО в процессе статического анализа. Для проведения качественного тестирования важно использовать широкий спектр вредоносных программ, который состоит из более чем тысячи файлов и документов. Некоторые специализированные организации предлагают такие коллекции. Однако использование только этих тестов не всегда дают объективную оценку результативности из-за возможности злоумышленников применять обфускацию и шифрование, а также комбинировать их с другими методами, которые позволяют обходить статический анализ.

2) *Динамическое тестирование (Тестирование с запуском).* Суть данной методики заключается в запуске вредоносных средств в виртуальной среде и их анализе системами защиты (также называется *эвристическим анализом*). Вместе со статическим анализом этот подход может быть результативным, но у него есть существенный недостаток: вредоносные программы могут представлять только часть цепочки атаки, и для их полноценной

работы могут потребоваться дополнительные условия и параметры, которые виртуальная среда может не предоставить. Также ВПО способно определять среду выполнения и не реализовывать свои вредоносные функции.

3) *Real-world test (RW)* [10]. Данная методика является наиболее сложной, так как представляет собой моделирование полного цикла заражения реальной системы и анализ реакции системами защиты. Такие тесты позволяют выявить различные проблемы, с которыми программное обеспечение безопасности может столкнуться при работе в реальных условиях с реальными угрозами. Однако главным недостатком является сложность создания лабораторной среды. Для достоверных результатов тестирования лабораториям приходится использовать физические компьютеры, а не виртуальные машины, что требует постоянного восстановления систем после каждого запуска ВПО.

Данные методики тестирования наиболее распространены и охватывают широкий спектр вредоносных программ [11], что позволяет проверить большое количество систем обнаружения и защиты от ВПО. Их сравнительная характеристика представлена в таблице 1.

Таблица 1. Сравнительная характеристика методов тестирования систем обнаружения и защиты от ВПО

№ п/п	Наименование методики	Достоинства	Недостатки
1	ODS-тестирование	– относительная простота реализации	– необходимо иметь большую базу ВПО – не позволяет оценивать системы обнаружения и защиты, основанные на эвристическом анализе
2	Динамическое тестирование	– позволяет оценивать системы обнаружения и защиты, основанные на поведенческом анализе	– необходимость создания виртуальной среды – виртуальная среда может не в полной степени создать условия для успешного развертывания ВПО
3	RW-тестирование	– позволяет выявить недостатки используемых в системах обнаружения и защиты алгоритмов статического и эвристического анализа – позволяет имитировать различные виды атак	– сложность создания лабораторной среды для проведения тестирования

Кроме того, имеются другие методы, которые направлены на проверку производительности, скорости реакции и других характеристик систем обнаружения и защиты от вредоносных программ. Однако, если необходимо протестировать систему, работающую в узконаправленной области (например, обнаружение веб-бэкдоров), требуются специализированные тесты.

3. Реакция СОВБ на различные наборы веб-бэкдоров. Существуют разные методы для выявления веб-бэкдоров, которые основываются на различных принципах, включая анализ файлов сайта, логов веб-сервера, HTTP-трафика и другие. При тестировании СОВБ необходимо учитывать эти принципы. В данной статье рассмотрим СОВБ, работающих по принципу анализа исходного кода веб-приложений. Многие методы обнаружения используют именно этот подход [8]. Одной из основных проблем при тестировании таких систем является недостаток качественных или полных наборов тестовых данных, а также игнорирование некоторыми исследователями возможных модификаций веб-бэкдоров, которые позволяют обойти системы защиты и обнаружения [4]. Для подтверждения этого, в качестве примера был рассмотрен один из методов обнаружения веб-бэкдоров, основанный на сверточной нейронной сети [12]. Алгоритм [13] производит анализ исходных кодов набора PHP-страниц, чтобы получить последовательность инструкций операционных кодов (опкодов). Затем используется *Word2vec* [14] для получения векторной карты для массивов опкодов и для массивов биграмм опкодов. Наконец, они служат двумя входами сверточной нейронной сети, которая осуществляет обнаружение. Пример того, как выглядят опкоды языка PHP, представлен на рисунке 1. Похожие идеи использовали также исследователи в работе [15].

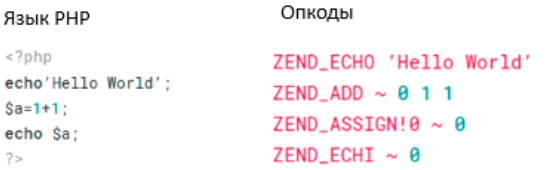


Рис. 1. Опкоды языка PHP

Исследователь утверждает, что алгоритм достигает точности (*accuracy*) 98,4%, что подтверждается нами при проверке на тестовом наборе данных, который использовался исследователем для тестирования.

Были выбраны 6 различных веб-бэкдоров, которые подверглись анализу с помощью алгоритма обнаружения. Среди них следующие веб-бэкдоры:

1) Веб-загрузчик. Представляет собой обычную форму загрузки файлов через два POST параметра: первый параметр задает имя файла, который необходимо создать, а второй – текст в кодировке base64, который нужно внести в этой файл.

2) Веб-загрузчик. Загрузка файлов через два POST параметра: первый параметр – директория сохранения файла, второй – сам файл.

3) Веб-шелл, который принимает через POST параметр закодированную в *Base64* команду и выполняет через функцию *eval(system(“command”)).*

4) Веб-шелл, который принимает через GET параметр команду и выполняет через функцию *system(“command”).*

5) Веб-бэкдор, принимающий на вход IP-адрес и порт атакующего и реализующего *Reverse Shell* [16] к злоумышленнику.

6) Большой веб-шелл, который имеет графический интерфейс и выполняет различные операции. Его код представлен на *Github*[17].

Очевидно, что различных вариаций вэб-бэкдоров может быть множество, в данном случае были выбраны простые варианты, которые часто используются злоумышленниками.

В первом случае веб-бэкдоры были представлены в виде отдельных файлов и использовались в качестве входных данных для алгоритма обнаружения с целью прогнозирования. В итоге алгоритм допустил ошибку лишь в одном из шести случаев (для третьего варианта), определив его как легитимный файл (файл, не содержащий веб-бэкдора).

Во втором случае эти же веб-бэкдоры встраивались в легитимный файл. Тем самым легитимные файлы становились носителями вэб-бэкдоров, и должны помечаться СОВБ, как вэб-бэкдоры. Для примера был взят файл *wp-login.php* из популярного фреймфорка *WordPress* [18]. Важным условием являлось то, чтобы веб-страница и веб-бэкдор работали корректно, иначе алгоритм не сможет вычислить операционные коды скриптов. В результате из шести веб-бэкдоров алгоритм верно определил только в одном случае – для загрузчика 1. В остальных случаях алгоритм принял неправильное решение, указав, что это легитимные файлы.

Результаты классификации алгоритмом веб-бэкдоров представлены в таблице 2.

Таблица 2. Воздействие модификации веб-бэкдоров на результативность алгоритма обнаружения

	Номер веб-бэкдора					
	1	2	3	4	5	6
Веб-бэкдор отдельным файлом	+	+	-	+	+	+
Веб-бэкдор встраивается в легитимный файл	+	-	-	-	-	-

Отсюда можно сделать вывод, что незначительная манипуляция с кодом веб-бэкдора может существенно повлиять на результаты алгоритмов обнаружения. В данном случае веб-бэкдоры были внедрены в легитимные файлы сайта, а не являлись отдельными файлами. Также конкретные способы обхода средств обнаружения можно увидеть в документах [19 – 20]. Это поднимает вопрос о том, насколько точность алгоритма, оцененная на тестовом наборе данных (98.4%), отражает его способность обнаруживать 98.4% реально существующих веб-бэкдоров. Метрики точности (*accuracy, precision*) и полноты (*recall*) алгоритмов обнаружения зависят от конкретных наборов данных и поэтому не всегда являются надежными показателями способности алгоритма обнаруживать веб-бэкдоры, поскольку исследование возможных вариантов веб-бэкдоров не проводилось, а сам алгоритм (или СОВБ) не тестировался в реальных условиях.

***Примечание**. В англоязычной литературе метрики precision и accuracy имеют различные значения, определяемые формулами (5) и (6) соответственно, при этом перевод у них одинаковый – «точность», также accuracy иногда называют «меткостью».*

Тем самым для ответа на следующие вопросы: «как сравнить несколько систем обнаружения веб-бэкдоров и выбрать наиболее результативный?» и «какие наборы данных следует использовать для объективного расчета показателей результативности?», необходимо разработать методику, которую могут использовать специалисты и исследователи в области ИБ для объективной оценки результативности функционирования СОВБ.

4. Постановка задачи. Цель работы – разработка объективного подхода для оценки результативности функционирования СОВБ. Для этого требуется разработать методику оценивания результативности функционирования СОВБ. Ввиду того, что СОВБ могут работать по различным принципам (анализ HTTP-пакетов, анализ логов системы, анализа исходного кода веб-страниц и др.) выделим исходное ограничение, что оцениваемые системы работают на основе анализа

исходного кода веб-страниц. Также предполагается, что изначально известна среда функционирования СОВБ: операционная система (ОС) и аппаратные характеристики устройства, где развернута СОВБ, такие как объем оперативной памяти, модель процессора, объем видеопамяти (в случае использования СОВБ видеокарты), а также известна нагрузка на СОВБ (средний предполагаемый объем анализируемых файлов).

Согласно базовым понятиям теории и методологии внешнего проектирования целенаправленных процессов и целеустремленных систем [21] можно определить, что объективно результативность СОВБ проявляется в процессе их использования. *Результативность,* является основным свойством системы. Также к ряду основных свойств могут относить *ресурсоемкость* (затраты ресурсов системой) и *оперативность* (затраты времени, скорость реакции).

Количественной мерой результативности может выступать обобщённый показатель результативности системы, который характеризует результат ее функционирования при заданных характеристиках ее состояния и определённых внешних воздействиях. Он может быть представлен с помощью формулы (1):

$$E = \sum_{i=1}^{m} e_i w_i , \qquad (1)$$

где E – обобщенный показатель результативности СОВБ, e_i – значение i-го частного показателя действенности СОВБ, при этом $0 \le e_i \le 1$, w_i – весовой коэффициент i-го частного показателя и $\sum_{i=1}^{m} w_i = 1$, $i = \overline{1, m}$ – количество частных показателей действенности.

Выбор и расчет частных показателей действенности зависит от специфики работы системы. Для получения значений частных показателей СОВБ можно применять метод, основанный на проведении испытаний на тестовом стенде, который максимально приближен к реальным условиям эксплуатации системы [22].

Оценивание возможно только в системе, которой предъявлены требования. Целью оценивания является выработка суждения о СОВБ на основе полученных (измеренных) показателей. Такое суждение формируется с помощью определенных правил и принципов, которые выражены в форме критериев оценивания. Обоснованный выбор наиболее подходящих решений и средств осуществляется посредством

сравнения полученных в ходе испытаний результатов с заданными требованиями.

Оценивание результативности СОВБ можно разбить на следующие этапы:

1. Формирование критериев на основе требований к СОВБ. Требования задаются исследователем.
2. Составление сценария проведения эксперимента исследователем.
3. Создание лабораторной среды для проведения эксперимента, подготовка тестовых наборов данных исследователем.
4. Расчет (получение) значений частных показателей.
5. Анализ полученных значений частных показателей и принятие решения о пригодности СОВБ.
6. Получение значения обобщённого показателя результативности.
7. Анализ полученных результатов, сравнение СОВБ, принятие решений исследователем.

Далее рассмотрим этапы оценивания более подробно.

Этап 1.

Системы, предназначенные для обнаружения веб-бэкдоров, могут играть важную роль в обеспечении безопасности веб-приложений. Однако такие системы должны не только выполнять свою основную функцию, но и соответствовать определенным требованиям:

1) не нарушать работоспособность веб-приложений;
2) осуществлять обнаружение веб-бэкдоров за адекватное время;
3) не становиться основным потребителем ресурсов на сервере и не приводить к снижению производительности веб-приложений.

Для того, чтобы учесть эти требования, установим три группы частных показателей, каждая из которых отражает действенность, ресурсоемкость (затраты ресурсов) и оперативность (затраты времени) функционирования СОВБ. Все три группы частных показателей будем использовать, только для определения пригодности СОВБ на этапе 5. Для этапов 6-7 – вычисление интегрального показателя результативности, сравнение и выбор подходящей СОВБ – будут использоваться только частые показатели действенности.

Определим вектор частных показателей, который требуется вычислить (получить), как $Y_{\langle n \rangle} = \langle y_1, y_2, ..., y_n \rangle = \langle e_1, e_2, ..., e_{n1},$

$r_1, r_2, ..., r_{n2}, t_1, t_2, ..., t_{n3}\rangle$, где $e_1, e_2, ..., e_{n1}$ – показатели, характеризующие действенность, $r_1, r_2, ..., r_{n2}$ – показатели, характеризующие ресурсоемкость, $t_1, t_2, ..., t_{n3}$ – показатели, характеризующие оперативность функционирования СОВБ, а $n = n1 + n2 + n3$ – общее количество частных показателей.

При вычислении частных показателей действенности получаются точечные значения. Для того чтобы делать содержательные выводы, необходимо находить доверительный интервал, т.е. интервал, который с заданной вероятностью накрывает значение частного показателя. К факторам, влияющим на ширину доверительного интервала, относятся размер выборки, изменчивость выборки и доверительная вероятность нахождения показателя в данном интервале. Поэтому сформируем критерий достаточности тестового набора данных (наборы зараженных файлов и легитимных файлов) для оценки частных показателей действенности СОВБ (S). Пусть $\Delta_e^d = \langle \Delta_{e1}^d, \Delta_{e2}^d, ..., \Delta_{n1}^d \rangle$ – вектор допустимых предельных ошибок вычисления частных показателей действенности $e_1, e_2, ..., e_{n1}$, а $\Delta_e = \langle \Delta_{e1}, \Delta_{e2}, ..., \Delta_{en1} \rangle$ – вектор предельных ошибок вычисления частных показателей $e_1, e_2, ..., e_{n1}$, полученный после проведения эксперимента, исходя из объема тестового набора данных. Тогда критерий будет выглядеть следующим образом:

$$S : (\Delta_{e_i} \le \Delta_{e_i}^d) \cong U, i = 1, .., n1, \tag{2}$$

где $\cong$ – знак равносильности высказываний, U – достоверное событие (истинное высказывание), $n1$ – количество частных показателей действенности.

Вектор допустимых предельных ошибок Δ_e^d и доверительная вероятность являются исходными данными. Они задаются экспертными методами.

На основании требований к системе обнаружения можно выделить область допустимых значений частных показателей. Пусть $\{y_i^d\}$ – множество (область) допустимых значений показателя y_i. Или в векторной форме $\{Y_{\langle n \rangle}^d\} = \{\langle y_1^d, y_2^d, ..., y_n^d \rangle\}$. Тогда критерий пригодности для СОВБ (G) будет выглядеть следующим образом [21]:

$$G:(Y_{\langle n\rangle}\in\left\{Y^d_{\langle n\rangle}\right\})\cong U, \tag{3}$$

где $\left\{Y^d_{\langle n\rangle}\right\}$ – область допустимых значений частных показателей.

Для вычисления обобщенного показателя результативности по формуле (1) также необходимо задать вектор весовых коэффициентов частных показателей действенности $W_e=\left\langle w_{e_1}, w_{e2},..., w_{en1}\right\rangle$. Весовые коэффициенты W_e также задаются изначально экспертными методами на основе приоритетности тех или иных частных показателей действенности.

Этап 2.

Для получения значений показателей $y_1, y_2,..., y_n$ СОВБ необходимо разработать сценарий проведения эксперимента. В сценарии поясняется, как и каким образом будут получены значения частных показателей.

Этап 3.

На основе сценария проведения эксперимента создается лабораторный стенд, а также подготавливается наборов тестовых данных – легитимных файлов и веб-бэкдоров. Формирование наборов данных является одним из основных шагов для вычисления частных показателей действенности. Особое внимание уделяется формированию набора тестовых данных для различных вариантов веб-бэкдоров: $D^1_{m1}, D^2_{m2},..., D^k_{mk}$, где $D^i_{mi}=\left\{d^i_1, d^i_2,..., d^i_{mi}\right\}$, i указывает на вид веб-бэкдора, mi – количество различных вариантов для веб-бэкдора i-го вида.

Этапы 4, 5.

На данных этапах проводятся испытания и вычисляются частные показатели на основе проведенного эксперимента. Также рассчитывается вектор предельных ошибок частных показателей действенности Δ_e. Если критерий достаточности набора тестовых данных S (2) не выполняется, то принимается решение на формирование большего набора тестовых данных. Затем эксперимент проводится повторно. Если же критерий S выполняется, то на основе критерия пригодности G (3) исключаются те СОВБ, которые не соответствуют требованиям.

Этапы 6, 7.

С использованием весовых коэффициентов W_e рассчитывается результативность с помощью формулы (1) для каждой СОВБ, участвующей в исследовании и удовлетворяющей критериям S и G. Затем принимается решение о выборе наилучшей системы обнаружения, которая имеет наибольшее значение результативности.

Как видно из перечисленных этапов, процесс оценивания основывается на вычислении частных показателей $Y_{\langle n \rangle} = \langle y_1, y_2, ..., y_n \rangle$ с использованием тестовых наборов данных в ходе проведения эксперимента, а также на вычислении результативности E. Согласно уравнению (1) $0 \le E \le 1$, и чем выше значение E, тем результативнее работает СОВБ.

Далее сформируем вектор частных показателей для методики оценивания результативности СОВБ и рассмотрим процесс составления тестовых наборов данных для расчета частных показателей.

5. Формирование вектора частных показателей. Для того, чтобы всецело охватить совокупность требований к системам обнаружения, в предыдущем разделе был введен вектор $Y_{\langle n \rangle}$. Действенность системы обнаружения веб-бэкдоров отображена в элементах $e_1, e_2, ..., e_{n1}$. В данной работе были выбраны три показателя, представленные в выражениях (4-6). В их основе лежат следующие базовые переменные, которые присущи бинарному классификатору:

TP – истинно положительные классификации, т.е. случаи, когда система правильно обнаружила веб-бэкдор;

TN – истинно отрицательные классификации, т.е. случаи, когда система правильно определила отсутствие веб-бэкдора;

FP – ложноположительные классификации, т.е. случаи, когда система неправильно определила наличие веб-бэкдора (ошибка первого рода);

FN – ложноотрицательные классификации, т.е. случаи, когда система неправильно определила отсутствие веб-бэкдора (ошибка второго рода).

$$recall : e_1 = Re, \text{ где } Re = \frac{TP}{TP + FN}, \tag{4}$$

$$precision : e_2 = Pr, \text{ где } Pr = \frac{TP}{TP + FP}, \tag{5}$$

$$accuracy: e_3 = Ac, \text{где } Ac = \frac{TP + TN}{TP + TN + FP + FN}. \quad (6)$$

Частный показатель $e_1 = Re$ отражает долю веб-бэкдоров, обнаруженных СОВБ, от общего числа веб-бэкдоров в тестовом наборе данных. При этом $0 \le e_1 \le 1$.

Частный показатель $e_2 = Pr$ отражает долю объектов, классифицированных СОВБ как веб-бэкдоры, и при этом действительно являющимися таковыми. При этом $0 \le e_2 \le 1$.

Частный показатель $e_3 = Ac$ отражает долю правильных классификаций СОВБ. При этом $0 \le e_3 \le 1$.

Эксперимент проводятся на наборе легитимных файлов, а также на наборах веб-бэкдоров $D^1_{m1}, D^2_{m2}, ..., D^k_{mk}$, которые будут рассмотрены подробнее в следующем разделе.

Следующую группу показателей, характеризующую затраты ресурсов на работу СОВБ, получим из загруженности оперативной памяти, видеокарты и процессора на выполнение операций:

$$r_1 = V_{\text{оп}}, r_2 = L_{\text{цп}}, r_3 = V_{\text{гп}},$$

где $V_{\text{оп}}$ – максимальный объем оперативной памяти, потребляемый системой (в Мбайтах/Гбайтах), $L_{\text{цп}}$ – максимальная загруженность процессора данной системой (в процентах), $V_{\text{гп}}$ – максимальный объем видеопамяти, используемый системой (в Мбайтах/Гбайтах), за время функционирования СОВБ. Частный показатель r_3 актуален для систем, использующих машинное обучение для выполнения обнаружения.

Также немаловажным аспектом работы СОВБ является третья группа показателей, которая характеризует оперативность. При исследовании СОВБ, основанных на анализе содержимого веб-файлов, можно использовать один показатель $t_1 = \tau$, где τ – среднее время анализа содержимого файла. Его можно вычислить, разделив общее время анализа файлов на количество проанализированных файлов.

Следует отметить, что показатели ресурсоемкости ($r_1, r_2, ..., r_{n2}$) и оперативности ($t_1, t_2, ..., t_{n3}$) зависят от технических особенностей

программной среды, где развернута СОВБ, а также от нагрузки на СОВБ. Поэтому при расчете данных показателей важно указывать использованные технические особенности среды (версия ОС, количество оперативной памяти, видеопамяти, процессор) и объем нагрузки на СОВБ (количество проанализированных файлов).

Таким образом, вектор частных показателей для разрабатываемой методики будет иметь следующий вид: $Y_{\langle 7\rangle} = \langle e_1, e_2, e_3, r_1, r_2, r_3, t_1\rangle = \langle Re, Pr, Ac, V_{\text{оп}}, L_{\text{цп}}, V_{\text{гп}}, \tau\rangle$.

6. Подготовка тестовых наборов данных для расчета частных показателей

6.1. Классификация веб-бэкдоров, встраиваемых в исходный код веб-приложения. Одним из важных этапов при оценивании СОВБ является подготовка тестовых наборов.

Тестовые наборы данных для получения значений частных показателей ресурсоемкости и оперативности формируются за счет исходных данных (объем файлов анализируемого веб-приложения). Эти показатели характеризуют затраты ресурсов (оперативной памяти, процессора, видеокарты) и среднее время анализа файла. Поэтому наличие или отсутствие веб-бэкдоров в тестовых наборах данных не влияет на значение данных показателей.

Чтобы получить объективные значения частных показателей действенности, необходимо использовать различные варианты вэб-бэкдоров, которые могут быть внедрены злоумышленниками в исходный код веб-приложений. Для этого были изучены тестовые наборы, упомянутые в работах [15, 23 – 26]. Анализ данных тестовых наборов данных показал, что веб-бэкдоры, встроенные в исходный код веб-приложений, могут выполнять различные функции и делятся на 4 вида: выполнение команд операционной системы; выполнение команд языка программирования веб-приложения; загрузка файлов (других веб-бэкдоров); выполнение специализированных операций (к примеру выполнение Reverse Shell к злоумышленнику, создание веб-прокси и т.д.). Код веб-бэкдора не всегда содержится в одном файле, также злоумышленник может использовать различные методы для его сокрытия, такие как шифрование, обфускация или внедрение в легитимные файлы. Для выполнения различных функций злоумышленник может передавать разные значения в параметрах HTTP-запросов или не передавать ничего, если в коде веб-бэкдора уже содержатся все необходимые значения.

На основе этого анализа была разработана классификация веб-бэкдоров (рисунок 2).

Рис. 2. Классификация веб-бэкдоров для создания тестовых наборов данных

Каждый такой веб-бэкдор можно описать с помощью пяти независимых характеристик. К примеру, веб-бэкдор может обладать следующими свойствами:

1. дает возможность выполнять команды операционной системы (блок 1.2, рисунок 2);
2. код веб-бэкдора находится в нескольких файлах (блок 2.2);
3. веб-бэкдору передаются необходимые параметры в теле запроса HTTP (блок 3.1);
4. в коде веб-бэкдора используется обфускация (блок 4.1);
5. код веб-бэкдора не встраивается в легитимные файлы (блок 5.1).

Путем перебора всех возможных вариантов получается 64 вида различных веб-бэкдоров.

В рамках отдельной характеристики возможно объединение, например, веб-бэкдор может выполнять как команды языка программирования веб-приложения (1.1), так и операционной

системы (1.2). Однако в контексте тестовых наборов целесообразно не использовать такое объединение.

При формировании наборов данных воспользуемся выборкой один к одному: один веб-бэкдор к одному легитимному файлу. Это необходимо для того, чтобы одинаково учитывать реакцию СОВБ как на веб-бэкдоры, так и на легитимные файлы при последующем расчете показателей действенности, так как показатель $e_з = Ac$ (6) бесполезен в задачах с неравными классами [27]. Тем самым в выборке будет $N_{вб}$ веб-бэкдоров и $N_л$ легитимных файлов, при этом $N_{вб} = N_л = N$. В свою очередь, исходя из классификации, представленной на рисунке 2, пространство веб-бэкдоров можно разделить на 64 группы. Набор веб-бэкдоров будем формировать на основе типической выборки, пропорциональной объему типических групп, предполагая, что веб-бэкдоры равнозначно распределены по этим группам (имеют одинаковые объемы групп). Таким образом, при отборе существует два набора – веб-бэкдоры и легитимные файлы – одинаковые по объему. В свою очередь веб-бэкдоры разделяются на 64 группы. Например, если использовать отбор из 256 элементов – 128 из них легитимные элементы, а 128 – веб-бэкдоры. При этом вэб-бэкдоры представлены набором $D^1_{m1}, D^2_{m2}, ..., D^k_{mk}$, где $m1, m2, ..., mk = 2$ (по две реализации на каждый вид веб-бэкдора), а $k = 64$, т.е. $D^1_2, D^2_2, ..., D^{64}_2$, где $D^i_2 = \left\{d^i_1, d^i_2\right\}$.

Был создан набор PHP-файлов, который можно использовать для расчета показателей действенности СОВБ. Этот набор файлов доступен в репозитории на *Github* [28]. При принятии решения о выборе языка программирования был учтен факт, что PHP является наиболее распространенным языком для создания различных типов веб-приложений [29]. Для имитации веб-приложения был использован фреймворк *WordPress*. Также в репозитории содержатся примеры команд для проверки работоспособности веб-бэкдоров и создания запросов к ним. Так в начале подраздела был представлен веб-бэкдор с определёнными свойствами. Его реализацию можно увидеть на рисунке 3. Веб-бэкдор состоит из двух файлов: *25.php* и *25.1.php*. Оба этих файла обфусцированы, что затрудняет их анализ. Для реализации логики используется метод *goto*, который позволяет перескакивать между инструкциями. Обфускация достигается за счет применения множества неочевидных символов, сокращений и необычных названий переменных, что делает код менее читаемым и усложняет его понимание.

Файл 25.php

```
<?php
/*   ____________________________________________
     |  Obfuscated by YAK Pro - Php Obfuscator  2.0.14  |
     |    GitHub: https://github.com/pk-fr/yakpro-po    |
     |__________________________________________________|
*/
goto jhXrD;
gyobV: require "\x77\x70\x2d\160\x6f\x6c\x69\x6e\56\x70\x68\x70";
goto TH36a; jhXrD: require "\62\65\56\x31\56\160\150\x70"; goto iM_oz;
TH36a: wp_run($_POST["\167\x70\55\160\x6f\x6c\151\x6e"]); goto T1S31;
iM_oz: sis_create(); goto gyobV;
T1S31: unlink("\167\x70\x2d\x70\157\x6c\x69\x6e\x2e\x70\150\160");
```

Файл 25.1.php

```
<?php function sis_create(){$xafq_0=fopen('wp-polin.php','w');
$xemy_1="<?php function ";$iosx_2="wp_run(\$bd){pas";$eduj_3="sth";
$syeg_4="ru(\$bd);}?>";fputs($xafq_0,$xemy_1.$iosx_2.$eduj_3.$syeg_4);}?>
```

Рис. 3. Пример реализации веб-бэкдора

Таким образом, при разработке сценария эксперимента расчет можно осуществлять в два этапа. На первом этапе следует определить значения частных показателей оперативности и ресурсоемкости для всего доступного объема скриптов, основываясь на среднем предполагаемом объеме анализируемых файлов, независимо от наличия или отсутствия в них бэкдора. На втором этапе проводятся расчеты частных показателей действенности. Данные для анализа представляют собой легитимные файлы и веб-бэкдоры, сформированные по указанной выше выборке.

6.2. Определение требуемого объема выборки и предельных ошибок частных показателей действенности. Для получения объективной оценки тестирования рекомендуется расширить набор, который используется для расчета частных показателей действенности.

Необходимо отметить, что, согласно выборке, которая была предложена в подразделе 6.1, а также по определению переменных TP, TN, FP, FN из раздела 5, следует, что

$$TP + FN = N_{\text{вб}} = N, \tag{7}$$

$$TN + FP = N_{\text{л}} = N\,. \tag{8}$$

Отсюда получаем, что результат анализа легитимных объектов не влияет на значения TP и FN. Обозначим $w_{TP} = \frac{TP}{N_{\text{вб}}} = \frac{TP}{N}$, $w_{FN} = \frac{FN}{N_{\text{вб}}} = \frac{FN}{N}$ – соответственно доля TP и FN относительно общего числа веб-бэкдоров в выборке. Аналогично результат анализа веб-бэкдоров не влияет на значения TN и FP. Обозначим, $w_{TN} = \frac{TN}{N_{\text{л}}} = = \frac{TN}{N}$, $w_{FP} = \frac{FP}{N_{\text{л}}} = \frac{FP}{N}$ – соответственно доля TN и FP относительно общего числа легитимных объектов в выборке.

За счет вычисления доверительных интервалов можно с требуемой вероятностью ограничить значения $w^*_{TP}, w^*_{TN}, w^*_{FP}, w^*_{FN}$ – истинные значения долей TP, TN, FP, FN, соответственно.

Вычислим, сколько объектов необходимо исследовать, чтобы с вероятностью $P = 0.997$ и ошибкой не более $\Delta = 0.05$ определить значения $w^*_{TP}, w^*_{TN}, w^*_{FP}, w^*_{FN}$.

Определение необходимого количества веб-бэкдоров для оценки w^*_{TP} *(*w^*_{FN}*).*

Так как набор вэб-бэкдоров формируются за счет типической выборки, пропорциональной объему групп, то формула для расчета необходимой численности выборки в данном случае будет иметь вид [30]:

$$N_{\text{вб}} = \frac{z^2 \overline{\sigma_w^2}}{\Delta_w^2}, \tag{9}$$

где Δ_w^2 – квадрат предельной ошибки доли, $\overline{\sigma_w^2}$ – средняя из групповых дисперсий типических групп для $w_{TP}(w_{FN})$, z – коэффициент доверия, который определяется на основе табличных значений в зависимости от вероятности P.

Средняя из групповых дисперсий вычисляется по формуле:

$$\overline{\sigma_w^2} = \frac{\sum_{j=1}^{k} \sigma_j^2 N_j}{\sum_{j=1}^{k} N_j}, k = 64, \tag{10}$$

где N_j – количество элементов в j-й типической группе, k – количество типических групп, σ_j^2 – дисперсия выборочной доли в j-й типической группе.

Как упоминалось ранее, существует 64 типических групп (64 вида веб-бэкдоров). Дисперсия выборочной доли в j-й типической группе определяется по формуле [30]:

$$\sigma_j^2 = w_j(1-w_j), \tag{11}$$

где w_j – доля $TP(FN)$ в j-й типической группе.

Так как изначально неизвестен характер реакции СОВБ мы будем полагать дисперсию в каждой группе максимальной. Она достигается при $w_j = 0,5$ и равна $\sigma_j^2 = 0,25$ [31]. $\sum_{j=1}^{64} N_j = N$, а $N_j = N/64$. Тогда $\overline{\sigma_w^2} = 0,25$. Значение z табличное, которое при $P = 0,997$ равно 3.

Таким образом:

$$N_{\text{вб}} = \frac{3^2 \cdot 0,25}{0,05^2} = 900.$$

Так как используется 64 вида веб бэкдоров, то необходимо $\frac{900}{64} = 14,063 \approx 15$ веб-бэкдоров каждого вида (960 веб-бэкдоров): $D_{15}^1, D_{15}^2, ..., D_{15}^{64}$, где $D_{15}^i = \left\{d_1^i, d_2^i, ..., d_{15}^i\right\}$.

Определение необходимого количества легитимных файлов для оценки w_{FP}^* (w_{TN}^*).

Так как легитимные файлы не разделяются по типам, то формула для расчета необходимой выборки в данном случае соответствует собственно-случайному отбору и имеет вид:

$$N_{\text{л}} = \frac{z^2 \cdot \sigma_w^2}{\Delta_w^2}, \tag{12}$$

где σ_w^2 – дисперсия доли в выборке. Она вычисляется по формуле:

$$\sigma_w^2 = w(1-w), \tag{13}$$

где w – доля $FP(TN)$ в выборке легитимных файлов.

Также полагаем, что она максимальна, т.е. $\sigma_w^2 = 0,25$. Получаем:

$$N_{\text{л}} = \frac{3^2 \cdot 0,25}{0,05^2} = 900.$$

Для обеспечения равного количества веб-бэкдоров и легитимных файлов, требуется минимум 1920 элементов (960 веб-бэкдоров и 960 легитимных файлов), чтобы с вероятностью 0.997 и ошибкой не более $\Delta_{wTP} = \Delta_{wTN} = \Delta_{wFP} = \Delta_{wFN} = 0,05$ определить $w_{TP}^*, w_{TN}^*, w_{FP}^*, w_{FN}^*$. Таким образом, необходимо, как минимум, 15 пакетов по 128 элементов каждый. При этом условии будут выполняться следующие неравенства:

$$\begin{gathered} w_{TP} - \Delta_{wTP} \le w_{TP}^* \le w_{TP} + \Delta_{wTP}, \\ w_{TN} - \Delta_{wTN} \le w_{TN}^* \le w_{TN} + \Delta_{wTN}, \\ w_{FP} - \Delta_{wFP} \le w_{FP}^* \le w_{FP} + \Delta_{wFP}, \\ w_{FN} - \Delta_{wFN} \le w_{FN}^* \le w_{FN} + \Delta_{wFN}. \end{gathered} \tag{14}$$

Набора из 15 пакетов будет достаточно, чтобы оценить значения $w_{TP}^*, w_{TN}^*, w_{FP}^*, w_{FN}^*$, с вероятностью 0,997 и ошибкой не более 0,05. Однако если есть возможность предположить дисперсию долей, то возможно уменьшение объема выборки.

Если взять выборку веб-бэкдоров и легитимных файлов, то для w_{TP}^* и w_{FN}^* ошибка будет вычисляться по формуле (15), а для w_{FP}^*

и w_{TN}^{*} по формуле (16). Количество веб-бэкдоров и легитимных файлов равно $N_{\text{вб}} = N_{\text{л}} = N$.

$$\Delta_{w_{TP}} = \Delta_{w_{FN}} = z\sqrt{\frac{\overline{\sigma_w^2}}{N}}, \tag{15}$$

$$\Delta_{w_{TN}} = \Delta_{w_{FP}} = z\sqrt{\frac{\sigma_w^2}{N}}, \tag{16}$$

где σ_w^2 вычисляется по формуле (13), а $\overline{\sigma_w^2}$ по формуле (10).

Для малых объемов выборок, при невозможности вычислить $\overline{\sigma_w^2}$, можно ограничить это значение. Пусть количество веб-бэкдоров $N_{\text{вб}}$ кратно 64 (исходя из предыдущих рассуждений), тем самым в каждой подгруппе будет $\frac{N_{\text{вб}}}{64} = \frac{N}{64} = m$ веб-бэкдоров. Средняя выборочная доля *TP* (для *FN* проводятся аналогичные вычисления) равна:

$$\overline{w_{TP}} = \frac{\sum_{j=1}^{64} w_{TP_j} N_j}{\sum_{j=1}^{64} N_j} = \frac{m\sum_{j=1}^{64} w_{TP_j}}{N} = \frac{\sum_{j=1}^{64} w_{TP_j}}{64}, \tag{17}$$

где w_{TP_j} – доля *TP* в j-й подгруппе, N_j – количество веб-бэкдоров в каждой подгруппе (равно m).

Согласно формулам (10) и (11) получаем:

$$\begin{aligned}\overline{\sigma_w^2} &= \frac{\sum_{j=1}^{64} w_j(1-w_j)}{64} = \frac{\sum_{j=1}^{64} w_{TP_j}(1-w_{TP_j})}{64} = \\ &= \frac{\sum_{j=1}^{64} w_{TP_j}}{64} - \frac{\sum_{j=1}^{64} w_{TP_j}^2}{64} = \overline{w_{TP}} - \frac{\sum_{j=1}^{64} w_{TP_j}^2}{64}.\end{aligned} \tag{18}$$

С другой стороны:

$$\overline{w_{TP}} = \frac{TP}{N} = \frac{\sum_{j=1}^{64} TP_j}{N}, \tag{19}$$

$$\sigma_{\overline{w}}^2 = \overline{w_{TP}}(1-\overline{w_{TP}}) = \overline{w_{TP}} - \overline{w_{TP}}^2, \tag{20}$$

где $\sigma_{\overline{w}}^2$ – дисперсия средней выборочной доли, TP_j – количество TP соответственно в j-й типической группе.

Известно следующее математическое неравенство Коши-Буняковского [32]:

$$\frac{x_1 + x_2 + ...x_n}{n} \le \sqrt{\frac{x_1^2 + x_2^2 + ...x_n^2}{n}}. \tag{21}$$

Если принять, что $x_1, x_2, ..., x_n > 0$, то возведем обе части неравенства (21) в квадрат и получим:

$$\left(\frac{\sum_{j=1}^{n} x_j}{n}\right)^2 \le \frac{\sum_{j=1}^{n} x_j^2}{n}. \tag{22}$$

Теперь, если принять в (22) $n = 64, x_j = w_{TP_j}$, а также воспользоваться формулами (18), (20), то получаем:

$$\overline{w_{TP}}^2 = \left(\frac{\sum_{j=1}^{64} w_{TP_j}}{64}\right)^2 \le \frac{\sum_{j=1}^{64} w_{TP_j}^2}{64} \Rightarrow \sigma_{\overline{w}}^2 \ge \overline{\sigma_w^2}. \tag{23}$$

Тем самым вычислив $\sigma_{\overline{w}}^2$, можно ограничить значение средней групповой дисперсии типических групп $\overline{\sigma_w^2}$.

Зная доверительные интервалы для $w_{TP}^*, w_{TN}^*, w_{FP}^*, w_{FN}^*$, можно также ограничить значения Re^*, Ac^*, Pr^* (истинные значения Re, Ac, Pr соответственно).

Оценка значения Re^*.

Согласно выражениям (4), (7) Re через w_{TP} выражается так:

$$Re = \frac{TP}{TP + FN} = \frac{TP}{N} = w_{TP}. \quad (24)$$

Тогда $Re^* = w_{TP}^*$.

Отсюда, используя неравенства (14) получаем:

$$Re - \Delta_{w_{TP}} \le Re^* \le Re + \Delta_{w_{TP}}. \quad (25)$$

Оценка значения Ac^*.

Согласно выражениям (6-8) Ac через w_{TP} и w_{TN} выражается так:

$$Ac = \frac{TP + TN}{TP + TN + FP + FN} = \frac{TP + TN}{2N} = \frac{1}{2} w_{TP} + \frac{1}{2} w_{TN}. \quad (26)$$

Тогда $Ac^* = \frac{1}{2} w_{TP}^* + \frac{1}{2} w_{TN}^*$.

Согласно неравенствам (14):

$$Ac^* = \frac{1}{2} w_{TP}^* + \frac{1}{2} w_{TN}^* \le \frac{1}{2}(w_{TP} + \Delta_{w_{TP}}) + \frac{1}{2}(w_{TN} + \Delta_{w_{TN}}) =$$
$$= \frac{1}{2} w_{TP} + \frac{1}{2} w_{TN} + \frac{\Delta_{w_{TP}} + \Delta_{w_{TN}}}{2} = Ac + \frac{\Delta_{w_{TP}} + \Delta_{w_{TN}}}{2}.$$

Аналогично вычисляется с другой стороны.

Отсюда получается:

$$Ac - \frac{\Delta_{w_{TP}} + \Delta_{w_{TN}}}{2} \le Ac^* \le Ac + \frac{\Delta_{w_{TP}} + \Delta_{w_{TN}}}{2}. \quad (27)$$

В частном случае, если $\Delta_{w_{TP}} = \Delta_{w_{TN}} = \Delta$ выражение (27) принимает вид:

$$Ac - \Delta \le Ac^* \le Ac + \Delta .$$

Оценка значения Pr^*.

Согласно выражениям (5), (7-8), Pr через w_{FP} и w_{TN} выражается так:

$$Pr = \frac{TP}{TP + FP} = \frac{\frac{TP}{N}}{\frac{TP}{N} + \frac{FP}{N}} = \frac{w_{TP}}{w_{TP} + w_{FP}} = \frac{1}{1 + \frac{w_{FP}}{w_{TP}}}. \quad (28)$$

Тогда $Pr^* = \frac{1}{1 + \frac{w_{FP}^*}{w_{TP}^*}}$.

Согласно неравенствам (14), имеем:

$$Pr^* = \frac{1}{1 + \frac{w_{FP}^*}{w_{TP}^*}} \le \frac{1}{1 + \frac{w_{FP} - \Delta_{wFP}}{w_{TP} + \Delta_{wTP}}} = \frac{w_{TP} + \Delta_{wTP}}{w_{TP} + w_{FP} + \Delta_{wTP} - \Delta_{wFP}}. \quad (29)$$

$$Pr^* = \frac{1}{1 + \frac{w_{FP}^*}{w_{TP}^*}} \ge \frac{1}{1 + \frac{w_{FP} + \Delta_{wFP}}{w_{TP} - \Delta_{wTP}}} = \frac{w_{TP} - \Delta_{wTP}}{w_{TP} + w_{FP} + \Delta_{wFP} - \Delta_{wTP}}. \quad (30)$$

Неравенство (30) верно, при $w_{TP} > \Delta_{wTP}$.

Объединяя (29) и (30), получим:

$$\frac{w_{TP} - \Delta_{wTP}}{w_{TP} + w_{FP} + \Delta_{wFP} - \Delta_{wTP}} \le Pr^* \le \frac{w_{TP} + \Delta_{wTP}}{w_{TP} + w_{FP} + \Delta_{wTP} - \Delta_{wFP}}. \quad (31)$$

В частном случае, если ошибки $\Delta_{wFP} = \Delta_{wTP} = \Delta$ получим из (31) следующие выражения:

$$Pr^* \le \frac{w_{TP}+\Delta}{w_{TP}+w_{FP}} = \frac{w_{TP}}{w_{TP}+w_{FP}} + \frac{\Delta}{w_{TP}+w_{FP}} = Pr + \frac{\Delta}{w_{TP}+w_{FP}},$$

$$Pr^* \ge \frac{w_{TP}-\Delta}{w_{TP}+w_{FP}} = \frac{w_{TP}}{w_{TP}+w_{FP}} - \frac{\Delta}{w_{TP}+w_{FP}} = Pr - \frac{\Delta}{w_{TP}+w_{FP}}.$$

Как видно, ошибка Pr^* зависит от $w_{TP}+w_{FP}$. Эта зависимость проиллюстрирована в таблице 3.

Таблица 3. Зависимость ошибки Pr^* от $w_{TP}+w_{FP}$ при $\Delta_{wFP}=\Delta_{wTP}=\Delta$

$w_{TP}+w_{FP}$	Pr^*
≥ 1	$Pr-\Delta \le Pr^* \le Pr+\Delta$
≥ 0.5	$Pr-2\Delta \le Pr^* \le Pr+2\Delta$
≥ 0.1	$Pr-10\Delta \le Pr^* \le Pr+10\Delta$

Из таблицы 3 видно, что при малых значениях $w_{TP}+w_{FP}$ необходимо увеличивать выборку для снижения возможной ошибки.

Таким образом, исходя из того, какой по объему набор тестовых данных, можно с определённой долей вероятности найти доверительный интервал, в котором находятся истинные значения показателей действенности Re^*, Ac^*, Pr^*. Предельные ошибки выборки для каждого показателя $\Delta_e = \langle \Delta_{e1}, \Delta_{e2}.\Delta_{e3} \rangle$, исходя из (25), (27), (31) вычисляются следующим образом:

$$\Delta_{e1} = \Delta_{Re} = \Delta_{wTP}, \tag{32}$$

$$\Delta_{e2} = \Delta_{Pr} = max\left(\frac{w_{TP}+\Delta_{wTP}}{w_{TP}+w_{FP}+\Delta_{wTP}-\Delta_{wFP}} - Pr; Pr - \frac{w_{TP}-\Delta_{wTP}}{w_{TP}+w_{FP}+\Delta_{wFP}-\Delta_{wTP}} \right), \tag{33}$$

$$\Delta_{e3} = \Delta_{Ac} = \frac{\Delta_{wTP}+\Delta_{wTN}}{2}. \tag{34}$$

На основе рассуждений, представленных в разделах 4-6, сформируем этапы методики оценивания результативности функционирования СОВБ. Они представлены на рисунке 4.

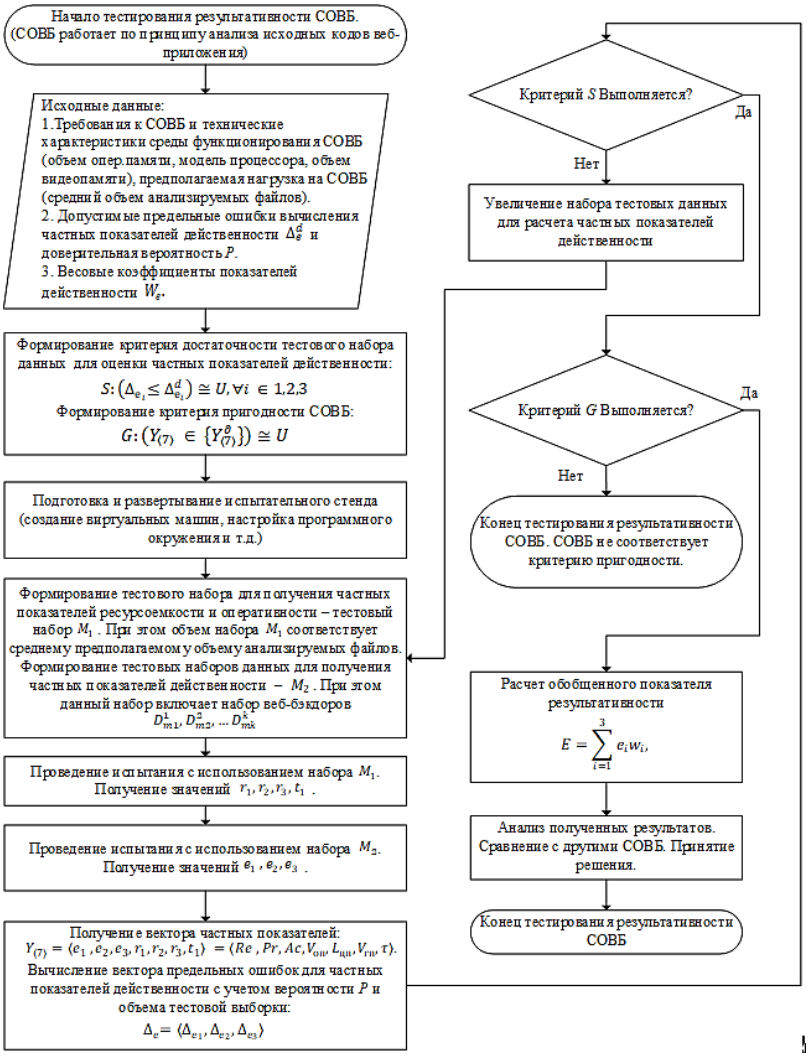


Рис. 4. Этапы методики оценивания результативности функционирования СОВБ

7. Проведение эксперимента и апробация методики. Перед тем, как полностью привести все шаги эксперимента, рассмотрим пример расчета предельных ошибок частных показателей $\Delta_e = \langle \Delta_{e1}, \Delta_{e2}, \Delta_{e3} \rangle$, исходя из объема тестовых наборов данных.

Пример вычисления $\Delta_e = \left\langle \Delta_{e_1}, \Delta_{e_2}, \Delta_{e_3} \right\rangle$.

В ходе тестирования алгоритма, представленного в разделе 3 настоящей статьи, использовался набор данных, состоящий из 128 объектов (64 легитимных файлов и 64 веб-бэкдоров). Данный алгоритм показал результаты, представленные в таблице 4.

Таблица 4. Результат проверки алгоритма обнаружения на тестовом наборе данных

		Прогноз наличия веб-бэкдора СОВБ	
		-	+
Фактическое наличие веб-бэкдора	-	62	2
	+	37	27

Исходя из формул (4-6) получаем частные показатели действенности: $e_1 = Re = 0,422$; $e_2 = Pr = 0,931$; $e_3 = Ac = 0,695$.

Вычислим предельные ошибки показателей. Коэффициент доверия z будем вычислять для $P = 0,997$.

Согласно (15) и (23):

$$\Delta_{w_{TP}} = \Delta_{w_{FN}} = z\sqrt{\frac{\overline{\sigma_w^2}}{64}} \le z\sqrt{\frac{\sigma_{\overline{w}}^2}{64}} = z\sqrt{\frac{\overline{w}\left(1-\overline{w}\right)}{64}} = 3\sqrt{\frac{\frac{27}{64}\cdot\frac{37}{64}}{64}} \approx 0,185.$$

Отсюда, согласно неравенствам (14):

$$0,422 - 0,185 \le w_{TP}^* \le 0,422 + 0,185,$$

$$0,578 - 0,185 \le w_{FN}^* \le 0,578 + 0,185.$$

Согласно (16) для w_{FP}^* и w_{TN}^*:

$$\Delta_{w_{FP}} = \Delta_{w_{TN}} = z\sqrt{\frac{\sigma_w^2}{64}} = z\sqrt{\frac{w(1-w)}{64}} = 3\sqrt{\frac{\frac{62}{64}\cdot\frac{2}{64}}{64}} \approx 0,065 .$$

Отсюда, согласно неравенствам (14):

$$0,031 - 0,065 \le w_{FP}^* \le 0,031 + 0,065 ,$$

$$0,969 - 0,065 \le w_{TN}^* \le 0,969 + 0,065 .$$

Отсюда, согласно (25), (27), (31):

$$0,422 - 0,185 \le Re^* \le 0,422 + 0,185 ,$$
$$0,711 \le Pr^* \le 1 ,$$
$$0,695 - 0,125 \le Ac^* \le 0,695 + 0,125 .$$

Тогда вектор ошибок для частных показателей действенности, согласно (32-34) будет выглядеть так: $\Delta_e = \langle 0,185; 0,220; 0,125 \rangle$.

Проведение эксперимента.

Необходимо вычислить обобщённые показатели результативности для трех СОВБ – *WebShellKiller, WEBSHELL.PUB, CloudWalker* [33], и выбрать наилучшее средство на основе сравнения этих показателей. Предъявлены следующие требования:

1) СОВБ должно работать в ОС *Ubuntu 20.04.5* с 8 Гб ОЗУ и 4-ядерным процессором *Intel core i7-10750.* Веб-приложение, которое анализирует СОВБ, основано на фреймворке WordPress. При этом:

– Среднее время анализа файла СОВБ не должно превышать 50 мс.

– Максимальный объем оперативной памяти, потребляемый средством, не должен превышать 500 Мб.

– Максимальный объем видеопамяти, потребляемый средством, не должен превышать 500 Мб.

– Максимальная загруженность процессора во время работы средства не должна превышать 30%.

2) Предельная ошибка для каждого вычисленного частного показателя действенности не должна превышать 0.05. Вероятность

нахождения частных показателей в пределах доверительного интервала равна $P = 0,997$.

3) Каждый из частных показателей действенности равнозначен (это равносильно тому, что весовые коэффициенты w_{e_i} для трех частных показателей равны 1/3).

Исходя из исходных требований, получаем, что область допустимых значений частных показателей:

$$\begin{aligned}\left\{Y^d_{\langle 7\rangle}\right\} = \left\{y^d_1 = \{0 \le Re \le 1\}, y^d_2 = \{0 \le Pr \le 1\}, y^d_3 = \{0 \le Ac \le 1\},\right.\\ \left. y^d_4 = \{V_{\text{оп}} \le 500\text{Мб}\}, y^d_5 = \{L_{\text{цп}} \le 30\%\}, y^d_6 = \{L_{\text{гп}} \le 500\text{Мб}\}, y^d_7 = \{\tau \le 50\text{мс}\}\right\}.\end{aligned} \quad (35)$$

Вектор допустимых предельных ошибок частных показателей действенности:

$$\Delta^d_e = \left\langle \Delta^d_{e1}, \Delta^d_{e2}, \Delta^d_{e3} \right\rangle = \left\langle 0,05; 0,05; 0,05 \right\rangle. \quad (36)$$

Вектор весовых коэффициентов для частных показателей действенности, в виду их равнозначности:

$$W_e = \left\langle w_{e1}, w_{e2}, w_{e3} \right\rangle = \left\langle \frac{1}{3}; \frac{1}{3}; \frac{1}{3} \right\rangle. \quad (37)$$

Тестирование СОВБ проводится следующим образом:

1 этап. Вычисление показателей e_1, e_2, e_3 на наборе легитимных файлов и веб-бэкдоров (с учетом выборки, представленной в подразделе 6.1).

2 этап. Вычисление показателей r_1, r_2, r_3 и t_1 на всем объеме файлов фреймворка WordPress.

Для первого этапа используется 96 файлов, содержащих веб-бэкдоры, которые были перечислены ранее [28] (96 файлов потому, что 32 веб-бэкдора состоят из одного файла и 32 веб-бэкдора – из двух файлов), а также 64 файлов, которые не содержат веб-бэкдоров. В качестве незараженных файлов использовались файлы из того же фреймворка *WordPress*. В процессе расчета частных показателей действенности веб-бэкдоры, которые состоят из двух файлов, рассматриваются как единый объект. Таким образом, для обнаружения такого веб-бэкдора достаточно определить его наличие хотя бы

в одном из двух файлов. В итоге получилось 128 объектов для тестирования СОВБ.

СОВБ проводит анализ файлов и выдает результат о наличии или отсутствии веб-бэкдора в каждом из них. После чего можно вычислить показатели $e_1 = Re, e_2 = Pr, e_3 = Ac$ и предельные ошибки выборки $\Delta_e = \langle \Delta_{e1}, \Delta_{e2}.\Delta_{e3} \rangle$.

Для второго этапа использовался набор из всех файлов PHP, из которых состоит сайт на фреймворке *WordPress*, с добавлением также всех созданных веб-бэкдоров. В конечном итоге всего получилось 1191 файл.

Испытательный стенд представляет собой виртуальную машину *Ubuntu 20.04.5* с 8 Гб ОЗУ и 4-ядерным процессором *Intel core i7-10750H*.

На первом этапе были получены TP, TN, FP, FN. С помощью них были вычислены значения частных показателей действенности Re, Pr, Ac с помощью формул (4-6). Также были вычислены доверительные интервалы каждого из показателей. Это можно увидеть в таблице 5. Предельные ошибки частных показателей действенности, исходя из объема выборки, равны:

$$\begin{aligned} \Delta_e^1 &= \langle 0{,}146; 0{,}516; 0{,}124 \rangle, \\ \Delta_e^2 &= \langle 0{,}184; 0{,}276; 0.138 \rangle, \\ \Delta_e^3 &= \langle 0{,}124; 0{,}566; 0{,}117 \rangle, \end{aligned} \tag{38}$$

где 1 – *WebShellKiller,* 2 – *WEBSHELL.PUB,* 3 – *CloudWalker.*

На втором этапе каждая из СОВБ получала на вход набор данных из 1191 файла. Полученные значения показателей ресурсоемкости и оперативности также представлены в таблице 5. (Показатель $V_{\text{гп}}$ для каждой системы равен 0, потому что ни одна из систем не использует видеопамять во время работы.)

На основании измеренных значений получаем вектора частных показателей для СОВБ:

$$\begin{aligned} Y^1 &= \langle 0{,}188; 0{,}706; 0{,}555; 35\text{Мб}; 20\%; 0\text{Мб}; 3.8\text{мс} \rangle, \\ Y^2 &= \langle 0{,}406; 0{,}867; 0{,}672; 58\text{Мб}; 24\%; 0\text{Мб}; 1.9\text{мс} \rangle, \\ Y^3 &= \langle 0{,}125; 0{,}571; 0{,}516; 156\text{Мб}; 39\%; 0\text{Мб}; 40{,}4\text{мс} \rangle. \end{aligned} \tag{39}$$

Таблица 5. Результаты тестирования СОВБ

Показатели	**WebShellKiller**	**WEBSHELL.PUB**	**CloudWalker**
TP	12	26	8
FP	5	4	6
TN	59	60	58
FN	52	38	56
Частные показатели действенности с доверительными интервалами			
$e_1(Re)$	0,188; $0,042 \le Re^* \le 0,334$	0,406; $0,222 \le Re^* \le 0,590$	0,125; $0,001 \le Re^* \le 0,249$
$e_2(Pr)$	0,706; $0,190 \le Pr^* \le 1$	0,867; $0,591 \le Pr^* \le 1$	0,571; $0,005 \le Pr^* \le 1$
$e_3(Ac)$	0,555; $0,431 \le Ac^* \le 0,679$	0,672; $0,534 \le Ac^* \le 0,810$	0,516; $0,399 \le Ac^* \le 0,633$
Частные показатели ресурсоемкости			
$r_1(V_{\text{оп}})$	35 Мб	58 Мб	156 Мб
$r_2(V_{\text{цп}})$	20%	24%	39%
$r_3(V_{\text{гп}})$	0 Мб	0 Мб	0 Мб
Частные показатели оперативности			
$t_1(\tau)$	3,8 мс	1,9 мс	40,4 мс

Однако полученные значения не соответствуют критерию достаточности тестовых наборов веб-бэкдоров (2). Каждая предельная ошибка частного показателя действенности (38) больше, чем 0.05. Тем самым объем тестовых данных для вычисления показателей частных показателей действенности необходимо увеличить. Затем проводить эксперимент нужно заново. Однако в целях демонстрации продолжим вычисления, предполагая, что критерий достаточности тестовых наборов все-таки выполнен.

Примечание

Как можно видеть, возможные ошибки частных показателей действенности (38) получились достаточно большими, потому что в качестве набора данных использовался всего один пакет – 128 объектов. Увеличение объема тестовых данных (легитимных файлов и веб-бэкдоров), согласно выборке, предложенной в подразделе 6.1, приведет к сужению диапазона доверительных интервалов, и тем самым к уменьшению предельных ошибок частных показателей.

Исходя из области допустимых значений (35) очевидно, что третья СОВБ *CloudWalker* исключается из эксперимента, так как $L_{\text{цп}} = 39\% > 30\%$.

На основе вектора весовых коэффициентов (37) и формулы (1) вычислим обобщенные значения показателей результативности для первого и второго СОВБ:

$$E^1 = \frac{1}{3}0,188 + \frac{1}{3}0,706 + \frac{1}{3}0,555 \approx 0,483,$$
$$E^2 = \frac{1}{3}0,406 + \frac{1}{3}0,867 + \frac{1}{3}0,672 \approx 0,648. \quad (40)$$

Как видно из (40) второе средство обладает более высоким значением результативности. Поскольку частные показатели действенности находятся в одном диапазоне значений от 0 до 1, в идеальном СОВБ обобщенный показатель результативности будет иметь значение 1. Это достигается, когда все частные показатели действенности равны 1. Таким образом, для заданных условий WEBSHELL.PUB лучше всего подходит в качестве СОВБ, однако его интегральная результативность не слишком высока (0.648, при максимальном значении равном 1).

8. Заключение. Многие исследователи производят оценку СОВБ только на основе собственных наборов данных, что делает эту оценку не полностью объективной, что было показано в работе [4]. В настоящей статье предложена методика, позволяющая производить объективное оценивание результативности функционирования СОВБ. В методике выделены три группы частных показателей, используемых для оценки СОВБ: действенность, ресурсоемкость (затраты ресурсов) и оперативность (затраты времени) функционирования. Частные показатели ресурсоемкости и оперативности непосредственно не участвуют в вычислении обобщенного показателя результативности, однако они используются при определении пригодности СОВБ. Частные показатели действенности используются для расчета показателя результативности СОВБ.

Для формирования тестовых данных была разработана классификация веб-бэкдоров, встроенных в исходный код веб-приложений. На основе объема тестового набора данных, полученного с помощью специальной выборки, представленной в разделе 6, вычисляются доверительные интервалы частных показателей действенности и соответствующие предельные ошибки этих значений. Таким образом, в методике также предусмотрен критерий достаточности тестовых наборов данных. Объективность оценивания результативности функционирования СОВБ заключается в том, что для создания набора тестовых данных применяется обобщенная классификация веб-бэкдоров, встроенных в исходный код веб-приложений, а также рассчитываются доверительные интервалы для значений частных показателей действенности.

Разработанная методика применима для СОВБ, которые работают на основе анализа исходного кода веб-страниц. Для ее использования необходимы определенные исходные данные, такие как допустимые предельные ошибки частных показателей действенности и вероятность их нахождения в доверительном интервале, а также весовые коэффициенты частных показателей действенности, которые определяются экспертными методами. В результате применения методики вычисляется обобщенный показатель результативности, который зависит от весовых коэффициентов частных показателей. Если частные показатели равнозначны, весовые коэффициенты принимают значение 1/3. Обобщенный показатель результативности варьируется от 0 до 1, при этом максимальное значение «1» указывает на то, что данное СОВБ при заданных условиях является максимально результативным и способно обнаружить любой веб-бэкдор, встроенный в исходный код веб-приложения.

Литература

1. Актуальные киберугрозы: итоги 2022 года. URL: https://www.ptsecurity.com/ru-ru/research/analytics/cybersecurity-threatscape-2022/ (дата обращения: 02.12.2023).
2. Albalawi M.M., Aloufi R.B., Alamrani N.A., Albalawi N.N., Aljaedi O.A., Alharbi A.R. Website Defacement Detection and Monitoring Methods: A Review // Electronics. 2022. vol. 11(21). DOI: /10.3390/electronics11213573.
3. Кибербезопасность в 2023–2024 гг.: тренды и прогнозы. Часть третья. URL: https://www.ptsecurity.com/ru-ru/research/analytics/kiberbezopasnost-v-2023-2024-gg-trendy-i-prognozy-chast-tretya/#id2 (дата обращения: 02.02.2024).
4. Боровков В.Е., Ключарёв П.Г. Методы защиты веб-приложений от злоумышленников // Вопросы кибербезопасности. 2023. № 5(57). С. 89–99.
5. ГОСТ Р ИСО 9000-2015. Системы менеджмента качества. Основные положения и словарь // М.: Стандартинформ. 2018.
6. Sam L.T., Aurelien F. Backdoors: Definition, Deniability and Detection // Research in Attacks, Intrusions, and Defenses (RAID 2018). 2018. pp. 92–113.
7. Киселев А.Н. Подход к обнаружению вредоносного программного обеспечения web-shell на основе анализа сетевого трафика web-инфраструктуры // Труды Военно-космической академии имени А.Ф. Можайского. 2021. № 677. С. 143–152.
8. Ma M., Han L., Zhou C. Research and application of artificial intelligence based webshell detection model: A literature review // arXiv preprint arXiv.2405.00066. 2024.
9. Omer A. Performance Comparison of Static Malware Analysis Tools Versus Antivirus Scanners To Detect Malware // 1 Performance Comparison of Static Malware Analysis Tools Versus Antivirus Scanners To Detect Malware. 2017. pp. 1–6.
10. Real-World Protection Test July-October 2022. URL: https://www.av-comparatives.org/tests/real-world-protection-test-july-october-2022/ (дата обращения: 02.12.2023)
11. Лысенко А.В., Кожевникова И.С., Ананьин Е.В. Анализ методов обнаружения вредоносных программ // Молодой ученый. 2016. № 21(125). С. 758–761.

12. Fu J, Li L., Wang Y. Webshell Detection Based on Convolutional Neural Network // Journal of Zhengzhou University (Natural Science Edition). 2019. vol. 51(2). pp. 1–8.
13. WebShell detection based on semantic features. URL: https://litheory.github.io/publication/webshell-detection-based-on-semantic-features/ (дата обращения: 11.01.2024).
14. Word2vec. URL: https://www.tensorflow. org/text/tutorials/word2vec (дата обращения: 11.01.2024).
15. Pan Z., Chen Y., Chen Y., Shen Y., Guo X. Webshell Detection Based on Executable Data Characteristics of PHP Code // Wireless Communications and Mobile Computing. 2021. no. 1. pp. 1–12. DOI: 10.1155/2021/5533963.
16. Kaushik K., Aggarwal S., Mudgal S., Saravgi S., Mathur V. A novel approach to generate a reverse shell: Exploitation and Prevention // International Journal of Intelligent Communication, Computing, and Networks. 2021. vol. 2. DOI: 10.51735/ijiccn/001/33.
17. p0wnyshell – Single-file PHP Shell. URL: https://github.com/flozz/p0wny-shell (дата обращения: 12.01.2024).
18. WordPress. URL: http://wordpress.com (дата обращения: 12.01.2024).
19. Obfuscation Techniques in MARIJUANA Shell «Bypass». URL: https://blog.sucuri.net/2020/12/obfuscation-techniques-in-marijuana-shell-bypass.html (дата обращения: 20.01.2024).
20. Keeping Web Shells under Cover (Web Shells Part 3). URL: https://www.acunetix.com/blog/articles/keeping-web-shells-undercover-an-introduction-to-web-shells-part-3/ (дата обращения: 21.01.2024).
21. Петухов Г.Б., Якунин В.И. Методологические основы внешнего проектирования целенаправленных процессов и целеустремленных систем. М.: АСТ. 2006. 504 с.
22. Гаценко О.Ю., Мирзабаев А.Н., Самонов А.В. Методы и средства оценивания качества реализации функциональных и эксплуатационно-технических характеристик систем обнаружения и предупреждения вторжений нового поколения // Вопросы кибербезопасности. 2018. № 2(26). С. 24–32.
23. Zhu T., Weng Z., Fu L., Ruan L. A Web Shell Detection Method Based on Multiview Feature Fusion // Applied Sciences. 2020. vol. 10(18). DOI: 10.3390/app10186274.
24. Pu A., Feng X.., Zhang Y., Wan X., Han J., Huang C. BERT-Embedding-Based JSP Webshell Detection on Bytecode Level Using XGBoost // Security and Communication Networks. 2022. pp. 1–11. DOI: 10.1155/2022/4315829.
25. Nguyen N., Le V., Phung V., Du P. Toward a Deep Learning Approach for Detecting PHP Webshell // SoICT 2019: Proceedings of the Tenth International Symposium on Information and Communication Technology. 2019. pp. 514–521. DOI: 10.1145/3368926.3369733.
26. Zhang H., Liu M., Yue Z., Xue Z., Shi Y., He X. A PHP and JSP Web Shell Detection System with Text Processing Based on Machine Learning // 2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom). 2020. pp. 1584–1591.
27. Оценка качества в задачах классификации и регрессии. URL: https://neerc.ifmo.ru/wiki/index.php?title=Оценка_качества_в_задачах_классификации_и_регрессии (дата обращения: 15.11.2023).
28. DS_WBDT. URL: https://github.com/scienceMGtech (дата обращения: 20.10.2023).
29. Zhao J., Lu J., Wang X., Zhu K., Yu L. WTA: A Static Taint Analysis Framework for PHP Webshell // Applied Sciences. 2021. vol. 11(16). DOI: 10.3390/app11167763.
30. Ниворожкина Л.И. и др. Статистические методы анализа данных // РИОР, 2016. 333 с.

31. Илышев А.М. Общая теория статистики. Учебник для студентов вузов, обучающихся по специальностям экономики и управления. М.: ЮНИТИ. 2012. 535 с.
32. Соловьев Ю.П. Неравенства. М.: МЦНМО, 2005. 16 с.
33. WebShell Scan Detection and Killing Tools. URL: https://cloud.tencent.com/developer/article/1745883 (дата обращения: 27.02.2024).

Боровков Владислав Евгеньевич — аспирант кафедры, кафедра «информационной безопасности», Московский Государственный Технический Университет имени Н.Э. Баумана. Область научных интересов: машинное обучение, глубокое обучение, информационная безопасность, оценка безопасности компьютерных систем. Число научных публикаций — 14. vbscience@yandex.ru; 2-я Бауманская улица, 5/4, 105005, Москва, Россия; р.т.: +7(499)263-6936.

Ключарёв Петр Георгиевич — д-р техн. наук, профессор кафедры, кафедра «информационной безопасности», Московский Государственный Технический Университет имени Н.Э. Баумана. Область научных интересов: криптография, теоретическая информатика, дискретная математика, информационная безопасность. Число научных публикаций — 75+. pk.iu8@yandex.ru; 2-я Бауманская улица, 5/4, 105005, Москва, Россия; р.т.: +7(499)263-6936.

Денисенко Денис Игоревич — независимый исследователь информационной безопасности, разработчик программного обеспечения. Область научных интересов: информационная безопасность, анализ защищённости информационных систем, высоконагруженные приложения, безопасность приложений. Число научных публикаций — 2. researchDD_journal@mail.ru; 2-я Бауманская улица, 5/4, 105005, Москва, Россия; р.т.: +7(499)263-6936.

DOI 10.15622/ia.24.1.6

V. BOROVKOV, P. KLYUCHAREV, D. DENISENKO

TECHNIQUE FOR ASSESSING THE EFFECTIVENESS OF THE FUNCTIONING OF WEB BACKDOOR DETECTION SYSTEMS

Borovkov V., Klyucharev P., Denisenko D. **Technique for Assessing the Effectiveness of the Functioning of Web Backdoor Detection Systems.**

Abstract. Currently, there is a significant increase in information security incidents related to attacks on web resources. Obtaining unauthorized access to web resources remains one of the main methods of penetration into corporate networks of organizations and expanding the capabilities of intruders. In this regard, many studies are aimed at developing web backdoor detection systems (WBDS), but there is a need to assess the effectiveness of these systems. The purpose of this study is to develop an objective approach to assess the effectiveness of the WBDS functioning. In this work, it was found that the effectiveness of web backdoor detection systems is objectively manifested in the process of their use, therefore, testing of such systems should be carried out in conditions as close as possible to real ones. In this regard, the article proposes a new technique for assessing the effectiveness of WBDS. It is based on the calculation of three groups of specific indicators characterizing the potency, resource intensity and responsiveness of the detection tool, as well as the calculation of a generalized effectiveness indicator. Based on an analysis of research in this area, a classification of web backdoors embedded by an attacker into the source code of web applications has been developed. This classification is used when generating test datasets to calculate specific potency indicators. The developed methodology is applicable to tools that work based on the analysis of the source code of web pages. Additionally, its use requires a number of initial data, such as permissible maximum errors of frequent potency indicators and the probability of them being within the confidence interval, as well as weighting coefficients of specific potency indicators, which are selected by expert methods. This work may be useful for information security specialists and researchers who want to conduct a more objective assessment of their WBDS.

Keywords: cybersecurity, web vulnerabilities, web backdoors, web shells, machine learning, testing methods and tools.

References

1. Aktual'nye kiberugrozy: itogi 2022 goda [Current cyber threats: the results of 2022]. Available at: https://www.ptsecurity.com/ru-ru/research/analytics/cybersecurity-threatscape-2022/ (accessed: 02.12.2023). (In Russ.).
2. Albalawi M.M., Aloufi R.B., Alamrani N.A., Albalawi N.N., Aljaedi O.A., Alharbi A.R. Website Defacement Detection and Monitoring Methods: A Review. Electronics. 2022. vol. 11(21). DOI: 10.3390/electronics11213573.
3. Kiberbezopasnost v 2023–2024 gg.: trendy i prognozy. Chast tretia [Cybersecurity in 2023-2024: trends and forecasts. Part Three]. Available at: https://www.ptsecurity.com/ru-ru/research/analytics/kiberbezopasnost-v-2023-2024-gg-trendy-i-prognozy-chast-tretya/#id2 (accessed: 02.02.2024). (In Russ.).
4. Borovkov V., Klyucharev P. [Methods of protecting web applications from intruders]. Voprosy kiberbezopasnosti – Issues of cybersecurity. 2023. no. 5(57). pp. 89–99. (In Russ.).
5. GOST R ISO 9000-2015. [Quality management systems. Basic provisions and vocabulary] // M.: Standartinform. 2018. (In Russ.).

6. Sam L.T., Aurelien F. Backdoors: Definition, Deniability and Detection. Research in Attacks, Intrusions, and Defenses (RAID 2018). 2018. pp. 92–113.
7. Kiselev A.N. [An approach to the detection of malicious web-shell software based on the analysis of network traffic of the web infrastructure]. Trudy Voenno-kosmicheskoj akademii imeni A.F. Mozhajskogo – Proceedings of the Military Space Academy named after A.F. Mozhaisky. 2021. no. 677. pp. 143–152. (In Russ.).
8. Ma M., Han L., Zhou C. Research and application of artificial intelligence based webshell detection model: A literature review. arXiv preprint arXiv.2405.00066. 2024.
9. Omer A. Performance Comparison of Static Malware Analysis Tools Versus Antivirus Scanners To Detect Malware. 1 Performance Comparison of Static Malware Analysis Tools Versus Antivirus Scanners To Detect Malware. 2017. pp. 1–6.
10. Real-World Protection Test July-October 2022. Available at: https://www.av-comparatives.org/tests/real-world-protection-test-july-october-2022/ (accessed: 02.12.2023)
11. Lysenko A., Kozhevnikova I., Ananyin E. [Analysis of malware detection methods]. Molodoj uchenyj – Young Scientist. 2016. no. 21(125). pp. 758–761. (In Russ.).
12. Fu J, Li L., Wang Y. Webshell Detection Based on Convolutional Neural Network. Journal of Zhengzhou University (Natural Science Edition). 2019. vol. 51(2). pp. 1–8.
13. WebShell detection based on semantic features. Available at: https://litheory.github.io/publication/webshell-detection-based-on-semantic-features/ (accessed: 11.01.2024).
14. Word2vec. Available at: https://www.tensorflow. org/text/tutorials/word2vec (accessed: 11.01.2024).
15. Pan Z., Chen Y., Chen Y., Shen Y., Guo X. Webshell Detection Based on Executable Data Characteristics of PHP Code. Wireless Communications and Mobile Computing. 2021. no. 1. pp. 1–12. DOI: 10.1155/2021/5533963.
16. Kaushik K., Aggarwal S., Mudgal S., Saravgi S., Mathur V. A novel approach to generate a reverse shell: Exploitation and Prevention. International Journal of Intelligent Communication, Computing, and Networks. 2021. vol. 2. DOI: 10.51735/ijiccn/001/33.
17. p0wnyshell – Single-file PHP Shell. Available at: https://github.com/flozz/p0wny-shell (accessed: 12.01.2024).
18. WordPress. Available at: http://wordpress.com (accessed: 12.01.2024).
19. Obfuscation Techniques in MARIJUANA Shell «Bypass». Available at: https://blog.sucuri.net/2020/12/obfuscation-techniques-in-marijuana-shell-bypass.html (accessed: 20.01.2024).
20. Keeping Web Shells under Cover (Web Shells Part 3). Available at: https://www.acunetix.com/blog/articles/keeping-web-shells-undercover-an-introduction-to-web-shells-part-3/ (accessed: 21.01.2024).
21. Petukhov G., Yakunin V. Metodologicheskie osnovy vneshnego proektirovanija celenapravlennyh processov i celeustremlennyh sistem [Methodological foundations of external design of purposeful processes and purposeful systems]. Moscow: AST. 2006. 504 p. (In Russ.).
22. Gatsenko O., Mirzabaev A., Samsonov A. [Methods and tools for evaluating the quality of implementation of functional and operational and technical characteristics of intrusion detection and prevention systems of a new generation]. Voprosy kiberbezopasnosti – Cybersecurity issues. 2018. no. 2(26). pp. 24–32. (In Russ.).
23. Zhu T., Weng Z., Fu L., Ruan L. A Web Shell Detection Method Based on Multiview Feature Fusion. Applied Sciences. 2020. vol. 10(18). DOI: 10.3390/app10186274.

24. Pu A., Feng X.., Zhang Y., Wan X., Han J., Huang C. BERT-Embedding-Based JSP Webshell Detection on Bytecode Level Using XGBoost. Security and Communication Networks. 2022. pp. 1–11. DOI: 10.1155/2022/4315829.
25. Nguyen N., Le V., Phung V., Du P. Toward a Deep Learning Approach for Detecting PHP Webshell. SoICT 2019: Proceedings of the Tenth International Symposium on Information and Communication Technology. 2019. pp. 514–521. DOI: 10.1145/3368926.3369733.
26. Zhang H., Liu M., Yue Z., Xue Z., Shi Y., He X. A PHP and JSP Web Shell Detection System with Text Processing Based on Machine Learning. 2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom). 2020. pp. 1584–1591.
27. Ocenka kachestva v zadachax klassifikacii i regressii [Quality assessment in classification and regression tasks]. Available at: https://neerc.ifmo.ru/wiki/index.php?title=Оценка_качества_в_задачах_классификации_и_регрессии (accessed: 15.11.2023). (In Russ.).
28. DS_WBDT. Available at: https://github.com/scienceMGtech (accessed: 20.10.2023).
29. Zhao J., Lu J., Wang X., Zhu K., Yu L. WTA: A Static Taint Analysis Framework for PHP Webshell. Applied Sciences. 2021. vol. 11(16). DOI: 10.3390/app11167763.
30. Nivorozhkina L.I. et al. Statisticheskiye metody analiza dannykh [Statistical methods of data analysis]. Moscow: RIOR. 2016. 333 p. (In Russ.).
31. Ilyshev A.M. Obshhaja teorija statistiki. Uchebnik dlja studentov vuzov, obuchajushhihsja po special'nostjam jekonomiki i upravlenija [General theory of statistics. Textbook for university students studying economics and management]. Moscow: UNITY. 2012. 535 p. (In Russ.).
32. Solovyev Yu. Neravenstva [Inequalities]. Moscow: MTsNMO. 2005. 16 p. (In Russ.).
33. WebShell Scan Detection and Killing Tools. Available at: https://cloud.tencent.com/developer/article/1745883 (accessed: 27.02.2024).

Borovkov Vladislav — Postgraduate student, Department of information security, Bauman Moscow State Technical University. Research interests: machine learning, deep learning, information security, computer system security assessment. The number of publications — 14. vbscience@yandex.ru; 5/4, 2nd Baumanskaya St., 105005, Moscow, Russia; office phone: +7(499)263-6936.

Klyucharev Peter — Ph.D., Dr.Sci., Professor of the department, Department of information security, Bauman Moscow State Technical University. Research interests: cryptography, theoretical computer science, discrete mathematics, information security. The number of publications — 75+. pk.iu8@yandex.ru; 5/4, 2nd Baumanskaya St., 105005, Moscow, Russia; office phone: +7(499)263-6936.

Denisenko Denis — Independent information security researcher, software developer. Research interests: information security, security analysis of information systems, high-load applications, application security. The number of publications — 2. researchDD_journal@mail.ru; 5/4, 2nd Baumanskaya St., 105005, Moscow, Russia; office phone: +7(499)263-6936.

УДК 004.912, 004.62 **DOI** 10.15622/ia.24.1.7

М.Д. Кузнецов, Е.С. Новикова

КОРПУС ПОЛИТИК КОНФИДЕНЦИАЛЬНОСТИ ВЕБ-СЕРВИСОВ И УСТРОЙСТВ ИНТЕРНЕТА ВЕЩЕЙ ДЛЯ АНАЛИЗА ИНФОРМИРОВАННОСТИ СУБЪЕКТОВ ПЕРСОНАЛЬНЫХ ДАННЫХ

Кузнецов М.Д., Новикова Е.С. **Корпус политик конфиденциальности веб-сервисов и устройств Интернета Вещей для анализа информированности субъектов персональных данных.**

Аннотация. Информация о том, какие персональные данные собираются и обрабатываются различными устройствами и цифровыми сервисами, представлена в политиках конфиденциальности, однако, как показывают исследования, пользователи крайне редко их читают и, как следствие, не осознают, какие риски информационной безопасности, связанные с обработкой персональных данных, возникают. Решение проблемы повышения информированности субъектов персональных данных связано с разработкой методов поддержки принятия решений, которые представляют политики конфиденциальности в виде, более простом для понимания, например, в виде количественных оценок рисков и пиктограмм и позволяют принимать осознанные решения. Их разработка требует наличия структурированного и размеченного корпуса документов. В настоящей работе систематизируются корпусы политик конфиденциальности, находящиеся в открытом доступе, показываются их отличительные характеристики, такие как год создания, объем и наличие аннотаций. Также представлено описание нового корпуса документов, написанных на русском языке, даются результаты структурного и семантического анализа собранных политик безопасности, и выполняется сравнение с корпусом политик конфиденциальности, написанных на английском языке. Показано, что описание сценариев хранения, сбора и обработки данных в документах на русском языке составляет всего 25% объема текста документа, что может говорить об отсутствии деталей о том, какие типы данных собираются, какие механизмы для сбора используются, и каковы сроки их хранения, что влияет на “прозрачность” использования персональных данных.

Ключевые слова: персональные данные, политики конфиденциальности, корпус документов, семантический анализ, латентное размещение Дирихле.

1. Введение. По мере роста уровня цифровизации современного общества увеличивается объем собираемых и обрабатываемых персональных данных, что в свою очередь приводит к росту уровня информационных угроз, связанных с их утечкой. Согласно аналитическому отчету компании InfoWatch [1] в мире наблюдается уверенное увеличение числа инцидентов, связанных с нарушением конфиденциальности данных с ограниченным доступом, в т.ч. персональных данных. Следует отметить, что причиной утечек конфиденциальной информации не всегда является действия внешнего или внутреннего злоумышленника. Так, например, в 2024 компания Avast, разрабатывающая решения в области информационной безопасности,

была оштрафована за передачу персональных данных своих пользователей третьим лицам без их согласия [2].

Риски утечки персональной информации также возникают в результате использования различных “умных” устройств и веб-сервисов. Согласно исследованиям, доля “умных” домохозяйств в мире [3] и в России [4] постоянно увеличивается, спросом пользуются как системы управления домом с голосовыми ассистентами, так и различные датчики безопасности и охранные системы, включая IP-камеры. Между тем, регулярно появляются отчеты о выявленных уязвимостях в программном коде устройств Интернета Вещей и управляющих сервисах [5], эксплуатация которых приводит к утечке разнородной информации, начиная с видео данных, заканчивая данными о потреблении электроэнергии домохозяйствами [6, 7]. Эти данные позволяют получить подробную информацию о распорядке дня, привычках и образе жизни домохозяйства, которая может быть использована третьими лицами в различных целях, например, для недобросовестного целевого маркетинга, вмешательства в личную жизнь и совершения преступных действий, связанных с финансами.

Для предотвращения неправомочного использования персональных данных в 2022 в РФ были приняты поправки к ФЗ “О Персональных данных” № 152-ФЗ, в которых были ужесточены требования к сбору, обработке, хранению и передаче персональных данных третьим лицам. В частности, в статье 5 главы 2 ФЗ № 152-ФЗ сформулированы требования к целям обработки таких данных, которые должны быть “конкретными, заранее определенными и законными”, а согласия на их обработку должны быть “информированными и сознательными”.

Информация о том, какие персональные данные собираются и обрабатываются устройствами, веб-сервисами и приложениями, обычно представлена в соглашениях на обработку персональных данных и/или в политиках конфиденциальности компаний разработчиков устройств. Однако, в большинстве случаев эти документы написаны на сложном юридическом языке, что не всегда позволяет пользователю их понять [8, 9]. Как следствие, они дают свое согласие на сбор, обработку и хранение персональных данных без четкого понимания того, как организован этот процесс и какие риски, связанные с обработкой данных, возникают. Таким образом, проблема обеспечения прозрачности политик безопасности, включающая в себя задачу повышения осведомленности владельцев данных о том, как используются их данные, является важной и тесно связана с разработкой методик, способов, инструментов и систем поддержки принятия решений (СППР), позволяющих пользователю

цифровых сервисов и услуг оценить целесообразность использования того или иного сервиса в контексте сбора и использования его персональных данных.

Проблема повышения уровня информированности пользователей устройств Интернета Вещей активно исследуется во всем мире [10–12], и в настоящее время предложены различные решения по анализу политик конфиденциальности, написанных на естественном языке [11–16]. Они позволяют выявить и охарактеризовать в структурированном виде различные сценарии использования персональных данных, такие как сбор, хранение, передача персональных данных и т.д. Однако, разработанные методы анализа документов предложены для политик конфиденциальности, написанных на английском языке, проблема анализа и исследования политик конфиденциальности на русском языке не проработана.

В данной статье авторы представляют структурный и семантический анализ нового корпуса русскоязычных политик конфиденциальности веб-сервисов, которые доступны на территории России, и выполняют его сравнительный анализ с другим корпусом политик конфиденциальности, написанных на английском языке. Таким образом, научно-практический вклад авторов статьи заключается в:

1. сравнительном анализе существующих корпусов политик конфиденциальности, которые могут быть использованы для разработки методов их анализа на основе глубокого машинного обучения;
2. создании нового корпуса политик конфиденциальности, написанных на русском языке;
3. выполнении структурного и семантического анализа нового корпуса политик конфиденциальности и сравнения его характеристик с корпусом политик конфиденциальности на английском языке.

Полученный корпус политик конфиденциальности послужит основой для разработки автоматизированных методов обработки и анализа пользовательских соглашений, которые могут применяться пользователями при принятии решений, касающихся управления персональными данными при выборе цифровых сервисов и устройств, которые выполняют обработку персональных данных.

Статья организована следующим образом. В разделе 2 обсуждаются основные направления исследований в области анализа текстов политик конфиденциальности и выполняется сравнительный анализ корпусов политик конфиденциальности, находящихся в открытом доступе. В разделе 3 представлена методика сбора политик конфиденциальности, которая была использована для создания нового корпуса. В разделах 4

и 5 представлены результаты структурного и семантического анализа собранного корпуса политик конфиденциальности, а также выполняется его сравнительный анализ с корпусом политик конфиденциальности для устройств Интернета Вещей, написанных на английском языке [17]. В разделе 6 обсуждаются полученные результаты и формулируются дальнейшие направления исследований.

2. Методы анализа политик конфиденциальности. Согласно ФЗ "О Персональных данных" № 152-ФЗ персональными данными являются данные, относящиеся "прямо или косвенно определенному или определяемому физическому лицу (субъекту персональных данных)". К ним относятся как общедоступные данные, такие как фамилия и отчество субъекта, его возраст, образование, электронная почта и телефон, так и биометрические, специальные и иные персональные данные. Следует отметить, что в Российском законодательстве нет четкого определения, какие персональные данные следует относить к категории "иные", однако в мировой практике [10, 18] к персональным данным относятся в том числе данные, которые позволяют уникально идентифицировать устройства пользователя, например, IP- и MAC-адреса устройств, цифровой отпечаток браузера и т.д., поскольку если указать свои персональные данные, например, ФИО или телефон, цифровой отпечаток устройства позволяет уникально идентифицировать пользователя и отслеживать его поведение в сети.

Анализ существующих подходов к анализу политик конфиденциальности позволил авторам выделить два основных подхода к решению данной проблемы. В рамках первого подхода решается задача построения формализованного представления различных сценариев использования персональных данных [12, 19–21]. Под сценариями использования персональных данных понимается деятельность, связанная с обработкой персональных данных, включая их сбор, хранение и передачу третьим лицам. Формальное представление таких сценариев использования может служить основой как для определения правил обработки персональных данных [22], так и для оценки рисков нарушения их конфиденциальности [12].

В рамках второго подхода выполняется анализ политик конфиденциальности, написанных на естественном языке, целью разрабатываемых методов и моделей является повышение "прозрачности" и понятности документов для пользователей. В [15, 23, 24] обсуждается проблема сбора политик конфиденциальности и предлагается схема аннотирования, которая отражает основные характеристики различных сценариев использования персональных данных. В [15] представлен

подход к автоматическому определению различных типов персональных данных, таких как электронная почта, контактный телефон, адрес, геолокация, которые упоминаются в тексте политик конфиденциальности, разработанных для Android-приложений. В [25] авторы разработали подход к определению вариантов отказа от использования персональных данных, представленных в тексте политики конфиденциальности. В [26, 27] авторы исследуют проблему неоднозначности и общности политики безопасности и предлагают основанный на онтологии подход к уменьшению нечеткости терминов политики безопасности путем установления семантических отношений между ними.

Наиболее часто используемым корпусом политик конфиденциальности является набор данных OPP-115 [23], который был разработан в рамках проекта Usable Privacy Policy [28] (UPP). Он включает в себя 115 политик конфиденциальности веб-сайтов, которые были собраны с помощью сервиса Amazon Alexa [29], который отражает актуальность и популярность веб-сайтов и публичной веб-директории DMOZ.org, содержащей ссылки на веб-сайты различных категорий, внесенных туда реальными пользователями, что может привести к попаданию в директорию нерелевантных веб-сайтов. Несомненным преимуществом этого корпуса документов является наличие аннотаций и схемы аннотирования, разработанной ее авторами. Она включает в себя различные сценарии использования персональных данных и информацию об экспертах, выполняющих аннотирование текстов. Каждая политика аннотировалась несколькими экспертами, что позволило получить более 20 000 аннотаций, отражающих различные аспекты использования персональных данных. В [30] была продемонстрирована связь между разработанной схемой аннотаций и принципами Общего регламента ЕС о защите данных (GDPR).

Другим аннотированным корпусом политик конфиденциальности является набор данных APP-350 [24], состоящий из политик конфиденциальности приложений, размещенных на площадке Google Play [31]. Авторы не предоставляют подробного описания того, как он был собран, однако можно предположить, что для его создания использован программный интерфейс сервиса Google Play, который предоставляет широкие возможности для сбора необходимых данных. Набор данных MAPS [24] представляет собой расширение корпуса политик APP-350. Он также сформирован на основе политик конфиденциальности приложений, представленных на платформе Google Play, и состоит из более 1 миллиона документов. Однако, в отличие от APP-350, он не содержит аннотаций.

В [32] представлен корпус политик конфиденциальности, отличительной чертой которого является период его формирования: сбор документов осуществлялся на протяжении более 20 лет. Таким образом, он состоит из более чем миллиона документов, которые обновлялись и изменялись в течение этого периода времени. Его авторы также разработали программный инструмент, который извлекает различные фрагменты, такие как n-граммы, именованные сущности, URL-адреса, чтобы оценить, как содержание политики безопасности меняется со временем, и показали, что с течением времени политики конфиденциальности становятся все более сложными для понимания. Основным источником данных для сбора данного набора послужил сервис Amazon Alexa.

Для тестирования методов анализа политик конфиденциальности на больших объемах данных Р. Заим и К. Барбер [33] собрали набор политик конфиденциальности, написанных на разных языках и собранных для более чем 1,5 млн. веб-сайтов. В качестве исходной точки сбора документов они использовали DMOZ. Следует отметить, что в ходе исследования собранного набора данных, авторы также показали, какие категории сайтов чаще всего не имеют политики конфиденциальности. В настоящее время ресурс DMOZ недействителен и заменен аналогичным проектом Curlie [34], который построен на базе проектов Open Directory Project (ODP) и DMOZ.

В [35] представлен набор данных PrivaSeer, состоящий из более чем миллиона политик безопасности, написанных на английском языке. Его авторы оценили уровень сходства между документами, провели тесты на их читаемость, проанализировали наличие различных сценариев использования персональных данных с помощью поиска ключевых фраз и слов, также был выполнен его семантический анализ с помощью методов тематического моделирования.

В корпусе политик IoTDataset [17] представлены политики конфиденциальности, разработанные специально для устройств Интернета Вещей. Сбор политик осуществлялся путем анализа продуктов на площадках интернет-торговли Amazon [36] и Walmart [37]. Авторы рассматривали следующие типы умных устройств: “умные весы”, “умные часы”, “умный браслет” и пр. Были проанализированы результаты поисковых запросов для первых 30-ти страниц. Было выявлено, что только 23% производителей умных устройств имеют свой официальный сайт, и чуть более половины из них имеют собственную политику конфиденциальности. Всего было получено 798 документов, после исключения политик, длины которых в символах не превышали

1 000, осталось 592 документа. Следует отметить, что ручной анализ “коротких” документов показал, что они появляются либо из-за того, что у некоторых производителей на сайте пустая страница с политикой конфиденциальности, либо она отсутствует. Позже предложенный авторами подход к сбору политик конфиденциальности был использован в [38] для создания набора политик в рамках проекта PrivacyLens, в котором уже содержится более 1200 документов.

Перечисленные выше корпусы политик конфиденциальности активно используются в исследовательских проектах, посвященных анализу политик конфиденциальности, написанных на естественном языке, однако ни один из них не содержит документы, написанные на русском языке. В таблице 1 приведен сравнительный анализ корпусов политик конфиденциальности, рассмотренных выше. Они активно используются в исследовательских проектах, посвященных анализу политик конфиденциальности, в частности в рамках проекта Polisis [39] разработан сервис, который позволяет визуализировать сценарии использования персональных данных, извлеченные из политик конфиденциальности, написанных на английском языке, а в проекте Pribot [39] решена задача создания чат бота, который отвечает на вопросы по политикам. Аналогичных сервисов, позволяющих анализировать политики конфиденциальности на русском языке, нет, поэтому решаемая в данной работе задача по созданию корпуса политик конфиденциальности на русском языке, может послужить основой для разработки подобных решений.

3. Методика сбора политик конфиденциальности. Для сбора данного корпуса документов была адаптирована методика, предложенная в [17]. Ее выбор обусловлен следующими факторами. Во-первых, в большинстве работ несмотря на то, что приводится источник документов, информации, касающейся аспектов практической реализации, недостаточно для разработки программного инструмента. Во-вторых, многие источники данных, например проект Curlie [34], не позволяют собрать достаточного числа необходимых документов, поскольку русскоязычный сегмент Интернета в них представлен скудно. Данные, хранимые в веб-директориях, довольно редко обновляются, это делается волонтерами, поэтому гарантий получения актуальных данных нет. Кроме того, необходимость собрать пользовательские соглашения требует выполнения дополнительных действий с найденными страницами сайтов.

Таблица 1. Сравнительный анализ наборов данных политик безопасности

#	Название	Количество элементов	Источник данных	Аннотирование	Особенности
1	OPP-115 [23] (2016)	115	Amazon Alexa	Да	Исследование политики в рамках проекта “Usable Privacy Policy”. Аннотировано квалифицированными юристами, собственный метод аннотирования.
2	MAPS [24] (2019)	> 1 млн.	Google Play	Нет	
3	APP-350 [24] (2019)	350	MAPS	Да	
4	Princeton-Leuven Longitudinal Privacy Policy Dataset [32] (2021)	> 1 млн.	Amazon Alexa	Нет	Предназначен для оценки изменений в политике конфиденциальности с течением времени. Содержит политики за последние 20 лет, авторы также представили краулер.
5	A Large Publicly Available Corpus of Website Privacy Policies Based [33] (2020)	> 1.5 млн.	DMOZ	Нет	Формирование набора данных для дальнейших исследований. Для генерации использовался DMOZ, крупнейший сетевой каталог.
6	PrivaSeer: Corpus of Web Privacy Policies [35] (2020)	> 1 млн.	Данные Common Crawl, собранные с 2008 года	Нет	Часть проекта PrivaSeer, поисковой системы по политикам конфиденциальности. Разработана собственная методика создания набора данных, включающая сборщик документов, механизмы фильтрации и методы классификации и дедупликации.
7	PolicyQA: A Reading Comprehension Dataset for Privacy Policies [14] (2020)	25 017	OPP-115	Да	Часть проекта PrivacyCheck. Состоит из 25 017 примеров объяснения языка политики безопасности, дает ответы на 714 вопросов о политиках конфиденциальности.
8	PrivacyQA [40] (2019)	1 750	Google Play	Да	Состоит из 1 750 вопросов для 35 политик конфиденциальности на английском языке. Вопросы представлены категориями из схемы аннотирования к набору OPP-115.
9	IoTDataset [17] (2021)	592	Amazon, Walmart, Google, и производители IoT устройств	Нет	Предназначен для анализа политик конфиденциальности устройств Интернета Вещей. Создан на основе политик безопасности производителей IoT-устройств.
10	PrivacyLens (2023)	> 1 200	Amazon, Walmart, и производители IoT устройств	Нет	Собран в рамках проекта PrivacyLens. Предназначен для анализа политик конфиденциальности устройств Интернета Вещей. Создан на основе политик безопасности производителей IoT-устройств, позволяет сравнивать версии политик конфиденциальности.
11	PPinRussian dataset (данная работа)	9 051	mail.ru top, rambler top	Нет	Предназначен для анализа политик конфиденциальности на русском языке. Создан на основе политик конфиденциальности веб-сервисов.

Методика, разработанная авторами данной работы и детально описанная в [17], позволяет решить две задачи:

1. определение источников данных, удовлетворяющих заданным условиям отбора;

2. очистка документов от излишней HTML-разметки.

Она была успешно использована при создании корпуса политик конфиденциальности на английском языке IoTDataset, разработанных специально для IoT устройств, и применена в проекте PrivacyLens других авторов [38]. Однако, в отличие от оригинальной методики, исходной точкой для сбора данных послужили ссылки на веб-сервисы, полученные с платформ интернет-аналитики от компаний Mail.ru [41] и Rambler [42]. Ссылки были собраны по следующим категориям: “World”, “State”, “Business”, “House”, “Cars”, “Internet”, “Job”, “Computers”, “Rest”, “Culture”, “Science”, “WapSites”, “Sport”, “Mysterious”, “Industry”, “References”, “MassMedia”, “Humor”.

Таким образом, использованная методика состоит из следующих шагов:

– сбор гиперссылок на веб-страницы на платформе интернет-аналитики;

– поиск политик конфиденциальности на сайте веб-страниц;

– загрузка политик конфиденциальности в HTML-формате;

– очистка и подготовка политик конфиденциальности к дальнейшему анализу.

Под очисткой и подготовкой политик конфиденциальности понимается удаление HTML-разметки из текста документа и добавление Markdown-разметки для сохранения структуры текста.

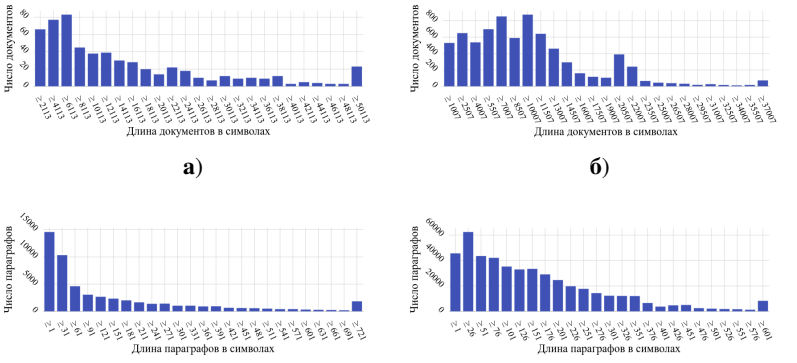
В ходе формирования набора данных было проанализировано 25 568 веб-сайтов, получено 21 585 (84%) необработанных политик конфиденциальности. После очистки текстов политик от HTML-разметки из набора данных были исключены документы длиной менее 1 000 символов. В результате было собрано 9 051 документов, полученный набор документов был назван PPinRussian. Следует отметить, что в ходе сбора политик конфиденциальности были выявлены случаи, когда вместо текста политики был размещен массивный рекламный блок или появлялось сообщение об отсутствии такого домена с предложением о его покупке/аренде.

4. Структурные особенности текстов политик конфиденциальности. Часто, чтобы определить методы и алгоритмы для дальнейшего анализа текстов, необходимо выполнить структурный анализ собранного корпуса документов. Под структурным анализом документов авторы понимают исследование значений длин документов, параграфов в символах, анализ наличия таких структурных элементов

форматирования, как списки, таблицы, разделы, заголовки. Понимание того, как организован текст в политиках конфиденциальности, позволяет более точно извлекать и связывать фрагменты текстов, описывающих различные аспекты сценариев использования персональных данных.

В данном разделе представлены результаты структурного анализа собранного корпуса политик конфиденциальности и выполнено его сравнение с набором политик конфиденциальности IoTDataset, написанных на английском языке.

На рисунке 1 представлены распределения длин политик и параграфов в символах.



а) **б**)

в) **г**)

Рис. 1. Гистограммы распределения длин: а, б) документов; в, г) параграфов для англоязычного и русскоязычного датасетов соответственно

Можно заметить, что наиболее распространенная длина документа в англоязычном корпусе составляет 7 000-8 000 символов. Это соответствует 3.9-4.5 стандартным страницам по 1 800 символов. Наиболее распространенными являются короткие параграфы длиной до 200 символов, состоящие из 2-3 предложений. В русскоязычном корпусе распределение длин документов отличается, наиболее часто встречаемыми являются документы длиной 8 000-9 000 символов, что соответствует 4.5-5 страницам, а также политики длиной 3 000-5 000 символов. Длины параграфов в корпусах также существенно различаются, причиной этому служат лингвистические особенности языков – в английском языке слова короче, предложения состоят из меньшего числа слов. Для политик конфиденциальности на русском языке характерны более длинные формулировки, особенно это верно для

вступительной части документа, где перечисляются различные названия нормативных документов, на основании которых производится обработка персональных данных, а также приводятся используемые в политике термины и определения.

Также в корпусах политик безопасности было проанализировано распределение таких структурных элементов текста, как таблицы, упорядоченные и неупорядоченные списки, параграфы и заголовки. На рисунке 2 показано распределение этих элементов в текстах политик конфиденциальности на русском и английском языках. Представленные данные позволяют сделать вывод о структуре политик безопасности. Так, обычно политики безопасности состоят из заголовков и параграфов. Наиболее часто используемыми элементами структуры помимо заголовков и параграфов являются нумерованные или ненумерованные списки и их элементы, в то время как таблицы используются редко. Русскоязычный документ в среднем состоит из 34.1 параграфов (85.1% документа), 3.8 заголовков (9.5% документа), 0.9 нумерованных списков (2.2% документа), 1.1 ненумерованных списков (2.7% документа) и 0.2 таблиц (0.5% документа). Документ на английском языке будет иметь похожую структуру: чуть меньше параграфов (31.5 параграфов или 44.9% документа), значительно больше заголовков – 33 заголовка (47.2% документа), 0.7 нумерованных списков (1% документа), 4.4 ненумерованных списков (6.3% документа) и 0.5 таблиц (0.7% документа).

Однако следует отметить, что такое большое число заголовков в английских документах может быть связано со сложностью подсчета данного элемента структурирования текста из-за большой вариативности HTML-разметки, используемой для их создания. Почти все сайты используют свой способ задания разделов, свои собственные правила нумерации разделов, заголовков и списков. На некоторых сайтах списки и заголовки нумеруются с помощью HTML-разметки, на других – нумерация задается вручную. В частности, компания Huawei предоставляет более 50 политик конфиденциальности для своих сервисов [43] и, хотя визуально все они имеют одинаковую структуру, для создания заголовков используется до 16 различных вариантов разметки HTML. Таким образом, даже в рамках одной компании не существует единой конвенции для оформления политик безопасности и их структурной компоновки, поэтому авторы посчитали заголовками строки длиной менее 100 символов и не содержащие маркеров “list item” (маркер элемента списка).

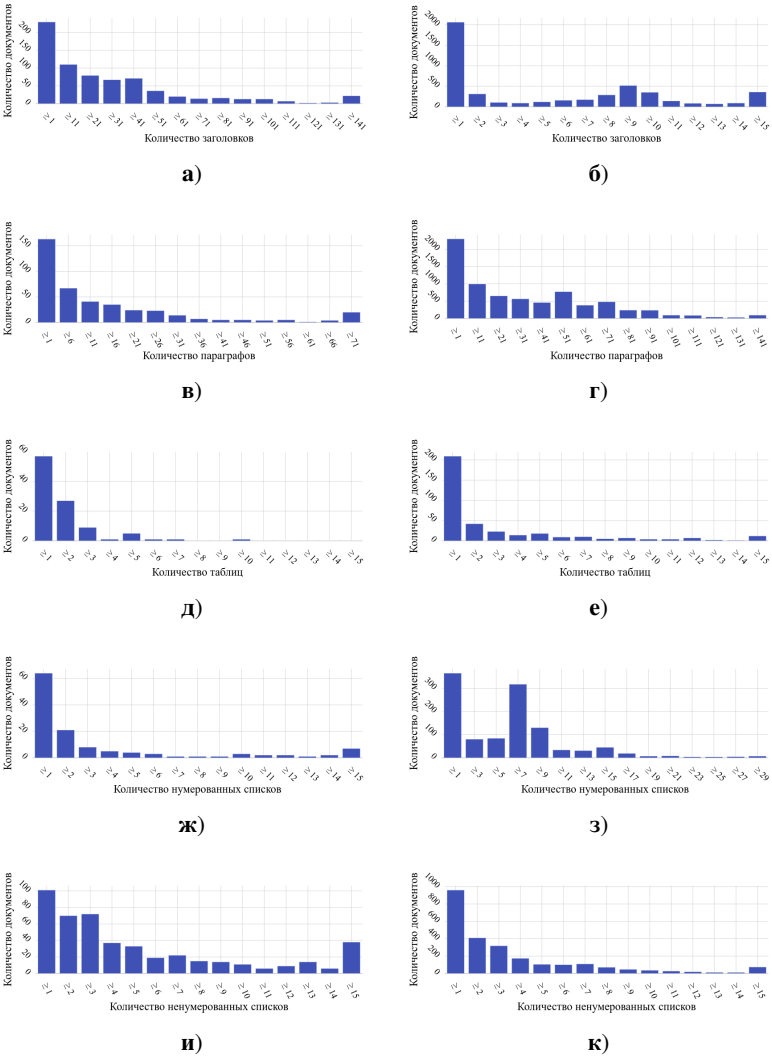


Рис. 2. Гистограммы распределения элементов структурирования данных в тексте политики безопасности: а, б) заголовки; в, г) параграфы; д, е) таблицы; ж, з) нумерованные списки; и, к) ненумерованные списки для англоязычного и русскоязычного датасетов соответственно

Очевидно, что такой подход не дает точных результатов, поскольку короткие параграфы, состоящие из одной строки, такие как контактная информация производителей, также относятся к заголовкам.

5. Семантический анализ корпусов политик конфиденциальности. При семантическом анализе политик конфиденциальности наибольший интерес представляет получение информации о том, какие характеристики сценариев сбора и обработки персональных данных описаны в документах, например, какие типы персональных данных собираются, в каких целях они собираются, на каких законных основаниях выполняется их обработка, каковы сроки их хранения, каким образом организована их защита и т.д.

В настоящей работе целью семантического анализа является извлечение тем, представленных в документе, и определение ключевых слов для них. Такие слова могут быть использованы для определения того, какие сценарии обработки персональных данных представлены в политике конфиденциальности.

Извлечение тем из корпуса политик конфиденциальности осуществлялось с помощью латентного размещения Дирихле (Latent Dirichlet Allocation, LDA), которое позволяет представить документ в виде множества тем, описанных комбинацией ключевых слов [44]. Для применения этого метода каждая политика была разбита на множество параграфов. Кроме того, было сделано предположение, что каждый параграф может содержать описание одного сценария использования персональных данных, т.е. выбиралась тема, аффилиация текста с которой была более некоторого заданного порога θ. Таким образом, выполненный анализ включал следующие шаги:

1. извлечение текстов параграфов из документов,
2. генерирование тематических моделей на основе анализа всего набора параграфов,
3. определение возможных характеристик сценариев использования персональных данных в соответствии с полученными семантическими моделями.

Перед применением LDA каждый параграф был преобразован в вектор слов. Затем были удалены стоп-слова, после чего была выполнена лемматизация текста с помощью библиотеки Python NLTK [45]. Исследования в [11] показали, что метрика TF-IDF (Term Frequency – Inverse Document Frequency) позволяет извлечь более подробную информацию о сценариях использования данных, так как эта модель векторизации текста присваивает большие веса словам, которые

встречаются реже, соответственно она позволяет более точно определить особенности содержания текста.

Число тем задается вручную, и данный параметр определяет качество тематического моделирования. Качество порождаемых моделей может быть оценено с помощью показателя когерентности порожденных тематических моделей. Когерентность показывает, насколько неслучайно слова, являющиеся значимыми для темы, появляются в тексте. В исследовании была использована метрика c_v, рассчитывающаяся как среднее от косинусных сходств s_{cos} векторов слов ($w_{n,k}$) в тексте и векторов, отражающих темы (w_k^*) (формула 1):

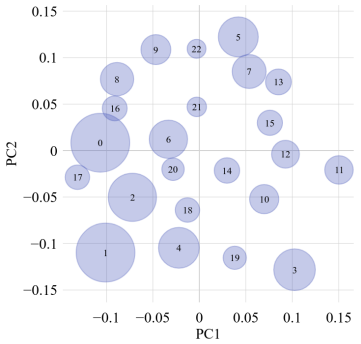
$$c_v = \frac{\sum_{k=1}^{K} \sum_{n=1}^{N} s_{cos}(w_{n,k}, w_k^*)}{N \cdot K}, \quad (1)$$

где N – количество слов, а K – количество тем. Результаты оценки приведены на рисунках 3(а) и 3(б). Оптимальное число семантических тем, на котором значение когерентности достигает максимума, для русскоязычного корпуса документов и англоязычного корпуса документов оказалось разным. Для русскоязычного корпуса PPinRussian оно равно 44, а для англоязычного IoTDataset – 23. На рисунках 3(в) и 3(г) представлена визуализация тематических кластеров параграфов, полученных в результате применения LDA и метода главных компонент (Principal Component Analysis, PCA). Радиус круга отражает число параграфов в кластере, т.е. чем больше параграфов в кластере, тем больше радиус. Кроме того, на рисунках 3(д) и 3(е) можно наблюдать кластеризацию с помощью алгоритма стохастического вложения соседей с t-распределением (t-distributed Stochastic Neighbor Embedding, t-SNE) по каждому документу с точки зрения их семантики, при этом можно заметить схожесть многих из них, что может быть связано с использованием шаблонов или составом аспектов в целом.

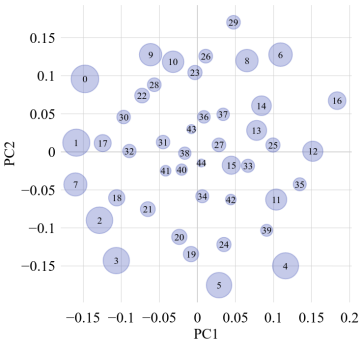
Очевидно, что и для корпуса русскоязычных документов и англоязычных документов есть кластеры, которые отличаются большим числом элементов в них, кроме того, они хорошо отделимы от других. Анализ таких кластеров показал, что к ним отнесены параграфы, которые содержат единообразную информацию, представленную одним и тем же набором слов, на основе которых может быть определена тема.

а)

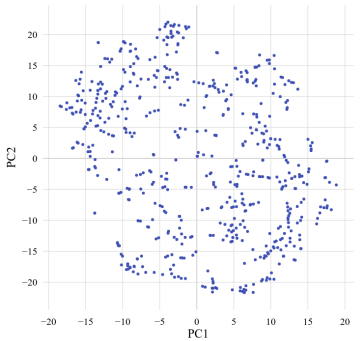
б)



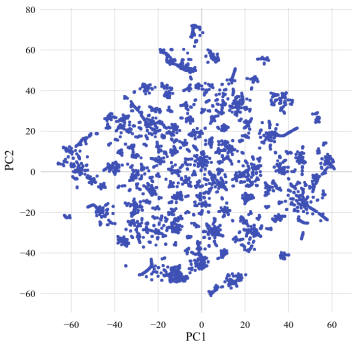
в)



г)



д)



е)

Рис. 3. Зависимость когерентности модели от количества выявляемых тем: а) датасет англоязычных политик; б) датасет русскоязычных политик; в) визуализация тематических кластеров для англоязычных документов IoTDataset; г) русскоязычных документов PPinRussian; д) визуализация семантики по каждому документу для IoTDataset; е) PPinRussian

Выделение тем, описанных характерными словами и их весами, в соответствии с показателем когерентности позволяет получить наибольшее количество таких тем, при этом среди них не наблюдается очевидных повторений. При получении тем с их количеством, заданным ниже, чем рекомендуемое в соответствии с когерентностью, произойдет утрата менее явных тем, отражающих специфические аспекты сценариев использования персональных данных. Кроме того, некоторые обобщенные темы имеют законченный смысл именно по совокупности включенных в них тем, но не обязательно пересекающихся по составу лексем, образующих их. Еще одной причиной формирования обобщенных тем является попытка получить осмысленное и описывающее некоторый аспект политики значение, которое в результате работы алгоритма LDA представляется в виде списка не связанных в единую языковую конструкцию лексем.

В таблице 2 в качестве примера представлены ключевые слова для тем 0, 1 и 2, определенные для русскоязычного корпуса документов.

Таблица 2. Ключевые слова и их веса для тем 0, 1 и 2

№	Тема 0		Тема 1		Тема 2	
	Вес	**Лексема**	**Вес**	**Лексема**	**Вес**	**Лексема**
1	0.031	сбор	0.037	электронный	0.011	ip-адрес
2	0.026	использование	0.033	адрес	0.010	выявление
3	0.021	хранение	0.027	почта	0.010	данный
4	0.021	персональный	0.016	телефон	0.008	персональный
5	0.018	предоставление	0.015	e-mail	0.008	статистика
6	0.017	операция	0.014	доставка	0.008	информация
7	0.016	данный	0.012	номер	0.007	сайт
8	0.016	средство	0.011	товар	0.007	законность
9	0.015	обновление	0.010	пользователь	0.007	пользователь
10	0.015	изменение	0.008	сайт	0.007	оператор
11	0.015	передача	0.007	информация	0.007	проблема
12	0.015	накопление	0.007	заказ	0.007	свой
13	0.014	уточнение	0.006	письмо	0.007	посетитель
14	0.014	систематизация	0.006	отправка	0.007	решение
15	0.013	удаление	0.006	информирование	0.007	цель
16	0.013	действие	0.005	посредством	0.006	использоваться
17	0.013	уничтожение	0.005	уведомление	0.006	технический
18	0.013	извлечение	0.005	услуга	0.006	осуществлять
19	0.012	обезличивание	0.005	мы	0.006	проводить
20	0.012	пользователь	0.005	связываться	0.005	актуализация

Из него следует, что тема 0 посвящена общим вопросам по сбору, обработке и использованию персональных данных, в теме 1 детализируются типы собираемых данных – электронный адрес, почта, телефон, а также кратко представлены возможные цели использования – доставка и информирование. В теме 2 также представлены типы

собираемых данных – IP-адрес, который в первую очередь используется для сбора статистики. Данные выводы получены путем ручного анализа параграфов, отнесенных к этим темам. Примерами других тем, представленных в документе являются особенности реализации обратной связи и уведомлений (тема 4), разрешение споров (тема 6), распространение персональных данных, в т.ч. их трансграничная передача (темы 17 и 39), цели использования персональных данных (тема 38), маркетинговые и новостные рассылки (тема 20), защита персональных данных (тема 42). Наиболее нетривиальной задачей оказалось определение семантических тем для кластеров небольшого объема, расположенных в центре графика рассеивания на рисунке 3(г). Выявленные ключевые слова достаточно общие, поэтому определить особенность той или иной темы не представлялось возможным, кроме того, параграфы, отнесенные к данным кластерам не имели четко выраженных общих концепций. В таблице 3 представлены примеры таких кластеров.

Таблица 3. Ключевые слова и их веса для тем 38, 41 и 42

№	Тема 38		Тема 41		Тема 42	
	Вес	**Лексема**	**Вес**	**Лексема**	**Вес**	**Лексема**
1	0.022	персональный	0.012	персональный	0.024	неправомерный
2	0.022	обработка	0.010	данный	0.018	защита
3	0.021	данный	0.010	пользователь	0.014	мера
4	0.016	цель	0.009	состояние	0.013	сайт
5	0.015	автоматизированный	0.009	сайт	0.013	случайный
6	0.013	техника	0.008	информация	0.013	копирование
7	0.013	вычислительный	0.008	категория	0.013	организационный
8	0.012	средство	0.008	оператор	0.013	персональный
9	0.010	помощь	0.007	данные	0.013	необходимый
10	0.010	несовместимый	0.007	получать	0.012	технический
11	0.009	информация	0.007	заказ	0.012	данный
12	0.007	данные	0.007	уведомление	0.011	информация
13	0.007	передача	0.007	обработка	0.011	администрация
14	0.007	допускаться	0.006	случай	0.011	принимать
15	0.006	заранее	0.006	субъект	0.011	пользователь
16	0.006	пользователь	0.005	лицо	0.010	блокирование
17	0.006	использование	0.005	услуга	0.010	доступ
18	0.006	сбор	0.005	товар	0.010	уничтожение
19	0.006	сайт	0.005	специальный	0.010	иной
20	0.006	получать	0.005	возврат	0.010	распространение

Поскольку многие темы оказались схожи между собой, было принято решение объединить их в группы, которые явно характеризуют особенности использование персональных данных. Были сформулированы следующие обобщенные темы: (1) термины и определения политики, (2) сбор трекинговых персональных данных, (3) сбор, обработка и хранение

персональных данных, (4) распространение и уничтожение персональных данных, (5) изменение персональных данных, (6) разрешение споров, (7) уведомление, маркетинг и персонализация, (8) цели обработки персональных данных, (9) защита персональных данных, (10) правовые основания обработки, (11) обновление политики безопасности, (12) согласие пользователя на обработку персональных. Объединение кластеров выполнялось с учетом извлеченных ключевых слов тем и их близости в пространстве проекций, после чего было проанализировано распределение обобщенных тем в корпусе документов. Для этого каждый параграф был отнесен к одному кластеру, если вероятность принадлежности была более или равна $\theta = 0.5$, если вероятность была ниже, то такой параграф исключался из дальнейшего анализа. Далее для параграфа определялась обобщенная группа. Результаты семантического анализа корпусов документов представлены в таблице 4.

Таблица 4. Распределение обобщенных тем в PPinRussian корпусе политик конфиденциальности на русском языке

№	**Обобщенная тема**	**Процент, %**
1	Термины и определения политики	12.1
2	Сбор трекинговых персональных данных	9.8
3	Сбор обработка и хранение персональных данных	20.4
4	Передача третьим лицам и уничтожение персональных данных	11.7
5	Изменение персональных данных	3.5
6	Разрешение споров	1.8
7	Уведомление, маркетинг, персонализация	4.4
8	Цели обработки персональных данных	9.4
9	Защита персональных данных	6.2
10	Правовые основания обработки персональных данных	12
11	Обновление политики безопасности	3.1
12	Согласие пользователя на обработку персональных данных	5.6

На рисунке 4 подробно показано распределение семантических тем в политиках конфиденциальности. Как и в предыдущем случае для каждого параграфа определялся тематический кластер, если порог принадлежности для этого кластера был $\theta \geqslant 0.5$, и обобщенная тема. Далее каждая политика конфиденциальности представлялась в виде вектора, содержащего число параграфов заданной темы, после чего все документы были разделены на кластеры в соответствии с распределением тем в текстах с помощью алгоритма кластеризации k-means. Для построения графика на рисунке 4 все политики были упорядочены сначала по вычисленным кластерам, а затем по количеству параграфов в документе. Таким образом, ось x соответствует номеру документа в упорядоченном списке всех документов, и, следовательно,

ширина столбца диаграммы пропорциональна количеству документов в соответствующем кластере. Ось y показывает число параграфов, отнесенных к каждой обобщенной теме. Таким образом, построенная диаграмма показывает среднее количество параграфов в кластере документов с заданной темой, а цветные участки каждого столбца отражают количество и соотношение тем в каждом конкретном кластере.

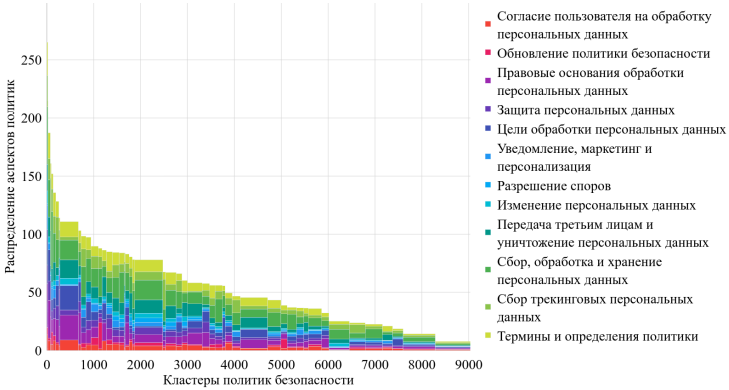


Рис. 4. Результат кластеризации русскоязычных политик безопасности с учетом распределения параграфов по обобщенным темам

Аналогичным образом были проанализированы темы англоязычного корпуса IoTDataset. Число семантических кластеров в этом корпусе почти в 2 раза меньше по сравнению с русскоязычным, что, возможно, объясняется как особенностями языка, так и тем, что в состав данного корпуса входят только политики для устройств Интернета Вещей. В большинстве случаев определить тему по ключевым словам было достаточно просто, исключение составили лишь некоторые кластеры, например кластеры 5 и 17. Примеры ключевых слов для некоторых тематических кластеров представлены в таблице 5. Например, в теме 3 обсуждаются права особой аудитории – граждан США в Калифорнии, которые защищены дополнительными законодательными актами в области обработки персональных данных, в теме 4 представлены меры безопасности по защите персональных данных, тема 6 описывает особенности обработки персональных данных специальных категорий пользователей – несовершеннолетних. В отличие от этих тем, в темах 5 и 17, приведенных в таблице 6, нет ключевых слов с ярко выраженным вкладом.

Таблица 5. Ключевые слова и их веса для тем 3, 4 и 5]

№	Тема3		Тема4		Тема5	
	Вес	Лексема	Вес	Лексема	Вес	Лексема
1	0.023	california	0.031	security	0.013	information
2	0.013	privacy	0.012	term	0.011	home
3	0.010	resident	0.011	data	0.008	may
4	0.009	right	0.009	information	0.008	access
5	0.009	notice	0.008	use	0.007	use
6	0.008	information	0.007	personal	0.007	personal
7	0.007	service	0.007	privacy	0.007	way
8	0.007	state	0.007	right	0.007	cooky
9	0.006	policy	0.006	condition	0.006	data
10	0.006	use	0.006	service	0.006	u
11	0.006	may	0.006	legal	0.005	service
12	0.005	united	0.006	sale	0.005	collect
13	0.005	cooky	0.005	b	0.005	send
14	0.005	personal	0.005	described	0.005	product
15	0.005	product	0.005	policy	0.005	provide
16	0.005	purpose	0.004	technical	0.005	email
17	0.004	website	0.004	notice	0.005	following
18	0.004	change	0.004	de	0.005	purchase
19	0.004	law	0.003	may	0.005	message
20	0.004	consumer	0.003	product	0.004	order

Таблица 6. Ключевые слова и их веса для тем 6 и 17

№	Тема 6		Тема 17	
	Вес	Лексема	Вес	Лексема
1	0.014	address	0.008	information
2	0.014	child	0.007	contract
3	0.009	information	0.007	data
4	0.008	childrens	0.006	personal
5	0.008	privacy	0.006	last
6	0.007	service	0.006	order
7	0.007	policy	0.006	service
8	0.006	website	0.005	payment
9	0.006	personal	0.005	privacy
10	0.006	site	0.005	performance
11	0.006	age	0.005	use
12	0.006	data	0.005	collect
13	0.005	may	0.005	may
14	0.005	u	0.004	policy
15	0.005	collect	0.004	u
16	0.005	name	0.004	site
17	0.005	use	0.004	transaction
18	0.005	ip	0.004	access
19	0.005	e-mail	0.004	name
20	0.005	term	0.004	website

Множества ключевых слов этих тем достаточно похожи, отличие заключается в том, что в теме 5 упоминаются данные, позволяющие

отслеживать действия пользователей в сети (cookies), а в теме 17 включены слова “платежи” (payment) и “транзакции” (transaction), что позволяет предположить, что речь идет о финансовых данных.

Для англоязычного корпуса также были сформулированы обобщенные темы: (1) вопросы пользователя по политике безопасности, (2) сторонние веб-сайты, (3) особая аудитория, (4) защита персональных данных, (5) сбор персональных данных, (6) сбор данных, позволяющих отслеживать поведения пользователя на веб-сайте, (7) распространение персональных (передача третьим лицам), (8) хранение персональных данных, (9) персонализация и маркетинг, (10) изменение политики безопасности. На рисунке 5 представлено распределение обобщенных тем в корпусе IoTDataset, а в таблице 7 показан результат кластеризации политик безопасности с учетом распределения параграфов по обобщенным темам.

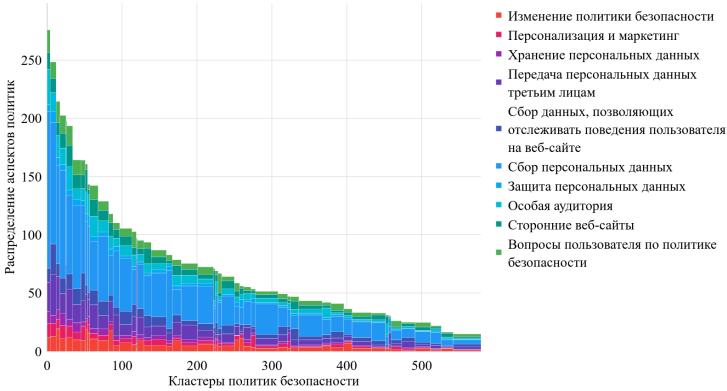


Рис. 5. Результат кластеризации англоязычных политик безопасности с учетом распределения параграфов по обобщенным темам

Очевидно, что сформированные обобщенные темы для корпуса политик на русском языке достаточно сильно отличаются от обобщенных тем корпуса политик на английском языке. Отлично также их количественное присутствие в документах. Некоторые темы присутствуют в обоих корпусах, в частности, это темы, характеризующие сбор персональных данных, данных по отслеживанию действий пользователя на веб-сайте, а также передаче третьим лицам. В обоих корпусах присутствует информация о том, какие данные собираются в

целях персонализации или маркетинга, какие меры осуществляются для защиты персональных данных.

Таблица 7. Распределение обобщенных тем в IoTDataset корпусе политик конфиденциальности на английском языке

№	Обобщенная тема	Процент, %
1	Вопросы пользователя по политике безопасности	7.5
2	Сторонние веб-сайты	7.1
3	Особая аудитория	7.7
4	Защита персональных данных	3.4
5	Сбор персональных данных	39.4
6	Сбор данных, позволяющих отслеживать поведение пользователя на веб-сайте	8.4
7	Передача персональных данных третьим лицам	11.6
8	Хранение персональных данных	3.3
9	Персонализация и маркетинг	4.6
10	Изменение политики безопасности	7

Ключевые отличия заключаются в том, что в политиках на русском языке отдельно прописываются положения по урегулированию споров и механизмов получения согласия на использование персональных данных. Особое внимание уделяется вопросам уничтожения персональных данных (до 10% от общего числа параграфов), а также прописывается ответственность пользователя за изменение персональных данных. Следует также отметить наличие достаточно объемной преамбулы в документах, где разъясняются общие термины и определения (более 10% от общего числа параграфов). В англоязычных политиках отдельно указываются особенности обработки данных, которые принадлежат субъектам, имеющим специальную категорию, например, несовершеннолетние или граждане стран, на территории которых действуют дополнительные законодательные акты в области защиты персональных данных. Прописываются особенности хранения персональных данных, часто с указанием конкретной длительности, а также детально описываются механизмы уведомления субъектов персональных данных в случае изменения политик безопасности (7% от общего числа параграфов). В политиках на английском языке значительную часть документа (39% от общего числа параграфов) так же занимает описание, каким образом осуществляется сбор данных, детализируются их типы. Такие отличия в политиках, в первую очередь, связаны с законодательными актами, действующими в Российской Федерации и/или на территориях других стран. Например, последние изменения в ФЗ “О персональных данных” № 152-ФЗ в части передачи информации третьим лицам и наличия письменного согласия субъекта персональных данных нашли отражение в политиках безопасности,

данные темы представлены достаточно объемно, занимая суммарно до 15% от объема текста (в параграфах), при этом описание сценариев хранения, сбора и обработки данных занимает всего 25% объема текста документа, что может говорить об отсутствии деталей о том, какие типы данных собираются, какие механизмы для сбора используются, и каковы сроки их хранения, что влияет на “прозрачность” и понятность самих текстов документов.

6. Заключение. Политика конфиденциальности – это официальный способ информирования пользователей о том, какие персональные данные собираются и как эти данные обрабатываются. Обычно, эти политики вызывают затруднения при чтении и понимании. В результате пользователи пропускают их и не понимают, кто и как использует их персональные данные и каковы риски нарушения их конфиденциальности. Таким образом, задача автоматизированного анализа политик безопасности, написанных на естественном языке, и их представления в прозрачной форме является весьма актуальной. Это особенно важно в настоящее время в связи с требованиями правовых документов к прозрачности обработки персональных данных с одной стороны, и стремительным развитием Интернета вещей и веб-сервисов с другой стороны. Каждый день люди используют множество “умных” устройств и сервисов, которые собирают большое количество разнообразных персональных данных, включая такие чувствительные из них, как данные о здоровье и биометрические данные, и не учитывают связанные с этим риски нарушения их конфиденциальности.

В настоящее время исследователи предложили различные подходы на основе машинного обучения для анализа политик безопасности, написанных на естественном языке, и представления их в прозрачной форме. Применение таких подходов требует использования аннотированных наборов данных политик безопасности для обучения модели анализа с целью анализа особенностей сценариев сбора и обработки данных. Авторы данной статьи провели сравнительный анализ существующих наборов данных и выделили их отличительные особенности, такие как год создания, объем и наличие аннотаций.

В статье представлены основные характеристики сформированного корпуса документов, включая результаты семантического моделирования. Семантический анализ выявил различия в темах, представленных в созданном наборе данных, по сравнению с набором данных IoTDataset. Например, были выявлены темы, связанные с правовыми основаниями сбора, терминами и определениями и разрешением споров между сторонами. Последнее означает, что при использовании для обучения

моделей анализа с соответствующими аннотациями этот набор данных может повысить точность обнаружения и рассуждений о различных аспектах сценариев использования персональных данных, включая аспекты, связанные с обязательствами обработчиков данных.

Будущие исследования будут включать в себя разработку автоматизированной проверки собранных документов, дальнейшее автоматизированное обнаружение различных аспектов использования персональных данных и расчет рисков конфиденциальности, связанных с использованием устройств или веб-сайтов.

Литература

1. Исследование утечек информации в отраслях за три года. URL: https://www.infowatch.ru/analytics/analitika/issledovaniye-utechek-informatsii-v-otraslyakh-za-tri-goda (дата обращения 20.05.2024).
2. Американские власти оштрафовали Avast за распространение персональных данных пользователей. URL: https://xakep.ru/2024/02/26/avast-ftc (дата обращения 20.05.2024).
3. Number of Internet of Things (IoT) connections worldwide from 2022 to 2023, with forecasts from 2024 to 2033. URL: https://www.statista.com/statistics/1183457/iot-connected-devices-worldwide (дата обращения 20.05.2024).
4. Самодолов А.П., Самодолова О.А., Николаенко Е.В. Особенности развития “умных домов” в России // Вестник ЮУрГУ. Серия: Строительство и архитектура. 2021. Т. 21. № 2. С. 78–85.
5. Отчет об уязвимостях в устройствах Интернета Вещей. URL: https://www.cnet.com/home/security/your-home-security-camera-could-be-hacked-so-treat-it-that-way (дата обращения 20.05.2024).
6. Mitigating Smart Meter Security Risk: A Privacy-preserving Approach. URL: https://eepower.com/technical-articles/mitigating-smart-meter-security-risk-a-privacy-preserving-approach/ (дата обращения 20.05.2024).
7. Alanazi F., Kim J., Cotilla-Sanchez E. Load Oscillating Attacks of Smart Grids: Vulnerability Analysis // IEEE Access. 2023. vol. 11. pp. 36538–36549. DOI: 10.1109/access.2023.3266249.
8. Steinfeld N. “I agree to the terms and conditions”: (How) do users read privacy policies online? An eye-tracking experiment // Computers in Human Behavior. 2016. vol. 55. part B. pp. 992–1000. DOI: 10.1016/j.chb.2015.09.038.
9. Karegar F., Pettersson J.S., Fischer-Hubner S. The Dilemma of User Engagement in Privacy Notices: Effects of Interaction Modes and Habituation on User Attention // ACM Transactions on Privacy and Security (TOPS). 2020. vol. 23. no. 1. pp. 1–38. DOI: 10.1145/3372296.
10. Регламент Европейского регулирования персональных данных. URL: http://data.europa.eu/eli/reg/2016/679/oj (дата обращения 20.05.2024).
11. Harkous H., Fawaz K, Lebret R, Schaub F, Shin KG, Aberer K. Polisis: automated analysis and presentation of privacy policies using deep learning // Proceedings of the 27th USENIX Security Symposium. 2018. pp. 531–548.
12. Novikova E., Doynikova E., Kotenko I. P2Onto: Making Privacy Policies Transparent // Computer Security, CyberICPS SECPRE ADIoT 2020, Proceedings of the International Workshop on Attacks and Defenses for Internet-of-Things. 2020. pp. 235–252.

13. Kuznetzov M., Novikova E. Towards application of text mining techniques to the analysis of the privacy policies // Proceedings of the 10th Mediterranean Conference on Embedded Computing. 2021. pp. 1–4. DOI: 10.1109/meco52532.2021.9460130.
14. Ahmad W., Chi J., Tian Y., Chang K.-W. PolicyQA: A Reading Comprehension Dataset for Privacy Policies // Proceedings of the Findings of the Association for Computational Linguistics (EMNLP). 2020. pp. 743–749.
15. Harkous H., et al. Polisis: automated analysis and presentation of privacy policies using deep learning // Proceedings of the 27th USENIX Conference on Security Symposium. 2018. pp. 531–548.
16. Zaeem R.N., German R.L., Barber K.S. PrivacyCheck: Automatic Summarization of Privacy Policies Using Data Mining // ACM Transactions on Internet Technology. 2018. vol. 18. vol. 4. DOI: 10.1145/3127519
17. Kuznetsov M., et al. Privacy Policies of IoT Devices: Collection and Analysis // Sensors. 2022. vol. 22. no. 5. DOI: 10.3390/s22051838.
18. Правила защиты конфиденциальности детей в Интернете. URL: https://www.ftc.gov/legal-library/browse/rules/childrens-online-privacy-protection-rule-coppa (дата обращения 20.05.2024).
19. Palmirani M., Martoni M., Rossi A., Bartolini C., Robaldo L. Legal ontology for modelling GDPR concepts and norms // Legal Knowledge and Information Systems. Amsterdam: IOS Press. 2018. vol. 313. pp. 91–100. DOI: 10.3233/978-1-61499-935-5-91.
20. Pandit H.J., O'Sullivan D., Lewis D. An Ontology Design Pattern for Describing Personal Data in Privacy Policies // 9th Workshop on Ontology Design and Patterns. 2018. vol. 2195. pp. 29–39.
21. Oltramari A., Piraviperumal D., Schaub F., Wilson S., Cherivirala S., Norton T.B., Russel N.C., Story P., Reidenberg, Sadeh N. PrivOnto: a semantic framework for the analysis of privacy policies // Semantic Web. 2018. vol. 9. no. 2. pp. 185–203.
22. Cano-Benito J., Cimmino A., Garcia-Castro R. Toward the ontological modeling of smart contracts: A solidity use case // IEEE Access. 2021. vol. 9. pp. 140156–140172. DOI: 10.1109/access.2021.3115577.
23. Wilson Ah., et al. The Creation and Analysis of a Website Privacy Policy Corpus // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2016. pp. 1330–1340. DOI: 10.18653/v1/P16-1126.
24. Zimmeck S., et al. MAPS: scaling privacy compliance analysis to a million apps // In Proceedings on Privacy Enhancing Technologies 2019. vol. 3. pp. 66–86. DOI: 10.2478/popets-2019-0037.
25. Kumar V.H., Iyengar R., Nisal N., Feng Y., Habib H., Story P., Cherivirala S., Nagan M., Cranor L., Wilson S., Schaud F., Sadeh N. Finding a Choice in a Haystack: Automatic Extraction of Opt-Out Statements from Privacy Policy Text // Proceedings of The Web Conference. 2020. pp. 1943–1954. DOI: 10.1145/3366423.3380262.
26. Hosseini M.B., Heaps J., Slavin R., Niu J., Breaux T. Ambiguity and Generality in Natural Language Privacy Policies // IEEE 29th International Requirements Engineering Conference (RE). 2021. pp. 70–81. DOI: 10.1109/RE51729.2021.00014.
27. Hosseini M.B., Breaux T., Slavin R., Niu J., Wang X. Analyzing Privacy Policies through Syntax-Driven Semantic Analysis of Information Types // Information and Software Technology Journal. 2021. vol. 138. DOI: 10.1016/j.infsof.2021.106608.
28. Веб-страница проекта Usable Privacy Policy. URL: https://usableprivacy.org (дата обращения 21.05.2024).
29. Веб-сайт Amazon Alexa. URL: https://www.alexa.com (дата обращения 22.05.2024).

30. Poplavska E., Norton T.B., Wilson S., Sadeh N. From Prescription to Description: Mapping the GDPR to a Privacy Policy Corpus Annotation Scheme // Proceedings of the 33rd International Conference on Legal Knowledge and Information Systems. 2020. pp. 243–246.
31. Веб-сайт сервиса Google Play. URL: https://play.google.com/store (дата обращения 24.05.2024).
32. Amos R., Acar G., Kshirsagar M., Narayanan A., Mayer J. Privacy Policies over Time: Curation and Analysis of a Million-Document Dataset // Proceedings of the Web Conference. 2021. pp. 2165–2176. DOI: 10.1145/3442381.3450048.
33. Zaeem R.N., Barber K.S. A Large Publicly Available Corpus of Website Privacy Policies Based on DMOZ // In Proceedings of the Eleventh ACM Conference on Data and Application Security and Privacy. 2021. pp. 143–148. DOI: 10.1145/3422337.3447827.
34. Веб-директория Curlie. URL: https://curlie.org (дата обращения 26.05.2024).
35. Srinath M., Wilson S., Giles C. Privacy at Scale: Introducing the PrivaSeer Corpus of Web Privacy Policies // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. 2021. pp. 6829–6839. DOI: 10.18653/v1/2021.acl-long.532.
36. Веб-сайт Amazon. URL: https://www.amazon.com (дата обращения 26.05.2024).
37. Веб-сайт Walmart. URL: https://www.walmart.com/ (дата обращения 28.05.2024).
38. Hamid A., Samidi H.R., Finin T., Pappachan P., Yus R. PrivacyLens: A Framework to Collect and Analyze the Landscape of Past, Present, and Future Smart Device Privacy Policies // arXiv pradprint arXiv.2308.05890. 2023.
39. Polisis. URL: https://www.epfl.ch/labs/lsir/polisis/ (дата обращения 02.06.2024).
40. Ravichander A., Black A., Wilson S., Norton T., Sadeh N. Question Answering for Privacy Policies: Combining Computational and Legal Perspectives // Proceedings of 2019 Conference on Empirical Methods in Natural Language Processing. 2019. pp. 4947–4958. DOI: 10.18653/v1/D19-1500.
41. Веб-сайт аналитической площадки Mail.ru Top. URL: https://top.mail.ru (дата обращения 02.06.2024).
42. Веб-сайт аналитической площадки Rambler Top-100. URL: https://top100.rambler.ru (дата обращения 02.06.2024).
43. Политика безопасности компании Huawei. URL: https://www.huawei.com/eu/privacy-policy (дата обращения 02.06.2024).
44. Blei D., Ng A., Jordan M. Latent Dirichlet Allocation // Journal of Machine Learning Research. 2003. vol. 3. pp. 993–1022.
45. Веб-сайт библиотеки NLTK. URL: https://www.nltk.org (дата обращения 02.06.2024).

Кузнецов Михаил Дмитриевич — младший научный сотрудник, лаборатория проблем компьютерной безопасности, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук (СПб ФИЦ РАН). Область научных интересов: компьютерные технологии, методы онтологического моделирования и формализации текста. Число научных публикаций — 23. mkuznetsov7991@gmail.com; 14 линия В.О., 39, 199178, Санкт-Петербург, Россия; р.т.: +7(911)830-0669.

Новикова Евгения Сергеевна — канд. техн. наук, старший научный сотрудник, лаборатория проблем компьютерной безопасности, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук (СПб ФИЦ РАН). Область научных интересов: безопасность информационных систем, обнаружение аномалий методами машинного обучения, конфиденциальность данных. Число научных публикаций — 60. novikova@comsec.spb.ru; 14 линия В.О., 39, 199178, Санкт-Петербург, Россия; р.т.: +7(812)328-7181.

DOI 10.15622/ia.24.1.7

M. KUZNETSOV, E. NOVIKOVA

CORPUS OF PRIVACY POLICIES FOR WEB SERVICES AND INTERNET OF THINGS DEVICES FOR ANALYZING THE AWARENESS OF PERSONAL DATA SUBJECTS

Kuznetsov M.D., Novikova E.S. **Corpus of privacy policies for web services and Internet of Things devices for analyzing the awareness of personal data subjects.**

Abstract. Information about what personal data is collected and processed by various devices and digital services is presented in privacy policies, however, as studies show, users rarely read them and, as a result, do not realize which data security risks associated with the processing of personal data arise. The solution to the problem of increasing the awareness of personal data subjects is associated with the development of decision support methods that present privacy policies in a form that is easier to understand, for example, in the form of quantitative risk assessments and pictograms. Their development requires a structured and marked-up corpus of documents. This paper systematizes the corpora of privacy policies that are in the open access and shows their distinctive characteristics, such as the year of creation, volume and presence of annotations. A description of a new corpus of documents written in Russian is also presented, the results of a structural and semantic analysis of the collected security policies are given, and a comparison with the corpus of privacy policies written in English is made. It has been shown that the description of scenarios for storing, collecting and processing data in documents in Russian accounts for only 25% of the volume of the document text, which may indicate a lack of details about what types of data are collected, what mechanisms are used for collection, and what are the storage periods, which affects the "transparency" of the use of personal data.

Keywords: personal data, privacy policies, document corpus, semantic analysis, Latent Dirichlet allocation.

References

1. Issledovanie utechek informacii v otrasljah za tri goda [Study of information leaks in industries over three years]. Avaliable at: https://www.infowatch.ru/analytics/analitika/issledovaniye-utechek-informatsii-v-otraslyakh-za-tri-goda (accessed 20.05.2024). (In Russ.).
2. Amerikanskie vlasti oshtrafovali Avast za rasprostranenie personal'nyh dannyh pol'zovatelej [American authorities fined Avast for distributing user personal data]. Available at: https://xakep.ru/2024/02/26/avast-ftc (accessed 20.05.2024). (In Russ.).
3. Number of Internet of Things (IoT) connections worldwide from 2022 to 2023, with forecasts from 2024 to 2033. Available at: https://www.statista.com/statistics/1183457/iot-connected-devices-worldwide (accessed 20.05.2024).
4. Samodolov A.P., Samodolova O.A., Nikolaenko E.V. [Features of the development of "smart homes" in Russia]. Vestnik JuUrGU. Serija: Stroitel'stvo i arhitektura – Bulletin of SUSU. Series: Construction and architecture. 2021. vol. 21. no. 2. pp. 78–85. (In Russ.).
5. Internet of Things Device Vulnerability Report. Available at: https://www.cnet.com/home/security/your-home-security-camera-could-be-hacked-so-treat-it-that-way (accessed 20.05.2024).
6. Mitigating Smart Meter Security Risk: A Privacy-preserving Approach. Available at: https://eepower.com/technical-articles/mitigating-smart-meter-security-risk-a-privacy-preserving-approach/ (accessed 20.05.2024).

7. Alanazi F., Kim J., Cotilla-Sanchez E. Load Oscillating Attacks of Smart Grids: Vulnerability Analysis. IEEE Access. 2023. vol. 11. pp. 36538–36549. DOI: 10.1109/access.2023.3266249.
8. Steinfeld N. “I agree to the terms and conditions”: (How) do users read privacy policies online? An eye-tracking experiment. Computers in Human Behavior. 2016. vol. 55. part B. pp. 992–1000. DOI: 10.1016/j.chb.2015.09.038.
9. Karegar F., Pettersson J.S., Fischer-Hubner S. The Dilemma of User Engagement in Privacy Notices: Effects of Interaction Modes and Habituation on User Attention. ACM Transactions on Privacy and Security (TOPS). 2020. vol. 23. no. 1. pp. 1–38. DOI: 10.1145/3372296.
10. General data protection regulation. Available at: http://data.europa.eu/eli/reg/2016/679/oj (accessed 20.05.2024).
11. Harkous H., Fawaz K, Lebret R, Schaub F, Shin KG, Aberer K. Polisis: automated analysis and presentation of privacy policies using deep learning. Proceedings of the 27th USENIX Security Symposium. 2018. pp. 531–548.
12. Novikova E., Doynikova E., Kotenko I. P2Onto: Making Privacy Policies Transparent. Computer Security, CyberICPS SECPRE ADIoT 2020, Proceedings of the International Workshop on Attacks and Defenses for Internet-of-Things. 2020. pp. 235–252.
13. Kuznetzov M., Novikova E. Towards application of text mining techniques to the analysis of the privacy policies. Proceedings of the 10th Mediterranean Conference on Embedded Computing. 2021. pp. 1–4. DOI: 10.1109/meco52532.2021.9460130.
14. Ahmad W., Chi J., Tian Y., Chang K.-W. PolicyQA: A Reading Comprehension Dataset for Privacy Policies. Proceedings of the Findings of the Association for Computational Linguistics (EMNLP). 2020. pp. 743–749.
15. Harkous H., et al. Polisis: automated analysis and presentation of privacy policies using deep learning. Proceedings of the 27th USENIX Conference on Security Symposium. 2018. pp. 531–548.
16. Zaeem R.N., German R.L., Barber K.S. PrivacyCheck: Automatic Summarization of Privacy Policies Using Data Mining. ACM Transactions on Internet Technology. 2018. vol. 18. vol. 4. DOI: 10.1145/3127519.
17. Kuznetsov M., et al. Privacy Policies of IoT Devices: Collection and Analysis. Sensors. 2022. vol. 22. no. 5. DOI: 10.3390/s22051838.
18. Children’s Online Privacy Protection Rule. Available at: https://www.ftc.gov/legal-library/browse/rules/childrens-online-privacy-protection-rule-coppa (accessed 20.05.2024).
19. Palmirani M., Martoni M., Rossi A., Bartolini C., Robaldo L. Legal ontology for modelling GDPR concepts and norms. Legal Knowledge and Information Systems. Amsterdam: IOS Press. 2018. vol. 313. pp. 91–100. DOI: 10.3233/978-1-61499-935-5-91.
20. Pandit H.J., O’Sullivan D., Lewis D. An Ontology Design Pattern for Describing Personal Data in Privacy Policies. 9th Workshop on Ontology Design and Patterns. 2018. vol. 2195. pp. 29–39.
21. Oltramari A., Piraviperumal D., Schaub F., Wilson S., Cherivirala S., Norton T.B., Russel N.C., Story P., Reidenberg, Sadeh N. PrivOnto: a semantic framework for the analysis of privacy policies. Semantic Web. 2018. vol. 9. no. 2. pp. 185–203.
22. Cano-Benito J., Cimmino A., Garcia-Castro R. Toward the ontological modeling of smart contracts: A solidity use case. IEEE Access. 2021. vol. 9. pp. 140156–140172. DOI: 10.1109/access.2021.3115577.
23. Wilson Ah., et al. The Creation and Analysis of a Website Privacy Policy Corpus. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2016. pp. 1330–1340. DOI: 10.18653/v1/P16-1126.

24. Zimmeck S., et al. MAPS: scaling privacy compliance analysis to a million apps. In Proceedings on Privacy Enhancing Technologies 2019. vol. 3. pp. 66–86. DOI: 10.2478/popets-2019-0037.
25. Kumar V.H., Iyengar R., Nisal N., Feng Y., Habib H., Story P., Cherivirala S., Nagan M., Cranor L., Wilson S., Schaud F., Sadeh N. Finding a Choice in a Haystack: Automatic Extraction of Opt-Out Statements from Privacy Policy Text. Proceedings of The Web Conference. 2020. pp. 1943–1954. DOI: 10.1145/3366423.3380262.
26. Hosseini M.B., Heaps J., Slavin R., Niu J., Breaux T. Ambiguity and Generality in Natural Language Privacy Policies. IEEE 29th International Requirements Engineering Conference (RE). 2021. pp. 70–81. DOI: 10.1109/RE51729.2021.00014.
27. Hosseini M.B., Breaux T., Slavin R., Niu J., Wang X. Analyzing Privacy Policies through Syntax-Driven Semantic Analysis of Information Types. Information and Software Technology Journal. 2021. vol. 138. DOI: 10.1016/j.infsof.2021.106608.
28. Usable Privacy Policy website. Available at: https://usableprivacy.org (accessed 21.05.2024).
29. Amazon Alexa website. Available at: https://www.alexa.com (accessed 22.05.2024).
30. Poplavska E., Norton T.B., Wilson S., Sadeh N. From Prescription to Description: Mapping the GDPR to a Privacy Policy Corpus Annotation Scheme. Proceedings of the 33rd International Conference on Legal Knowledge and Information Systems. 2020. pp. 243–246.
31. Google Play website. Available at: https://play.google.com/store (accessed 24.05.2024).
32. Amos R., Acar G., Kshirsagar M., Narayanan A., Mayer J. Privacy Policies over Time: Curation and Analysis of a Million-Document Dataset. Proceedings of the Web Conference. 2021. pp. 2165–2176. DOI: 10.1145/3442381.3450048.
33. Zaeem R.N., Barber K.S. A Large Publicly Available Corpus of Website Privacy Policies Based on DMOZ. In Proceedings of the Eleventh ACM Conference on Data and Application Security and Privacy. 2021. pp. 143–148. DOI: 10.1145/3422337.3447827.
34. Curlie web directory. Available at: https://curlie.org (accessed 26.05.2024).
35. Srinath M., Wilson S., Giles C. Privacy at Scale: Introducing the PrivaSeer Corpus of Web Privacy Policies. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. 2021. pp. 6829–6839. DOI: 10.18653/v1/2021.acl-long.532.
36. Amazon website. Available at: https://www.amazon.com (accessed 26.05.2024).
37. Walmart website. Available at: https://www.walmart.com/ (accessed 28.05.2024).
38. Hamid A., Samidi H.R., Finin T., Pappachan P., Yus R. PrivacyLens: A Framework to Collect and Analyze the Landscape of Past, Present, and Future Smart Device Privacy Policies. arXiv pradprint arXiv.2308.05890. 2023.
39. Polisis. Available at: https://www.epfl.ch/labs/lsir/polisis/ (accessed 02.06.2024).
40. Ravichander A., Black A., Wilson S., Norton T., Sadeh N. Question Answering for Privacy Policies: Combining Computational and Legal Perspectives. Proceedings of 2019 Conference on Empirical Methods in Natural Language Processing. 2019. pp. 4947–4958. DOI: 10.18653/v1/D19-1500.
41. Web analytics platform Mail.ru Top. Available at: https://top.mail.ru (accessed 02.06.2024).
42. Web analytics platform Rambler Top-100. Available at: https://top100.rambler.ru (accessed 02.06.2024).
43. Huawei privacy policy. Available at: https://www.huawei.com/eu/privacy-policy (accessed 02.06.2024).

44. Blei D., Ng A., Jordan M. Latent Dirichlet Allocation. Journal of Machine Learning Research. 2003. vol. 3. pp. 993–1022.

45. NLTK library website. Available at: https://www.nltk.org (accessed 02.06.2024).

Kuznetsov Mikhail — Junior researcher, Laboratory of computer security problems, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS). Research interests: computer technologies, methods of ontological modeling and text formalization. The number of publications — 23. mkuznetsov7991@gmail.com; 39, 14-th Line V.O., 199178, St. Petersburg, Russia; office phone: +7(911)830-0669.

Novikova Evgenia — Ph.D., Senior researcher, Laboratory of computer security problems, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS). Research interests: privacy and personal data security, privacy-preserving computations, and machine learning-based anomaly and intrusion detection. The number of publications — 60. novikova@comsec.spb.ru; 39, 14-th Line V.O., 199178, St. Petersburg, Russia; office phone: +7(812)328-7181.

УДК 004.056.53 **DOI** 10.15622/ia.24.1.8

М.В. Евсюков

СУБЪЕКТОЗАВИСИМЫЙ МЕТОД ОБНАРУЖЕНИЯ АТАК НА БИОМЕТРИЧЕСКОЕ ПРЕДЪЯВЛЕНИЕ В СИСТЕМАХ РАСПОЗНАВАНИЯ ДИКТОРА НА ОСНОВЕ ОБНАРУЖЕНИЯ АНОМАЛИЙ

Евсюков М.В. **Субъектозависимый метод обнаружения атак на биометрическое предъявление в системах распознавания диктора на основе обнаружения аномалий.**

Аннотация. Основная тенденция, присущая современным исследованиям в области обнаружения атак на биометрическое предъявление, заключается в том, что в большинстве работ применяется субъектонезависимый подход. Тем не менее, существует ряд исследований, свидетельствующих о перспективности применения субъектозависимого подхода, который подразумевает использование информации о предполагаемой личности субъекта для увеличения точности обнаружения спуфинга. В связи с этим, цель данной работы – реализация субъектозависимого метода обнаружения атак на биометрическое предъявление в системах распознавания диктора на основе обнаружения аномалий, а также его экспериментальная оценка применительно к задаче обнаружения синтезированной речи и преобразованного голоса. Для извлечения признаков используются искусственные нейронные сети, предобученные для задач обнаружения атак на биометрическое предъявление, распознавания диктора и распознавания звуковых паттернов. В качестве классификаторов применяется ряд моделей обнаружения аномалий, каждая из которых обучается на подлинных данных целевого диктора. Экспериментальная оценка предложенного метода с использованием набора данных ASVspoof 2019 LA показывает, что лучшая субъектозависимая система обнаружения атак на биометрическое предъявление, использующая нейронную сеть, предобученную для распознавания дикторов, обеспечивает EER (Equal Error Rate, равный процент ошибок) равный 4.74%. Данный результат свидетельствует о том, что признаки, извлечённые сетями, предобученными для распознавания диктора, содержат полезную информацию для обнаружения атак на биометрическое предъявление. Кроме того, предложенный метод позволил увеличить точность трёх базовых систем ОАБП, предназначенных для обнаружения синтезированного голоса. При проведении экспериментов с двумя базовыми системами на наборе данных ASVspoof 2019 LA улучшение EER составило 7.1% и 9.2%, а min t-DCF – 4.6%, относительно исходного результата. При проведении экспериментов с третьей базовой системой на наборе данных ASVspoof 2021 LA улучшение EER составило 3.9% относительно исходного результата с незначительным улучшением min t-DCF.

Ключевые слова: субъектозависимый подход, обнаружение спуфинга, обнаружение атак на биометрическое предъявление, биометрические системы, голосовая биометрия, трансфер обучения, обнаружение аномалий.

1. Введение. Современные методы распознавания диктора демонстрируют высокую точность при обработке подлинного человеческого голоса [1], однако их главным недостатком является уязвимость атакам на биометрическое предъявление [2]. Под термином «атака на биометрическое предъявление» (АБП) понимается предъявление биометрической системе скопированного,

сгенерированного, преобразованного или искажённого сигнала биометрической характеристики с целью вмешательства в процесс её функционирования [3]. Термин «спуфинг-атака» является синонимом термина «атака на биометрическое предъявление». В связи с высокой актуальностью угрозы АБП обнаружение атак на биометрическое предъявление (ОАБП) является важнейшим направлением исследований, а подсистема ОАБП является необходимой составной частью современных голосовых биометрических систем [3].

В то время как первые исследования в области ОАБП опирались на использование статистических моделей [4] и конструирование признаков [5], развитие методов машинного обучения повлекло за собой распространение глубоких нейронных сетей [6], что в свою очередь позволило создать сквозные системы (также известные как интегральные [7]), которые принимают на вход необработанное аудио без предварительного извлечения признаков [8]. В настоящий момент наиболее актуальными задачами в рассматриваемой области являются противодействие дипфейк-атакам, состязательным атакам на системы распознавания диктора и системы ОАБП, а также разработка систем распознавания диктора, обладающих встроенной защитой от данных видов атак. Именно на решение перечисленных задач была направлена недавняя конференция ASVspoof 5 [9].

Видное место на международных конференциях и конкурсах, посвящённых обнаружению АБП, направленных против голосовых биометрических систем, занимают работы, выполненные российскими учёными. Система, предложенная в работе [10], заняла второе место на конкурсе ASVspoof 2015, посвящённому обнаружению синтеза речи и преобразования голоса. Система, предложенная в работе [11], заняла первое место на конкурсе ASVspoof 2017, посвящённому обнаружению АБП, направленных против голосовых биометрических систем, использующих повторное воспроизведение. Системы, предложенные в работе [12] заняли первое место в конкурсе ASVspoof 2019 в секции Logical Access (обнаружение синтеза речи и преобразования голоса) среди одиночных систем и второе место среди систем-ансамблей, а также третье место в секции Physical Access (обнаружение повторного воспроизведения) среди одиночных систем и второе место среди систем-ансамблей. Системы, предложенные в работе [13], заняли первое место на конкурсе ASVspoof 2021 в секциях Logical Access и Deepfake, а также третье место в секции Physical Access.

Основная тенденция, присущая современным исследованиям в области ОАБП, заключается в том, что в большинстве работ применяется субъектонезависимый подход. Это означает, что

создатели систем ОАБП обучают модель машинного обучения на большом наборе данных, который содержит примеры голосов разных людей. Обученная таким образом модель ОАБП способна отличать подлинный голос от АБП, независимо от личности диктора, даже для дикторов, голоса которых не включены в обучающий набор данных.

Несмотря на то, что модель ОАБП, как правило, обучается с использованием субъектонезависимого подхода, системы ОАБП обычно функционируют во взаимодействии с системами верификации диктора, которые обладают информацией о предполагаемой личности субъекта. Существуют исследования, которые демонстрируют, что применение этой информации в рамках системы ОАБП позволяет повысить её точность.

Например, в работе [14], посвящённой обнаружению атак повтором, проанализировано влияние разнообразия дикторов в наборе данных на распределения голосовых признаков. Данное исследование приводит экспериментальное обоснование того, что подлинный голос и примеры АБП, использующие повторное воспроизведение, проще отличить друг от друга в случае распределений голосовых признаков одного диктора, чем в случае распределений множества дикторов. Кроме того, авторы работы [14] создают субъектозависимые системы ОАБП путём адаптации моделей смеси гауссовых распределений и нейронных сетей для конкретных дикторов, используя подлинные и сфабрикованные данные. В результате разработанные субъектозависимые системы при прочих равных демонстрируют бóльшую точность, чем их субъектонезависимые аналоги. Похожее исследование для обнаружения синтезированного голоса представлено в работах [15, 16]. Основное отличие данного исследования от работ [14 – 16] заключается в том, что в качестве классификаторов используется набор методов обнаружения аномалий.

В другой работе [17] обучается свёрточная нейронная сеть xResNet при помощи функции потерь OC-Softmax [18] для задачи субъектонезависимого обнаружения синтезированного голоса. Далее данная сеть используется для извлечения признаков, на которых обучается PLDA-модель (Probabilistic Linear Discriminant Analysis, вероятностный линейный дискриминантный анализ) [19], которая также выполняет субъектонезависимое ОАБП. Затем глобальная PLDA-модель адаптируется для целевых дикторов при помощи их подлинных данных.

Структура системы ОАБП, рассматриваемая данной работе, похожа на структуру, описанную в [17]. В обеих работах рассматриваются субъектозависимые модели ОАБП, обученные на

признаках, для извлечения которых используются глубокие нейронные сети. Однако отличие данной работы от [17] заключается в том, что в качестве классификатора применяется не PLDA, а набор моделей обнаружения аномалий, для обучения которых используются только подлинные данные. Другое отличие от работы [17] заключается в том, что в данной работе для обучения и оценки моделей используется набор данных ASVspoof 2019 LA, в то время как в работе [17] применяется набор данных, полученный путём объединения нескольких баз данных. Тем не менее, использование субъектозависимого подхода не позволило авторам работы [17] дополнительно улучшить значение EER, по сравнению с предложенной ими субъектонезависимой системой.

В работе [20] предлагается субъектозависимый вариант модели ОАБП AASIST (Audio Anti-Spoofing using Integrated Spectro-Temporal Graph Attention Networks) [21], а также рассматриваются различные способы инкорпорирования специфичной информации о голосе целевого диктора в систему ОАБП. В результате субъектозависимый вариант системы демонстрирует бóльшую точность, чем субъектонезависимый. Отличие данной работы от [20] заключается в том, что, в то время как система, предложенная в [20] является информированной о личности субъекта (speaker-aware) и способна обрабатывать голоса разных дикторов, системы, предложенные в данной работе и в работах [14, 16, 17], предназначены для обработки голоса конкретного целевого диктора.

Работы [14, 16, 17, 20] являются примерами успешного увеличения точности ОАБП за счёт использования информации о личности диктора, который проходит верификацию. Однако в статье [22], в которой исследуется обнаружение атак повтором при распознавании по геометрии лица, был реализован другой способ добавления субъектозависимой информации в модель ОАБП.

В условиях практического применения разработчику системы ОАБП доступна информация о личности предполагаемого диктора, а также образцы его подлинного голоса, которые использовались для регистрации в системе биометрического распознавания или были собраны в процессе её функционирования. Целесообразно исходить из предположения, что примеры сфабрикованных данных для произвольного диктора отсутствуют, поскольку самостоятельная генерация примеров АБП силами разработчика системы не только существенно повышает трудоёмкость создания системы ОАБП, но и не позволяет обеспечить достаточное разнообразие видов спуфинга, в связи с непрерывным появлением новых угроз. Перечисленные

ограничения привели авторов работы [22] к использованию моделей обнаружения аномалий для субъектозависимого ОАБП.

Подход к бинарной классификации, известный как обучение с одним классом или обнаружение аномалий, целесообразно использовать в том случае, когда один из классов, называемый положительным или целевым, хорошо характеризуется экземплярами в обучающих данных, а для другого класса, именуемого нецелевым или отрицательным, данные полностью отсутствуют, немногочисленны или не образуют статистически репрезентативной выборки генерального распределения отрицательного класса [23]. Механизм функционирования методов обнаружения аномалий позволяет учесть тот факт, что положительный класс более полно представлен в обучающем наборе данных, чем отрицательный [24]. В связи с этим в ходе обучения положительные примеры данных используются как основные, а отрицательные – как вспомогательные, позволяющие уточнить решающую границу. В случае использования методов обнаружения аномалий применительно к задаче ОАБП, в качестве положительного (нормального) класса выступают примеры подлинных данных, а в качестве отрицательного (аномального) – примеры данных, используемых для проведения АБП. Различные методы обучения с одним классом ранее были реализованы в рамках исследований, посвящённых обнаружению АБП, направленных против голосовых биометрических систем, и продемонстрировали высокую точность, выйдя на лидирующие позиции конкурсов ASVspoof 2015 [25] и ASVspoof 2019 [18].

Несмотря на это, в большинстве современных исследований задача ОАБП рассматривается как задача классификации с несколькими классами, что подразумевает равноправное использование подлинных и сфабрикованных обучающих данных. Основное преимущество данного подхода заключается в том, что он позволяет достигнуть высокой точности противодействия известным видам АБП. С другой стороны, его основной недостаток заключается в том, что модель, обученная таким образом, не обладает достаточной обобщающей способностью против неизвестных атак. Вследствие этого эффективность противодействия системы алгоритмам спуфинга, которые не представлены в обучающем наборе данных, оказывается недостаточно высокой.

Таким образом, существует ряд научных работ, свидетельствующих об эффективности использования субъектозависимого подхода [14 – 17, 20] и методов обнаружения аномалий [18, 25] применительно к обнаружению АБП, направленных

против голосовых биометрических систем. Кроме того, в работе [22] продемонстрированы преимущества совместного использования данных техник применительно к задаче обнаружения АБП, направленных против систем биометрического распознавания по геометрии лица. Тем не менее, совместное использование субъектозависимого подхода и методов обнаружения аномалий ранее не было исследовано применительно к обнаружению АБП, направленных против систем распознавания диктора.

2. Постановка задачи. Основная цель данной работы – реализация субъектозависимого метода ОАБП в системах распознавания диктора на основе обнаружения аномалий и его экспериментальная оценка применительно к задаче обнаружения синтезированного голоса.

Для достижения поставленной цели необходимо решить следующие задачи:

– описание системы ОАБП, построенной в соответствии с субъектозависимым методом ОАБП в системах распознавания диктора на основе обнаружения аномалий;

– оценка эффективности предлагаемого метода при извлечении признаков с использованием искусственных нейронных сетей, предобученных для решения различных задач в области обработки речи и звука;

– оценка эффективности применения различных моделей обнаружения аномалий в рамках систем, построенных в соответствии с предлагаемым методом.

3. Предлагаемый метод обнаружения атак на биометрическое предъявление. Структура системы ОАБП, построенной в соответствии с субъектозависимым методом ОАБП в системах распознавания диктора на основе обнаружения аномалий, представлена на рисунке 1.

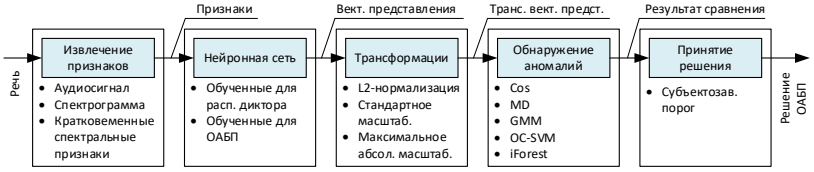


Рис. 1. Структура системы ОАБП, построенной в соответствии с субъектозависимым методом ОАБП в системах распознавания диктора на основе обнаружения аномалий

Для извлечения голосовых признаков используются предобученные искусственные нейронные сети, в связи с высокой эффективностью, которую они демонстрируют применительно к задаче обнаружения АБП, направленных против голосовых биометрических систем [6]. Поскольку разные нейронные сети обрабатывают разные голосовые признаки в качестве входных данных, выбор алгоритма извлечения признаков определяется требованиями конкретной модели машинного обучения. Далее в данном исследовании признаки, извлечённые при помощи нейронных сетей, называются векторными представлениями, чтобы подчеркнуть их отличие от признаков, для извлечения которых используются другие вычислительные методы.

Извлечённые при помощи искусственной нейронной сети векторные представления обрабатываются с использованием методов обнаружения аномалий. Известно, что точность некоторых методов обнаружения аномалий может быть повышена путём предварительного применения методов трансформации к обрабатываемым векторным данным [26]. В связи с этим, в рамках исследования предлагаемого метода оценивается влияние таких преобразований, как l2-нормализация [27], стандартное масштабирование [26], максимальное абсолютное масштабирование [26] и метод главных компонент [28], на точность ОАБП.

На этапе регистрации диктора субъектозависимая модель обнаружения аномалий обучается на трансформированных векторных представлениях с использованием примеров подлинного голоса целевого диктора. Кроме того, в соответствии с процедурой, продемонстрированной в экспериментальной части работы, вычисляется субъектозависимое пороговое значение. На этапе применения системы модель обнаружения аномалий применяется к трансформированным векторным представлениям для оценки степени подлинности предъявленных данных. Затем степень подлинности сравнивается с пороговым значением для получения решения классификации фрагмента речи: подлинный или АБП.

Процесс обучения системы ОАБП с использованием субъектонезависимого подхода представлен на рисунке 2. Процесс обучения системы ОАБП с применением субъектозависимого метода ОАБП в системах распознавания диктора на основе обнаружения аномалий представлен на рисунке 3. Из рисунка 2 видно, что субъектонезависимый подход предполагает использование как подлинных, так и сфабрикованных данных различных дикторов для обучения глобальной модели ОАБП. В то же время, как показано на

рисунке 3, субъектозависимый метод ОАБП в системах распознавания диктора на основе обнаружения аномалий предполагает использование только подлинных обучающих данных и создание собственной модели ОАБП для каждого целевого диктора.

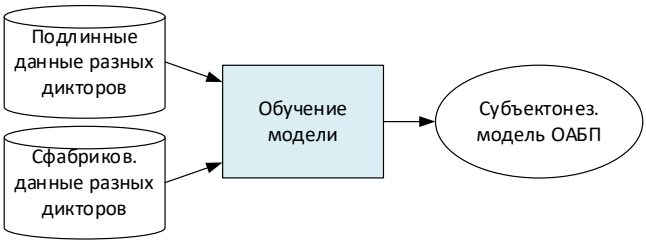


Рис. 2. Процесс обучения системы ОАБП с применением субъектонезависимого подхода

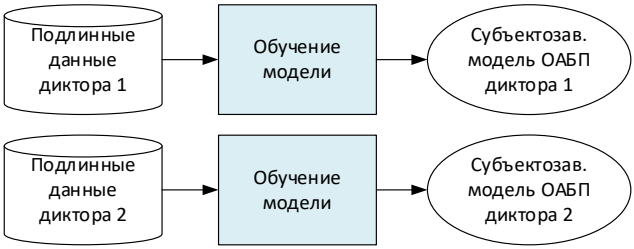


Рис. 3. Процесс обучения системы ОАБП с применением субъектозависимого метода ОАБП в системах распознавания диктора на основе обнаружения аномалий

3.1. Искусственные нейронные сети. В рамках данной работы для извлечения голосовых признаков используются три группы предобученных искусственных нейронных сетей:

– нейронные сети, предобученные для задачи ОАБП;

– нейронные сети, предобученные для задач идентификации и верификации диктора;

– нейронные сети, предобученные для задачи распознавания звуковых паттернов.

3.1.1. Нейронные сети, предобученные для задачи ОАБП. Эксперименты, в которых для извлечения признаков используются нейронные сети, предобученные для задачи обнаружения синтезированного голоса, проводятся для того, чтобы проверить возможность применения предлагаемого метода для увеличения точности существующих субъектонезависимых систем ОАБП.

Информация об используемых глубоких нейронных сетях, предобученных для задачи обнаружения синтезированного голоса, представлена в таблице 1.

Таблица 1. Используемые глубокие нейронные сети, предобученные для обнаружения синтезированного голоса

Исследование	Модель	Длительность фрагмента речи, сек.	Размер векторных представлений	Тестовый набор данных	EER, %
[8]	Res-TSSDNet	6.4	128, 64, 32	ASVspoof 2019 LA [30]	1.64
	Inc-TSSDNet		128, 64, 32		4.04
[29]	wav2vec 2.0 + AASIST	4	512	ASVspoof 2021 LA [31]	0.82 (7.65)

Все рассматриваемые сети, предобученные для задачи обнаружения синтезированного голоса, принимают необработанные фрагменты речи в качестве входных данных. Для модели, предложенной в работе [27], авторы заявляют EER равный 0.82%. Однако в открытом доступе имеется только версия модели, EER которой составляет 7.65%. Именно она используется в данной работе.

3.1.2. Нейронные сети, предобученные для задачи распознавания диктора.

Выводы, полученные в работе [32] свидетельствуют о существовании потенциала к переносу знаний от задачи распознавания диктора к задаче ОАБП при использовании методов многоцелевого обучения. В связи с этим в рамках данного исследования проводится серия экспериментов с искусственными нейронными сетями, предобученными для распознавания диктора, чтобы проверить, позволит ли применение методов обнаружения аномалий к векторным представлениям таких сетей найти решающую границу, обеспечивающую надёжное ОАБП. В случае получения успешных результатов экспериментов будет продемонстрирована возможность применения одной нейронной сети для распознавания диктора и ОАБП, что является предпосылкой для существенного снижения вычислительной нагрузки на биометрическую систему.

Информация об используемых глубоких нейронных сетях, предобученных для задачи распознавания диктора, представлена в таблице 2. Данные сети были реализованы и обучены в рамках исследований, направленных на автоматический поиск оптимальной сетевой архитектуры [33], распознавание диктора в условиях, отличных от лабораторных [34], и сквозное распознавание с использованием необработанного аудио [35, 36].

Таблица 2. Используемые глубокие нейронные сети, предобученные для задачи распознавания диктора

Исследование	Модель	Задача	Тип входных данных	Длительность фрагмента речи, сек.	Размер векторных представлений	Набор данных	EER, %
[33]	AutoSpeech	Идент.	Спектрог.	3	2048	VoxCeleb1 [37]	8.95
		Вериф.					
	ResNet18	Идент.			512		12.30
		Вериф.					
	ResNet34	Идент.					11.99
		Вериф.					
[34]	Thin ResNet VLAD	Идент.	Спектрог.	3	512	VoxCeleb2 [38]	3.22
[35]	SincNet	Идент.	Аудио	Произвольная	2048	LibriSpeech [39]	0.96
[36]	RawNet3	Идент.	Аудио	3	256	VoxCeleb1 &2 [37, 38]	0.89

3.1.3. Нейронные сети, предобученные для задачи распознавания звуковых паттернов. В работе [22] была продемонстрирована эффективность использования искусственных нейронных сетей, предобученных для задачи распознавания образов, применительно к ОАБП при защите биометрических систем распознавания по геометрии лица. В связи с этим, в рамках данной работы исследуется возможность создания субъектозависимых систем ОАБП, которые используют нейронные сети, предобученные для распознавания звуковых паттернов, с целью извлечения признаков.

В таблице 3 представлена информация об используемых в данном исследовании глубоких нейронных сетях, предобученных для задачи распознавания звуковых паттернов.

Таблица 3. Используемых глубокие нейронные сети, предобученные для задачи распознавания звуковых паттернов

Модель	Входные данные	Размер векторных представлений	mAP (показатель точности)	AUC (показатель точности)
Cnn14_16k	Логмел-спектрограмма 16 кГц	2048	0.427	0.973
Cnn14	Логмел-спектрограмма 32 кГц	2048	0.412	0.969
Cnn14_emb32	Логмел-спектрограмма 32 кГц	32	0.364	0.958
ResNet22	Логмел-спектрограмма 32 кГц	2048	0.430	0.973
Wavegram_Logmel _Cnn14	Логмел-спектрограмма 32 кГц	2048	0.439	0.973

Работа [40] является наиболее объёмным исследованием, посвящённым обучению искусственных нейронных сетей для решения задачи распознавания звуковых паттернов. В рамках неё применяется набор данных AudioSet [41], включающий в себя более 5000 часов аудиозаписей, которые являются примерами 527 различных типов звуков. Важной частью работы [40] является анализ возможности использования сетей, предобученных для задачи распознавания звуковых паттернов, применительно к другим задачам в области обработки звука посредством применения такой техники как перенос обучения [42].

В качестве входных данных рассматриваемые модели используют логмел-спектрограммы, извлечённые из аудиозаписей продолжительностью 3 секунды.

3.2. Методы обнаружения аномалий. Методы обнаружения аномалий предназначены для того, чтобы на этапе использования системы ОАБП оценить степень подлинности векторных представлений, извлечённых из предъявленного фрагмента речи. При этом применяемые в данном исследовании методы подразумевают необходимость регистрации диктора, которая, в зависимости от конкретного метода, может принимать форму обучения модели, оценки параметров распределения вероятностей или вычисления эталона для сравнения.

Исходя из того, что на практике для регистрации диктора используется ограниченное количество данных, в этом исследовании применяются простые методы обнаружения аномалий, которые включают в себя меры расстояния в пространстве признаков, вероятностные модели и поверхностные модели машинного обучения.

3.2.1. Косинусное сходство. Косинусное сходство (cos) – мера, отражающая степень подобия двух ненулевых векторов, численно равная косинусу угла между ними [43].

Косинусное сходство векторов $\vec{a}$ и $\vec{b}$ может быть рассчитано по формуле 1 [43]:

$$\cos\theta = \frac{\vec{a}\cdot\vec{b}}{\left|\vec{a}\right|\left|\vec{b}\right|}, \quad (1)$$

где θ – угол между векторами $\vec{a}$ и $\vec{b}$.

На этапе регистрации вычисляется среднее значение векторных представлений, рассчитанных для обучающего набора подлинных данных диктора, которое затем используется в качестве эталона для

сравнения. На этапе применения системы вычисляется косинусное подобие между полученным ранее эталоном и векторным представлением предъявляемого фрагмента речи, которое используется в качестве степени подлинности примера тестовых данных.

3.2.2. Расстояние Махаланобиса. Расстояние Махаланобиса (Mahalanobis Distance, MD) – мера расстояния между точкой и многомерным нормальным распределением. Расстояние Махаланобиса отличается от евклидового расстояния между точкой и средним значением некоторого распределения вероятностей инвариантностью к масштабированию, а также тем, что при вычислении расстояния Махаланобиса учитываются существующие корреляции между параметрами случайной величины [44].

Расстояние Махаланобиса между точкой x и распределением вероятностей F может быть рассчитано по формуле 2 [44]:

$$MD(x,F) = \sqrt{(x-\mu)^T S^{-1}(x-\mu)}\,, \qquad (2)$$

где μ – математическое ожидание распределения вероятностей F, S – матрица ковариации распределения вероятностей F.

На этапе регистрации, исходя из предположения, что векторные представления фрагментов речи диктора распределены по нормальному закону, с использованием тренировочного набора данных оцениваются математическое ожидание и матрица ковариации распределения вероятностей. На этапе применения системы аппроксимированные параметры распределения используются для вычисления расстояния Махаланобиса между распределением и векторным представлением тестовых данных, которое используется в качестве степени фальсифицированности примера тестовых данных.

3.2.3. Машина опорных векторов с одним классом. Машина опорных векторов (Support Vector Machine, SVM) – модель бинарной классификации, которая предусматривает обучение с учителем. Её обучение заключается в аппроксимации параметров гиперплоскости, которая разделяет различные классы данных, так, чтобы расстояние от каждого класса до неё было максимальным [45].

В рамках данного исследования применяется разновидность машины опорных векторов, предназначенная для обнаружения аномалий, которая называется «машина опорных векторов с одним классом» (One-Class SVM, OC-SVM) [45]. Её основная особенность заключается в том, что при её обучении используются данные только положительного класса.

На этапе регистрации происходит обучение машины опорных векторов с одним классом с использованием подлинных данных диктора. На этапе применения системы в качестве степени подлинности высказывания используется расстояние со знаком в пространстве векторных представлений от точки, соответствующей фрагменту речи, до разделяющей многомерной поверхности.

В ходе экспериментов выявлено, что оптимальные результаты достигаются при использовании значений гиперпараметров по умолчанию ($\nu = 0.5$, тип ядра – радиальная базисная функция).

3.2.4. Модель смеси гауссовых распределений. Модель смеси гауссовых распределений (Gaussian Mixture Model, GMM) – вероятностная модель, которая аппроксимирует многомерное распределение вероятностей при помощи взвешенной суммы конечного набора нормальных распределений [46].

Плотность вероятности модели смеси гауссовых распределений может быть представлена формулой 3 [47]:

$$p(x) = \sum_{i=1}^{M} w_i p_i(x), \tag{3}$$

где $p_i(x)$ – плотность вероятности i-го компонента смеси, M – количество компонентов смеси, w_i – вес i-го компонента смеси. При этом веса удовлетворяют ограничению $\sum_{i=1}^{M} w_i = 1$.

В свою очередь, плотность вероятности i-го компонента смеси может быть представлена формулой 4 [47]:

$$p_i(x) = \frac{1}{(2\pi)^{D/2} \left|S_i\right|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_i)^T (S_i)^{-1}(x-\mu_i)\right\}, \tag{4}$$

где D – количество измерений, μ_i – математическое ожидание i-го компонента смеси, S_i – матрица ковариации i-го компонента смеси.

На этапе регистрации набор подлинных данных диктора используется для обучения модели смеси гауссовых распределений. На этапе применения системы в качестве степени подлинности экземпляра данных используется логарифм правдоподобия того, что соответствующая тестовому фрагменту речи точка принадлежит аппроксимируемому распределению.

В связи с ограниченным объёмом обучающего набора данных в большинстве экспериментов использовалась модель с одним компонентом и полной матрицей ковариации. Были испытаны такие способы инициализации модели как *k*-средних и случайная пятикратная инициализация, однако точность ОАБП при их использовании не отличалась. Кроме того, была протестирована возможность применения диагональной матрицы ковариации, однако её использование привело к ухудшению результатов.

3.2.5. Модель изолирующего леса. Модель изолирующего леса (Isolation Forest, iForest) – модель машинного обучения, которая использует бинарные деревья для обнаружения аномалий. Принцип её функционирования основан на предположении о том, что аномальные точки проще отделить от остальных данных, чем нормальные. Чтобы изолировать экземпляр данных, алгоритм рекурсивно генерирует разделяющие гиперплоскости, случайным образом выбирая атрибут, а также его значение для разделения точек на две части [48].

На этапе регистрации выполняется обучение модели на подлинных данных диктора. На этапе применения системы в качестве степени подлинности экземпляра данных используется глубина дерева изоляции, т.е. количество гиперплоскостей, которые необходимо провести, чтобы отделить выбранную точку от всех остальных.

В связи с ограниченным количеством обучающих данных, наибольшая эффективность модели в ходе экспериментов была достигнута при использовании значений гиперпараметров по умолчанию (100 деревьев, подвыборка не выполняется).

4. Методы проведения экспериментов

4.1. Данные. В рамках экспериментальной части работы используется набор данных ASVspoof 2019 LA, который состоит из трёх подмножеств, предназначенных для обучения, разработки и тестирования моделей машинного обучения. Информация о количестве данных в подмножествах набора данных ASVspoof 2019 LA представлена в таблице 4 (учитываются только те дикторы, для которых имеются примеры подлинных и сфабрикованных данных) [30].

Таблица 4. Количество данных в подмножествах набора данных ASVspoof 2019 LA

Подмножество набора данных	Дикторы-мужчины			Дикторы-женщины		
	Число дикторов	Число записей для каждого диктора		Число дикторов	Число записей для каждого диктора	
		Подлин.	Сфабр.		Подлин.	Сфабр.
Обучение	8	132	1176	12	127	1116
Разработка	4	140	1848	6	154	2484
Тестирование	21	68	936	27	146	1638

Одна из особенностей экспериментального исследования субъектозависимых систем ОАБП заключается в том, что необходимо наличие обучающих и тестовых данных для каждого диктора. Поскольку подмножества набора данных ASVspoof 2019 LA не имеют общих дикторов [30], была реализована специальная процедура разделения данных каждого диктора на обучающие и тестовые, представленная на рисунке 4.

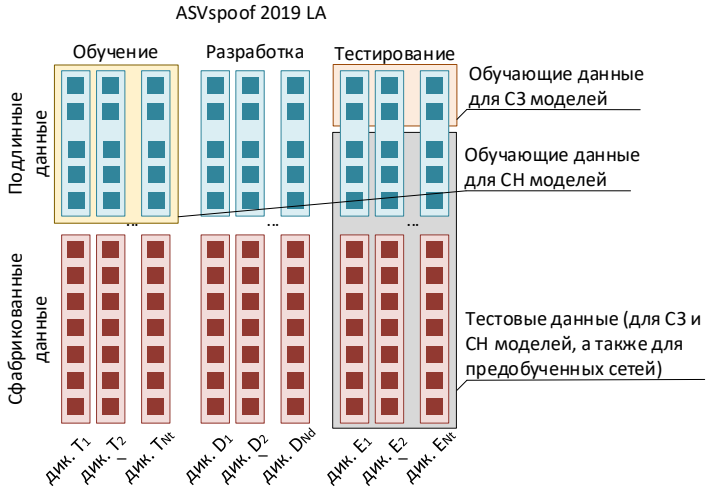


Рис. 4. Используемая процедура разделения данных на обучающие и тестовые («СЗ» – субъектозависимый, «СН» – субъектонезависимый, «дик.» – диктор)

Для обучения и тестирования субъектозависимых моделей используется тестовое подмножество набора данных ASVspoof 2019 LA, поскольку оно содержит наибольшее разнообразие примеров АБП. При этом в рамках исследования используются данные только тех дикторов, для которых имеются как подлинные, так и сфабрикованные данные. При обучении субъектозависимой модели для каждого диктора выделяются 30 подлинных фрагментов речи в качестве обучающего набора данных. Остальные подлинные данные, а также все сфабрикованные данные, используются для тестирования моделей. Для обучения субъектонезависимых моделей используется тренировочное подмножество набора данных ASVspoof 2019 LA. Экспериментальная оценка субъектонезависимых и субъектозависимых моделей проводится на одном и том же множестве фрагментов речи.

Кроме того, при исследовании предлагаемого метода применительно к системе ОАБП, описанной в работе [29], для оценки точности применяется тестовое (evaluation) подмножество набора данных ASVspoof 2021 LA, которое содержит 14816 примеров подлинных и 133360 примеров сфабрикованных аудио, распределённых неравномерно по 68 различным дикторам [31]. Как и в случае с набором данных ASVspoof 2019 LA, в рамках исследования используются данные только тех дикторов, для которых имеются подлинные и сфабрикованные данные (таких дикторов 48). Схема использования тестового подмножества набора данных ASVspoof 2021 LA аналогична схеме использования тестового подмножества набора данных ASVspoof 2019 LA, представленной на рисунке 4.

4.2. Показатели точности. Для исследования точности некоторого класса субъектозависимых систем ОАБП обучается набор идентичных систем, принадлежащих одному классу, т.е., использующих одинаковые метод обнаружения аномалий, нейронную сеть и набор алгоритмов трансформации. В качестве основного показателя точности используется среднее значение EER (Equal Error Rate, равный процент ошибок) [49] для систем исследуемого класса.

В качестве дополнительного показателя точности систем, использующих искусственные нейронные сети, предобученные для задачи обнаружения синтезированного голоса, используется среднее значение min t-DCF (минимальное значение тандемной функции стоимости обнаружения) для систем исследуемого класса, процедура вычисления которой представлена в статье [50]. При этом в текущем исследовании при вычислении min t-DCF для каждого конкретного диктора используются результаты попыток верификации, в которых он является целевым субъектом, предоставленные организаторами конкурсов ASVspoof 2019 [30] и ASVspoof 2021 [31].

Чтобы исключить влияние способа разделения подлинных данных диктора на обучающие и тестовые, эксперименты проводятся 21 раз для каждого класса систем с случайным разделением данных. Для оценки доверительных интервалов полученных значений показателей точности используется формула 5 [51]:

$$a_{CI} = \bar{a} \pm t_{0.05,20} \frac{s}{\sqrt{n}}, \quad (5)$$

где a_{CI} – 95%-ный доверительный интервал значения показателя точности, $\bar{a}$ – точечная оценка значения показателя точности (среднее

значение, полученное в ходе испытаний), $t_{0.05,20}$ – критическое значение t-распределения Стьюдента для двустороннего доверительного интервала с $\alpha = 0.05$ и 20 степенями свободы (равно 2.086), s – исправленное выборочное среднеквадратичное отклонение, n – количество экспериментов (равно 21).

4.3. Детали реализации. Экспериментальная часть исследования реализована на языке программирования Python, с применением ряда библиотек машинного обучения.

Нейронная сеть Thin ResNet VLAD [34] реализована с использованием библиотеки Keras. Остальные предобученные нейронные сети реализованы с использованием фреймворка PyTorch. При извлечении векторных представлений применялся графический процессор NVIDIA GeForce RTX 3060 GPU. Для реализации методов обнаружения аномалий использовалась библиотека Scikit-learn. При обучении и применении моделей обнаружения аномалий использовался процессор AMD Ryzen 5 5600X 6-Core 3.70 GHz.

Для оценки эффективности применения различных комбинаций методов обнаружения аномалий, искусственных нейронных сетей и алгоритмов трансформации исследовано более 570 субъектозависимых классов систем ОАБП. Для каждого класса обучено и протестировано 48 субъектозависимых систем ОАБП (в соответствии с количеством дикторов в тестовом подмножестве набора данных ASVspoof 2019 LA). Каждая из этих систем была обучена и протестирована 21 раз с различным случайным разделением данных на подлинные и тестовые. Среднее время обучения и оценки точности одной субъектозависимой системы ОАБП (для одного диктора, с единственным разбиением данных) составило 5.3 секунд.

Кроме того, были обучены 92 базовых субъектонезависимых системы ОАБП, использующих методы обнаружения аномалий.

5. Результаты экспериментов

5.1. Субъектозависимые системы, использующие нейронные сети, предобученные для распознавания диктора. В таблице 5 представлены средние EER субъектозависимых систем ОАБП, использующих методы обнаружения аномалий и искусственные нейронные сети, предобученные для задачи распознавания диктора. Для каждой комбинации нейронной сети и метода обнаружения аномалий был протестирован набор классов систем ОАБП, использующих разные трансформации, с целью определения их оптимальной конфигурации. В каждой ячейке таблицы отражён лучший результат, продемонстрированный системами ОАБП, использующими нейронную сеть, соответствующую строке таблицы

и метод обнаружения аномалий, соответствующий столбцу. Кроме того, для каждой субъектозависимой системы также приведён EER, который продемонстрировала субъектонезависимая система, использующая идентичные нейронную сеть, метод обнаружения аномалий и набор трансформаций.

Таблица 5. Средние EER (%) систем ОАБП, использующих нейронные сети, предобученные для задачи распознавания диктора, при исследовании на наборе данных ASVspoof 2019 LA («СЗ» – субъектозависимый, «СН» – субъектонезависимый)

Исследование	Модель	cos		GMM		iForest		MD		OC-SVM	
		СЗ	СН	СЗ	СН	СЗ	СН	СЗ	СН	СЗ	СН
[33]	AutoSpeech (иден.)	17.17	26.74	13.34	26.32	26.01	41.83	22.74	27.30	15.02	34.50
	AutoSpeech (вериф.)	16.20	25.04	13.17	23.16	23.88	40.72	22.49	25.81	14.90	24.59
	ResNet18 (иден.)	**4.74**	18.10	5.17	16.30	11.55	18.89	9.56	16.77	**5.06**	17.61
	ResNet18 (вериф.)	**4.79**	18.49	**4.93**	15.89	13.87	20.73	9.81	16.29	5.26	17.99
	ResNet34 (иден.)	8.21	21.42	8.64	22.43	12.83	24.16	13.47	23.47	8.61	21.73
	ResNet34 (вериф.)	6.29	19.17	7.31	18.41	10.86	19.70	12.09	19.33	6.92	19.12
[35]	SincNet	34.68	55.65	32.03	51.30	35.09	52.86	32.79	56.19	34.40	55.48
[36]	RawNet3	15.71	36.81	13.94	28.67	18.75	40.64	21.56	31.12	14.11	36.55
[34]	Thin ResNet VLAD	24.95	46.79	24.54	42.10	25.72	48.20	26.55	41.79	24.36	49.09

Анализ таблицы 5 показывает, что точность субъектозависимых систем ОАБП во всех случаях превосходит точность аналогичных субъектонезависимых систем. Кроме того, можно сделать вывод, что точность ОАБП в большей степени зависит от используемой предобученной нейронной сети, чем от метода обнаружения аномалий. Наилучший результат продемонстрировали системы ОАБП, использующие сети ResNet18, предобученные в рамках исследования [33].

Среди классов систем ОАБП, использующих одинаковую нейронную сеть, системы, в рамках которых применяются такие методы обнаружения аномалий как косинусное подобие, модель смеси гауссовых распределений и машина опорных векторов с одним классом, продемонстрировали наилучшие результаты. Применение леса изоляции и расстояния Махаланобиса приводит к ухудшению точности ОАБП, по сравнению с другими методами. При этом,

предположительного ввиду ограниченного количества данных, лес изоляции продемонстрировал наименьшую точность. По всей видимости, обработка векторных представлений размерностью 512 и более представляет существенную сложность для леса изоляции, поскольку, согласно результатам исследования [48], данный метод подвержен проблеме проклятия размерности (curse of dimensionality).

Что касается влияния трансформаций на точность ОАБП, согласно результатам экспериментов, косинусное подобие показывает наилучшие результаты при отсутствии трансформаций. Остальные методы обеспечивают наилучшие результаты при использовании l2-нормализации. Вопреки тому, что для машин опорных векторов часто рекомендуется применение масштабирования [26], его использование совместно с машинами опорных векторов с одним классом не позволило улучшить точность ОАБП. Использование метода главных компонент в качестве алгоритма трансформации в большинстве случаев приводило к заметному снижению точности. Вероятно, это связано с тем, что уменьшение размерности приводит к потере существенной части информации, необходимой для ОАБП.

В связи ограниченным объёмом обучающих данных, в ходе проведения экспериментов с субъектозависимыми системами, использующими модель смеси гауссовых распределений, наибольшая точность была достигнута при количестве компонентов смеси равным одному. Однако, поскольку субъектонезависимые системы обучаются на большем количестве данных, при их реализации использовались модели смеси гауссовых распределений с полной матрицей ковариации и количеством компонентов равным 1, 4, 16, 32 и 64. Для каждой из таких систем в таблице отражён наилучший результат. В большинстве случаев использование модели смеси гауссовых распределений с одним компонентом позволило обеспечить наименьший EER. Однако для сети ResNet34, предобученной для идентификации, и сети SincNet лучший результат был достигнут при использовании смеси с 64 компонентами, а для сети ResNet34, предобученной для верификации – при использовании смеси с 4 компонентами.

Четыре субъектозависимые системы ОАБП продемонстрировали EER равный 5.06% и менее с наибольшей предельной ошибкой 95%-ного доверительного интервала равной 0.09%. Для сравнения, в работе [8] представлены две передовые сквозные системы обнаружения синтезированного голоса, использующие нейронные сети Inc-TSSDNet и Res-TSSDNet, обученные на наборе данных ASVspoof 2019 LA, которые

демонстрируют EER 3.75% и 1.64%, соответственно. В то время как они обеспечивают более высокую точность, в ходе разработки систем, рассматриваемых в данном разделе, не использовались примеры АБП ни при обучении искусственных нейронных сетей, ни при регистрации дикторов (обучении субъектозависимых моделей обнаружения аномалий), что фактически делает все виды АБП неизвестными.

Наиболее точный класс субъектозависимых систем ОАБП, в рамках которого применяется сеть ResNet18, предобученная для задачи идентификации, и косинусное подобие продемонстрировал EER равный 4.74% (±0.07% на уровне значимости $\alpha = 0.05$). На рисунке 5 показаны значения EER, полученные для разных дикторов, в результате одной из итераций обучения и оценки систем данного класса.

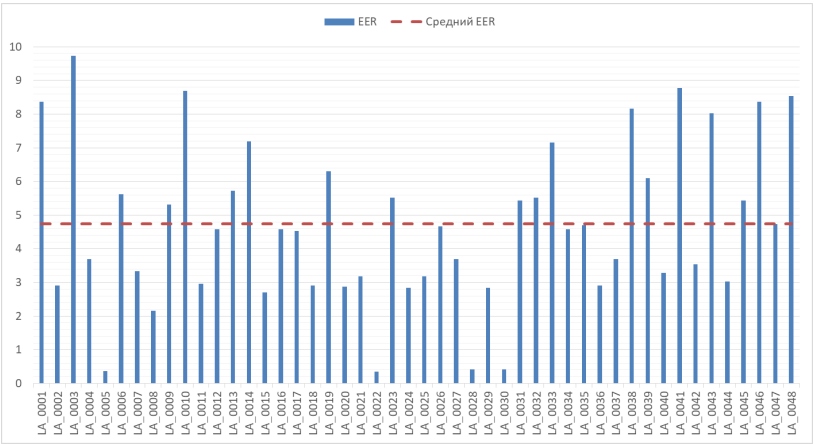


Рис. 5. Значения EER (%), полученные для разных дикторов, в результате одной из итераций обучения и оценки класса субъектозависимых систем ОАБП на наборе данных ASVspoof 2019 LA, использующих сеть ResNet18, предобученную для идентификации, и косинусное подобие

Приведённые наблюдения свидетельствуют о том, что некоторые искусственные нейронные сети, предобученные для задачи распознавания диктора, могут быть использованы для обнаружения синтезированного голоса. С практической точки зрения, полученные результаты свидетельствуют о возможности создания биометрических систем, в которых одна и та же искусственная нейронная сеть применяется для распознавания диктора и ОАБП.

5.2. Субъектозависимые системы, использующие нейронные сети, предобученные для задачи ОАБП. Таблица 6 аналогична таблице 5, представленной в предыдущем разделе. Основное отличие заключается том, что в данном разделе рассматриваются субъектозависимые системы ОАБП, которые используют нейронные сети, обученные для обнаружения синтезированного голоса в рамках исследований [8, 29]. В связи с этим, данные сети, без применения методов обнаружения аномалий, используются в качестве базовых систем ОАБП при сравнении значений показателей точности на наборе тестовых данных, проиллюстрированном на рисунке 4 (значения показателей точности для субъектонезависимых систем ОАБП, использующих методы обнаружения аномалий, не приводятся). Кроме того, для каждой системы ОАБП приведён не только EER, но также и min t-DCF, полученный в ходе испытаний.

Таблица 6. Средние значения показателей точности систем ОАБП, использующих нейронные сети, предобученные для обнаружения синтезированного голоса

Набор данных	Модель	Базовая система		cos		GMM		iForest		MD		OC-SVM	
		EER	min t-DCF	EER	min t-DCF	EER	min t-DCF	EER	min t-DCF	EER	min t-DCF	EER	min t-DCF
ASVspoof 2019 LA	Inc-TSSDNet	3.31	0.090	3.11	0.088	3.16	0.089	3.42	0.096	4.11	0.114	**3.09**	0.086
	Res-TSSDNet	1.42	0.068	**1.30**	0.065	1.84	0.079	1.87	0.087	2.89	0.089	1.36	0.067
ASVspoof 2021 LA	wav2vec 2.0 + AASIST	7.44	0.363	**7.16**	0.362	13.9	0.541	10.6	0.439	17.9	0.641	7.31	0.363

Поскольку исследуемые сети, предложенные в работе [8], имеют 4 полносвязных слоя, для каждой комбинации нейронной сети и метода обнаружения аномалий были исследованы не только различные комбинации трансформаций, но и возможность извлечения векторных представлений размеров 32, 64 и 128 значений. В ячейках таблицы отражены наилучшие результаты, полученные субъектозависимой системой, использующей указанные нейронную сеть и метод обнаружения аномалий.

Наблюдения касательно эффекта применения трансформаций на точность ОАБП, приведённые в предыдущем разделе, справедливы для экспериментов, результаты которых представлены в данном разделе.

Три класса субъектозависимых систем ОАБП, использующие сеть Inc-TSSDNet [8] для извлечения признаков, при тестировании на наборе данных ASVspoof 2019 LA [30] превзошли результат базовой системы, не использующей методы обнаружения аномалий. В рамках

данных систем применяется косинусное подобие, модель смеси гауссовых распределений и машина опорных векторов с одним классом в качестве классификаторов. Система, использующая машину опорных векторов с одним классом совместно с l2-нормализацией и обрабатывающая векторные представления размерности 64, продемонстрировала наилучший результат, относительно улучшив EER базовой системы на 7.1%, а min t-DCF – на 4.6%.

В то время как 95%-ный доверительный интервал для EER рассматриваемой субъектозависимой системы составил $3.09\% \pm 0.027\%$, 95%-ный доверительный интервал для EER базовой системы, использующей сеть Inc-TSSDNet, составил $3.31\% \pm 0.025\%$. Поскольку данные доверительные интервалы не пересекаются, приведённые результаты статистически значимы при $\alpha = 0.05$.

В то же время, 95%-ный доверительный интервал для min t-DCF рассматриваемой субъектозависимой системы составил 0.086 ± 0.0007. 95%-ный доверительный интервал для базовой системы, использующей сеть Inc-TSSDNet, составил 0.090 ± 0.0005. Поскольку данные доверительные интервалы не пересекаются, приведённые результаты статистически значимы при $\alpha = 0.05$.

Два класса субъектозависимых систем ОАБП, использующих сеть Res-TSSDNet [8], при тестировании на наборе данных ASVspoof 2019 LA [30] превзошли результат базовой системы, не использующей методы обнаружения аномалий. В качестве методов обнаружения аномалий в рамках данных систем используются косинусное подобие и машина опорных векторов с одним классом. В рамках субъектозависимых систем ОАБП обрабатываются векторные представления размерностью 64 и не используются алгоритмы трансформации. Система, использующая косинусное подобие, показывает наилучший результат, относительно превосходя EER базовой системы на 9.2%, а min t-DCF – на 4.6%.

В то время как 95%-ный доверительный интервал для EER рассматриваемой субъектозависимой системы ОАБП составил $1.30\% \pm 0.027\%$, 95%-ный доверительный интервал для EER соответствующей базовой системы составил $1.42\% \pm 0.046\%$. Поскольку данные доверительные интервалы не перекрываются, приведённые результаты статистически значимы при $\alpha = 0.05$.

В то же время, 95%-ный доверительный интервал для min t-DCF рассматриваемой субъектозависимой системы составил 0.065 ± 0.0004. 95%-ный доверительный интервал для базовой системы, использующей сеть Res-TSSDNet, составил 0.068 ± 0.0006. Поскольку

данные доверительные интервалы не пересекаются, приведённые результаты статистически значимы при $\alpha = 0.05$.

Два класса субъектозависимых систем ОАБП, использующих комбинацию сетей wav2vec 2.0 и AASIST [29] для извлечения признаков, при тестировании на наборе данных ASVspoof 2021 LA [31] превзошли результат базовой системы, не использующей методы обнаружения аномалий. В качестве методов обнаружения аномалий в рамках данных систем используются косинусное подобие (без применения трансформаций) и машина опорных векторов с одним классом (с применением l2-нормализации). Система, использующая косинусное подобие, показывает наилучший результат, относительно превосходя EER базовой системы на 3.9%.

В то время как 95%-ный доверительный интервал для EER рассматриваемой субъектозависимой системы ОАБП составил $7.16\% \pm 0.039\%$, 95%-ный доверительный интервал для EER соответствующей базовой системы составил $7.44\% \pm 0.059\%$. Поскольку данные доверительные интервалы не перекрываются, приведённые результаты статистически значимы при $\alpha = 0.05$. При этом уменьшение min t-DCF, по сравнению с базовой системой, незначительно.

Меньший относительный прирост точности в случае тестирования на наборе данных ASVspoof 2021 LA объясняется наличием искажений в тестовых данных, обусловленных их передачей по сетям связи, и применением различных кодеков [31].

Чтобы рассмотреть, как улучшения EER распределены по разным дикторам, на рисунке 6 продемонстрированы разности (улучшения) между EER базовой субъектонезависимой системы, и EER наилучшего класса субъектозависимых систем ОАБП, использующего сеть Res-TSSDNet, полученные для разных дикторов в результате одной итерации обучения и оценки систем данного класса. Как видно из рисунка 6, точность ОАБП удалось улучшить для 42 из 48 дикторов.

Таким образом, применение моделей обнаружения аномалий совместно с субъектозависимым подходом позволило увеличить точность нейронных сетей, используемых для обнаружения синтеза речи и преобразования голоса.

5.3. Субъектозависимые системы, использующие нейронные сети, предобученные для задачи распознавания звуковых паттернов. В таблице 7 представлены средние EER субъектозависимых систем ОАБП, использующих методы обнаружения аномалий и нейронные сети, предобученные для задачи

распознавания звуковых паттернов. В отличие от таблиц 5 и 6, в таблице 7 приведён EER только для субъектозависимых систем. Представленные результаты свидетельствуют о том, что сети, предобученные для задачи обнаружения звуковых паттернов в рамках исследования [40] не позволяют обеспечить эффективное извлечение признаков для дальнейшей обработки в рамках субъектозависимой системы обнаружения синтезированного голоса.

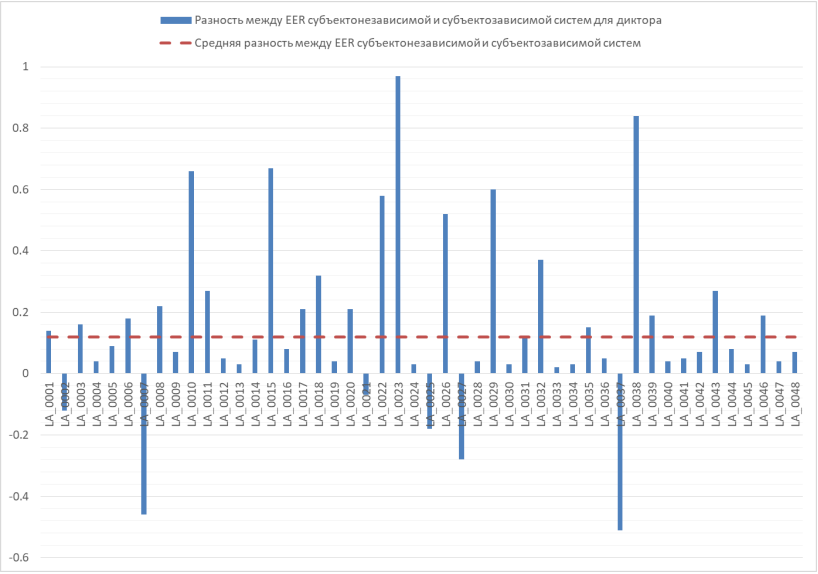


Рис. 6. Разности (улучшения) в п.п. между EER базовой субъектонезависимой системы, и EER наилучшего класса субъектозависимых систем ОАБП, использующего сеть Res-TSSDNet, полученные для разных дикторов в результате одной итерации обучения и оценки на наборе данных ASVspoof 2019 LA

Таблица 7. Средние EER (%) субъектозависимых систем ОАБП, использующих нейронные сети, предобученные для задачи распознавания звуковых паттернов, при использовании набора данных ASVspoof 2019 LA

Модель	cos	GMM	iForest	MD	OC-SVM
Cnn14	19.33	18.34	18.53	18.63	18.34
Cnn14_16k	23.13	20.21	19.79	23.45	21.22
Cnn14_emb32	24.99	29.49	29.47	30.28	24.38
ResNet22	22.84	24.75	22.67	26.37	21.78
Wavegram_Logmel_Cnn14	21.36	21.04	20.92	22.39	19.53

5.4. Пороговое значение. Методы обнаружения аномалий позволяют вычислить степень подлинности экземпляра тестовых данных. Однако для качественной работы системы ОАБП необходимо найти порог, который разделяет значения степени подлинности на подлинные и сфабрикованные таким образом, чтобы обеспечить приемлемое соотношение между количеством ложноположительных и ложноотрицательных ошибок [49].

В случае субъектонезависимой системы ОАБП существует прямолинейная процедура определения порогового значения. Вычисляются степени подлинности экземпляров тестовых данных и находится пороговое значение, которое обеспечивает требуемое соотношение процента ложных принятий (False Acceptance Rate, FAR) и процента ложных отказов (False Rejection Rate, FRR) [49].

Описанная выше процедура вычисления порогового значения неприменима для систем, построенных в соответствии с субъектозависимым методом ОАБП в системах распознавания диктора на основе обнаружения аномалий. Это связано с тем, что, во-первых, в работе [22] было продемонстрировано, что использование субъектозависимых пороговых значений совместно с субъектозависимыми моделями ОАБП обеспечивает более качественные результаты, по сравнению с использованием глобального порога, и, во-вторых, целесообразно полагать, что в сценарии практического применения субъектозависимой системы ОАБП примеры спуфинг-атак для конкретного диктора отсутствуют. В связи с этим, данном разделе статьи приводится пример выбора субъектозависимого порогового значения для систем ОАБП, построенных в соответствии с предлагаемым методом, в практическом сценарии применения.

С целью реализации данного примера для трёх дикторов из тестового подмножества набора данных ASVspoof 2019 LA были обучены идентичные субъектозависимые системы ОАБП. Данные системы используют сеть ResNet18, предобученную для идентификации диктора в рамках исследования [33], косинусное подобие и не используют алгоритмы трансформации.

Из-за отсутствия примеров АБП для конкретного диктора при выборе порогового значения возможно ориентироваться только на подлинные данные, которые могут быть использованы для вычисления FRR.

Предлагаемая процедура определения порогового значения для конкретной системы ОАБП в практическом сценарии применения состоит из двух шагов:

1. Определение FRR, который соответствует пороговому значению, задающему приемлемый средний FAR для класса субъектозависимых систем ОАБП (используются тестовые подлинные и сфабрикованные данные различных дикторов);

2. Выбор порогового значения для конкретной системы ОАБП, которое соответствует определённому ранее FRR (используются подлинные данные целевого диктора, которые не применялись для обучения модели ОАБП).

На рисунке 7 представлены кривые компромисса обнаружения и ошибок [52] для исследуемых систем.

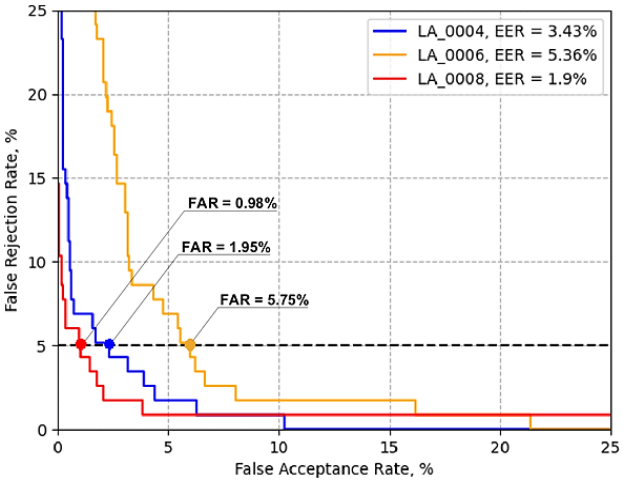


Рис. 7. Кривые компромисса обнаружения и ошибок для исследуемых систем

В качестве целевого значения FRR в данном примере выбрано 5%. Соответствующие значения FAR для выбранных дикторов, представлены на рисунке 7. В случае использования порогового значения, которое соответствует FRR равному 5%, для всех дикторов тестового подмножества набора данных ASVspoof 2019 LA, среднее значение FAR составит 4.29%.

Таким образом, знание вида кривой компромисса обнаружения и ошибок, характерной для некоторого класса субъектозависимых систем ОАБП, использующих обнаружение аномалий, позволяет определять субъектозависимое пороговое значение, регулируя FRR.

6. Заключение. В данной работе представлена реализация субъектозависимого метода ОАБП в системах распознавания диктора на основе обнаружения аномалий и проведена его экспериментальная

оценка применительно к задаче обнаружения синтезированного голоса. При этом в качестве классификатора использовались методы обнаружения аномалий, а в качестве инструмента извлечения голосовых признаков – предобученные искусственные нейронные сети.

Исследована возможность применения нейронных сетей, обученных для задачи распознавания диктора, с целью построения субъектозависимых систем ОАБП. Несмотря на то, что системы, использующие такие сети, уступают в точности передовым системам ОАБП, их производительность заслуживает внимания, поскольку примеры АБП не использовались ни при обучении данных искусственных нейронных сетей, ни при обучении моделей обнаружения аномалий. EER лучшей системы такого рода составил 4.74% при применении специального протокола испытаний с использованием набора данных ASVspoof 2019 LA. Представленные результаты подтверждают, что векторные представления сетей, предобученных для распознавания диктора, содержат ценную для задачи обнаружения синтезированного голоса информацию. Кроме того, они указывают на возможность разработки биометрической системы, которая использует одну искусственную нейронную сеть для распознавания диктора и ОАБП.

Предложенный метод позволил улучшить точность искусственных нейронных сетей, предобученных для задачи обнаружения синтезированного голоса, без их повторного обучения и без внесения каких-либо изменений в их архитектуру и параметры. При проведении экспериментов с двумя базовыми системами [8] на наборе данных ASVspoof 2019 LA [30] удалось улучшить EER на 7.1% и 9.2%, а min t-DCF – на 4.6% относительно исходных результатов. При проведении экспериментов с третьей базовой системой [29] на наборе данных ASVspoof 2021 LA [31] удалось улучшить EER на 3.9% относительно исходного результата с незначительным улучшением min t-DCF.

В то же время, применение искусственных нейронных сетей, предобученных для задачи распознавания звуковых паттернов, проявило себя как неэффективный способ извлечения признаков для задачи обнаружения синтезированного голоса.

Полученные результаты свидетельствуют о перспективности применения субъектозависимого подхода для увеличения точности современных систем ОАБП, однако требуются дальнейшие исследования для того, чтобы обеспечить более значительный прирост точности (в особенности на сложных данных, содержащих различные

искажения, примеры которых представлены в наборе данных ASVspoof 2021 LA).

В связи с этим, в ходе дальнейшей работы планируется исследовать возможные преимущества от применения субъектонезависимых примеров сфабрикованных обучающих данных, влияние количества субъектозависимых обучающих данных на прирост точности ОАБП, целесообразность использования методов аугментации для повышения качества субъектозависимых обучающих данных, а также перспективность обучения искусственных нейронных сетей с учётом субъектозависимого подхода.

Литература

1. Bai Z., Zhang X.-L. Speaker Recognition Based on Deep Learning: An overview // Neural Networks. 2021. vol. 140. pp. 65–99. DOI: 10.1016/j.neunet.2021.03.004.
2. Wang X., Yamagishi J. A Practical Guide to Logical Access Voice Presentation Attack Detection // Frontiers in Fake Media Generation and Detection. Singapore: Springer. 2022. pp. 169–214. DOI: 10.1007/978-981-19-1524-6_8.
3. ГОСТ Р 58624.1-2019. Информационные технологии. Биометрия. Обнаружение атаки на биометрическое предъявление. Часть 1. Структура. М.: Стандартинформ, 2019. 16 с.
4. Chettri B., Sturm B.L. A Deeper Look at Gaussian Mixture Model Based Anti-Spoofing Systems // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2018. pp. 5159–5163. DOI: 10.1109/ICASSP.2018.8461467.
5. Wei L., Long Y., Wei H., Li Y. New Acoustic Features for Synthetic and Replay Spoofing Attack Detection // Symmetry. 2022. vol. 14. no. 2. DOI: 10.3390/sym14020274.
6. Balamurali B.T., Lin K.E., Lui S., Chen J.-M., Herremans D. Toward Robust Audio Spoofing Detection: A Detailed Comparison of Traditional and Learned Features // IEEE Access. 2019. vol. 7. pp. 84229–84241. DOI: 10.1109/ACCESS.2019.2923806.
7. Марковников Н.М., Кипяткова И.С. Аналитический обзор интегральных систем распознавания речи // Труды СПИИРАН. 2018. № 3(58). С. 77–110. DOI: 10.15622/sp.58.4.
8. Hua G., Teoh A.B.J., Zhang H. Towards End-To-End Synthetic Speech Detection // IEEE Signal Processing Letters. 2021. vol. 28. pp. 1265–1269. DOI: 10.1109/LSP.2021.3089437.
9. Wang X., Delgado H., Tak H., Jung J., Shim H., Todisco M., Kukanov I., Liu X., Sahidullah M., Kinnunen T., Evans N., Lee K.A., Yamagishi J. ASVspoof 5: Crowdsourced Speech Data, Deepfakes, and Adversarial Attacks at Scale // arxiv preprint: arXiv:2408.08739v1. 2024.
10. Novoselov S., Kozlov A., Lavrentyeva G., Simonchik K., Shchemelinin V. STC Anti-spoofing Systems for the ASVspoof 2015 Challenge // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2016. pp. 5475–5479. DOI: 10.1109/ICASSP.2016.7472724.
11. Lavrentyeva G., Novoselov S., Malykh E., Kozlov A., Kudashev O., Shchemelinin V. Audio Replay Attack Detection with Deep Learning Frameworks // Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech. 2017. pp. 82–86. DOI: 10.21437/Interspeech.2017-360.
12. Lavrentyeva G., Novoselov S., Tseren A., Volkova M., Gorlanov A., Kozlov A. STC Antispoofing Systems for the ASVspoof2019 Challenge // Proceedings of the Annual

Conference of the International Speech Communication Association, Interspeech. 2019. pp. 1033–1037. DOI: 10.21437/Interspeech.2019-1768.
13. Tomilov A., Svishchev A., Volkova M., Chirkovskiy A., Kondratev A., Lavrentyeva G. STC Antispoofing Systems for the ASVspoof2021 Challenge // Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech. 2021. pp. 61–67. DOI: 10.21437/ASVSPOOF.2021-10.
14. Suthokumar G., Sriskandaraja K., Sethu V., Ambikairajah E., Li H. An Analysis of Speaker Dependent Models in Replay Detection // APSIPA Transactions on Signal and Information Processing. 2020. vol. 9. no. 1. DOI: 10.1017/ATSIP.2020.9.
15. Евсюков М.В., Путято М.М., Макарян А.С. Исследование различимости подлинного и синтезированного голоса дикторов // Вопросы кибербезопасности. 2024. № 2(60). С. 44–52. DOI: 10.21681/2311-3456-2024-2-44-52.
16. Евсюков М.В., Путято М.М., Макарян А.С., Черкасов А.Н. Оценка точности субъектозависимого подхода к обнаружению синтезированного голоса // Вестник Воронежского государственного университета. Серия: Системный анализ и информационные технологии. 2024. № 1. С. 77–93. DOI: 10.17308/sait/1995-5499/2024/1/77-93.
17. Castan D., Rahman M.H., Bakst S., Cobo-Kroenke C., McLaren M., Graciarena M., Lawson A. Speaker-Targeted Synthetic Speech Detection // Proc. of The Speaker and Language Recognition Workshop (Odyssey 2022). 2022. pp. 62–69. DOI: 10.21437/Odyssey.2022-9.
18. Zhang Y., Jiang F., Duan Z. One-Class Learning Towards Synthetic Voice Spoofing Detection // IEEE Signal Processing Letters. 2021. vol. 28. pp. 937–941. DOI: 10.1109/LSP.2021.3076358.
19. Brummer N., Swart A., Mosner L., Silnova A., Plchot O., Stafylakis T., Burget L. Probabilistic Spherical Discriminant Analysis: An Alternative to PLDA for length-normalized embeddings // Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech. 2022. pp. 1446–1450. DOI: 10.21437/Interspeech.2022-731.
20. Liu X., Sahidullah M., Lee K.A., Kinnunen T. Speaker-Aware Anti-spoofing // Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech. 2023. pp. 2498–2502. DOI: 10.21437/Interspeech.2023-1323.
21. Jung J.W. Heo H.S., Tak H., Shim H.J., Chung J.S., Lee B.J., Yu H.J., Evans N. AASIST: Audio Anti-Spoofing Using Integrated Spectro-Temporal Graph Attention Networks // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2022. pp. 6367–6371. DOI: 10.1109/ICASSP43922.2022.9747766.
22. Fatemifar S., Arashloo S.R., Awais M., Kittler J. Client-Specific Anomaly Detection for Face Presentation Attack Detection // Pattern Recognition. 2020. vol. 112. no. 8. DOI: 10.1016/j.patcog.2020.107696.
23. Seliya N., Zadeh A.A., Khoshgoftaar T.M. A Literature Review on One-Class Classification and its Potential Applications in Big Data // Journal of Big Data. 2021. vol. 8. no. 1. DOI: 10.1186/s40537-021-00514-x.
24. Khan S., Madden M. A Survey of Recent Trends in One Class Classification // Artificial Intelligence and Cognitive Science, Lecture Notes in Computer Science. 2009. vol. 6206. pp. 188–197. DOI: 10.1007/978-3-642-17080-5_21.
25. Villalba J., Miguel A., Ortega A., Lleida E. Spoofing Detection with DNN and One-Class SVM for the ASVspoof 2015 Challenge // Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech. 2015. pp. 2067–2071. DOI: 10.21437/interspeech.2015-468.

26. Amorim L.B.V., Cavalcanti G.D.C., Cruz R.M.O. The Choice of Scaling Technique Matters for Classification Performance // Applied Soft Computing. 2023. vol. 133. DOI: 10.1016/j.asoc.2022.109924.
27. Wang C., Xu R., Xu S., Meng W., Zhang X. CNDesc: Cross Normalization for Local Descriptors Learning // IEEE Transactions on Multimedia. 2022. vol. 99. DOI: 10.1109/TMM.2022.3169331.
28. Dorabiala O., Aravkin A.Y., Kutz J.N. Ensemble Principal Component Analysis // IEEE Access. 2024. vol. 12. pp. 6663–6671. DOI: 10.1109/ACCESS.2024.3350984.
29. Tak H., Todisco M., Wang X., Jung J., Yamagishi J., Evans N. Automatic Speaker Verification Spoofing and Deepfake Detection Using Wav2vec 2.0 and Data Augmentation // Proc. of The Speaker and Language Recognition Workshop (Odyssey 2022). 2022. pp. 112–119. DOI: 10.21437/Odyssey.2022-16.
30. Wang X. et al. ASVspoof 2019: A Large-Scale Public Database of Synthesized, Converted and Replayed Speech // Computer Speech & Language. 2020. vol. 64. DOI: 10.1016/j.csl.2020.101114.
31. Yamagishi J., Wang X., Todisco M., Sahidullah M., Patino J., Nautsch A., Liu X., Lee K.A., Kinnunen T., Evans N., Delgado H. ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection // Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech. 2021. pp. 47–54. DOI: 10.21437/asvspoof.2021-8.
32. Ge W., Tak H., Todisco M., Evans N. On the Potential of Jointly-Optimised Solutions to Spoofing Attack Detection and Automatic Speaker Verification // Proceedings of the 6th International Conference, IberSPEECH. 2022. pp. 51–55. DOI: 10.21437/iberspeech.2022-11.
33. Ding S., Chen T., Gong X., Zha W., Wang Z. AutoSpeech: Neural Architecture Search for Speaker Recognition // Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech. 2020. pp. 916–920. DOI: 10.21437/Interspeech.2020-1258.
34. Xie W., Nagrani A., Chung J.S., Zisserman A. Utterance-Level Aggregation for Speaker Recognition in the Wild // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2019. pp. 5791–5795. DOI: 10.1109/ICASSP.2019.8683120.
35. Ravanelli M., Bengio Y. Speaker Recognition from Raw Waveform with SincNet // IEEE Spoken Language Technology Workshop (SLT). 2018. pp. 1021–1028. DOI: 10.1109/SLT.2018.8639585.
36. Jung J.W., Kim Y., Heo H.S., Lee B.-J., Kwon Y., Son Chung J.S. Pushing the Limits of Raw Waveform Speaker Recognition // Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech. 2022. pp. 2228–2232. DOI: 10.21437/Interspeech.2022-126.
37. Nagraniy A., Chungy J.S., Zisserman A. VoxCeleb: A large-scale speaker identification dataset // Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech. 2017. pp. 2616–2620. DOI: 10.21437/Interspeech.2017-950.
38. Chung J.S., Nagrani A., Zisserman A. VoxCeleb2: Deep Speaker Recognition // Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech. 2018. pp. 1086–1090. DOI: 10.21437/Interspeech.2018-1929.
39. Panayotov V., Chen G., Povey D., Khudanpur S. LibriSpeech: An ASR Corpus Based on Public Domain Audio Books // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2015. pp. 5206–5210. DOI: 10.1109/ICASSP.2015.7178964.

40. Kong Q., Cao Y., Iqbal T., Wang Y., Wang W., Plumbley M.D. PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition // IEEE/ACM Transactions on Audio Speech and Language Processing. 2020. vol. 28. pp. 2880–2894. DOI: 10.1109/TASLP.2020.3030497.
41. Gemmeke G.F., Ellis D.P.W., Freedman D., Jansen A., Lawrence W., Moore R.C. Audio Set: An Ontology and Human-Labeled Dataset for Audio Events // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2017. pp. 776–780. DOI: 10.1109/ICASSP.2017.7952261.
42. Hosna A., Merry E., Gyalmo J., Alom Z., Aung Z., Azim M.A. Transfer Learning: A Friendly Introduction // Journal of Big Data. 2022. vol. 9. no. 1. DOI: 10.1186/s40537-022-00652-w.
43. Januzaj Y., Luma A. Cosine Similarity – A Computing Approach to Match Similarity Between Higher Education Programs and Job Market Demands Based on Maximum Number of Common Words // International Journal of Emerging Technologies in Learning. 2022. vol. 17. no. 12. pp. 258–268. DOI: 10.3991/ijet.v17i12.30375.
44. Ghorbani H. Mahalanobis Distance and its Application for Detecting Multivariate Outliers // Facta Universitatis, Series: Mathematics and Informatics. 2019. vol. 34. no. 3. pp. 583–595. DOI: 10.22190/fumi1903583g.
45. Alegre F., Amehraye A., Evans N. A One-Class Classification Approach to Generalised Speaker Verification Spoofing Countermeasures Using Local Binary Patterns // IEEE 6th International Conference on Biometrics: Theory, Applications and Systems (BTAS). 2013. pp. 1–8. DOI: 10.1109/BTAS.2013.6712706.
46. Scrucca L. Entropy-Based Anomaly Detection for Gaussian Mixture Modeling // Algorithms. 2023. vol. 16. no. 4. DOI: 10.3390/a16040195.
47. Reynolds D.A., Quatieri T.F., Dunn R.B. Speaker Verification Using Adapted Gaussian Mixture Models // Digital Signal Processing: A Review Journal. 2000. vol. 10. no. 1-3. pp. 19–41. DOI: 10.1006/dspr.1999.0361.
48. Liu F.T., Ting K.M., Zhou Z.H. Isolation forest // Proceedings of the Eighth IEEE International Conference on Data Mining (ICDM). 2008. pp. 413–422. DOI: 10.1109/ICDM.2008.17.
49. Hao B., Hei X. Voice Liveness Detection for Medical Devices // Design and Implementation of Healthcare Biometric Systems. 2019. pp. 109–136. DOI: 10.4018/978-1-5225-7525-2.ch005.
50. Kinnunen T., Lee K.A., Delgado H., Evans N., Todisco M., Sahidullah M., Yamagishi J., Reyonolds D.A. t-DCF: a Detection Cost Function for the Tandem Assessment of Spoofing Countermeasures and Automatic Speaker Verification // Proc. The Speaker and Language Recognition Workshop (Odyssey 2018), 2018. pp. 312–319.
51. Hazra A. Using the Confidence Interval Confidently // Journal of Thoracic Disease. 2017. vol. 9. no. 10. DOI: 10.21037/jtd.2017.09.14.
52. Martin A., Doggington G., Kamm T., Ordowski M. Przybocki M. The DET curve in assessment of detection task performance // Proceedings of the 5th European Conference on Speech Communication and Technology, Eurospeech (ISCA). 1997. pp. 1895–1898. DOI:10.21437/Eurospeech.1997-504.

Евсюков Михаил Витальевич — аспирант, кафедра кибербезопасности и защиты информации, Кубанский государственный технологический университет. Область научных интересов: обнаружение атак на биометрическое предъявление, голосовая биометрия, машинное обучение, постквантовая криптография. Число научных публикаций — 23. michael.evsyukov@gmail.com; улица Московская, 2, 350072, Краснодар, Россия; р.т.: +7(861)255-0346.

DOI 10.15622/ia.24.1.8

M. EVSYUKOV

SPEAKER-SPECIFIC METHOD OF SPOOFING ATTACK DETECTION BASED ON ANOMALY DETECTION

Evsyukov M. **Speaker-Specific Method of Spoofing Attack Detection Based on Anomaly Detection.**

Abstract. Most research in the field of voice presentation attack detection relies on the speaker-independent approach. Nevertheless, several scientific works indicate that using the speaker-specific approach, which involves utilizing prior knowledge about the identity of the claimed speaker to enhance the accuracy of spoofing detection, is likely to be beneficial. Therefore, the goal of this work is to propose a speaker-specific method of spoofing attack detection based on anomaly detection and to evaluate its applicability to the detection of synthesized speech and converted voice. Artificial neural networks pre-trained for the tasks of spoofing detection, speaker recognition, and audio pattern recognition are used for feature extraction. A set of anomaly detection models are used as backend classifiers. Each of them is trained on bonafide data of a target speaker. The experimental evaluation of the proposed method on the ASVspoof 2019 LA dataset shows that the best speaker-specific spoofing detection system, which uses an anomaly detection model and a neural network pre-trained for the task of speaker recognition, achieves an EER of 4.74%. This result suggests that embeddings extracted by networks pre-trained for speaker recognition contain information that can be utilized for spoofing detection. In addition, the proposed method allowed to increase the accuracy of three baseline systems pre-trained for the task of spoofing detection. Experiments with two baseline systems on the ASVspoof 2019 LA dataset showed relative improvement in terms of EER by 7.1% and 9.2%, and in terms of min t-DCF by 4.6%. Experiments with the third baseline system on the ASVspoof 2021 LA dataset showed relative improvement in terms of EER by 3.9% without significant improvement of min t-DCF.

Keywords: speaker-specific approach, spoofing detection, presentation attack detection, biometric systems, voice biometrics, transfer learning, anomaly detection.

References

1. Bai Z., Zhang X.-L. Speaker Recognition Based on Deep Learning: An overview. Neural Networks. 2021. vol. 140. pp. 65–99. DOI: 10.1016/j.neunet.2021.03.004.
2. Wang X., Yamagishi J. A Practical Guide to Logical Access Voice Presentation Attack Detection. Frontiers in Fake Media Generation and Detection. Singapore: Springer. 2022. pp. 169–214. DOI: 10.1007/978-981-19-1524-6_8.
3. GOST R 58624.1-2019. Informacionnye tehnologii. Biometrija. Obnaruzhenie ataki na biometricheskoe predjavlenie. Chast' 1. Struktura [Information technology. Biometrics. Biometric presentation attack detection. Part 1. Framework]. M.: Gosstandart Rossii, 2019. (In Russ.).
4. Chettri B., Sturm B.L. A Deeper Look at Gaussian Mixture Model Based Anti-Spoofing Systems. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2018. pp. 5159–5163. DOI: 10.1109/ICASSP.2018.8461467.
5. Wei L., Long Y., Wei H., Li Y. New Acoustic Features for Synthetic and Replay Spoofing Attack Detection. Symmetry. 2022. vol. 14. no. 2. DOI: 10.3390/sym14020274.
6. Balamurali B.T., Lin K.E., Lui S., Chen J.-M., Herremans D. Toward Robust Audio Spoofing Detection: A Detailed Comparison of Traditional and Learned Features. IEEE Access. 2019. vol. 7. pp. 84229–84241. DOI: 10.1109/ACCESS.2019.2923806.

7. Markovnikov N., Kipyatkova I. An Analytic Survey of End-to-End Speech Recognition Systems. Trudy SPIIRAN – SPIIRAS Proceedings. 2018. vol. 3. no. 58. pp. 77-110. DOI: 10.15622/sp.58.4. (In Russ.).
8. Hua G., Teoh A.B.J., Zhang H. Towards End-To-End Synthetic Speech Detection. IEEE Signal Processing Letters. 2021. vol. 28. pp. 1265–1269. DOI: 10.1109/LSP.2021.3089437.
9. Wang X., Delgado H., Tak H., Jung J., Shim H., Todisco M., Kukanov I., Liu X., Sahidullah M., Kinnunen T., Evans N., Lee K.A., Yamagishi J. ASVspoof 5: Crowdsourced Speech Data, Deepfakes, and Adversarial Attacks at Scale. arxiv preprint: arXiv:2408.08739v1. 2024.
10. Novoselov S., Kozlov A., Lavrentyeva G., Simonchik K., Shchemelinin V. STC Anti-spoofing Systems for the ASVspoof 2015 Challenge. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2016. pp. 5475–5479. DOI: 10.1109/ICASSP.2016.7472724.
11. Lavrentyeva G., Novoselov S., Malykh E., Kozlov A., Kudashev O., Shchemelinin V. Audio Replay Attack Detection with Deep Learning Frameworks. Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech. 2017. pp. 82–86. DOI: 10.21437/Interspeech.2017-360.
12. Lavrentyeva G., Novoselov S., Tseren A., Volkova M., Gorlanov A., Kozlov A. STC Antispoofing Systems for the ASVspoof2019 Challenge. Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech. 2019. pp. 1033–1037. DOI: 10.21437/Interspeech.2019-1768.
13. Tomilov A., Svishchev A., Volkova M., Chirkovskiy A., Kondratev A., Lavrentyeva G. STC Antispoofing Systems for the ASVspoof2021 Challenge. Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech. 2021. pp. 61–67. DOI: 10.21437/ASVSPOOF.2021-10.
14. Suthokumar G., Sriskandaraja K., Sethu V., Ambikairajah E., Li H. An Analysis of Speaker Dependent Models in Replay Detection. APSIPA Transactions on Signal and Information Processing. 2020. vol. 9. no. 1. DOI: 10.1017/ATSIP.2020.9.
15. Evsyukov M.V., Putyato M.M., Makaryan A.S. [The Effect of Speaker Variability on Distinguishability of Bonafide and Synthetized Speech]. Voprosy kiberbezopasnosti – Cybersecurity issues. 2024. vol. 60. no. 2. pp. 44–52. DOI: 10.21681/2311-3456-2024-2-44-52. (In Russ.).
16. Evsjukov M.V., Putjato M.M., Makarjan A.S., Cherkasov A.N. [Assessing Accuracy of Speaker-Specific Approach to Logical Access Spoofing Detection]. Vestnik Voronezhskogo gosudarstvennogo universiteta. Serija: Sistemnyj analiz i informacionnye tehnologii – Proceedings of Voronezh State University. Series: Systems Analysis and Information Technologies. 2024. no. 1. pp. 77–93. DOI: 10.17308/sait/1995-5499/2024/l/77-93. (In Russ.).
17. Castan D., Rahman M.H., Bakst S., Cobo-Kroenke C., McLaren M., Graciarena M., Lawson A. Speaker-Targeted Synthetic Speech Detection. Proc. of The Speaker and Language Recognition Workshop (Odyssey 2022). 2022. pp. 62–69. DOI: 10.21437/Odyssey.2022-9.
18. Zhang Y., Jiang F., Duan Z. One-Class Learning Towards Synthetic Voice Spoofing Detection. IEEE Signal Processing Letters. 2021. vol. 28. pp. 937–941. DOI: 10.1109/LSP.2021.3076358.
19. Brummer N., Swart A., Mosner L., Silnova A., Plchot O., Stafylakis T., Burget L. Probabilistic Spherical Discriminant Analysis: An Alternative to PLDA for length-normalized embeddings. Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech. 2022. pp. 1446–1450. DOI: 10.21437/Interspeech.2022-731.

20. Liu X., Sahidullah M., Lee K.A., Kinnunen T. Speaker-Aware Anti-spoofing. Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech. 2023. pp. 2498–2502. DOI: 10.21437/Interspeech.2023-1323.
21. Jung J.W. Heo H.S., Tak H., Shim H.J., Chung J.S., Lee B.J., Yu H.J., Evans N. AASIST: Audio Anti-Spoofing Using Integrated Spectro-Temporal Graph Attention Networks. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2022. pp. 6367–6371. DOI: 10.1109/ICASSP43922.2022.9747766.
22. Fatemifar S., Arashloo S.R., Awais M., Kittler J. Client-Specific Anomaly Detection for Face Presentation Attack Detection. Pattern Recognition. 2020. vol. 112. no. 8. DOI: 10.1016/j.patcog.2020.107696.
23. Seliya N., Zadeh A.A., Khoshgoftaar T.M. A Literature Review on One-Class Classification and its Potential Applications in Big Data. Journal of Big Data. 2021. vol. 8. no. 1. DOI: 10.1186/s40537-021-00514-x.
24. Khan S., Madden M. A Survey of Recent Trends in One Class Classification. Artificial Intelligence and Cognitive Science, Lecture Notes in Computer Science. 2009. vol. 6206. pp. 188–197. DOI: 10.1007/978-3-642-17080-5_21.
25. Villalba J., Miguel A., Ortega A., Lleida E. Spoofing Detection with DNN and One-Class SVM for the ASVspoof 2015 Challenge. Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech. 2015. pp. 2067–2071. DOI: 10.21437/interspeech.2015-468.
26. Amorim L.B.V., Cavalcanti G.D.C., Cruz R.M.O. The Choice of Scaling Technique Matters for Classification Performance. Applied Soft Computing. 2023. vol. 133. DOI: 10.1016/j.asoc.2022.109924.
27. Wang C., Xu R., Xu S., Meng W., Zhang X. CNDesc: Cross Normalization for Local Descriptors Learning. IEEE Transactions on Multimedia. 2022. vol. 99. DOI: 10.1109/TMM.2022.3169331.
28. Dorabiala O., Aravkin A.Y., Kutz J.N. Ensemble Principal Component Analysis. IEEE Access. 2024. vol. 12. pp. 6663–6671. DOI: 10.1109/ACCESS.2024.3350984.
29. Tak H., Todisco M., Wang X., Jung J., Yamagishi J., Evans N. Automatic Speaker Verification Spoofing and Deepfake Detection Using Wav2vec 2.0 and Data Augmentation. Proc. of The Speaker and Language Recognition Workshop (Odyssey 2022). 2022. pp. 112–119. DOI: 10.21437/Odyssey.2022-16.
30. Wang X. et al. ASVspoof 2019: A Large-Scale Public Database of Synthesized, Converted and Replayed Speech. Computer Speech & Language. 2020. vol. 64. DOI: 10.1016/j.csl.2020.101114.
31. Yamagishi J., Wang X., Todisco M., Sahidullah M., Patino J., Nautsch A., Liu X., Lee K.A., Kinnunen T., Evans N., Delgado H. ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection. Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech. 2021. pp. 47–54. DOI: 10.21437/asvspoof.2021-8.
32. Ge W., Tak H., Todisco M., Evans N. On the Potential of Jointly-Optimised Solutions to Spoofing Attack Detection and Automatic Speaker Verification. Proceedings of the 6th International Conference, IberSPEECH. 2022. pp. 51–55. DOI: 10.21437/iberspeech.2022-11.
33. Ding S., Chen T., Gong X., Zha W., Wang Z. AutoSpeech: Neural Architecture Search for Speaker Recognition. Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech. 2020. pp. 916–920. DOI: 10.21437/Interspeech.2020-1258.
34. Xie W., Nagrani A., Chung J.S., Zisserman A. Utterance-Level Aggregation for Speaker Recognition in the Wild. IEEE International Conference on Acoustics,

Speech and Signal Processing (ICASSP). 2019. pp. 5791–5795. DOI: 10.1109/ICASSP.2019.8683120.
35. Ravanelli M., Bengio Y. Speaker Recognition from Raw Waveform with SincNet. IEEE Spoken Language Technology Workshop (SLT). 2018. pp. 1021–1028. DOI: 10.1109/SLT.2018.8639585.
36. Jung J.W., Kim Y., Heo H.S., Lee B.-J., Kwon Y., Son Chung J.S. Pushing the Limits of Raw Waveform Speaker Recognition. Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech. 2022. pp. 2228–2232. DOI: 10.21437/Interspeech.2022-126.
37. Nagraniy A., Chungy J.S., Zisserman A. VoxCeleb: A large-scale speaker identification dataset. Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech. 2017. pp. 2616–2620. DOI: 10.21437/Interspeech.2017-950.
38. Chung J.S., Nagrani A., Zisserman A. VoxCeleb2: Deep Speaker Recognition. Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech. 2018. pp. 1086–1090. DOI: 10.21437/Interspeech.2018-1929.
39. Panayotov V., Chen G., Povey D., Khudanpur S. LibriSpeech: An ASR Corpus Based on Public Domain Audio Books. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2015. pp. 5206–5210. DOI: 10.1109/ICASSP.2015.7178964.
40. Kong Q., Cao Y., Iqbal T., Wang Y., Wang W., Plumbley M.D. PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition. IEEE/ACM Transactions on Audio Speech and Language Processing. 2020. vol. 28. pp. 2880–2894. DOI: 10.1109/TASLP.2020.3030497.
41. Gemmeke G.F., Ellis D.P.W., Freedman D., Jansen A., Lawrence W., Moore R.C. Audio Set: An Ontology and Human-Labeled Dataset for Audio Events. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2017. pp. 776–780. DOI: 10.1109/ICASSP.2017.7952261.
42. Hosna A., Merry E., Gyalmo J., Alom Z., Aung Z., Azim M.A. Transfer Learning: A Friendly Introduction. Journal of Big Data. 2022. vol. 9. no. 1. DOI: 10.1186/s40537-022-00652-w.
43. Januzaj Y., Luma A. Cosine Similarity – A Computing Approach to Match Similarity Between Higher Education Programs and Job Market Demands Based on Maximum Number of Common Words. International Journal of Emerging Technologies in Learning. 2022. vol. 17. no. 12. pp. 258–268. DOI: 10.3991/ijet.v17i12.30375.
44. Ghorbani H. Mahalanobis Distance and its Application for Detecting Multivariate Outliers. Facta Universitatis, Series: Mathematics and Informatics. 2019. vol. 34. no. 3. pp. 583–595. DOI: 10.22190/fumi1903583g.
45. Alegre F., Amehraye A., Evans N. A One-Class Classification Approach to Generalised Speaker Verification Spoofing Countermeasures Using Local Binary Patterns. IEEE 6th International Conference on Biometrics: Theory, Applications and Systems (BTAS). 2013. pp. 1–8. DOI: 10.1109/BTAS.2013.6712706.
46. Scrucca L. Entropy-Based Anomaly Detection for Gaussian Mixture Modeling. Algorithms. 2023. vol. 16. no. 4. DOI: 10.3390/a16040195.
47. Reynolds D.A., Quatieri T.F., Dunn R.B. Speaker Verification Using Adapted Gaussian Mixture Models. Digital Signal Processing: A Review Journal. 2000. vol. 10. no. 1-3. pp. 19–41. DOI: 10.1006/dspr.1999.0361.
48. Liu F.T., Ting K.M., Zhou Z.H. Isolation forest. Proceedings of the Eighth IEEE International Conference on Data Mining (ICDM). 2008. pp. 413–422. DOI: 10.1109/ICDM.2008.17.

49. Hao B., Hei X. Voice Liveness Detection for Medical Devices. Design and Implementation of Healthcare Biometric Systems. 2019. pp. 109–136. DOI: 10.4018/978-1-5225-7525-2.ch005.
50. Kinnunen T., Lee K.A., Delgado H., Evans N., Todisco M., Sahidullah M., Yamagishi J., Reyonolds D.A. t-DCF: a Detection Cost Function for the Tandem Assessment of Spoofing Countermeasures and Automatic Speaker Verification. Proc. The Speaker and Language Recognition Workshop (Odyssey 2018), 2018. pp. 312–319.
51. Hazra A. Using the Confidence Interval Confidently. Journal of Thoracic Disease. 2017. vol. 9. no. 10. DOI: 10.21037/jtd.2017.09.14.
52. Martin A., Doggington G., Kamm T., Ordowski M. Przybocki M. The DET curve in assessment of detection task performance. Proceedings of the 5th European Conference on Speech Communication and Technology, Eurospeech (ISCA). 1997. pp. 1895–1898. DOI:10.21437/Eurospeech.1997-504.

Evsyukov Mikhail — Postgraduate student, Department of cybersecurity and information protection, Kuban State Technological University. Research interests: presentation attack detection, voice biometrics, machine learning, postquantum cryptography. The number of publications — 23. michael.evsyukov@gmail.com; 2, Moskovskaya St., 350072, Krasnodar, Russia; office phone: +7(861)255-0346.

УДК 004.852 **DOI** 10.15622/ia.24.1.9

А.В. Пономарев, А.А. Агафонов

АНАЛИТИЧЕСКИЙ ОБЗОР МЕТОДОВ РАСПРЕДЕЛЕНИЯ ЗАДАЧ ПРИ СОВМЕСТНОЙ РАБОТЕ ЧЕЛОВЕКА И МОДЕЛИ ИИ

Пономарев А.В., Агафонов А.А. **Аналитический обзор методов распределения задач при совместной работе человека и модели ИИ.**

Аннотация. Во многих практических сценариях принятие решений исключительно моделью ИИ оказывается нежелательным или даже невозможным, и использование модели ИИ является лишь частью сложного процесса принятия решений, включающего и эксперта-человека. Тем не менее при создании и обучении моделей ИИ этот факт зачастую упускается – модель обучается для самостоятельного принятия решений, а это не всегда является оптимальным. В статье представлен обзор методов, позволяющих учесть совместную работу ИИ и эксперта-человека в процессе конструирования (в частности, обучения) систем ИИ, что более точно соответствует практическому применению модели, позволяет повысить точность решений, принимаемых системой «человек – модель ИИ», а также явно управлять другими важными параметрами системы (например, нагрузкой на человека). Обзор включает анализ современной литературы по заданной тематике по следующим основным направлениям: 1) сценарии взаимодействия человека и модели ИИ и формальные постановки задачи для повышения эффективности системы «человек – модель ИИ»; 2) методы для обеспечения эффективного функционирования системы «человек – модель ИИ»; 3) способы оценки качества совместной работы человека и модели ИИ. Сделаны выводы относительно достоинств, недостатков и условий применимости методов, выявлены основные проблемы существующих подходов. Обзор может быть полезен широкому кругу исследователей и специалистов, занимающихся применением ИИ для поддержки принятия решений.

Ключевые слова: искусственный интеллект, ответственный ИИ, поддержка принятия решений, человеко-машинное взаимодействие, эксперт-человек, распределение задач, совместная работа человека и ИИ, неопределенность модели, нейронные сети, классификатор, обучение с отказом, обучение с делегированием.

1. Введение. Современные решения, основанные на применении искусственного интеллекта (ИИ) в целом и глубоких нейронных сетей в частности, во многих задачах позволяют получать результаты близкие к тем, что могут быть получены человеком, при этом скорость работы и масштабируемость решений, основанных на ИИ, оказывается существенно выше, что обусловливает все более широкое их распространение. Тем не менее полная автоматизация возможна далеко не для всех задач. Среди основных сдерживающих факторов можно выделить следующие. Во-первых, система ИИ действует только на основе той информации, которая преобразована в цифровую форму и доступна системе (а также была использована при конструировании и/или обучении системы). Соответственно, в сложных предметных областях ИИ может столкнуться с неполнотой информации и, как

следствие, с деградацией качества решений, в то время как эксперт может предпринять шаги для выяснения дополнительных фактов. Во-вторых, вопросы ответственности систем ИИ еще не до конца проработаны, а цена ошибки в ряде случаев слишком высока. Все это приводит к тому, что, несмотря на развитие ИИ, в очень многих практических сценариях ИИ работает (и в обозримой перспективе будет работать) совместно с экспертом-человеком, однако при создании и обучении систем ИИ этот факт зачастую упускается. В статье представлен обзор методов учета такой перспективы совместной работы ИИ и человека в процессе конструирования (в частности, обучения) систем ИИ, что позволяет повысить качество решений, принимаемых системой «человек – модель ИИ» [1, 2].

В статье представлены основные результаты аналитического обзора методов в области распределения задач при совместной работе человека и модели ИИ. Проблема совместной работы человека и модели ИИ (или машинного обучения), с одной стороны, довольно интенсивно исследуется в последнее время (причем предлагаются принципиально различные подходы, отличающиеся как особенностями самого сотрудничества, так и решениями по его организации), с другой – имеет довольно богатую историю, которую можно начинать с т.н. обучения с отказом (англ. learning with rejection, rejection learning, learning to reject, learning with abstention, selective prediction) [3, 4]. Подобная задача рассматривается и в российских публикациях, так, например, в [5] предлагается метод отказа от предсказания для задачи непараметрической регрессии. Кроме того, авторы [5] используют словосочетание «делегировать эксперту» для обозначения ситуации, в которой решение задачи передается человеку, если неуверенность модели оказывается высокой. Поэтому в ходе данного обзора термин «делегирование» будет использоваться для обозначения подобных сценариев.

Ввиду большого разнообразия подходов к совместной работе, представляется не вполне целесообразным вмещать их все в одну статью, поэтому данная статья ограничивается рассмотрением проблемы совместной работы и набора решений по ее организации, удовлетворяющих следующим условиям:

– Задан четко определенный класс задач, которые могут решаться независимо как человеком (экспертом), так и моделью ИИ. Примерами такого класса задач может быть диагностика определенного заболевания по медицинским снимкам, принятие решения о выдаче кредита на основе кредитной истории потенциального заемщика и т.п. Таким образом, с задачей можно

связать набор признаков, конкретные значения которых соответствуют экземпляру задачи (обрабатываемому образцу).

– И человек, и модель ИИ могут совершать ошибки. Более того, эффективность (точность) решения задачи из рассматриваемого класса может варьироваться в зависимости от экземпляра задачи (как при ее решении моделью ИИ, так и человеком). Данному условию удовлетворяет большое количество задач, возникающих на практике – действительно, для большинства моделей ИИ есть «сложные» и «простые» образцы (или даже области пространства признаков).

Перечисленным условиям не удовлетворяют, например, работы по определению состава смешанных команд в рамках социо-киберфизических систем [6 – 8], потому что в них речь не идет об обработке однородных задач, принадлежащих одному классу. Этим условиям также не удовлетворяет и постановка, типичная для обучения с отказом, потому что в ней не рассматривается возможность ошибки человека [9]. Тем не менее, перечисленным условиям удовлетворяет множество важных с практической точки зрения задач, что обусловливает актуальность обзора, результаты которого представлены в статье.

Цель статьи состоит в том, чтобы сформировать систематическое изложение ключевых вопросов и современных методов распределения задач между человеком и моделью ИИ (в рамках их совместной работы), что было бы полезно как практикам в области построения систем с элементами ИИ, так и исследователям, позволяя им сориентироваться в палитре существующих методов и определить возможные направления развития. На данный момент подобных обзоров не было обнаружено. Так, близкий по тематике обзор [9] посвящен исключительно обучению с отказом, где не рассматривается возможность ошибки эксперта, в [10] рассматривается ряд методов обучения с делегированием, но статья не претендует на полноту освещения, в [11] рассматривается широкий набор потенциальных сценариев симбиоза человека и ИИ, но достаточно поверхностно. При выполнении данного обзора авторы опирались на методологию систематического обзора литературы [12]. Особенность реализации этой методологии в данном случае связана с перегруженностью ключевых слов, по которым можно идентифицировать искомые публикации, поэтому формирование выборки статей осуществлялось на основе графа цитирования знаковых публикаций, а не отбором по ключевым словам. В ходе исследования осуществлялся поиск ответов на следующие вопросы:

1) Какие рассматриваются сценарии взаимодействия человека и модели ИИ, и, соответственно, какие предлагаются формальные постановки задач для повышения эффективности системы «человек – модель ИИ»?

2) Какие предлагаются методы для обеспечения эффективного функционирования системы «человек – модель ИИ»?

3) Как производится оценка качества совместной работы человека и модели ИИ? В частности, какие применяются специфические метрики для оценки эффективности подобных систем.

Статья структурирована в соответствии с рассматриваемыми вопросами следующим образом. В разделе 2 описана методика проведения обзора, разделы 3-5 представляют результаты ответов на основные вопросы исследования, описывая выявленные постановки задачи совместной работы, конкретные методы обеспечения эффективности и подходы к оценке качества. В заключении подводятся итоги обзора и выявляются наиболее перспективные направления будущих исследований.

2. Методика проведения обзора. Важными характеристиками, определяющими качество обзора литературы, являются, с одной стороны, представительность, то есть соответствующее направление исследований должно быть достаточно полным образом представлено в статьях, включаемых в обзор, с другой – воспроизводимость («идеалом» которой является получение аналогичных результатов любым другим исследователем, осуществляющим обзор на схожую тему). Первая характеристика, как правило, достигается использованием поиска по ключевым словам в достаточно представительных реферативных базах данных (Scopus, Web of Science, РИНЦ), вторая – следованием той или иной методологии проведения обзора и формированием протокола, позволяющего проследить и воспроизвести шаги исследования (например [12, 13]).

При проведении данного обзора авторы придерживались методологии систематического обзора литературы [12] с некоторыми несущественными корректировками, вызванными особенностью области проведения анализа. Эти корректировки связаны, в первую очередь, с определением множества статей, подвергающихся анализу. Традиционная реализация методологии предполагает, что формируется определенный набор ключевых слов, которые используются для отбора статей в одной или нескольких библиографических базах данных. Однако, как уже отмечалось, данная область – совместная работа человека и ИИ – является очень разнообразной, и в ее рамках сосуществует множество принципиально

различных форм и моделей совместной работы, выделить интересующую интерпретацию только на основе ключевых слов оказывается проблематичным, поскольку в данной области отсутствует устоявшаяся терминология. Поэтому для формирования множества исследуемых публикаций был использован граф цитирований. Был выделен набор публикаций (т.н. «ядро»), в которых впервые был предложен и исследован рассматриваемый вариант совместной работы. Затем сформировано множество публикаций, которые ссылаются хотя бы на одну из статей «ядра». В качестве базы цитирований был выбран Google Scholar из-за своего широкого охвата и относительной оперативности индексирования.

Основные шаги обзора и характеристики промежуточных результатов показаны на рисунке 1. К «ядру» были отнесены 4 статьи, в которых рассматриваемый сценарий совместной работы человека и модели ИИ либо был рассмотрен впервые, либо были сделаны важные практические или теоретические замечания относительно построения подобных систем [1, 2, 14, 15]. Три статьи из данного перечня были опубликованы в материалах высокорейтинговых конференций (A* по данным CORE), одна – препринт ArXiv. На момент формирования перечня все они имели более 100 цитирований в Google Scholar (с момента публикации первой из них прошло 5 лет).

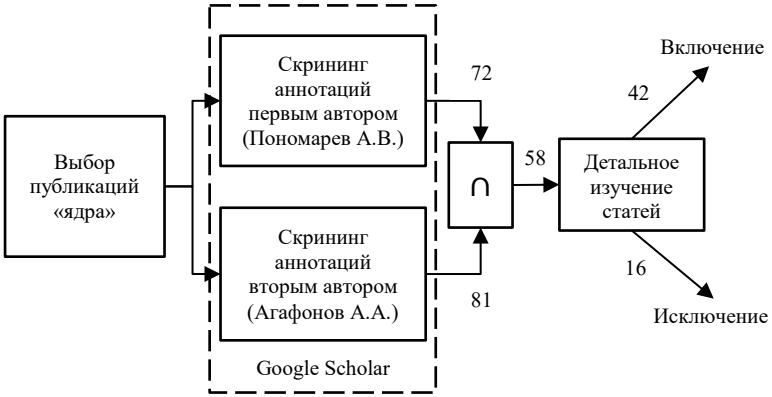


Рис. 1. Порядок проведения обзора

Авторы данного обзора провели независимый скрининг аннотаций всех статей, опубликованных по 2023 г. включительно и цитирующих хотя бы одну из статей «ядра», по данным Google Scholar (всего около 500). Задачей скрининга был отбор статей для

дальнейшего, более детального, изучения. При проведении скрининга отбирались статьи, удовлетворяющие хотя бы одному из следующих критериев: 1) предлагается оригинальный метод; 2) производится сопоставление методов (экспериментальное или теоретическое); 3) предлагается методология сопоставления методов; 4) обзорная статья. В результате каждым из авторов обзора был получен список статей, потенциально подходящих для дальнейшего изучения.

Было сформировано пересечение данных списков, в которое вошли статьи, признанные относящимися к исследуемому сценарию обоими авторами обзора, всего таких статей оказалось 58. В ходе детального изучения еще 16 из них были исключены (часть из них при детальном изучении не удовлетворяла критериям отбора, часть оказалась версиями одной статьи, но под разными названиями). Таким образом, в статье представлены результаты, основанные на структуризации 42 статей на заданную тематику [1, 2, 14 – 53].

Среди отобранных статей большая часть (30 статей) – публикации на конференциях достаточно высокого уровня (CORE A* и A) – AAAI Conference on Artificial Intelligence, NeurIPS, IJCAI, ICML и другие. Другая многочисленная группа – препринты статей, опубликованные на ArXiv. В списке отобранных статей оказалось всего две статьи, опубликованные в журналах: Proceedings of the National Academy of Sciences и Frontiers in Digital Health.

На рисунке 2 приведено распределение отобранных публикаций по годам. Видно, что интерес к данной проблеме постепенно возрастает. Об этом же свидетельствует и статистика источников публикаций – на данный момент среди источников превалируют передовые издания, широкого распространения описываемые методы еще не получили.

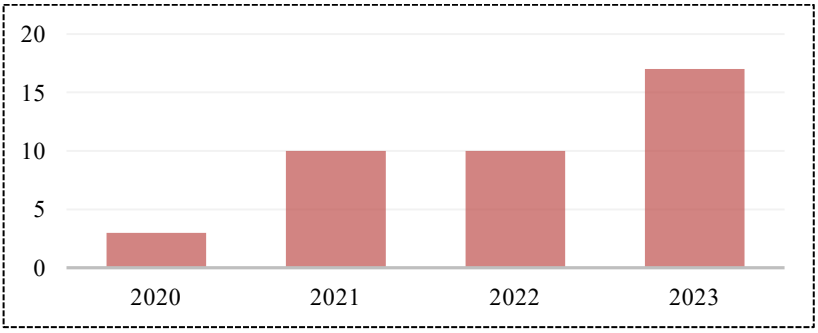


Рис. 2. Распределение количества публикаций по годам

3. Сценарии взаимодействия и формальные постановки. В данном разделе характеризуются основные разновидности и постановки задач, которые различаются как особенностями взаимодействия между человеком и моделью ИИ, так и преследуемой целью (находящей отражение в целевой функции либо в функции потерь). Ведущую роль в структуризации формальных постановок играет сценарий взаимодействия между человеком и моделью ИИ. В рамках каждого из сценариев, в свою очередь, выделяются различные постановки.

Основные критерии, по которым целесообразно структурировать существующие методы распределения задач при совместной работе человека и модели ИИ, представлены на рисунке 3 жирным шрифтом. Каждый конкретный метод может быть позиционирован посредством выбора одной или более категорий по каждому из критериев. При этом следует заметить, что часть критериев относятся к постановке задачи (и рассматриваются в данном разделе статьи), а другая часть («Метод обеспечения совместной работы» и «Тип структуры распределения задач») относятся к пространству решений и рассматриваются в разделе 4.

3.1. Сценарии взаимодействия. Сценарий взаимодействия между человеком и моделью ИИ определяет характер принимаемых решений, последовательность активизации участников системы и доступную им информацию. Можно выделить три вида сценариев: делегирование, последовательная обработка и параллельная обработка.

Под делегированием понимается такой способ организации взаимодействия человека и модели ИИ, когда каждый рассматриваемый образец назначается для обработки либо человеку, либо модели ИИ [18, 19, 27, 29, 31 – 33, 35, 37 – 41, 43, 45, 62]. Этот сценарий является, пожалуй, наиболее часто рассматриваемым в литературе – именно такой сценарий реализуется в рамках обучения с отказом (раздела машинного обучения, в котором хоть и не рассматривается профиль ошибок человека, но уже ставится задача построения классификатора, который бы «воздерживался» от предсказания при недостаточной уверенности), а также в рамках т.н. обучения с делегированием (learning to defer). Идея, обусловливающая востребованность подобного сценария, заключается в том, что при наличии большого количества экземпляров, обработка всех их человеком может быть чересчур дорогостоящей (или требовать слишком большого времени), а обработка моделью – слишком неточной.

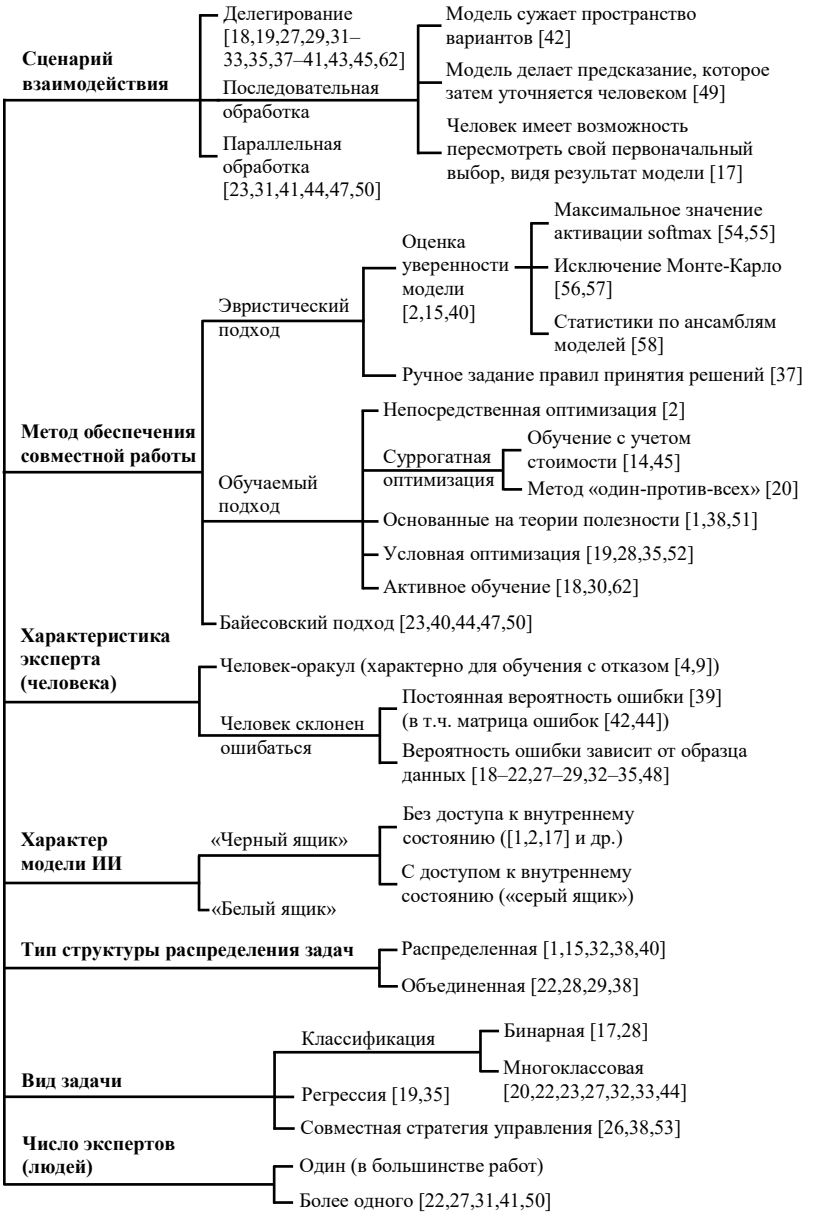


Рис. 3. Методы распределения задач при совместной работе человека и ИИ

Таким образом, в моделях, ориентированных на делегирование, как правило, решается задача поиска определенного компромисса между стоимостью обращения к эксперту (снижение которой достигается назначением образцов модели) и точностью решения задачи (повышение которой, как правило, достигается назначением образцов человеку). Следует, однако, подчеркнуть, что человек в подобных моделях далеко не всегда воспринимается как «оракул», способный дать абсолютно точный ответ (что характерно для более ранних работ в области обучения с отказом), вместо этого при делегировании зачастую учитывается вероятность получения верного ответа от модели и от человека в той или иной области пространства признаков.

Особенностью последовательной обработки является то, что образец поочередно обрабатывается и моделью, и человеком, причем между этими двумя действиями происходит и передача информации. В литературе описано несколько различных сценариев, относящихся к последовательной обработке, однако в большинстве из них итоговое решение остается за человеком, выходные данные модели ИИ используются им для повышения качества принятия решения. Можно выделить следующие разновидности последовательной обработки:

– Модель сужает пространство вариантов, человек выбирает из оставшихся [42]. Данная постановка тесно связана с т.н. conformal prediction. Основной мотив здесь – снизить сложность классификации для человека.

– Модель делает предсказание (возможно, сопровождаемое внутренней оценкой уверенности), а человек, на основе анализа предсказания модели и анализа самого образца, выносит окончательное решение [49]. Это позволяет одновременно и снизить сложность классификации для человека, и потенциально повысить точность.

– Сначала выбор делает человек, потом ему демонстрируется выбор модели и дается возможность пересмотреть решение [17]. Этот вариант характеризуется большей нагрузкой на человека, поскольку ему в любом случае приходится принимать решение, а иногда еще и пересматривать его, но потенциально позволяет повысить качество принятия решений по сравнению с предыдущей разновидностью, так как человек оказывается сильнее вовлечен в решение задачи.

Варьируя последовательность активации участников и характер передаваемой информации в рамках последовательной схемы, можно получить значительное многообразие конкретных сценариев, которые будут отличаться своими свойствами, удобством для человека.

В целом, подобные сценарии характерны для случаев, когда экземпляров решаемых задач не очень много, и гораздо важнее принять верное решение, нежели снизить нагрузку на человека. При разработке и анализе подобных сценариев акцент делается на фактическую эффективность работы человека при наличии той или иной информации, полученной от модели. Исследования здесь носят в значительной степени эмпирический характер и смыкаются с исследованиями в области эффективных человеко-машинных интерфейсов.

Наконец, параллельная обработка предполагает, что каждый экземпляр задачи обрабатывается независимо и человеком, и моделью ИИ, а затем производится автоматическое слияние полученных результатов [23, 31, 41, 44, 47, 50]. Здесь человек также должен обрабатывать все образцы, то есть в подобных методах речь идет не о снижении нагрузки на человека или стоимости, а преследуется преимущественно цель повышения качества принятия решений системой человек – модель ИИ.

3.2. Характеристики и число экспертов (людей). Как уже было указано во введении, в статье рассматривается только такая постановка задачи совместной работы модели ИИ и человека, в которой допускается возможность ошибки человека. При этом, несмотря на общее допущение о возможности ошибки, в разных методах делаются различные предположения относительно характера таких ошибок. Можно говорить о модели ошибок человека, и предположение о структуре этой модели является одной из важных характеристик рассматриваемых в статье методов распределения задач.

Простейшим подобного рода допущением является постоянная вероятность ошибки [51] или, в случае задачи многоклассовой классификации, матрица ошибок, соответствующая эксперту [42, 44].

Более правдоподобным и часто используемым, но и более сложным допущением о поведении эксперта является допущение зависимости вероятности ошибки от образца [18 – 22, 27 – 29, 32 – 35, 48]. То есть, предполагается, что в признаковом пространстве могут быть области, «простые» для данного эксперта, а могут быть «трудные». Причем, в ситуации, когда экспертов несколько, «простые» и «трудные» области разных экспертов могут различаться. Подобное допущение влечет за собой необходимость прямого или косвенного обучения модели, предсказывающей точность эксперта в каждой области признакового пространства, что и делается в большинстве методов.

Другим аспектом, относящимся к экспертам, является их количество. В большинстве статей рассматривается ситуация, в которой есть одна модель ИИ и один эксперт, однако есть и работы, в которых допускается, что экспертов может быть много, причем они могут различаться по своим знаниям и компетенциям. При этом необходимо не только определить то, должен ли образец быть обработан моделью или экспертом, но и выбрать одного (или нескольких) из экспертов [22, 27, 31, 41, 50].

3.3. Вид задачи. В подавляющем большинстве работ рассматривается совместное решение задачи классификации – бинарной [17, 28] или многоклассовой [20, 22, 23, 27, 32, 33, 44].

Совместное решение задачи регрессии рассматривается всего в двух статьях: [19, 35].

Вместе с тем, есть и работы, где речь идет о формировании совместной стратегии управления [26, 38, 53], например, управление осуществляется автоматически (моделью), но в некоторые моменты (в некотором состоянии) оказывается выгодно передать его человеку-эксперту.

3.4. Характер модели ИИ. Класс моделей ИИ, используемых для решения задачи, может накладывать определенные ограничения на метод обеспечения совместной работы. Так, некоторые методы ориентированы на определенные классы моделей (например, SVM) [28], в других – делаются минимальные допущения о характере модели – например, она может быть «черным ящиком», что характерно для большинства случаев.

Можно выделить две разновидности модели «черного ящика»: без доступа к внутреннему состоянию, с доступом к внутреннему состоянию (т.н. «серый ящик»). Первая разновидность характерна тем, что пользователь (или другая модель) может наблюдать только результат модели. Во втором случае появляется возможность использования внутренних представлений модели (например, скрытых слоев нейронной сети) для их последующего анализа или аппроксимации модели.

Модель «белого ящика» предполагает, что процесс и логика принятия решения доступна, и, кроме результата модели, можно видеть то, что привело к его получению.

4. Методы обеспечения совместной работы. Можно выделить три группы методов обеспечения совместной работы: обучаемые, эвристические и байесовские. Первые две группы особенно часто используются в сценарии делегирования, последняя же – наиболее характерна для сценария параллельной обработки. Обучаемые

подходы включают такие методы, где предлагается обучение специальной модели, принимающей решение о том, кто должен обрабатывать образец – человек или модель ИИ. В эвристических методах определяется правило, в соответствии с которым осуществляется делегирование. Простейшим и широко используемым видом подобных правил являются правила, основанные на оценке неопределенности модели. Эвристические методы широко распространены в области обучения с отказом, их адаптация для случая с «неидеальным» человеком зачастую производится путем обучения прокси-модели, позволяющей оценить надежность классификации образца человеком. В этом случае решающее правило просто назначает образец модели ИИ или человеку в зависимости от того, у кого оказывается выше оценка надежности [16].

Потенциальным преимуществом эвристических методов является отсутствие необходимости обучения модели делегирования, однако при допущении «неидеальности» человека, а особенно, зависимости вероятности правильного результата от образца (наличии областей сильной и слабой экспертизы) это преимущество сводится на нет тем фактом, что для эвристических методов требуется получить прокси-модель эксперта, обучение которой требует достаточно много данных о реальных действиях человека.

Распределение задач в ходе совместной работы человека и модели ИИ предполагает принятие двух решений – формирование целевого класса (в случае классификации) на основе признаков образца и определение того, какой из участников системы (человек или модель ИИ) должен обрабатывать заданный образец. Эти решения могут приниматься раздельно (разными моделями) или одновременно (одной моделью), таким образом, сама структура распределения задач может быть либо распределенной (несколько моделей), либо объединенной (одна модель).

Как правило, распределенной структуре соответствует раздельное обучение, то есть обучение системы происходит в два этапа [1, 15, 32, 38, 40]. На первом этапе обучается модель для решения «целевой задачи» классификации без учета возможности делегирования. Для этого не используются ни метки, характеризующие решение задачи человеком, ни специальные функции потерь. На втором этапе обучается модель делегирования, принимающая решение о том, должен ли образец обрабатываться моделью (обученной на первом этапе) или человеком.

При объединенной структуре (характерно для совместного обучения) модель обучается и решению целевой задачи,

и вспомогательной (например, принятие решения о делегировании) с использованием набора данных, включающего результаты обработки образцов человеком-экспертом [22, 28, 29, 38].

Сопоставлению и теоретическому исследованию совместного и раздельного обучения моделей посвящена статья [18]. Основным преимуществом раздельного обучения является его большая универсальность – на первом этапе обучение происходит стандартным образом, не требуя результатов эксперта. Это значит, что раздельные методы могут быть применимы и к моделям, обучение которых не контролируется (полученным от третьих лиц). Достоинством (и основным мотивом развития) подходов, предполагающих совместное обучение, является то, что модель классификации может «фокусироваться» на разделимых регионах, обеспечивая лучшую классификацию в них, при этом изначально уделяя меньше «внимания» регионам, в которых классы оказываются плохо разделимы, оставляя их для человека. Подобный подход особенно хорошо применим к моделям с достаточно низкой выразительной способностью, поскольку позволяет лучше управлять тем, как разделяющая поверхность расположена в пространстве признаков (и выбрать «наилучшей» ту область признакового пространства, где размещение разделяющей поверхности оказывается наиболее целесообразным) [18].

Серьезной проблемой раздельного обучения является то, что этот подход не позволяет модели классификации подстраиваться под область компетенции эксперта. Ограничения раздельного обучения можно проиллюстрировать рисунком 4.

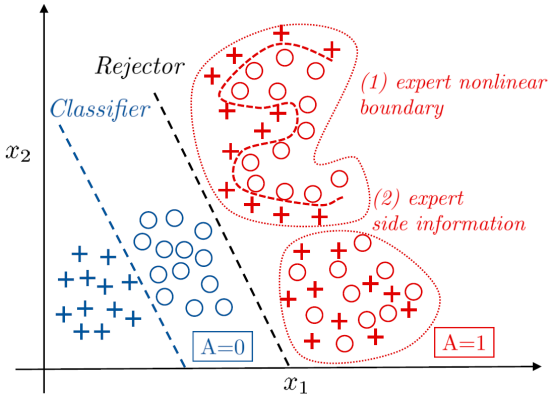


Рис. 4. Иллюстрация преимуществ совместного обучения (из [14])

Скажем, если распределение классов в пространстве признаков выглядит так, как на рисунке 4, то на первом шаге будет очень сложно обучить модель, однако если обучать одновременно и целевую модель (Classifier), и модель делегирования (Rejector), то целевая может оказаться очень простой (линейной), как и модель делегирования.

4.1. Оценка уверенности модели. Одним из простейших эвристических подходов к распределению задач между моделью и экспертом является подход, основанный на оценке уверенности модели. Этот подход в значительной мере унаследован из обучения с отказом, где также применяется достаточно широко. Общая идея заключается в том, чтобы обучить сначала модель для решения целевой задачи, обеспечивающую не только формирование целевой метки, но и соответствующего ей показателя уверенности. А затем установить диапазоны значений уверенности, при которых экземпляр должен перенаправляться эксперту. Смысл в том, чтобы перенаправлять эксперту те образцы, применительно к которым уверенность модели оказывается достаточно низкой.

Существует несколько способов получения уверенности модели, основные из них:

– Максимальное значение многопеременной логистической функции активации (Softmax Response). Применяется обычно к нейросетевым моделям многоклассовой классификации, выход которых формируется с помощью многопеременной логистической функции активации («софтмакс»). Соответственно, на максимальное значение такого выходного слоя может быть установлен порог – если максимальное значение оказывается ниже порога, то сеть «отказывается» от предсказания в пользу эксперта [54, 55];

– Исключение Монте-Карло (Monte-Carlo Dropout, MC-dropout) [56, 57] – оценка уверенности посредством подсчета статистики предсказаний нескольких прямых распространений с дропаутом. Здесь, в частности, используется интерпретация дропаута как техники ансамблирования, собирающей разные сети с разделяемыми весами в один ансамбль. Однако это требует большого количества прямых распространений (сотни), что может быть достаточно затратно.

– Использование статистик по ансамблям моделей [58].

Простой подход, основанный на уверенности моделей и порогах, предложен в [2]. Для обученной модели бинарной классификации устанавливаются два порога: t_0 и t_1. Если предсказание модели оказывается меньше t_0, то формируется отрицательный результат; если больше t_1, то положительный; если же предсказание

оказывается между t_0 и t_1, то образец передается эксперту. Каждая пара порогов оценивается на основе совместной функции потерь, выбирается такая пара, для которой значение функции потерь на обучающем множестве минимально.

Более точный подход из этой группы предлагается, например, в [15] – для каждого образца оценивается уверенность модели, неопределенность при классификации образца экспертом, а потом для модели назначаются те образцы, для которых разница между этими неопределенностями оказывается наибольшей. Для оценки неопределенности при классификации образца экспертом (а ее нужно выполнить до назначения и без реальных оценок экспертов) используется нейронная сеть, на вход которой подаются эмбеддинги объектов, а на выходе – признак несогласия нескольких экспертов [59]. Сеть обучается на наборе данных, для которых есть экспертные оценки. Предложенный подход, учитывающий разницу между обозначенными неопределенностями, используется и в [40], однако, в отличие от [15], здесь рассматривается ряд байесовских методов для вычисления неопределенности модели, а не эвристический подход, основанный на уверенности модели глубокого обучения (нейронной сети).

Общим достоинством всех этих методов является то, что они позволяют использовать существующие модели и добавлять к ним возможность перенаправления эксперту.

4.2. Суррогатные функции потерь. Основным инструментом для обучения моделей, учитывающих возможность переадресации задачи человеку-эксперту, является определение специальной функции потерь, учитывающей наличие экспертных меток. Данные функции потерь учитывают стоимость обращения к эксперту и основываются на сопоставлении вероятности ошибки имеющейся модели и вероятности ошибки человека-эксперта. Составленное таким образом соотношение не всегда оказывается легко оптимизируемым, и с этой целью оно заменяется более удобным в работе приближением, поэтому составляемые таким образом функции потерь часто называются «суррогатными» [1, 14, 20, 27].

Важными аспектами, учитываемыми при разработке и исследовании суррогатных функций потерь являются:

– Так называемая «консистентность» (consistency) по Байесу. Консистентная суррогатная функция потерь – это такая функция потерь, минимизация которой согласуется с оптимальным байесовским классификатором.

– Ведет ли использование функции потерь к получению хорошо калиброванных классификаторов [20]. Так, в [27] показано, что предложенная авторами функция потерь ведет к получению калиброванных классификаторов, а функция потерь, предложенная ранее в [2] – не ведет.

Конструирование суррогатных функций потерь наиболее распространено при решении задачи делегирования. Формальная постановка задачи следующая. Пусть $\mathcal{X}$ – пространство признаков, $\mathcal{Y} = \mathcal{M}$ – пространство меток и меток, даваемых экспертами (K классов), $\mathcal{D} = \{x_n, y_n, m_n\}_{n=1}^{N}$ – набор данных для обучения. То есть, каждый образец набора данных снабжен не только целевой меткой y_n, но и меткой, полученной от эксперта m_n. Целью является обучение двух моделей: классификатора $h: \mathcal{X} \to \mathcal{Y}$ и функции делегирования $r: \mathcal{X} \to \{0,1\}$ (называемой также в литературе rejector). При $r(x) = 0$ окончательное решение принимает классификатор, иначе – эксперт.

«Естественная» функция потерь для обучения этой пары моделей записывается следующим образом [14]:

$$\mathcal{L}_{nat}(h, r) = \mathbb{E}_{x,y,m}[\ell(x, y, h(x))\mathbb{I}_{[r(x)=0]} + \ell_{\exp}(x, y, m)\mathbb{I}_{[r(x)=1]}]. \quad (1)$$

Здесь $\ell(\cdot)$ – функция потерь классификатора, а $\ell_{\exp}(\cdot)$ – функция потерь эксперта. Возможные дополнительные расходы, связанные с привлечением эксперта, могут быть учтены прямо в $\ell_{\exp}$, таким образом, значение этой функции может быть ненулевым даже в случае правильного прогноза. Однако эти дополнительные расходы необходимо выразить в терминах ошибок, что может быть довольно сложно на практике. В любом случае, непосредственная минимизация подобной «естественной» функции оказывается практически невозможной, в первую очередь, в силу дискретного характера функции делегирования r.

В [2] предлагается адаптация $\mathcal{L}_{nat}$, допускающая непосредственную оптимизацию градиентными методами (записано для одного образца):

$$L(x_i, y_i, m_i, h, r) = (1 - r(x_i))\ell(y_i, h(x_i)) + r(x_i)\ell(y_i, h(m_i)). \quad (2)$$

Здесь следует обратить внимание на два важных отличия от $\mathcal{L}_{nat}$. Во-первых, функция делегирования не является бинарной, что, в частности, позволяет использовать эту функцию потерь с градиентными методами, во-вторых, авторы [2] не используют отдельную функцию потерь для экспертной классификации, $\ell(\cdot)$ здесь – это бинарная кросс-энтропия (речь идет о бинарной классификации), поэтому дополнительные расходы на привлечение эксперта никак не учитываются, и речь, по всей видимости, идет просто о максимизации точности. Авторы также описывают несколько тонкостей в обучении модели, среди которых следует выделить следующие: 1) r может зависеть не только от признаков объекта, но и от результата обработки объекта основной моделью $h(x)$, 2) может быть целесообразно ограничить распространение градиента стратегии распределения по h, чтобы h оставался хорошей моделью на всей области значений $\mathcal{X}$ и не происходило деградации качества в тех областях, где целесообразно привлечение эксперта.

Однако напрямую это выражение оптимизировать тяжело, поэтому в [14] предложена суррогатная (но консистентная) функция потерь, основанная на многопеременной логистической функции. Авторы рассматривают задачу многоклассовой классификации (с K классами) и предлагают свести задачу совместной классификации к задаче *cost sensitive learning* (CSS, обучение с учетом стоимости) на расширенном наборе классов. Расширенный набор классов формируется добавлением еще одного класса, означающего перенаправление образца эксперту. Переход осуществляется следующим образом. Для каждого образца вводится понятие стоимости классификации его как каждого из классов ($c(i)$ – стоимость классификации образца как принадлежащего i-тому классу, $i \in \{1,\ldots,K+1\}$). Для образца (x, y) $c(i)$ определяется как $\ell(x, y, \hat{y})$ для классов $\{1,\ldots,K\}$ и как $\ell_{\exp}(x, y, m)$ для дополнительного класса, соответствующего передаче образца эксперту. Авторы предлагают следующую консистентную функцию потерь:

$$L_{CE}(g_1,\ldots,g_{K+1},x,c(1),\ldots,c(K+1)) = \\ = -\sum_{i=1}^{K+1}\left(\max_{j\in[K+1]} c(j) - c(i)\right)\log\left(\frac{\exp(g_i(x))}{\sum_k \exp(g_k(x))}\right). \quad (3)$$

Здесь g_i – это выходы модели (авторы предполагают, что это нейронная сеть). Основу данной целевой функции составляет

многопеременная логистическая функция («софтмакс»), применяемая к выходам модели, поэтому данная функция потерь также называется «софтмакс-параметризацией».

В статье [20] показано, что модели, обученные с помощью софтмакс-параметризации не является калиброванными, поэтому предложен другой вариант суррогатной функции потерь, т.н. «один-против-всех» (one-vs-all, OvA):

$$L_{OvA}(g_1,\ldots,g_{K+1},x,y,m)=\phi[g_y(x)]+\sum_{y'\in Y,y'\neq y}\phi[-g_{y'}(x)]+ \\ +\phi[-g_{K+1}(x)]+\mathbb{I}[m=y](\phi[g_{K+1}(x)]-\phi[-g_{K+1}(x)]), \tag{4}$$

где ϕ – бинарная суррогатная функция потерь (например, логистическая функция). Неформально, логика этой функции потерь заключается в следующем: первое слагаемое обеспечивает повышение выходного значения для правильного класса (g_y), второе слагаемое (оператор суммирования) обеспечивает понижение выходного значения для ошибочных классов, третье и четвертое слагаемое в комплексе обеспечивают повышение выходного значения для выхода модели, связанного с перенаправлением эксперту (g_{K+1}), если ответ эксперта для данного примера правильный, и понижение значения этого выхода, если ответ эксперта неверный.

Сами классификатор и функция делегирования устроены в любом случае одинаково:

$$r(x)=\mathbb{I}[g_{K+1}(x)\geq\max_k g_k(x)], \\ h(x)=\arg\max_{k\in\{1,\ldots,K\}} g_k(x). \tag{5}$$

В статьях [27, 33] данные функции обобщены на случай нескольких экспертов – такие модели не просто принимают решение передать ли образец эксперту, но и какому именно эксперту его передать.

В статье [39] показано, что модели делегирования, обученные с помощью существующих суррогатных потерь (CSS и OvA), могут быть склонны к недообучению в тех случаях, когда обращение к экспертам влечет за собой дополнительную стоимость. В связи с этим, предлагается способ ретроспективной коррекции суррогатных потерь как для CSS, так и для OvA.

В статье [45] делается успешная попытка улучшения подхода [14] для его использования в сочетании с конкретными людьми в рамках распределения задач. Предлагаемое улучшение, заключающееся в тонкой настройке, повышает общую точность системы «человек – модель ИИ». Для этого модель эксперта сначала обучается с использованием агрегированных человеческих меток, а затем – с использованием меток, полученных от конкретных людей.

В [52] демонстрируется, что существующие подходы не всегда могут совместно оптимизировать классификатор и модель делегирования с низкой ошибкой неправильной классификации (даже в том случае, если существуют линейный классификатор и соответствующая модель делегирования, обеспечивающие безошибочную классификацию). Для решения этой проблемы задача делегирования рассматривается как задача смешанного целочисленного линейного программирования и предлагается новая консистентная суррогатная функция потерь, которая обеспечивает лучшую эмпирическую производительность по сравнению с существующими суррогатными подходами.

4.3. Методы, основанные на теории полезности. В ряде работ [1, 38, 51] для конструирования задачи оптимизации используется теория полезности. Достоинством такого подхода является то, что само назначение стоимости ошибки, стоимости обращения к эксперту может быть оценено напрямую из знаний предметной области. Однако на практике модель, построенная в терминах теории полезности не всегда напрямую оказывается хороша для использования при обучении стратегии делегирования, поэтому она может быть адаптирована, заменена суррогатной функцией, во многом с использованием тех же идей, что изложены в предыдущем подразделе.

Так, авторы [1] предлагают следующую формулировку задачи делегирования в терминах теории полезности:

$$\arg\max_{r,h} \mathbb{E}_{(x,y,m)\sim P}\left[r(x)(u(y,m)-c)+(1-r(x))(u(y,h(x)))\right]. \qquad (6)$$

Здесь $u(y,m)$ – полезность ответа эксперта m при верном ответе y, $u(y,h(x))$ – полезность результата модели при верном ответе y, c – стоимость обращения к эксперту. Можно отметить, что это выражение очень похоже на $\mathcal{L}_{nat}$, поскольку оба выражают основную идею делегирования.

Авторы [1] также предлагают целую палитру методов обучения h и r, выделяя дискриминативные подходы, в которых эти функции обучаются непосредственно отображению признаков в решения без построения промежуточных вероятностных моделей различных компонентов системы, и вероятностные подходы, основанные на стоимости информации.

Фиксированный дискриминативный подход заключается в том, что сначала обучается модель h (любым известным методом), а затем – модель r с использованием сформированной функции ожидаемой полезности (подход аналогичен предложенному в [15]).

Объединенный дискриминативный подход предполагает совместное обучение h и r. Авторы также сталкиваются с тем, что непосредственная оптимизация затруднительна, поэтому для поиска модели h и стратегии делегирования r используют следующую суррогатную функцию потерь:

$$\ell(y, r(x)m + (1 - r(x))h(x)) + cr(x). \quad (7)$$

В ходе работы для принятия решений о направлении задачи эксперту авторы аппроксимируют идеализированное предсказание с использованием меры уверенности модели, $\max(h(x))$. Запрос эксперту посылается тогда, когда $(1 - r(x))\max(h(x)) < r(x)$. То есть, запрос посылается эксперту, если $r(x)$ имеет большое значение либо если неопределенность предсказания высока.

В выделяемом авторами [1] подходе, основанном на стоимости информации, предполагается независимое обучение трех вероятностных моделей: модели распределения меток при условии известных значений признаков $(p_\alpha(y \mid x))$; модели ответа эксперта при условии известных значений признаков $(p_\beta(y \mid x))$ и модели распределения меток при условии известных признаков и ответов экспертов $(p_\gamma(y \mid m, x))$. Для построения моделей авторы предлагают использовать нейронные сети с последующей вероятностной калибровкой методом Платта [60]. Во время выполнения эти вероятности используются для оценки ожидаемой полезности обращения к эксперту.

В [51] на основе теории полезности производится формализация последовательного подхода к сотрудничеству, когда каждый образец обрабатывается сначала моделью, а потом эксперт, зная результат работы модели, принимает решение о том, стоит ли просто принять

его или детально исследовать образец и выполнить классификацию самостоятельно. Авторы составляют матрицу выигрышей (таблица 1), где под метарешением понимается решение эксперта о том, стоит ли доверять модели, обработка образца связана с затратой усилий $\lambda > 0$. Само же решение – это результат системы ИИ-человек, и он может быть либо правильным (этому случаю соответствует максимальная полезность 1), либо неправильным (чему соответствует стоимость ошибки $\beta \geq 1$).

Таблица 1. Матрица полезности (из [51])

Метарешение\Решение	**Правильно**	**Неправильно**
Принять (Accept, A)	1	$-\beta$
Решать самому (Solve, S)	$1-\lambda$	$-\beta-\lambda$

Оптимальный классификатор в такой постановке должен максимизировать ожидаемую полезность:

$$h^* = \arg\max_h \mathbb{E}_{x,y}[U(m,d)], \tag{8}$$

где m – функция, в соответствии с которой принимается метарешение (принять или решать самостоятельно), а d – итоговое решение.

Опираясь на требование к оптимальному классификатору и допущение о рациональности человека (и, следовательно, определенную стратегию принятия решений), авторы записывают общее выражение для ожидаемой полезности и предлагают оптимизировать его непосредственно в ходе градиентного спуска. Авторы столкнулись с тем, что непосредственная оптимизация оказалась затруднительна, поскольку при случайной инициализации модель «неуверенна», а значит, решать для всех образцов должен человек, и в этой области нет градиентов для обучения модели, поэтому они начали с модели, обученной для решения задачи без человека.

Схожие модели, основанные на теории полезности, используются и при рассмотрении процессов совместной работы. Модель на основе теории полезности здесь, как правило, сочетается с уравнениями Беллмана [26, 53].

В отличие от большинства работ, которые рассматривают задачу обучения с учителем (например, классификацию), статья [38] фокусируется на проблеме выбора оптимальной стратегии делегирования в контексте обратной связи типа «бандит», при которой

вознаграждение и результаты зависят от всех предыдущих действий человека. Это требует оценки альтернативных вариантов действий и выбора тех действий, которые приведут к наибольшему ожидаемому вознаграждению. Например, образец данных может представлять пациента, в отношении которого агент (человек, принимающий решение, или модель ИИ) может предпринять какое-либо действие (один из методов лечения) и затем получить соответствующее вознаграждение (эффект от лечения). Для нахождения стратегии делегирования авторы максимизируют средневзвешенное вознаграждение человека и модели, подобно тому, как было показано в формуле (6). При этом рассматривается как случай раздельного обучения модели делегирования и модели принятия решений, так и их совместного обучения.

4.4. Условная оптимизация. В предыдущих подразделах в процессе обучения модели, управляющей совместным решением задач, нагрузка на эксперта учитывалась опосредованно – в виде штрафа в совместной целевой функции за обработку образца экспертом или отдельного слагаемого в функции полезности. Однако в некоторых случаях подобные веса назначить сложно и, более того, может существовать физическое ограничение на количество образцов, которые могут обрабатываться экспертом [19, 28, 35].

В статье [35] предлагается формальная постановка для решения задачи регрессии при наличии такого ограничения, а в [28] – классификации. Так, в [28] рассматриваются классификаторы на основе отступа (margin) между классами (например, SVM). Пусть $\mathcal{V}$ – обучающее множество, S – часть обучающего множества, которая при обучении будет передана экспертам, n – ограничение сверху на количество элементов в S. Тогда распределение образцов между моделью h_θ и экспертом сводится к тому, чтобы выбрать некоторое множество обучающих образцов $S \in \mathcal{V}$, которые будут передаваться эксперту ($|S| \le n$), и построить решающую поверхность (decision boundary), разделяющую векторы признаков в подмножестве обучающего множества $S^c = \mathcal{V} \setminus S$. Целевая функция может быть записана так:

$$\min_{S,\theta} \sum_{i \in V \setminus S} \ell(h_\theta(x_i), y_i) + \sum_{i \in S} c(x_i, y_i), \quad s.\,t.\ |S| \le n, \tag{9}$$

где $c(x_i, y_i)$ – ошибка человека на образце (human error per sample).

Интересно, что n устанавливается относительно обучающего множества, причем, во-первых, фактически экспертные метки все равно должны быть известны для всех образцов обучающего множества (чтобы выбрать те, которые целесообразно исключить при обучении модели); во-вторых, на практике гораздо большую ценность играет ограничение количества задач, назначаемых эксперту во время выполнения. Здесь авторы опираются на то, что $\mathcal{V}$ является представительной выборкой из исходного распределения, и выбранные для назначения эксперту образцы задают область пространства признаков, в которую попадает приблизительно $n/|\mathcal{V}|$ образцов как обучающего, так и тестового множеств.

Общее выражение минимизации конкретизируется для случая SVM, и показано, что в этом случае выбор S образцов может быть осуществлен жадным алгоритмом.

Для принятия решения о том, стоит ли назначать новый (не виденный ранее) образец человеку или обрабатывать его моделью (во время вывода) предлагается обучить еще одну модель $\pi(d \mid x)$. Модель обучается на основе набора данных $\{(x_i, d_i\}_{i \in \mathcal{V}}$, где x_i – признаки объектов (те же самые, как и в основной задаче – обучении с делегированием), а $d_i = +1$, если $i \in S^*$ (фактическое множество образцов, назначенных эксперту, в результате решения задачи условной оптимизации) и $d_i = -1$ в противном случае. Считается, что эта модель хорошо аппроксимирует распределение $p(x)\pi(d = -1 \mid x)$ (распределение объектов, хорошо классифицируемых основной моделью).

В [19] для решения подобной задачи оптимизации предлагается градиентный алгоритм, который итеративно оптимизирует классификатор в образцах, где он превосходит человека в обучающей выборке, а затем обучает модель делегирования, чтобы предсказать, у человека или у модели ИИ будет более высокая ошибка на уровне каждого образца. Авторы показывают, что алгоритм гарантированно находит прогнозирующие модели и политики делегирования, с учетом ограничения на число элементов в S.

4.5. Активное обучение. Получение образцов, размеченных экспертом, как правило, достаточно трудоемкий и затратный процесс, в большинстве же подходов, рассмотренных ранее, предполагалось, что для всего обучающего множества присутствуют экспертные метки. Часть этих меток может оказаться избыточной, поэтому перспективным подходом является применение различных методов

и техник, позволяющих снизить зависимость от экспертной разметки. Одним из таких методов является активное обучение – модель запрашивает разметку именно тех образцов, которые оказываются наиболее полезными с точки зрения построения модели ошибок эксперта. Так, в статье [18] предложен алгоритм делегирования, основанный на активном обучении. Алгоритм включает два этапа:

– на первом запускается стандартный алгоритм активного обучения (например, CAL [61]) для пространства $\mathcal{D}$, чтобы получить функцию f несоответствия предсказаний эксперта и эталонных ответов с ошибкой не более ϵ;

– на втором этапе данные размечаются этой функцией $\hat{f}$, и на основе этих данных обучается пара классификатор-модель делегирования.

В статье [62] предлагается трехэтапный подход к сокращению количества экспертных прогнозов, необходимых для обучения алгоритмов делегирования. Он включает в себя следующие шаги (этапы):

1. Обучение модели встраивания (embedding model) с метками (ground truth), которые используются для извлечения представлений признаков.

2. Представления признаков служат исходными данными для обучения модели прогнозирования экспертных знаний (expertise predictor model), чтобы аппроксимировать возможности эксперта-человека.

3. Модель прогнозирования экспертных знаний генерирует искусственные экспертные прогнозы для экземпляров, не размеченных экспертом-человеком.

Затем для обучения алгоритмов делегирования можно использовать как человеческие, так и искусственные экспертные прогнозы. Таким образом, в отличие от [18], здесь не требуется итеративное выявление образцов, для которых запрашиваются прогнозы экспертов-людей. Вместо этого, учитывая небольшое количество прогнозов экспертов-людей, алгоритм учится выводить искусственные прогнозы для неразмеченных образцов в обучающем наборе данных.

Другая работа [30], рассматривающая возможность активного обучения, посвящена онлайн-прогнозированию консенсуса группы экспертов. Предполагается, что консенсус экспертов-людей определяет исключительно метку образца, которую необходимо предсказать. Поскольку запрос полного консенсуса может быть затратным, авторы динамически оценивают консенсус на основе

частичной обратной связи, анализируя уверенность экспертов и модели ИИ. Авторы ищут компромисс между стоимостью обращения к экспертам и точностью классификации. Таким образом, цель работы состоит в том, чтобы максимально повысить точность предсказания консенсуса при ограниченном «бюджете» на аннотации экспертов.

4.6. Слияние данных. В рамках подхода параллельной обработки можно выделить группу методов т.н. слияния данных. Как отмечалось ранее, основная цель слияния данных состоит в том, чтобы повысить точность принятия решений системой «человек – модель ИИ». Наиболее распространенные методы этой группы сосредоточены на комбинировании предсказаний модели ИИ с метками, предсказанными людьми. При этом в процессе классификации может участвовать как один эксперт-человек [23, 44, 47], так и множество экспертов [31, 41, 50].

Для комбинирования предсказаний чаще всего используется байесовская статистика, предполагающая, что вероятность, которая отражает степень доверия событию, может изменяться в зависимости от некоторой дополнительной информации. Так, в статье [44] рассматривается задача многоклассовой классификации изображений, где решения по категориальной классификации независимо принимают один эксперт-человек (предсказывают только метку) и одна модель классификации, прогнозирующая распределение по всем возможным меткам (классам). Для объединения предсказаний используется вероятностный подход, при котором условное распределение по предсказываемым меткам может быть учтено с помощью правила Байеса следующим образом:

$$p(y \mid h(x), m(x)) \propto p(h(x) \mid y, m(x)) p(y \mid m(x)), \tag{10}$$

где $x \in \mathcal{X}$ – образец набора данных; $y \in \mathcal{Y}$ – истинная метка; $h(x) \in \mathcal{Y}$ – метка, предсказанная человеком; $m(x) \in \mathbb{R}^K$ – нормированный вектор вероятности, выводимый моделью ИИ (K – число классов).

Важно заметить, что далее авторы делают допущение об условной независимости $h(x)$ и $m(x)$ при y, в соответствии с которым приведенное выше выражение может быть преобразовано к следующему виду:

$$p(y \mid h(x), m(x)) \propto p(h(x) \mid y) p(y \mid m(x)). \tag{11}$$

Слагаемое $p(h(x)\,|\,y)$ можно интерпретировать как калиброванные вероятности на уровне класса. $p(h(x)\,|\,y)$ параметризуется матрицей ошибок эксперта h, которая обозначается как φ и содержит элементы $\varphi_{ij} = p(h(x) = i\,|\,y = j)$. Второе слагаемое $p(y\,|\,m(x))$ можно интерпретировать как калиброванные вероятности на уровне образца. Однако вероятностный результат классификатора $m(x)$ может отличаться от $p(y\,|\,m(x))$. В связи с этим, авторы предлагают процедуру post-hoc калибровки, которая сопоставляет $m(x)$ с хорошо откалиброванными вероятностями с помощью т.н. калибровочной карты с параметрами θ. Вывод классификатора после применения такой калибровочной карты обозначается как $m^{\theta}(x)$.

Наконец, прогнозируемая вероятность класса j, учитывая, что человек предсказывает класс i, и модель создает вектор вероятности классов $m(x)$, будет определяться следующим выражением:

$$p(y = j\,|\,h(x) = i, m(x)) = \frac{\varphi_{ij} m_j^{\theta}(x)}{\sum_{k=1}^{K} \varphi_{ik} m_k^{\theta}(x)}. \qquad (12)$$

Хотя наиболее простой оценкой элементов матрицы ошибок является оценка максимального правдоподобия, при малом количестве человеческих меток данная оценка будет иметь большую дисперсию. Вместо этого авторы предлагают байесовский подход к включению априорной информации, однако в рамках данного обзора он не представляет большого интереса.

В [23] предлагается подход байесовского моделирования, с помощью которого формируется комбинированный прогноз, а также оценки скрытой корреляции между классификаторами. Эта корреляция отражает зависимости между показателями достоверности классификации людей и моделей ИИ. Рассматривается сценарий, в котором, кроме предсказанной метки, человек предоставляет свою степень уверенности («низкая», «средняя», «высокая»). Таким образом, в отличие от [44], предлагаемая байесовская модель оценивает корреляцию между уверенностью человека и модели ИИ и, кроме того, не опирается на предположение об условной независимости.

В [50] отмечается, что, хотя комбинированная модель в [44] обеспечивает гораздо большую точность, чем при независимой классификации человеком и моделью ИИ, она ограничивается

объединением предсказания лишь одного человека с результатами модели, что может существенно снизить точность комбинированного подхода, поскольку результат классификации зависит от точности конкретного человека. В данной же статье ([50]) предлагается подход к объединению решений множества людей с результатом модели ИИ. Кроме того, предлагается эффективный алгоритм поиска оптимальной подгруппы людей, чьи объединенные метки позволят получить наиболее точный результат классификации.

В работах [31, 41, 47] также рассматривается задача комбинирования предсказаний множества людей с результатами модели ИИ, однако в них также уделяется немало внимания аспекту делегирования. Поэтому более подробно данные методы рассматриваются в пункте 4.8.

4.7. Ручное конфигурирование границ принятия решений. Ручное конфигурирование границ принятия решений подразумевает под собой то, что человек принимает непосредственное участие в определении области признаков, для которой модель ИИ в дальнейшем может осуществлять предсказания. Если рассматриваемый образец не попадает в границы данной области признаков, то модель ИИ делегирует задачу человеку.

В рамках данного обзора была обнаружена всего одна работа [37], в которой рассматривается подобный подход применительно к задаче модерации контента. Авторы называют этот подход «условным делегированием».

Области, определяющие границы принятия решений модели, задаются с помощью набора правил на основе ключевых слов, созданного в результате совместной работы человека и модели ИИ перед развертыванием. Например, после проверки прогнозов модели по комментариям со словом «отсталый» человек может решить, что модель хорошо справляется с их выявлением, и установить «отсталый» в качестве правила условного делегирования. После развертывания комментарии, относящиеся к этим областям, т.е. содержащие любые ключевые слова, указанные пользователем, могут быть использованы для принятия окончательных мер, таких как скрытие или отправка на дальнейшую проверку.

4.8. Гибридные подходы. Под гибридными подходами следует понимать подходы, которые, так или иначе, сочетают в себе два или более ранее рассмотренных методов обеспечения совместной работы человека и ИИ. Соответственно, каждый из подобных подходов может в себе воплощать сразу несколько сценариев взаимодействия человека и ИИ.

В рамках данного обзора было обнаружено несколько работ, причем в каждой из них внимание уделяется гибридизации методов слияния данных и делегирования, т.е. все они реализуют сразу два сценария: параллельная обработка и делегирование.

Авторы в [47] выделяют ряд ограничений ранее рассмотренного подхода комбинирования предсказаний [44]. В частности, в качестве недостатка отмечается то, что для каждого образца требуются метки, полученные от человека, что может представлять трудности, если эти метки недоступны в достаточном количестве. Кроме того, при наличии значительного разрыва между точностью человека и модели ИИ, одно может преобладать над другим, например, комбинированная модель может начать полагаться на менее точных людей. Наряду с этим, авторы отмечают, что типовые подходы к делегированию полностью игнорируют результаты модели ИИ в том случае, когда задача адресуется человеку. Поэтому в [47] предлагается способ объединения двух подходов: обучение с делегированием [14] и комбинирование предсказаний [44]. Общая идея заключается в том, что откалиброванные выходные данные модели ИИ объединяются с метками, полученными от людей, только в том случае, если было принято решение делегировать задачу человеку.

В [31] предлагается расширение подхода обучения с делегированием, основанного на использовании суррогатных функций потерь, для случая множества экспертов. Так, принятие решения о классификации может быть делегировано одному или нескольким экспертам (при этом сама модель ИИ также рассматривается в качестве эксперта). Окончательным результатом здесь является совокупное решение выбранного подмножества экспертов. Для получения совокупного решения в статье рассматриваются несколько методов формирования весов экспертов.

В [41] рассматривается подход интеграции обучения с участием нескольких экспертов [63] и обучения с использованием зашумленных меток [64, 65], то есть предполагается, что истинные метки могут отсутствовать (часто характерно для реальных наборов данных). Предлагаемый подход оптимизирует систему «человек – модель ИИ», стремясь повысить точность классификации при минимизации затрат на обращение к эксперту-человеку, которые варьируются от 0 до M, где M – максимальное количество экспертов-людей.

5. Оценка качества и валидация методов совместной работы. Процедура оценки качества методов и алгоритмов, используемых при распределении задач между человеком и ИИ также имеет рад особенностей: во-первых, качество можно оценивать по

двум зачастую взаимоисключающим направлениям – точность итоговой модели и нагрузка на человека, во-вторых, помимо исходных данных и эталонного результата для оценки нужны еще данные о решении задач экспертом, что не всегда возможно, поэтому в исследованиях зачастую применяются различные способы моделирования ответов эксперта на основе эталонных результатов, которые также охарактеризованы в данном разделе.

5.1. Метрики и процедуры оценки качества

5.1.1. Один показатель. Как уже указывалось, при совместном выполнении задач, как правило, важна не только общая точность, но количество задач, выполняемых человеком. Однако в ряде постановок метрики качества учитывают только точностные характеристики итоговой модели (связанные с долей ошибок). В наибольшей степени это характерно для двух ситуаций:

– целью является получение наиболее надежной модели для ответственных сценариев применения;

– человек все равно оказывается вовлечен в принятие решения по всем задачам (в последовательном или параллельном сценариях) (например, [16, 17]).

В работах [16, 17, 22, 27, 32, 33] в качестве главной метрики оценки модели применяется только точность (в стандартном смысле), определяемая как отношение количества образцов, при которых предсказание итоговой модели совпало с эталонным результатом, к общему количеству образцов, использованных при оценке.

5.1.2. Несколько показателей. В большинстве статей (особенно, в тех случаях, когда рассматривается сценарий делегирования) оценка производится с помощью двух метрик, описывающих долю задач, решенных человеком, и общую точность системы. Данная система метрик во многом унаследована из области обучения с отказом. Трудоемкость для человека, обычно, характеризуется через метрику, называемую «покрытие» (англ. coverage), определяемую как доля образов, которые были обработаны автоматически моделью (не переданы на обработку эксперту). Покрытие, соответственно, может изменяться от 0 (когда все образцы были обработаны экспертом) до 1 (когда все образцы обработаны моделью).

Поведение модели совместной работы при определенных настройках модели делегирования на определенном тестовом наборе данных, таким образом, может быть охарактеризовано парой характеристик: покрытие и точность (или количество ошибок). При варьировании настроек модели могут быть построены кривые,

характеризующие баланс между покрытием и точностью, широко используемые при анализе различных методов [20, 28, 29, 36].

В статье [24] предложена оригинальная метрика DEV (deferred error volume), сочетающая точность и покрытие, которая определяется как площадь под кривой, образованной оценками качества для разных комбинаций вероятности делегирования и порога на делегирование.

5.1.3. Свертка (на основе теории полезности). В подходах, основанных на теории полезности, все значимые характеристики (ошибка модели, назначение эксперту и пр.) выражены в рамках единой функции полезности, поэтому в качестве основной метрики используется также значение функции полезности. Это может быть ожидаемая полезность [51, 53], вычисляемая с помощью вероятностных выходов модели, а может быть эмпирическая полезность [51], вычисляемая уже с учетом примененных порогов (делегирования и классификации).

Сюда же относятся и онлайн-модели, например, основанные на обучении с подкреплением. Для подобных формализаций естественно сводить задачу к поиску экстремума свертки функции полезности, и оценка производится либо с помощью значения этой функции [21, 26] (вознаграждения, полученного агентом, который осуществляет распределение задач), либо через сожаление (regret) [43], характеризующее поведение модели по сравнению с идеальной.

5.2. Наборы данных. Валидация рассматриваемого класса моделей совместной работы эксперта и модели ИИ требует наличия не только эталонных меток, но и экспертных (которые могут отличаться от эталонных, отражая неполноту знаний эксперта), потому что в некоторых случаях (в зависимости от результата модели делегирования, например) при выполнении результирующей модели может происходить обращение к эксперту. Заранее неизвестно, при обработке каких именно образцов такое обращение целесообразно, поэтому экспертные метки должны быть определены для всех. Существует относительно немного публичных наборов данных, содержащих такие метки, поэтому в части исследований используются синтетические экспертные метки, полученные в результате применения какой-либо модели ошибок к эталонным меткам. Впрочем, моделирование поведения эксперта оказывается полезным не только при полном отсутствии экспертных оценок, как правило, при исследовании метода оказывается важно как именно он ведет себя при различной надежности эксперта, и моделирование используется для имитации ответов с разной надежностью [20, 27, 28, 32, 33].

5.2.1. Наборы данных, содержащие экспертные оценки. Наборы, содержащие экспертные оценки, можно, в свою очередь, охарактеризовать с помощью трех важнейших признаков: количество задействованных экспертов, полнота разметки набора каждым экспертом, и количество экспертных мнений на образец.

Наиболее распространенной категорией таких наборов являются такие, где экспертов задействовано много, причем каждый размечает не все образцы, но на каждый образец приходится несколько экспертных оценок – зачастую подобные наборы являются результатом использования краудсорсинга, где участникам площадки, как правило, за определенное вознаграждение, предлагается провести классификацию образцов какого-либо публичного набора данных. Подобным образом на основе известных наборов данных в области компьютерного зрения CIFAR и ImageNet были получены вариации, содержащие экспертные метки: CIFAR-10H [66] и ImageNet-16H [23]. Набор данных Galaxy Zoo [67, 68] получен в рамках проекта гражданской науки по классификации изображений галактик на одноименной площадке.

Поскольку в таких наборах данных каждый эксперт размечает не все объекты, возможности моделирования точности отдельного эксперта оказываются очень ограничены, такие наборы больше подходят для некоторого «усредненного» моделирования человеческого взгляда на задачу. То, что для каждого образца присутствует несколько меток, позволяет построить распределение, связанное с образцом и использовать это распределение в качестве основы для модели эксперта [42, 48].

Набор данных CIFAR-10H применяется в работах [21, 42, 44, 48], ImageNet-16H в работах [23, 44], а GalaxyZoo – в [19, 20, 27].

Похожая схема также у набора, используемого в статье [22] – это набор данных рентгеновских снимков ChestX-ray8 [69, 70]. Аннотация каждого снимка содержит оценки трех специалистов (из 22 задействованных в разметке) и согласованную оценку.

5.2.2. Полностью синтетические наборы данных. В ряде публикаций используются полностью синтетические наборы данных [18, 21, 28, 35, 42, 43, 48, 51]. Как правило, валидация на таких наборах данных дополняет валидацию на реальных наборах (например, [28, 35, 43]), однако в отдельных случаях используются только синтетические наборы.

Специальным образом сконструированный набор позволяет смоделировать некоторую интуитивную ситуацию, послужившую толчком к созданию метода (касающуюся распределения признаков,

распределению точности оценок эксперта), и подтвердить эффективность метода, по крайней мере, для этой ситуации.

В [25] предлагается набор данных *Financial Fraud Alert Review* (FiFAR) – синтетический набор для обнаружения мошенничества с банковскими счетами, содержащий предсказания 50 (смоделированных) «экспертов», обладающих различными характеристиками.

5.2.3. Моделирование ответов эксперта. Моделирование ответов эксперта используется с двумя основными целями: во-первых, оценка того или иного метода на наборах данных, не имеющих оценок экспертов (имеющих только эталонные метки); во-вторых, для исследования влияния точности эксперта на результат работы системы в целом. Наличие модели позволяет управлять уровнем точности и, соответственно, производить контролируемый эксперимент.

Формирование экспертных меток варьируемой надежности для реальных наборов данных, в которых есть только одна (эталонная) метка применяется в работах [20, 22, 27, 28, 48].

Можно выделить несколько распространенных приемов моделирования эксперта ограниченной точности. В части таких приемов не предполагается, что в наборе данных есть экспертные метки – для каждого образца присутствует только одна эталонная.

Эксперт, выбирающий класс случайным образом [27]. Как правило, такая модель используется либо в качестве достаточно слабого базового решения при сравнении, либо в ситуации, когда экспертов может быть много, чтобы показать, что модель в состоянии идентифицировать такого эксперта и ограничить его влияние на итоговый результат.

Эксперт, обладающий специализацией (например, [27]). Широко используемая в задачах многоклассовой классификации модель. В этой модели все множество выходных классов разбивается на два подмножества: те, которые данный эксперт различает хорошо и те, которые он различает не так хорошо. Для каждого из подмножеств задается вероятность правильного предсказания эксперта. Легко заметить, что подобная модель может являться обобщением и случайного эксперта (для него область специализации связана с пустым множеством классов) и «всезнающего» (область специализации связана с множеством, включающим все классы).

В качестве моделирования экспертов разной квалификации применяются также нейронные сети разной выразительности [32, 33] – сеть с большим количеством параметров соответствует более квалифицированному эксперту.

Для наборов данных, в которых присутствуют метки людей, основным способом моделирования «силы» эксперта является использование вероятностной смеси случайного выбора и сэмплирования из эмпирического распределения, задаваемого метками, которые люди-эксперты назначили данному образцу [27]. То есть с некоторой вероятностью p выбирается одна из реальных меток, которые были присвоены образцу, а с вероятностью $(1-p)$ – случайная метка. Очевидно, максимальное значение параметра p соответствует некоторой средней точности эксперта, моделировать более точные данные таким образом не получится.

5.3. Онлайн-валидация. Позволяющим получать достоверные результаты, но довольно трудоемким и дорогостоящим способом оценки методов и моделей совместной работы является онлайн-оценка, проводимая, как правило, с помощью краудсорсинга [16, 17, 31, 34, 36, 37, 49].

Суть подобной оценки заключается в том, что авторы реализуют предлагаемый метод сотрудничества с помощью инструментов той или иной площадки краудсорсинга – Amazon Mechanical Turk [31, 34, 36, 37, 49] или Prolific [16, 17], привлекают участников эксперимента через площадку, а затем делают выводы об эффективности предложенного метода по тому, как сочетается точность (и, возможно, время выполнения задач) в контрольной группе и группе, работающей в рамках предложенного метода.

Следует отметить, что онлайн-валидация позволяет проводить оценку даже таких методов совместной работы, для которых не существует (и, по всей видимости, не может существовать) офлайн способов оценки, например, в силу того, что они опираются на некоторые психологические особенности, которые достаточно сложно моделировать. Так, именно онлайн-валидация проводится в статье [17], где авторы предлагают модифицировать выходные вероятности модели, применяя к ним определенное (обучаемое) монотонное преобразование для изменения восприятия этих вероятностей человеком. Очевидно, восприятие вероятности и его эффект на поведение человека можно оценить только с помощью онлайн-валидации.

6. Заключение. В статье представлен обзор современных публикаций, касающихся совместной работы модели ИИ и эксперта-человека при решении однотипных задач, сводящихся, преимущественно, к классификации образцов, описанных тем или иным образом (представленных в виде изображений, фрагментов текста, строк таблиц). Выделены основные разновидности постановки

задачи, описаны основные подходы, лежащие в основе организации совместной работы, и принятые методы оценки алгоритмов.

Описанные в статье разработки в данной области позволяют совместить опыт экспертов (а также, возможно, сторонние данные, доступные им) и высокую производительность моделей ИИ (как правило, машинного обучения). Причем, в отличие от хорошо исследованной области «обучения с отказом», здесь допускается несовершенство и ограниченность знаний эксперта, которая, в общем случае, может быть различной в различных областях признакового пространства, что, в целом, сочетается с множеством практических сценариев. Данные подходы позволяют снизить затраты и, как правило, повысить точность решения задач по сравнению с экстремальными случаями – когда все задачи выполняются либо моделью, либо экспертом.

Тем не менее можно выделить следующие ограничения, присущие значительной части рассмотренных методов:

– Одна из основных отличительных особенностей (учет неравномерности компетентности эксперта в разных областях признакового пространства) имеет и оборотную сторону – в той или иной форме, явно или неявно, необходимо построить либо модель компетентности эксперта, для чего требуется достаточно много данных, размеченных экспертом (сотни образцов), что может быть трудновыполнимо. Может допускаться определенный компромисс, заключающийся в том, что вместо моделирования каждого эксперта рассматривается один «коллективный эксперт», что снижает индивидуальную нагрузку на человека при формировании обучающего множества (и в некоторых случаях допускает применение краудсорсинга), но и неизбежно снижает точность модели делегирования и системы в целом.

– Повышение точности системы за счет учета компетентности эксперта достигается во многом посредством специализации модели, которая, в свою очередь, способствует принесению в жертву ее общности и устойчивости. Особенно это характерно для методов, где происходит совместное обучение моделей классификации и делегирования, в ходе которого основная модель классификации «концентрируется» только на тех областях признакового пространства, в которых решение принимать будет именно она, соответственно, такая модель становится существенно менее полезной во всех других сценариях (например, формирование рекомендаций). При раздельном обучении этот эффект не так заметен, но и показатели качества итогового решения оказываются чуть ниже.

– Со специализацией связана и проблема дрейфа распределений. В частности, «подстройка» модели под особенности конкретного эксперта может привести к тому, что при смене эксперта стратегия распределения окажется неудовлетворительной.

– Во многих методах (особенно основанных на введении суррогатной функции потерь) не учитываются ограничения на загрузку экспертов.

Развитие данной области, в значительной мере, связано с преодолением перечисленных ограничений. Кроме того, сюда можно добавить и следующие направления развития:

– Перспективным методом снижения нагрузки на эксперта при обучении модели совместной работы видится применение активного обучения. Подобный подход уже предложен в статьях [18, 30, 62] и достаточно хорошо себя зарекомендовал, но требует дальнейшего развития.

– На практике зачастую отсутствует возможность совместного обучения моделей классификации и делегирования, поскольку имеется готовая модель классификации (обычно доступная только в виде «черного ящика»), обученная на большом (и не всегда доступном) наборе данных, и для этой модели необходимо найти эффективную стратегию делегирования. В этом смысле могут быть перспективными методы, основывающиеся на анализе внутренних представлений модели классификации или на ее аппроксимации, если модель классификации представляет собой «черный ящик».

– Стандартизация экспериментальных исследований посредством создания программной библиотеки, содержащей реализации основных методов и наборов данных, существенно облегчит разработку новых методов и сопоставление их с существующими.

Отдельным важным направлением исследований, находящимся на пересечении искусственного интеллекта и человеко-машинного взаимодействия, является изучение влияния, которое оказывает наличие модели ИИ и особенностей протокола совместного принятия решений на поведение эксперта [46, 71 – 75].

Литература

1. Wilder B., Horvitz E., Kamar E. Learning to Complement Humans // IJCAI'20: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence. 2020. pp. 1526–1533.
2. Madras D., Pitassi T., Zemel R. Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer // Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS 2018). 2018. pp. 6150–6160.

3. Chow C.K. On Optimum Recognition Error and Reject Tradeoff // IEEE Trans. Inf. Theory. 1970. vol. 16. no. 1. pp. 41–46.
4. Cortes C., DeSalvo G., Mohri M. Learning with rejection // Algorithmic Learning Theory (ALT 2016). Lecture Notes in Computer Science. 2016. vol. 9925. pp. 67–82.
5. Алексеев А., Носков Ф., Панов М. Непараметрическая регрессия с возможностью отказа от предсказания // ИТиС 2022. Институт проблем передачи информации им. А.А. Харкевича РАН (Москва), 2022. С. 215–226.
6. Lyons J.B., Sycara K., Lewis M., Capiola A. Human–Autonomy Teaming: Definitions, Debates, and Directions // Frontiers in Psychology. 2021. vol. 12. DOI: 10.3389/fpsyg.2021.589585.
7. Shively R.J., Lachter J., Brandt S.L., Matessa M., Battiste V., Johnson W.W. Why Human-Autonomy Teaming? // Advances in Neuroergonomics and Cognitive Engineering (AHFE 2017). Cham: Springer, 2018. vol. 586. pp. 3–11.
8. Кильдеева С., Катасёв А., Талипов Н. Модели и методы прогнозирования и распределения заданий по исполнителям в системах электронного документооборота // Вестник Технологического университета. 2021. Т. 24. № 1. С. 79–85.
9. Hendrickx K., Perini L., Van der Plas D., Meert W., Davis J. Machine learning with a reject option: a survey // Machine Learning. 2024. vol. 113. no. 5. pp. 3073–3110.
10. Leitão D., Saleiro P. Human-AI Collaboration in Decision-Making: Beyond Learning to Defer // Workshop on Human-Machine Collaboration and Teaming, ICML. 2022.
11. Zahedi Z., Kambhampati S. Human-AI Symbiosis: A Survey of Current Approaches. arXiv preprint arXiv:2103.09990. 2021. DOI: 10.48550/arXiv.2103.09990.
12. Kitchenham B., Charters S. Guidelines for performing Systematic Literature Reviews in Software Engineering. Keele, Staffs: Kitchenham, 2007. 65 p.
13. Snyder H. Literature review as a research methodology: An overview and guidelines // Journal of business research. 2019. vol. 104. pp. 333–339.
14. Mozannar H., Sontag D. Consistent estimators for learning to defer to an expert // 37th International Conference on Machine Learning. 2020. pp. 7076–7087.
15. Raghu M., Blumer K., Corrado G., Kleinberg J., Obermeyer Z., Mullainathan S. The Algorithmic Automation Problem: Prediction, Triage, and Human Effort. arXiv preprint arXiv:1903.12220. 2019.
16. Ma S., Le Y., Wang X., Zheng C., Shi C., Yin M., Ma X. Who Should I Trust: AI or Myself? Leveraging Human and AI Correctness Likelihood to Promote Appropriate Trust in AI-Assisted Decision-Making // Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. New York, USA: ACM, 2023. pp. 1–19. DOI: 10.1145/3544548.3581058.
17. Vodrahalli K., Gerstenberg T., Zou J. Uncalibrated Models Can Improve Human-AI Collaboration // Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS 2022). 2022. vol. 35. pp. 4004–4016.
18. Charusaie M.-A., Mozannar H., Sontag D., Samadi S. Sample Efficient Learning of Predictors that Complement Humans // Proceedings of the 39th International Conference on Machine Learning. 2022. pp. 2972–3005.
19. Okati N., De A., Gomez-Rodriguez M. Differentiable Learning Under Triage // Advances in Neural Information Processing Systems. 2021. vol. 34. pp. 9140–9151.
20. Verma R., Nalisnick E. Calibrated Learning to Defer with One-vs-All Classifiers // Proceedings of the 39 th International Conference on Machine Learning. 2022. pp. 22184–22202.
21. Gao R., Maytal Saar-Tsechansky M., De-Arteaga M., Han L., Sun W., Kyung Lee M., Lease M.. Learning Complementary Policies for Human-AI Teams. arXiv preprint arXiv:2302.02944. 2023.

22. Hemmer P., Schellhammer S., Vössing M., Jakubik J., Satzger G. Forming Effective Human-AI Teams: Building Machine Learning Models that Complement the Capabilities of Multiple Experts // Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI-22). 2022. pp. 2478–2484. DOI: 10.24963/ijcai.2022/344.
23. Steyvers M., Tejeda H., Kerrigan G., Smyth P. Bayesian modeling of human–AI complementarity // Proceedings of the National Academy of Sciences (Proceedings of the National Academy of Sciences of the United States of America). 2022. vol. 119. no. 11. DOI: 10.1073/pnas.2111547119.
24. Lemmer S.J., Corso J.J. Evaluating and Improving Interactions with Hazy Oracles // Proceedings of the AAAI Conference on Artificial Intelligence. 2023. vol. 37. no. 5. pp. 6039–6047.
25. Alves J.V., Leitão D., Jesus S., Sampaio M., Saleiro P., Figueiredo M., Bizarro P. FiFAR: A Fraud Detection Dataset for Learning to Defer. arXiv preprint arXiv:2312.13218. 2023.
26. Straitouri E., Adish Singla A., Balazadeh Meresht V., Gomez-Rodriguez M. Reinforcement Learning Under Algorithmic Triage. arXiv preprint arXiv:2109.11328. 2021.
27. Verma R., Barrejón D., Nalisnick E. Learning to Defer to Multiple Experts: Consistent Surrogate Losses, Confidence Calibration, and Conformal Ensembles // Proceedings of The 26th International Conference on Artificial Intelligence and Statistics. 2023. pp. 11415–11434.
28. De A., Okati N., Zarezade A., Gomez Rodriguez M. Classification Under Human Assistance // The 35th AAAI Conference on Artificial Intelligence (AAAI-21). 2021. vol. 35(7). pp. 5905–5913.
29. Liu D.-X., Mu X., Qian C. Human Assisted Learning by Evolutionary Multi-Objective Optimization // Proceedings of the AAAI Conference on Artificial Intelligence. 2023. vol. 37. no. 10. pp. 12453–12461.
30. Showalter S., Boyd A., Smyth P., Steyvers M. Bayesian Online Learning for Consensus Prediction // Proceedings of The 27th International Conference on Artificial Intelligence and Statistics. 2024. vol. 238. pp. 2539–2547.
31. Keswani V., Lease M., Kenthapadi K. Towards Unbiased and Accurate Deferral to Multiple Experts // AIES 2021 – Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, USA: ACM, 2021. pp. 154–165.
32. Mao A. et al. Two-Stage Learning to Defer with Multiple Experts // NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems. 2023. pp. 3578–3606.
33. Mao A., Mohri M., Zhong Y. Principled Approaches for Learning to Defer with Multiple Experts // International Symposium on Artificial Intelligence and Mathematics (ISAIM 2024). 2024. pp. 107–135.
34. Noti G., Chen Y. Learning When to Advise Human Decision Makers // Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence. California: International Joint Conferences on Artificial Intelligence Organization, 2023. pp. 3038–3048.
35. De A., Koley P., Ganguly N., Gomez-Rodriguez M. Regression under human assistance // Proceedings of the 34th AAAI Conference on Artificial Intelligence. 2020. pp. 2611–2620.
36. Kobayashi M., Wakabayashi K., Morishima A. Human+AI Crowd Task Assignment Considering Result Quality Requirements // Proceedings of the AAAI Conf. Hum. Comput. Crowdsourcing. 2021. vol. 9. pp. 97–107.
37. Lai V., Carton S., Bhatnagar R., Liao Q.V., Zhang Y., Tan C. Human-AI Collaboration via Conditional Delegation: A Case Study of Content Moderation //

Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. 2022. pp. 1–18. DOI: 10.1145/3491102.3501999.
38. Gao R., Saar-Tsechansky M., De-Arteaga M., Han L., Lee M.K., Lease M. Human-AI Collaboration with Bandit Feedback // Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI 2021). 2021. pp. 1722–1728.
39. Narasimhan H., Jitkrittum W., Menon A.K., Rawat A., Kumar S.. Post-hoc Estimators for Learning to Defer to an Expert // Advances in Neural Information Processing Systems. 2022. vol. 35. pp. 29292–29304.
40. Popat R., Ive J. Embracing the uncertainty in human–machine collaboration to support clinical decision-making for mental health conditions // Frontiers in Digital Health. 2023. vol. 5. DOI: 10.3389/fdgth.2023.1188338.
41. Zhang Z., Wells K., Carneiro G. Learning to Complement with Multiple Humans (LECOMH): Integrating Multi-rater and Noisy-Label Learning into Human-AI Collaboration. arXiv preprint arXiv:2311.13172. 2023.
42. Straitouri E., Wang L., Okati N., Gomez Rodriguez M. Improving Expert Predictions with Conformal Prediction // Proceedings of the 40th International Conference on Machine Learning. 2023. pp. 32633–32653.
43. Gao R., Yin M. Confounding-Robust Policy Improvement with Human-AI Teams. arXiv preprint arXiv:2310.08824. 2023.
44. Kerrigan G., Smyth P., Steyvers M. Combining Human Predictions with Model Probabilities via Confusion Matrices and Calibration // Advances in Neural Information Processing Systems. 2021. vol. 34. pp. 4421–4434.
45. Raman N., Yee M. Improving Learning-to-Defer Algorithms Through Fine-Tuning // 1st Workshop on Human and Machine Decisions (WHMD 2021) at NeurIPS. 2021. 6 p.
46. Hemmer P., Westphal M., Schemmer M., Vetter S., Vossing M., Satzger G. Human-AI Collaboration: The Effect of AI Delegation on Human Task Performance and Task Satisfaction // Proceedings of the 28th International Conference on Intelligent User Interfaces. New York, NY, USA: ACM, 2023. pp. 453–463.
47. Gupta S. et al. Take Expert Advice Judiciously: Combining Groupwise Calibrated Model Probabilities with Expert Predictions // ECAI 2023. Front. Artif. Intell. Appl. 2023. vol. 372. pp. 956–963.
48. Babbar V., Bhatt U., Weller A. On the Utility of Prediction Sets in Human-AI Teams // Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence. California: International Joint Conferences on Artificial Intelligence Organization, 2022. pp. 2457–2463.
49. Mozannar H., Satyanarayan A., Sontag D. Teaching Humans When To Defer to a Classifier via Exemplars // Proceedings of the 36th AAAI Conf. Artif. Intell (AAAI 2022). 2022. vol. 36(5). pp. 5323–5331.
50. Singh S., Jain S., Jha S.S. On Subset Selection of Multiple Humans To Improve Human-AI Team Accuracy // Proceedings of the e 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023). 2023. pp. 317–325.
51. Bansal G., Nushi B., Kamar E., Horvitz E., Weld D.S. Is the Most Accurate AI the Best Teammate? Optimizing AI for Teamwork // Proceedings of the AAAI Conference on Artificial Intelligence. 2021. vol. 35(13). pp. 11405–11414.
52. Mozannar H., Lang H., Wei D., Sattigeri P., Das S., Sontag D. Who Should Predict? Exact Algorithms For Learning to Defer to Humans // Proceedings of the The 26th International Conference on Artificial Intelligence and Statistics (PLMR 2023). 2023. vol. 206. pp. 10520–10545.
53. Joshi S., Parbhoo S., Doshi-Velez F. Learning-to-defer for sequential medical decision-making under uncertainty. Trans. Mach. Learn. Res. 2021. vol. 2023.

54. Cordelia L.P., De Stefano S., Tortorella F., Vento M. A Method for Improving Classification Reliability of Multilayer Perceptrons // IEEE Trans. Neural Networks. 1995. vol. 6. pp. 1140–1147.
55. De Stefano C., Sansone C., Vento M. To reject or not to reject: that is the question – an answer in case of neural classifiers // IEEE Transactions on Systems, Man, and Cybernetics, Part C. 2000. vol. 30. pp. 84–94.
56. Gal Y., Ghahramani Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning // Proceedings of the 33rd International Conference on International Conference on Machine Learning (ICML 2016). 2016. vol. 48. pp. 1050–1059.
57. Geifman Y., El-Yaniv R. Selective classification for deep neural networks // Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems. 2017. pp. 4878–4887.
58. Lakshminarayanan B., Pritzel A., Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles // Adv. Neural Inf. Process. Syst. 2017. vol. 30. pp. 6403–6414.
59. Raghu M., Blumer K., Sayres R., Obermeyer Z., Kleinberg R., Mullainathan S., Kleinberg J. Direct Uncertainty Prediction with Applications to Healthcare. 2018. pp. 1–14.
60. Platt J.C. Using analytic QP and sparseness to speed training of support vector machines // Advances in neural information processing systems. 1999. pp. 557–563.
61. Cohn D., Atlas L., Ladner R. Improving Generalization with Active Learning // Mach. Learn. 1994. vol. 15. no. 2. pp. 201–221.
62. Hemmer P., Thede Д., Vössing M., Jakubik J., Kühl N. Learning to Defer with Limited Expert Predictions // Proceedings of the 37th AAAI Conf. Artif. Intell. AAAI 2023. 2023. vol. 37. pp. 6002–6011.
63. Goh H.W., Tkachenko U., Mueller J. CROWDLAB: Supervised learning to infer consensus labels and quality scores for data with multiple annotators // arXiv preprint arXiv:2210.06812. 2022.
64. Xiao R., Dong Y., Wang H., Feng L., Wu R., Chen G., Zhao J. ProMix: Combating Label Noise via Maximizing Clean Sample Utility // Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI). 2023. vol. 2023-Augus. pp. 4442–4450.
65. Garg A., Nguyen C., Felix R., Do T.-T., Carneiro G. Instance-Dependent Noisy Label Learning via Graphical Modelling // Proceedings of the 2023 IEEE Winter Conf. Appl. Comput. Vision (WACV 2023). 2023. pp. 2287–2297.
66. Peterson J., Battleday R., Griffiths T., Russakovsky O. Human uncertainty makes classification more robust // Proceedings of the IEEE Int. Conf. Comput. Vis. 2019. pp. 9616–9625. DOI: 10.1109/ICCV.2019.00971.
67. Lintott C.J., Schawinski K., Slosar A., Land K., Bamford S., Thomas D., Raddick D., Nichol R.C., Szalay A.S., Andreescu D., Murray P., Vandenberg J. Galaxy Zoo: Morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey // Monthly Notices of the Royal Astronomical Society. 2008. vol. 389. no. 3. pp. 1179–1189.
68. Kamar E., Hacker S., Horvitz E. Combining human and machine intelligence in large-scale crowdsourcing // Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2012). 2012. vol. 1. pp. 467–474.
69. Majkowska A., Mittal S., Steiner D.F., Reicher J.J., McKinney S.M., Duggan G.E., Eswaran K., Cameron Chen P.-H., Liu Y., Raju Kalidindi S., Ding A., Corrado G.S., Tse D., Shetty S. Chest radiograph interpretation with deep learning models:

Assessment with radiologist-adjudicated reference standards and population-adjusted evaluation // Radiology. 2020. vol. 294. no. 2. pp. 421–431.
70. Wang X., Peng Y., Lu L., Lu Z., Bagheri M., Summers R. ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases // Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017. pp. 3462–3471.
71. Salehi P., Chiou E., Mancenido M., Mosallanezhad A., Cohen M., Shah A. Decision Deferral in a Human-AI Joint Face-Matching Task: Effects on Human Performance and Trust // Proceedings of the Human Factors and Ergonomics Society. 2021. vol. 65. no. 1. pp. 638–642.
72. Bondi E., Koster R., Sheahan H., Chadwick M., Bachrach Y., Cemgil T., Paquet U., Dvijotham K.. Role of Human-AI Interaction in Selective Prediction // Proc. 36th AAAI Conf. Artif. Intell. AAAI 2022. 2022. vol. 36. pp. 5286–5294.
73. Collins K., Barker M., Espinosa Zarlenga M., Raman N., Bhatt U., Jamnik M., Sucholutsky I., Weller A., Dvijotham K. Human Uncertainty in Concept-Based AI Systems // AIES 2023: Proc. of the AAAI/ACM Conf. on AI, Ethics, and Society. 2023. pp. 869–889.
74. Donahue K., Gollapudi S., Kollias K. When Are Two Lists Better Than One?: Benefits and Harms in Joint Decision-Making // Proceedings of the AAAI Conf. Artif. Intell. 2024. vol. 38. no. 9. pp. 10030–10038.
75. Spitzer P., Holstein J., Hemmer P., Vössing M., Kühl N., Martin D., Satzger G. On the Effect of Contextual Information on Human Delegation Behavior in Human-AI collaboration. arXiv preprint arXiv:2401.04729. 2024.

Пономарев Андрей Васильевич — канд. техн. наук, доцент, старший научный сотрудник, лаборатория интегрированных систем автоматизации, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук (СПб ФИЦ РАН). Область научных интересов: коллективный интеллект, крауд-вычисления, рекомендательные системы, машинное обучение. Число научных публикаций — 100. ponomarev@iias.spb.su; 14-я линия В.О., 39, 199178, Санкт-Петербург, Россия; р.т.: +7(812)328-8071.

Агафонов Антон Александрович — младший научный сотрудник, лаборатория интегрированных систем автоматизации, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук (СПб ФИЦ РАН). Область научных интересов: объяснимый искусственный интеллект, человеко-машинное взаимодействие, прикладное машинное обучение. Число научных публикаций — 9. agafonov.a@spcras.ru; 14-я линия В.О., 39, 199178, Санкт-Петербург, Россия; р.т.: +7(812)328-8071.

Поддержка исследований. Работа выполнена при финансовой поддержке РНФ (проект № 24-21-00337).

DOI 10.15622/ia.24.1.9

A. PONOMAREV, A. AGAFONOV

ANALYTICAL REVIEW OF TASK ALLOCATION METHODS FOR HUMAN AND AI MODEL COLLABORATION

Ponomarev A., Agafonov A. **Analytical Review of Task Allocation Methods for Human and AI Model Collaboration.**

Abstract. In many practical scenarios, decision-making by an AI model alone is undesirable or even impossible, and the use of an AI model is only part of a complex decision-making process that includes a human expert. Nevertheless, this fact is often overlooked when creating and training AI models – the model is trained to make decisions independently, which is not always optimal. The paper presents a review of methods that allow taking into account the joint work of AI and a human expert in the process of designing (in particular, training) AI systems, which more accurately corresponds to the practical application of the model, allows to increase the accuracy of decisions made by the system "human – AI model", as well as to explicitly control other important parameters of the system (e.g., human workload). The review includes an analysis of the current literature on a given topic in the following main areas: 1) scenarios of interaction between a human and an AI model and formal problem statements for improving the efficiency of the "human – AI model" system; 2) methods for ensuring the efficient operation of the "human – AI model" system; 3) ways to assess the quality of human-model AI collaboration. Conclusions are drawn regarding the advantages, disadvantages, and conditions of applicability of the methods, as well as the main problems of existing approaches are identified. The review can be useful for a wide range of researchers and specialists involved in the application of AI for decision support.

Keywords: artificial intelligence, responsible AI, decision support, human-computer interaction, human expert, task allocation, human-AI collaboration, model uncertainty, neural networks, classifier, learning with rejection, learning to defer.

References

1. Wilder B., Horvitz E., Kamar E. Learning to Complement Humans. IJCAI'20: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence. 2020. pp. 1526–1533.
2. Madras D., Pitassi T., Zemel R. Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer. Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS 2018). 2018. pp. 6150–6160.
3. Chow C.K. On Optimum Recognition Error and Reject Tradeoff. IEEE Trans. Inf. Theory. 1970. vol. 16. no. 1. pp. 41–46.
4. Cortes C., DeSalvo G., Mohri M. Learning with rejection. Algorithmic Learning Theory (ALT 2016). Lecture Notes in Computer Science. 2016. vol. 9925. pp. 67–82.
5. Alekseev A., Noskov F., Panov M. Neparametricheskaja regressija s vozmozhnost'ju otkaza ot predskazanija [Non-parametric regression with reject option]. Sbornik trudov 46-j mezhdisciplinarnoj shkoly-konferencii IPPI RAN "Informacionnye tekhnologii i sistemy 2022" [Proceedings of the 46th international conference "Information technologies and systems" of IITP RAS]. 2022. pp. 215–226. (In Russ.).
6. Lyons J.B., Sycara K., Lewis M., Capiola A. Human–Autonomy Teaming: Definitions, Debates, and Directions. Frontiers in Psychology. 2021. vol. 12. DOI: 10.3389/fpsyg.2021.589585.

7. Shively R.J., Lachter J., Brandt S.L., Matessa M., Battiste V., Johnson W.W. Why Human-Autonomy Teaming? Advances in Neuroergonomics and Cognitive Engineering (AHFE 2017). Cham: Springer, 2018. vol. 586. pp. 3–11.
8. Kildeeva S., Katasev A., Talipov N. [Models and methods of forecasting and task assignment in electronic document management]. Vestnik Tekhnologicheskogo universiteta – Herald of technolgical university. 2021. vol. 24. no. 1. pp. 79–85. (In Russ.).
9. Hendrickx K., Perini L., Van der Plas D., Meert W., Davis J. Machine learning with a reject option: a survey. Machine Learning. 2024. vol. 113. no. 5. pp. 3073–3110.
10. Leitão D., Saleiro P. Human-AI Collaboration in Decision-Making: Beyond Learning to Defer. Workshop on Human-Machine Collaboration and Teaming, ICML. 2022.
11. Zahedi Z., Kambhampati S. Human-AI Symbiosis: A Survey of Current Approaches. arXiv preprint arXiv:2103.09990. 2021. DOI: 10.48550/arXiv.2103.09990.
12. Kitchenham B., Charters S. Guidelines for performing Systematic Literature Reviews in Software Engineering. Keele, Staffs: Kitchenham, 2007. 65 p.
13. Snyder H. Literature review as a research methodology: An overview and guidelines. Journal of business research. 2019. vol. 104. pp. 333–339.
14. Mozannar H., Sontag D. Consistent estimators for learning to defer to an expert. 37th International Conference on Machine Learning. 2020. pp. 7076–7087.
15. Raghu M., Blumer K., Corrado G., Kleinberg J., Obermeyer Z., Mullainathan S. The Algorithmic Automation Problem: Prediction, Triage, and Human Effort. arXiv preprint arXiv:1903.12220. 2019.
16. Ma S., Le Y., Wang X., Zheng C., Shi C., Yin M., Ma X. Who Should I Trust: AI or Myself? Leveraging Human and AI Correctness Likelihood to Promote Appropriate Trust in AI-Assisted Decision-Making. Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. New York, USA: ACM, 2023. pp. 1–19. DOI: 10.1145/3544548.3581058.
17. Vodrahalli K., Gerstenberg T., Zou J. Uncalibrated Models Can Improve Human-AI Collaboration. Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS 2022). 2022. vol. 35. pp. 4004–4016.
18. Charusaie M.-A., Mozannar H., Sontag D., Samadi S. Sample Efficient Learning of Predictors that Complement Humans. Proceedings of the 39th International Conference on Machine Learning. 2022. pp. 2972–3005.
19. Okati N., De A., Gomez-Rodriguez M. Differentiable Learning Under Triage. Advances in Neural Information Processing Systems. 2021. vol. 34. pp. 9140–9151.
20. Verma R., Nalisnick E. Calibrated Learning to Defer with One-vs-All Classifiers. Proceedings of the 39 th International Conference on Machine Learning. 2022. pp. 22184–22202.
21. Gao R., Maytal Saar-Tsechansky M., De-Arteaga M., Han L., Sun W., Kyung Lee M., Lease M.. Learning Complementary Policies for Human-AI Teams. arXiv preprint arXiv:2302.02944. 2023.
22. Hemmer P., Schellhammer S., Vössing M., Jakubik J., Satzger G. Forming Effective Human-AI Teams: Building Machine Learning Models that Complement the Capabilities of Multiple Experts. Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI-22). 2022. pp. 2478–2484. DOI: 10.24963/ijcai.2022/344.
23. Steyvers M., Tejeda H., Kerrigan G., Smyth P. Bayesian modeling of human–AI complementarity. Proceedings of the National Academy of Sciences (Proceedings of the National Academy of Sciences of the United States of America). 2022. vol. 119. no. 11. DOI: 10.1073/pnas.2111547119.

24. Lemmer S.J., Corso J.J. Evaluating and Improving Interactions with Hazy Oracles. Proceedings of the AAAI Conference on Artificial Intelligence. 2023. vol. 37. no. 5. pp. 6039–6047.
25. Alves J.V., Leitão D., Jesus S., Sampaio M., Saleiro P., Figueiredo M., Bizarro P. FiFAR: A Fraud Detection Dataset for Learning to Defer. arXiv preprint arXiv:2312.13218. 2023.
26. Straitouri E., Adish Singla A., Balazadeh Meresht V., Gomez-Rodriguez M. Reinforcement Learning Under Algorithmic Triage. arXiv preprint arXiv:2109.11328. 2021.
27. Verma R., Barrejón D., Nalisnick E. Learning to Defer to Multiple Experts: Consistent Surrogate Losses, Confidence Calibration, and Conformal Ensembles. Proceedings of The 26th International Conference on Artificial Intelligence and Statistics. 2023. pp. 11415–11434.
28. De A., Okati N., Zarezade A., Gomez Rodriguez M. Classification Under Human Assistance. The 35th AAAI Conference on Artificial Intelligence (AAAI-21). 2021. vol. 35(7). pp. 5905–5913.
29. Liu D.-X., Mu X., Qian C. Human Assisted Learning by Evolutionary Multi-Objective Optimization. Proceedings of the AAAI Conference on Artificial Intelligence. 2023. vol. 37. no. 10. pp. 12453–12461.
30. Showalter S., Boyd A., Smyth P., Steyvers M. Bayesian Online Learning for Consensus Prediction. Proceedings of The 27th International Conference on Artificial Intelligence and Statistics. 2024. vol. 238. pp. 2539–2547.
31. Keswani V., Lease M., Kenthapadi K. Towards Unbiased and Accurate Deferral to Multiple Experts. AIES 2021 – Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, USA: ACM, 2021. pp. 154–165.
32. Mao A. et al. Two-Stage Learning to Defer with Multiple Experts. NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems. 2023. pp. 3578–3606.
33. Mao A., Mohri M., Zhong Y. Principled Approaches for Learning to Defer with Multiple Experts. International Symposium on Artificial Intelligence and Mathematics (ISAIM 2024). 2024. pp. 107–135.
34. Noti G., Chen Y. Learning When to Advise Human Decision Makers. Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence. California: International Joint Conferences on Artificial Intelligence Organization, 2023. pp. 3038–3048.
35. De A., Koley P., Ganguly N., Gomez-Rodriguez M. Regression under human assistance. Proceedings of the 34th AAAI Conference on Artificial Intelligence. 2020. pp. 2611–2620.
36. Kobayashi M., Wakabayashi K., Morishima A. Human+AI Crowd Task Assignment Considering Result Quality Requirements. Proceedings of the AAAI Conf. Hum. Comput. Crowdsourcing. 2021. vol. 9. pp. 97–107.
37. Lai V., Carton S., Bhatnagar R., Liao Q.V., Zhang Y., Tan C. Human-AI Collaboration via Conditional Delegation: A Case Study of Content Moderation. Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. 2022. pp. 1–18. DOI: 10.1145/3491102.3501999.
38. Gao R., Saar-Tsechansky M., De-Arteaga M., Han L., Lee M.K., Lease M. Human-AI Collaboration with Bandit Feedback. Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI 2021). 2021. pp. 1722–1728.
39. Narasimhan H., Jitkrittum W., Menon A.K., Rawat A., Kumar S.. Post-hoc Estimators for Learning to Defer to an Expert. Advances in Neural Information Processing Systems. 2022. vol. 35. pp. 29292–29304.

40. Popat R., Ive J. Embracing the uncertainty in human–machine collaboration to support clinical decision-making for mental health conditions. Frontiers in Digital Health. 2023. vol. 5. DOI: 10.3389/fdgth.2023.1188338.
41. Zhang Z., Wells K., Carneiro G. Learning to Complement with Multiple Humans (LECOMH): Integrating Multi-rater and Noisy-Label Learning into Human-AI Collaboration. arXiv preprint arXiv:2311.13172. 2023.
42. Straitouri E., Wang L., Okati N., Gomez Rodriguez M. Improving Expert Predictions with Conformal Prediction. Proceedings of the 40th International Conference on Machine Learning. 2023. pp. 32633–32653.
43. Gao R., Yin M. Confounding-Robust Policy Improvement with Human-AI Teams. arXiv preprint arXiv:2310.08824. 2023.
44. Kerrigan G., Smyth P., Steyvers M. Combining Human Predictions with Model Probabilities via Confusion Matrices and Calibration. Advances in Neural Information Processing Systems. 2021. vol. 34. pp. 4421–4434.
45. Raman N., Yee M. Improving Learning-to-Defer Algorithms Through Fine-Tuning. 1st Workshop on Human and Machine Decisions (WHMD 2021) at NeurIPS. 2021. 6 p.
46. Hemmer P., Westphal M., Schemmer M., Vetter S., Vossing M., Satzger G. Human-AI Collaboration: The Effect of AI Delegation on Human Task Performance and Task Satisfaction. Proceedings of the 28th International Conference on Intelligent User Interfaces. New York, NY, USA: ACM, 2023. pp. 453–463.
47. Gupta S. et al. Take Expert Advice Judiciously: Combining Groupwise Calibrated Model Probabilities with Expert Predictions. ECAI 2023. Front. Artif. Intell. Appl. 2023. vol. 372. pp. 956–963.
48. Babbar V., Bhatt U., Weller A. On the Utility of Prediction Sets in Human-AI Teams. Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence. California: International Joint Conferences on Artificial Intelligence Organization, 2022. pp. 2457–2463.
49. Mozannar H., Satyanarayan A., Sontag D. Teaching Humans When To Defer to a Classifier via Exemplars. Proceedings of the 36th AAAI Conf. Artif. Intell (AAAI 2022). 2022. vol. 36(5). pp. 5323–5331.
50. Singh S., Jain S., Jha S.S. On Subset Selection of Multiple Humans To Improve Human-AI Team Accuracy. Proceedings of the e 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023). 2023. pp. 317–325.
51. Bansal G., Nushi B., Kamar E., Horvitz E., Weld D.S. Is the Most Accurate AI the Best Teammate? Optimizing AI for Teamwork. Proceedings of the AAAI Conference on Artificial Intelligence. 2021. vol. 35(13). pp. 11405–11414.
52. Mozannar H., Lang H., Wei D., Sattigeri P., Das S., Sontag D. Who Should Predict? Exact Algorithms For Learning to Defer to Humans. Proceedings of the The 26th International Conference on Artificial Intelligence and Statistics (PLMR 2023). 2023. vol. 206. pp. 10520–10545.
53. Joshi S., Parbhoo S., Doshi-Velez F. Learning-to-defer for sequential medical decision-making under uncertainty. Trans. Mach. Learn. Res. 2021. vol. 2023.
54. Cordelia L.P., De Stefano S., Tortorella F., Vento M. A Method for Improving Classification Reliability of Multilayer Perceptrons. IEEE Trans. Neural Networks. 1995. vol. 6. pp. 1140–1147.
55. De Stefano C., Sansone C., Vento M. To reject or not to reject: that is the question – an answer in case of neural classifiers. IEEE Transactions on Systems, Man, and Cybernetics, Part C. 2000. vol. 30. pp. 84–94.
56. Gal Y., Ghahramani Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. Proceedings of the 33rd International Conference on

International Conference on Machine Learning (ICML 2016). 2016. vol. 48. pp. 1050–1059.
57. Geifman Y., El-Yaniv R. Selective classification for deep neural networks. Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems. 2017. pp. 4878–4887.
58. Lakshminarayanan B., Pritzel A., Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. Adv. Neural Inf. Process. Syst. 2017. vol. 30. pp. 6403–6414.
59. Raghu M., Blumer K., Sayres R., Obermeyer Z., Kleinberg R., Mullainathan S., Kleinberg J. Direct Uncertainty Prediction with Applications to Healthcare. 2018. pp. 1–14.
60. Platt J.C. Using analytic QP and sparseness to speed training of support vector machines. Advances in neural information processing systems. 1999. pp. 557–563.
61. Cohn D., Atlas L., Ladner R. Improving Generalization with Active Learning. Mach. Learn. 1994. vol. 15. no. 2. pp. 201–221.
62. Hemmer P., Thede Д., Vössing M., Jakubik J., Kühl N. Learning to Defer with Limited Expert Predictions. Proceedings of the 37th AAAI Conf. Artif. Intell. AAAI 2023. 2023. vol. 37. pp. 6002–6011.
63. Goh H.W., Tkachenko U., Mueller J. CROWDLAB: Supervised learning to infer consensus labels and quality scores for data with multiple annotators. arXiv preprint arXiv:2210.06812. 2022.
64. Xiao R., Dong Y., Wang H., Feng L., Wu R., Chen G., Zhao J. ProMix: Combating Label Noise via Maximizing Clean Sample Utility. Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI). 2023. vol. 2023-Augus. pp. 4442–4450.
65. Garg A., Nguyen C., Felix R., Do T.-T., Carneiro G. Instance-Dependent Noisy Label Learning via Graphical Modelling. Proceedings of the 2023 IEEE Winter Conf. Appl. Comput. Vision (WACV 2023). 2023. pp. 2287–2297.
66. Peterson J., Battleday R., Griffiths T., Russakovsky O. Human uncertainty makes classification more robust. Proceedings of the IEEE Int. Conf. Comput. Vis. 2019. pp. 9616–9625. DOI: 10.1109/ICCV.2019.00971.
67. Lintott C.J., Schawinski K., Slosar A., Land K., Bamford S., Thomas D., Raddick D., Nichol R.C., Szalay A.S., Andreescu D., Murray P., Vandenberg J. Galaxy Zoo: Morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. Monthly Notices of the Royal Astronomical Society. 2008. vol. 389. no. 3. pp. 1179–1189.
68. Kamar E., Hacker S., Horvitz E. Combining human and machine intelligence in large-scale crowdsourcing. Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2012). 2012. vol. 1. pp. 467–474.
69. Majkowska A., Mittal S., Steiner D.F., Reicher J.J., McKinney S.M., Duggan G.E., Eswaran K., Cameron Chen P.-H., Liu Y., Raju Kalidindi S., Ding A., Corrado G.S., Tse D., Shetty S. Chest radiograph interpretation with deep learning models: Assessment with radiologist-adjudicated reference standards and population-adjusted evaluation. Radiology. 2020. vol. 294. no. 2. pp. 421–431.
70. Wang X., Peng Y., Lu L., Lu Z., Bagheri M., Summers R. ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017. pp. 3462–3471.
71. Salehi P., Chiou E., Mancenido M., Mosallanezhad A., Cohen M., Shah A. Decision Deferral in a Human-AI Joint Face-Matching Task: Effects on Human Performance

and Trust. Proceedings of the Human Factors and Ergonomics Society. 2021. vol. 65. no. 1. pp. 638–642.
72. Bondi E., Koster R., Sheahan H., Chadwick M., Bachrach Y., Cemgil T., Paquet U., Dvijotham K.. Role of Human-AI Interaction in Selective Prediction. Proc. 36th AAAI Conf. Artif. Intell. AAAI 2022. 2022. vol. 36. pp. 5286–5294.
73. Collins K., Barker M., Espinosa Zarlenga M., Raman N., Bhatt U., Jamnik M., Sucholutsky I., Weller A., Dvijotham K. Human Uncertainty in Concept-Based AI Systems. AIES 2023: Proc. of the AAAI/ACM Conf. on AI, Ethics, and Society. 2023. pp. 869–889.
74. Donahue K., Gollapudi S., Kollias K. When Are Two Lists Better Than One?: Benefits and Harms in Joint Decision-Making. Proceedings of the AAAI Conf. Artif. Intell. 2024. vol. 38. no. 9. pp. 10030–10038.
75. Spitzer P., Holstein J., Hemmer P., Vössing M., Kühl N., Martin D., Satzger G. On the Effect of Contextual Information on Human Delegation Behavior in Human-AI collaboration. arXiv preprint arXiv:2401.04729. 2024.

Ponomarev Andrew — Ph.D., Associate Professor, Senior researcher, Computer-aided integrated systems laboratory, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS). Research interests: collective intelligence, crowd computing, recommender systems, applied machine learning. The number of publications — 100. ponomarev@iias.spb.su; 39, 14-th Line V.O., 199178, St. Petersburg, Russia; office phone: +7(812)328-8071.

Agafonov Anton — Junior researcher, Computer-aided integrated systems laboratory, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS). Research interests: explainable artificial intelligence, human-computer interaction, applied machine learning. The number of publications — 9. agafonov.a@spcras.ru; 39, 14-th Line V.O., 199178, St. Petersburg, Russia; office phone: +7(812)328-8071.

Acknowledgements. This research is funded by the Russian Science Foundation (grant 24-21-00337).

УДК 004.032.26 **DOI** 10.15622/ia.24.1.10

А.Н. Голубинский, А.А. Толстых, М.Ю. Толстых

АВТОМАТИЧЕСКАЯ ГЕНЕРАЦИЯ АННОТАЦИЙ НАУЧНЫХ СТАТЕЙ НА ОСНОВЕ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ

Голубинский А.Н., Толстых А.А., Толстых М.Ю. **Автоматическая генерация аннотаций научных статей на основе больших языковых моделей.**

Аннотация. Предложена концепция автоматизации процесса аннотирования научных материалов (русскоязычных научных статей) и выполнена ее практическая реализация посредством технологий машинного обучения, дообучения больших языковых моделей. Обозначена актуальность корректного и рационального составления аннотаций, выделена проблематика, касающаяся установления баланса между затратами времени на аннотирование и обеспечением соблюдения ключевых требований к аннотации. Проанализированы основы аннотирования, представленные в семействе стандартов по информации, библиотечному и издательскому делу, приведены классификация аннотаций и требования к их наполнению и функционалу. Схемографически представлено существо и содержание процесса аннотирования, типовая структура объекта исследования. Проанализирован вопрос интеграции в процесс аннотирования цифровых технологий, особое внимание уделено преимуществам внедрения машинного обучения и технологий искусственного интеллекта. Кратко описан цифровой инструментарий, применяемый для генерации текста в приложениях обработки естественного языка. Отмечены его недостатки для решения поставленной в данной научной статье задачи. В исследовательской части обоснован выбор модели машинного обучения, применяемый для решения задачи условной генерации текста. Проанализированы существующие предобученные большие языковые модели и с учетом постановки задачи и имеющихся ограничений вычислительных ресурсов выбрана модель ruT5-base. Приведено описание датасета, включающего научные статьи из журналов, включенных в перечень рецензируемых научных изданий, в которых должны быть опубликованы основные научные результаты диссертаций на соискание ученых степеней кандидата и доктора наук. Охарактеризована методика разметки данных, основанная на работе токинезатора предобученной большой языковой модели, графически и таблично приведены численные характеристики распределений датасета и параметры конвейера обучения. Для оценки модели использована метрика качества ROUGE, для оценки результатов – метод экспертных оценок, включающий грамматику и логику в качестве базовых критериев. Качество автоматической генерации аннотаций сопоставимо с реальными текстами, отвечает требованиям информативности, структурированности и компактности. Статья может представлять интерес для аудитории ученых и исследователей, стремящихся оптимизировать свою научную деятельность в части интеграции в процесс написания статей инструментов цифровизации, а также специалистам, занимающимся обучением больших языковых моделей.

Ключевые слова: аннотация, генерация, большие языковые модели, цифровизация, машинное обучение.

1. Введение.

Научные публикации являются важным источником сведений и знаний в области академических и прикладных исследований и разработок. Когда научные материалы опубликованы, первая часть, с которой начинается ознакомление читателей, после

самого названия и сведений об авторах, – это, как правило, аннотация. Она представляет собой краткое изложение статьи, которое должно передавать емкое и лаконичное сообщение, являть сжатый обзор всей статьи и излагать ее суть.

Зачастую аннотация научной работы составляется в конце ее подготовки и оформления, когда у автора сформировалось четкое представление о существе, ходе и итогах исследования, уверенность в его завершенности [1, 2]. При этом автор готов обозначить корректную характеристику темы научной работы, ее проблемы, выделить объект и предмет, цели и задачи, а также указать результаты решения обозначенной проблемы в выбранной предметной области.

Корректно написанная аннотация может служить одновременно нескольким целям: позволить читателям оперативно понять суть научного материала, чтобы решить, ознакомиться ли с ним целиком; настроить внимание респондентов к тому, чтобы следить за ходом представления сведений, анализом и аргументацией в тексте научного исследования; помочь читателям запомнить ключевые аспекты научного материала.

Процесс аннотирования является важной задачей как для авторов-исследователей, так и других потребителей научного контента. При этом можно отметить сопутствующую проблематику, заключающуюся в отсутствии универсальных методов аннотирования, субъективности автора-составителя к реализации требований релевантности и точности аннотации, трудоемкости и времязатратности самого процесса, семантической неоднозначности результатов.

Наиболее разумным в данном контексте видится способ решения указанных затруднений, заключающийся в применении гибридных методов составления аннотации: комбинирование ручного аннотирования и автоматизированных методов, базирующихся на использовании современных цифровых технологий. Цифровой инструментарий, например, в виде технологий искусственного интеллекта, может выполнять первоначальное аннотирование, которое затем будет проверяться и корректироваться автором научной публикации, что позволит снизить временные затраты на составление текста и повысить качество результата.

2. Стандартизация и основы аннотирования. В России действует ряд стандартов, устанавливающих требования к содержанию, построению и оформлению текста аннотации. Ввиду тематической ориентированности данного исследования на аннотирование именно научных статей, рассмотрим стандарты из

семейства Системы стандартов по информации, библиотечному и издательскому делу [3 – 6].

В целом, положения [3] и [4] нормативных технических документов идентичны, за исключением некоторых особенностей. Ниже приведены результаты сравнительного анализа нормативных документов (рисунок 1), а также расширенная классификация аннотаций (рисунок 2).

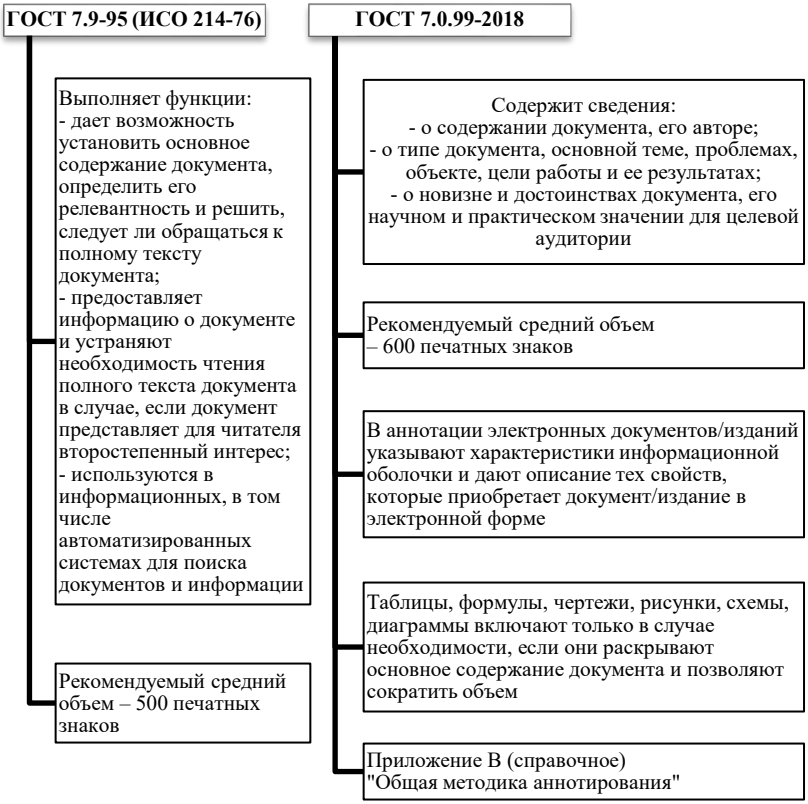


Рис. 1. Сравнительный анализ нормативных технических документов, регламентирующих построение и оформление текста аннотации (индикативного реферата) на документ (ключевые различия)

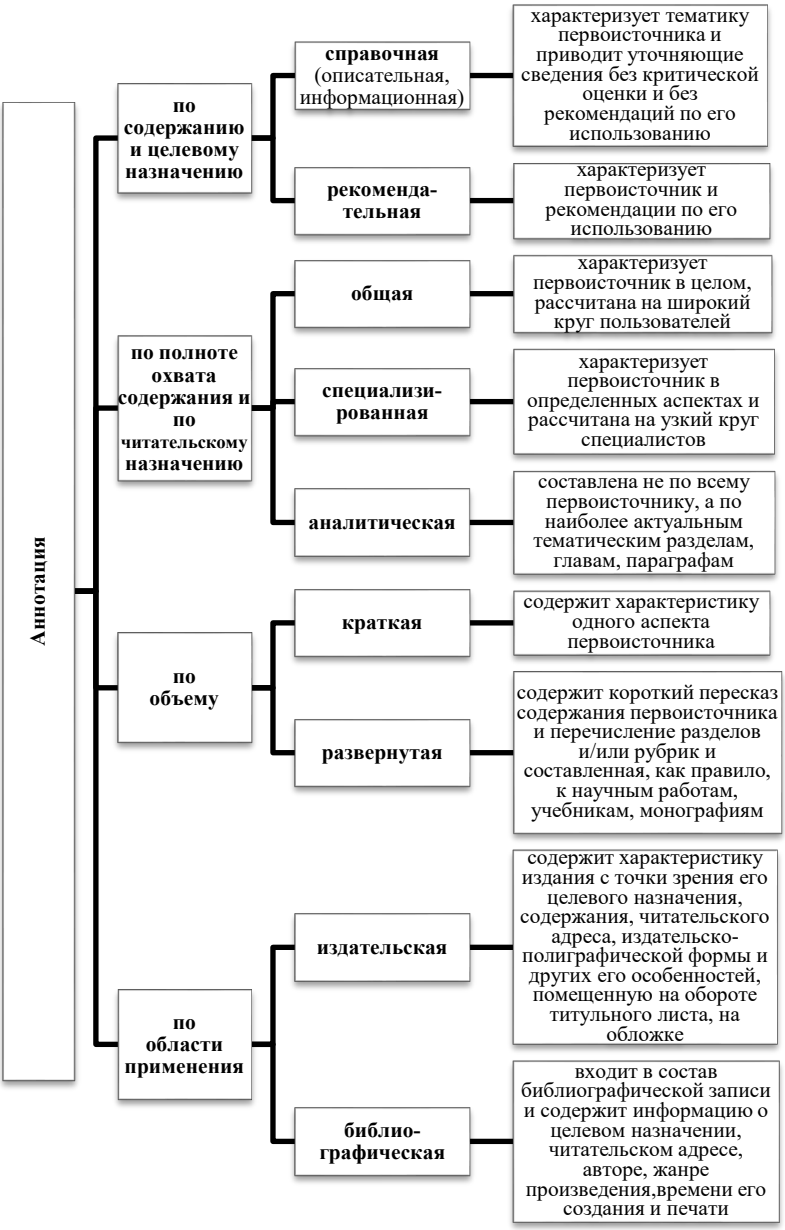


Рис. 2. Классификация и виды аннотаций по ГОСТ Р 7.0.99-2018

Под аннотацией на рисунке 2 понимается краткая характеристика первичного документа с точки зрения его назначения, содержания, вида, формы и других особенностей.

Таким образом, редакция [4] шире и детальней, кроме того, она содержит отдельное приложение, в котором раскрывается методика аннотирования, ключевые моменты которой для удобства восприятия приведены графически (рисунок 3).

Рис. 3. Основы методики аннотирования согласно требованиям нормативных технических документов

Можно также отметить, что аннотация должна представлять собой краткое отдельное резюме статьи, состоящее из нескольких предложений по каждому из следующих ключевых моментов в вопросных формах (рисунок 4).

Рис. 4. Типовая структура аннотации

Эмпирически установлены и логически обоснованы некоторые запреты в содержании аннотации. Так, например [7], не следует повторять текст самой статьи (исключить перенос предложений из основного текста научного материала), а также ее название. В тексте аннотации не должны приводиться таблицы, внутритекстовые сноски, обилие цифр. Следует избегать синтаксических конструкций, несвойственных языку научных и технических документов, нецелесообразно применять сложные грамматические конструкций, вводные слова, общие формулировки.

Стоит также отметить, что многие поисковые системы и библиографические базы данных используют аннотации вместе с заголовками научных статей для определения ключевых терминов в процессе индексации опубликованных научных трудов.

3. Интеграция цифровых технологий в процесс аннотирования. Цифровизация аннотирования представляет собой процесс применения цифровых технологий [8 – 10] для улучшения, автоматизации и ускорения процесса создания аннотаций и последующей работы с ними (рисунок 5).

Рис. 5. Основные аспекты цифровизации аннотирования

С технической стороны процесс аннотирования включает выделение ключевых элементов исходного материала (меток) и формирование фактически метаданных к тексту. Указанные процедуры могут быть автоматизированы и оптимизированы посредством использования современных передовых цифровых технологий – машинного обучения, в частности в приложениях обработки естественного языка (Natural language processing, NLP) и компьютерного зрения [11].

В динамично трансформирующемся ландшафте академических исследований инструменты на базе искусственного интеллекта совершают революцию в мастерстве написания текста. В России лидируют ChatGPT, YandexGPT2 [12], justGPT, GigaChat, которые предлагают авторам-исследователям эффективные способы свести обширные тексты научных материалов в краткие изложения, сэкономить время, улучшить качество контента и избежать плагиата. Также доступно использование приложений и расширений браузеров (например, Hypothesis, Kami), платформ для коллективной работы (например, Google Docs, Overleaf) и менеджеров/приложений для оркестрации процесса аннотирования в ручном формате (например, Zotero, Mendeley, Evernote).

В целом, функционал указанных сервисов заключается в автоматическом извлечении метаданных из добавленных текстов или файлов, настройке фильтрации для улучшения семантического поиска,

работе с библиографическими данными. Однако в своем большинстве они являются платными, не адаптированы под нюансы отечественного научного знания: высока вероятность некорректного извлечения метаданных из русских источников в отсутствии унифицированных международных идентификаторов; неполноценный перевод частей интерфейса и технической документации к сервисам усложняет их использование для русскоязычных пользователей. Более того, указанные сервисы являются закрытыми с точки зрения информации о моделях, датасетах и методике обучения, что не позволяет провести корректное сравнение с открытыми решениями.

Кроме того, применение указанного цифрового инструментария затрагивает вопросы этики научных исследований и академического мошенничества [13]. Последнее включает в себя преднамеренные попытки обмана и плагиат, фабрикацию данных, искажение исторических источников, подделку доказательств, заказ работ, выдачу чужих работ за свои, так называемую «двойную» сдачу материалов (например, одной и той же статьи в несколько редакций различных журналов), выборочное сокрытие нежелательных или неприемлемых результатов и кражу идей. Например, плагиат может появляться при генерации текстов с помощью больших языковых моделей, выдаваемых за оригинальные результаты. Выполнение работ на заказ упрощается за счёт автоматизации и сокращения времени на написание текста работы и обзора литературы. «Двойная» сдача материалов может реализовываться за счёт быстрого автоматического перефразирования словесных конструкций исходного материала без изменения семантической составляющей (гипотез, результатов экспериментов, выводов и т.д.). Научным и академическим сообществом отмечаются риски недобросовестного применения цифровых технологий в отношении развития научного знания, вместе с тем принимаются соответствующие меры реагирования в виде так называемых «карательных» (применение строгих санкций к нарушителям) и ценностных (разработка и внедрение этических кодексов, пропаганда этичного научного поведения и формирования честной академической среды). Очевидно соблюдение баланса между использованием аппарата цифровой трансформации и интеллектуальной авторской деятельностью.

Таким образом, применение передовых цифровых решений может способствовать обеспечению ясности, грамматической точности и актуальности текста, в том числе аннотации, удовлетворяя широкий спектр академических потребностей.

4. Исследовательская часть. Выбор модели обучения. Задача аннотирования научных статей является задачей условной генерации текста, т.е. создания последовательности слов (символов) на основе заданного контекста, тематики или условий [14 – 21].

Как правило, условная генерация текста может быть реализована с помощью двух базовых подходов. Первый заключается в использовании предопределенного шаблона для генерации текста на основе различных входных данных. Например, используя предопределенный шаблон, искусственный интеллект может сгенерировать определенное описание продукта на основе типа продукта, его характеристик и преимуществ. Второй способ использует метод неконтролируемого обучения, называемый глубоким обучением, который изучает нюансы языковых структур и функционирует с условием наличия больших объемов входных данных. Данный алгоритм более гибкий, может генерировать более естественный язык по сравнению с подходом на основе шаблонов.

Условная генерация текста имеет широкие практические применения в различных отраслях. К основным областям использования относятся: создание тематического контента (статьи, описания), поддержка клиентов с использованием чат-ботов; перевод (предоставление оперативных и точных интерпретаций посредством анализа входного языка и применения соответствующей языковой структуры и правил использования) и др.

В настоящее время существует несколько предобученных больших языковых моделей (large language model, LLM), предназначенных для условной генерации, в связи с чем решение поставленной в работе задачи сводится к дообучению (finetuning) одной из предобученных LLM. В работе [14] было представлено семейство LLM, предобученных на корпусе текстов, б**о**льшая часть которого являлась текстами на русском языке.

В наборе LLM [14] описаны следующие модели для условной генерации из семейства LLM T5: ruT5-base и ruT5-large. Выбор данного семейства моделей обусловлен ограниченностью вычислительных ресурсов: из открытых источников известно, что модели с большим числом параметров, например Llama 2 [15], GPT-3 [16] и подобные, показывают более высокие результаты, однако требуют гораздо больше вычислительных мощностей для обработки запросов (например, дообучение Llama 2 требует около 112 GB GPU в режиме fp32, GPT-3 – около 80 GB GPU, в то время как ruT5-base – около 18 GB GPU). Кроме того, рассмотренные модели имеют тенденцию создавать текст, который чрезмерно повторяется или не

отражает нюансы человеческого языка, так как обучены на корпусе, преимущественно, содержащего тексты на английском языке. В рамках эксперимента были доступны 16 GB GPU. В связи с этим была выбрана модель ruT5-base, содержащая 222×10^6 параметров (весов), является моделью трансформера [20] для русского языка, состоит из энкодера и декодера, решает задачу генерацию текстов, а также может быть обучена на широком списке NLP-задач.

Описание датасета и методики разметки данных. Для решения задачи аннотирования научных статей необходимо подготовить соответствующий датасет: пары «текст статьи» – «аннотация». Для уменьшения размера датасета принято решение ограничения предметной области научных статей: экономка и юриспруденция, педагогика, а также технические науки в контексте правоохранительной деятельности.

В качестве репозитория научных статей выбраны выпуски за последние 5 лет следующих журналов: Вестник Московского университета МВД России имени В.Я. Кикотя, Вестник Краснодарского университета МВД России, Вестник Воронежского института МВД России. Издания являются научно-практическими журналами, освещающими актуальные проблемы образовательного процесса, общественных, технических (информационных) и гуманитарных наук. Их корреспондентами являются как именитые ученые, так и молодые авторы: ученые деятели, преподаватели, студенты (курсанты и слушатели) высших учебных заведений, научно-педагогические кадры, практические работники и служащие правоохранительных органов, интересующиеся актуальными проблемами научного знания, участвующие в процессе обмена информацией и ведения конструктивного научного диалога.

Первоначально размечено 825 научных статей из 22 томов. После первичной обработки датасета, заключающейся в объединении статей, аннотаций и метаданных (название журнала, название статьи), исключено 15 статей (2%), аннотации к которым отсутствуют. Таким образом, исходный датасет для обучения составляет 810 пар «текст статьи» – «аннотация».

Для дальнейшего контроля хода обучения исходный датасет разбит на обучающую и валидационную выборки в соотношении 80/20 (обучающая часть – 648 пар; валидационная – 162 пары).

На рисунке 6 приведено распределение длин статей (в символах, включая пунктуационные знаки) в датасете, на рисунке 7 – распределение длин аннотаций.

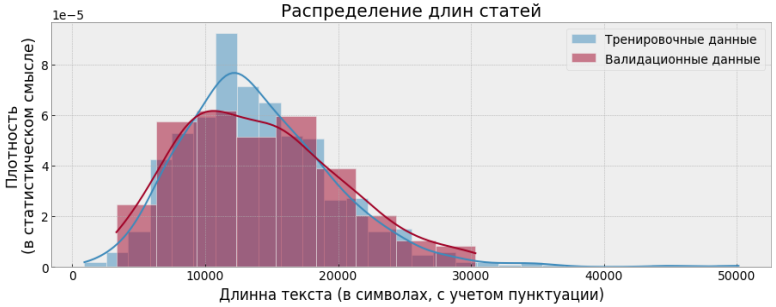


Рис. 6. Распределение длин статей в датасете

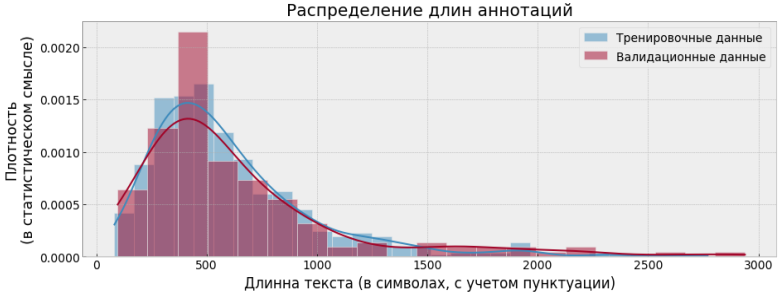


Рис. 7. Распределение длин аннотаций в датасете

Из рисунков 6 и 7 видно, что основная часть статей содержит не боле 25 000 символов, а аннотаций – 1 500 (оценка по 95% квантилю). Численные характеристики распределений (для всего датасета) приведены в таблице 1.

Таблица 1. Численные характеристики распределений датасета

Тип	**25% квантиль**	**Медиана**	**75% квантиль**	**95% квантиль**	**Математическое ожидание**
Текст	10197,5	13196,5	17356,5	24120,7	14043,6
Аннотация	349,7	510,5	741,2	1325,9	602,0

Подобное распределение обусловлено стандартизированным требованиями к размеру аннотации, о которых говорилось ранее, применяемыми научными журналами, в которых были размещены статьи. На рисунке 8 приведены наиболее часто встречающиеся слова в тексте статей, на рисунке 9 – в аннотациях.

Рис. 8. Наиболее частые слова в текстах статей

Рис. 9. Наиболее частые слова в аннотациях статей

Анализ наиболее частых слов в статьях и аннотациях к ним из датасета (рисунок 8 и 9) позволяет говорить об общей направленности текстов и совпадает с тематическими рубрикаторами исследуемых журналов.

Следующим этапом подготовки данных для обучения LLM является выбор максимального размера входных данных в токенах. Под токенами понимается результат алгоритма представления языковой сущности (слово, часть слова или отдельный символ) в виде целого числа (с добавлением нуля, $\mathbb{N} \cap \{0\}$), сам алгоритм в свою очередь называется «токенизатором» [17]. Для дообучения LLM необходимо использовать тот же токинизатор (алгоритм преобразования текста в численное представление) [17]. Токены вступают фундаментальными единицами информации, которые модели обрабатывают и производят. Эффективность модели часто

можно проследить по тому, насколько хорошо происходит преобразование токенов.

Благодаря использованию токинезатора предобученной LLM, получены распределения длин текстов и аннотаций, численные распределения которых приведены в таблице 2.

Таблица 2. Распределения длин текстов и аннотаций в токенах

Тип	Текст	Аннотация
40% квантиль	2596	103
50% квантиль	2805	103
60% квантиль	2807	103
70% квантиль	3343	103
80% квантиль	3334	118

Исходя из того, что на располагаемых вычислительных мощностях (16 GB GPU) не представляется возможным дообучить LLM со входом (текст) более 3000 токенов (требует более 16 GB GPU в режиме fp16), принято решение использовать методику ограничения длины входа (отбрасывания всех токенов после 3000). Верхняя граница длины входа LLM определяется с одной стороны ограниченностью вычислительных ресурсов, доступных для обучения, а с другой – статистическим распределением длин текстов в исходном датасете.

Для выхода модели применяется 128 токенов (недостающие токены заменяются специальным токеном <pad> [17]). Такое количество обеспечивает удовлетворительный размер аннотации научной статьи (3-4 предложения) и позволяет наиболее эффективно использовать имеющиеся данные для обучения. Под эффективностью в данном случае понимается то, что длинны аннотаций в датасете достаточно близко укладываются в 128 токенов (нет больших последовательностей <pad>), кратность степени двойки обусловлена архитектурой (следующий размер выхода – 256 токенов).

Конвейер обучения. Время обучения модели составляет 53 часа. Ее численные параметры приведены ниже:

- темп обучения (learning rate): 2×10^{-5};
- затухание весов (weights decay): 0.01;
- максимальная длинна входа (в токенах): 3000;
- максимальная длинна аннотации (в токенах): 128;
- параметры архитектуры T5 взяты без изменений из [14].

Мониторинг обучения происходит с помощью метрики ROUGE (Recall-Oriented Understudy for Gisting Evaluation). Впервые данная метрика была предложена в [18], она является специализированной метрикой для задачи автоматической аннотации текстов. ROUGE основана на измерении пересечения между выходом модели (результатом условной генерации) и целевыми аннотациями, написанными человеком. Иными словами, производится подсчет совпадений слов и словосочетаний в сгенерированном тексте и в целевом, кроме того, метрика не чувствительна к регистру. Более высокие баллы (близкие к 1) указывают на эффективность с точки зрения сохранения ключевой информации из исходного текста при создании аннотации. На рисунке 10 приведены график изменения ROUGE при обучении модели.

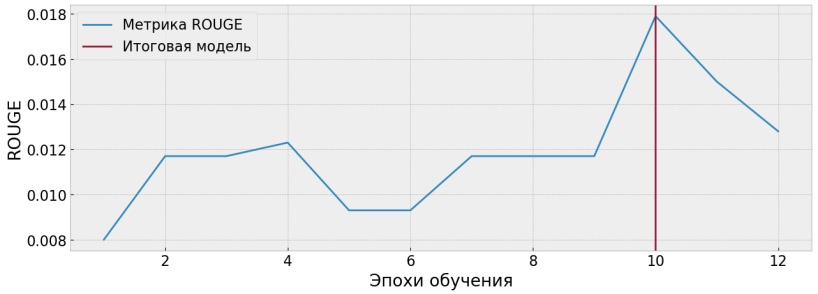


Рис. 10. Динамика дообучения модели

Алгоритм работы выбранной метрики заключается следующем. На этапе предварительной обработки сгенерированные аннотации анализируются для устранения любого шума или нерелевантной информации (например, знаков препинания, стоп-слов), которые могут помешать процессу оценки. Далее выполняется извлечение признаков, таких как n-граммы и прочие показатели сходства, которые получают как из сгенерированного системой текста, так и из исходных аннотаций, что обеспечивает основу для сравнения двух текстов. После этого с использованием различных методов, таких как статистика совместной встречаемости n-грамм, вычисление коэффициентов перекрытия слов, выполняется расчет оценок сходства путем сравнения признаков, извлеченных из сгенерированного моделью текста с признаками из исходных аннотаций.

Заключительными этапами алгоритма являются агрегация оценок сходства, нормализация и интерпретация. При агрегации отдельные оценки сходства, полученные для каждого типа признаков,

объединяются для получения единой оценки ROUGE, представляющей общую эффективность сгенерированного моделью текста аннотации. Окончательный балл ROUGE часто нормируется на отрезок [0, 1], при этом более высокие баллы указывают на более высокую эффективность модели с точки зрения сохранения ключевой информации из исходного текста.

Помимо метрики ROUGE, каждую эпоху проводится субъективная оценка на основе 5 сгенерированных пар «текст – аннотация», что позволяет дополнительно верифицировать результаты с точки зрения экспертной оценки. В таблице 3 приведена динамика данной оценки с 10 эпохи обучения, значимым являлся балл за логику (непротиворечивость и соответствие аннотации тексту исходной статьи), баллы усреднены по парам и экспертам.

Таблица 3. Динамика изменения оценки качества генерации в процессе обучения сети с 10 эпохи обучения

Эпохи	**Баллы**
10 эпоха	8,0 баллов
11 эпоха	7,8 баллов
12 эпоха	7,5 баллов

Методика оценки результатов. Оценка обучения модели – важный процесс, необходимый для подтверждения ее результативности и эффективности, качества и производительности. Профессиональное сообщество в сфере технологий искусственного интеллекта и машинного обучения не согласовало унифицированные требования к типовой методологической базе оценки моделей. В основном это связано с проблемами формализации измерений качества семантической информации.

Формальная оценка качества облегчается с помощью структурированного инструментария на основе эмпирических данных, уточненных экспертным консенсусом.

Для верификации результатов обучения модели, разработанной в данном исследовании, используется экспертная оценка [22]. В качестве субъектов экспертирования выступают обучающиеся старших курсов ведомственного вуза, у которых имеется опыт выполнения научных исследований, участия в научно-представительских мероприятиях, написания научных статей по тематикам из собранного датасета, что свидетельствует о приемлемом уровне экспертной компетентности и сопоставляется с

идентифицируемыми задачами оценки и измеримостью результатов. Оценка проводилась по валидационной выборке.

Отметим, что в данном контексте толкование дефиниции «эксперт» относится к пониманию лица как деятельного субъекта, включенного в механизмы принятия решений. Обучающиеся, обладающие правами и возможностями принятия заключений относительно вопросов экспертирования, могут не являться специалистами и профессионалами в оцениваемой области, но будут реализовывать ролевые экспертные роли согласно условиями и критериям процесса оценивания результатов обучения модели. Такой подход также реализует концепцию студентоориентированности [23], актуальную для отечественной системы образования, при которой понимание студента сводится не просто к его идентификации как штатного участника образовательного процесса, а индивидуального субъекта, продуцирующего систему рефлексии в рамках единого общественно значимого процесса воспитания и обучения в интересах человека, семьи, общества и государства.

Предлагается методика оценки эффективности модели на основе двух критериев: оценка грамматики и оценка логики. Каждой аннотации эксперты выставляют оценку по 10 бальной шкале. Под грамматикой в данном контексте понимаются любые синтаксические, грамматические и иные ошибки, позволяющие идентифицировать аннотацию как сгенерированную. Например, грамматической будет являться ошибка изменения алфавита посередине слова (кириллица / латиница), некорректное написание слов, отсутствие пробелов и т. д.

В качестве примера, в приведенной ниже аннотации (автоматически сгенерированной) полужирным шрифтом выделены грамматические ошибки:

Рассматривается личность преступника как базовый элемент криминалистической характеристики преступлений, **совершаемых террористической направленности**, с точки зрения надлежащего субъекта преступления, а также его мотиваци**ю**, целеполагани**е**.

В приведенной аннотации отсутствует согласованность в спряжении слов, ошибки в окончаниях. Данная аннотация была оценена 18 экспертами в среднем в 6,1 балл по показателю грамматика.

Второй параметр оценки – логика, показывающая семантическую корректность аннотации, а также соответствие

аннотации тексту исходной статьи. Для оценивания данного параметра в распоряжении экспертов имеются исходные тексты статей.

В оценке принимает участие 51 эксперт, каждый из которых оценивает 40 пар «статья-аннотация». Эксперты осведомлены, что каждый из наборов содержит реальные аннотации. Для каждого эксперта составлен набор из 20 реальных и 20 сгенерированных аннотаций, перемешанных в случайном порядке. Информация об источнике конкретной аннотации (реальная / сгенерированная) экспертам недоступна. Вместе с тем общий набор данных для оценки содержит всего 200 пар «статья-аннотация».

Таким образом, каждую пару «статья-аннотация» оценивают, в среднем, 20 экспертов. Подобный подход позволяет минимизировать единичные ошибки экспертов, дает более точную оценку качества сгенерированных текстов.

Ниже приведены сгенерированные аннотации (сохранена орфография и пунктуация) с высокой средней оценкой экспертов (выше 8,5 баллов).

В статье рассматриваются предпосылки возникновения и развития азартных игр, а также особенности их организации и проведения. Анализируются законодательные акты Российской Федерации, регулирующие деятельность по организации и проведению игорных заведений. Формулируются предложения по совершенствованию законодательства в данной сфере.

Рассматриваются вопросы защиты прав и свобод человека и гражданина. Формулируются предложения по совершенствованию механизма реализации конституционных обязанностей граждан РФ. Предлагается классификация способов защиты гражданских прав по материально-правовым основаниям: репрессивные, пресекательные, восстановительные и компенсационные.

В статье рассмотрен вопрос создания математической модели поддержки процесса временного перераспределения трудовых ресурсов в проектно-ориентированных организационных системах. Предложен алгоритм, позволяющий автоматизировать предварительный отбор кандидатов на роли в проектах, выполняемых в проектных организациях. Предложены алгоритмы распределения участников по проектам при достаточно большом количестве проектов и составляющих их операций.

Исследуются проблемы, возникающие при реализации административного надзора за лицами, судимыми за насильственные преступления против половой неприкосновенности несовершеннолетних. Выявлены основные направления деятельности участкового уполномоченного полиции в сфере профилактики преступлений и других правонарушений.

В статье рассмотрены некоторые области деятельности органов внутренних дел по использованию технологии больших данных (Big Data). Рассмотрены некоторые проблемы и вызовы, которые могут быть затруднины при использовании технологий больших данных.

Как видно из сгенерированных текстов в примерах, присутствуют незначительные орфографические ошибки, которые могут быть исправлены в автоматическом режиме с помощью компьютерных программ, осуществляющих проверку заданного текста на предмет наличия в нем орфографических, пунктуационных, а также стилевых ошибок.

Следует отметить, что несмотря на орфографические ошибки, семантическое содержание приведенных примеров аннотаций является корректным.

Оценка результатов. Результаты оценки модели до дообучения неудовлетворительные: по критерию «грамматика» – 7,8±1,33 балла, логика – 1,52±0,79 балла. Для более удобного восприятия информации распределение оценок экспертов представлено в виде диаграммы «ящик с усами» (boxplot) на рисунке 11 [24].

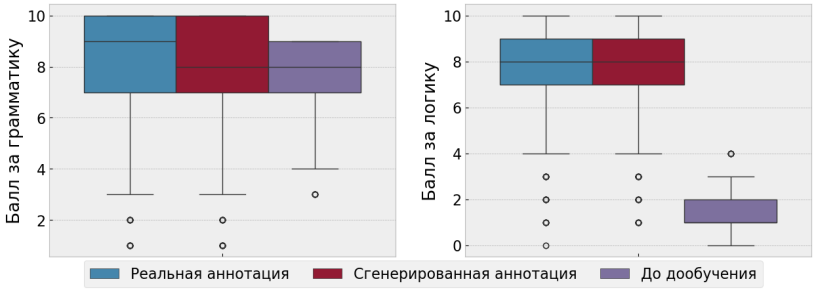


Рис. 11. Диаграмма «ящик с усами» для оценок экспертов

Из анализа рисунка 10 можно сделать следующий вывод: распределения оценок для реальных и сгенерированных аннотаций практически неотличимы.

Для подтверждения выдвинутого тезиса о неразличимости распределений проведён статистический тест Колмогорова-Смирнова для гипотезы о том, что выборки взяты из одного распределения вероятностей [25]. Для оценок грамматики p-value составляет 0,842; для оценок логики – 0,941. Таким образом, статистический тест подтверждает факт статистической неразличимости оценок качества реальных и сгенерированных аннотаций.

На рисунке 12 приведена альтернативная визуализация оценок экспертов.

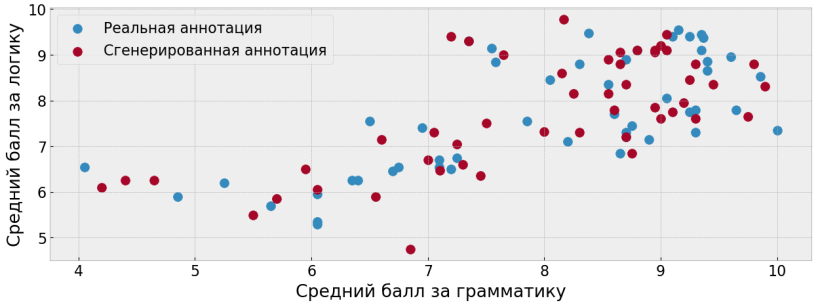


Рис. 12. Скаттерограмма оценок экспертов, усреднённых для каждой пары «статья-аннотация»

На рисунке 12 цветом обозначен источник аннотации – реальная или сгенерированная. Точки разных цветов сильно перемешаны, что не позволяет провести четкую классификацию в данных координатах. Данный факт позволяет говорить о сопоставимости качества сгенерированных и реальных аннотаций.

Рисунок 13 визуализирует результаты в разрезе по областям науки. В валидационных данных содержались 6 областей науки: юриспруденция, педагогика, информационная безопасность, психология, социология, история.

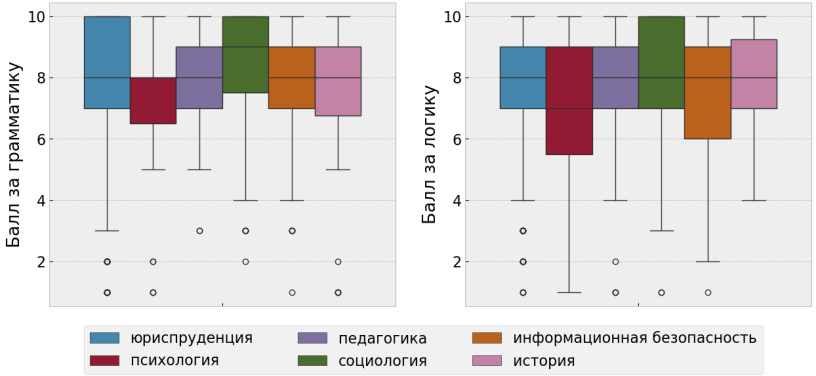


Рис. 13. Диаграмма «ящик с усами» для оценок экспертов в разрезе по областям науки

Из рисунка видно, что модель генерирует аннотации к статьям из различных областей науки с одинаковым качеством. Таким образом, для расширения области применения модели целесообразно обогащать датасет парами «текст-аннотация» из различных областей науки. Однако, вопрос поведения модели при постепенном расширении датасета необходимо исследовать отдельно – существует ли граница, после которой разнообразие предметных областей начнёт ухудшать качество генерации? Данный вопрос является темой дальнейших исследований.

На рисунке 14 приведены результаты оценки модели на дополнительных тестовых данных. Тестовые данные были собраны после обучения, они не учувствовали ни в обучении, ни в валидации. Объем дополнительных тестовых данных – 112 пар «аннотация – текст», источники, методика сбора и оценки аналогичны источникам и методикам для тренировочных и валидационных данных.

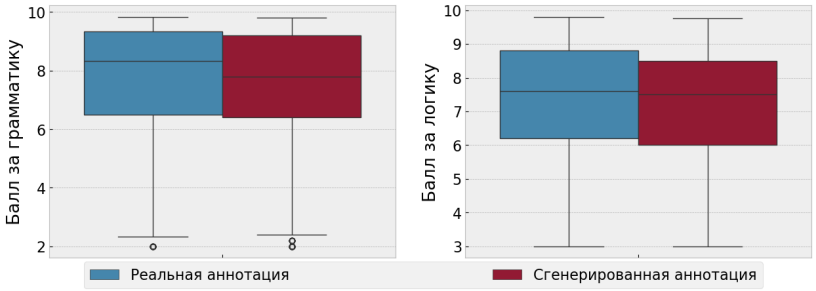


Рис. 14. Диаграмма «ящик с усами» для оценок экспертов на тестовых данных

Оценка на новых данных практически не отличается от оценки на валидации. Таким образом, можно говорить о достижении сетью возможности обобщения, а не только заучивания обучающих данных.

5. Заключение и выводы. В данной статье проработан вопрос совершенствования процесса аннотирования научных статей, имеющий высокую актуальность ввиду очевидной необходимости оптимизации способов составления кратких характеристик первичных научных документов с учетом их особенностей (назначения, содержания, формы). Необходимость также подтверждается увеличением объемом научных сведений, при которых значение аннотации определяется одной из ее функций – помощь исследователям в оперативном нахождении релевантного научного материала и извлечение ключевой информации. Аннотирование научных статей также способствует структурированию знаний,

выделению ключевых идей и результатов, а в условиях общего возрастания объемов научной информации позволяет проводить систематизацию и управление научными знаниями посредством навигационных репозиториев и баз данных.

Автоматизация процесса составления аннотации в современных реалиях должна производиться на базе цифровой трансформации. Так, технология блокчейн может быть интегрирована в процесс составления аннотации для обеспечения прозрачности, защиты интеллектуальной собственности и гарантии подлинности и целостности представленных в научных трудах сведений. Инструментарий предобученных больших языковых моделей позволяет значительно сократить время, необходимое для анализа больших массивов научных текстов, повысить точность и корректность извлечения информации.

В данном научном исследовании решена задача автоматической генерации аннотаций к научным статьям. Качество генерации сопоставимо с реальными аннотациями, а также отвечает требованиям информативности, структурированности и компактности, которые отмечены в действующих стандартах по издательскому оформлению статей в печатных и электронных научных, периодических и продолжающихся сборниках. Сгенерированные тексты согласуются с типовой структурой аннотации: содержат справочную информацию, цель, описание подходов, результатов и краткие выводы. Они реализуют задачу по отображению существенных признаков содержания научных трудов, позволяющих выявить их научное, теоретическое или практическое значение для целевой аудитории, новизну, отличить конкретный научный материал от других, аналогичных по тематике и целевому назначению, представляют информацию о достоинствах статей. Также полученные тексты аннотаций выполняют установленные стандартами функции: дают возможность установить основное содержание документа, определить его релевантность; предоставляют базовые сведения о научной статье, устраняют необходимость чтения полного текста документа; могут быть использованы в системах поиска документов и информации.

Основным элементом решения является собранный и размеченный датасет, который позволяет провести дообучение базовой языковой модели. Датасет состоял из 825 научных материалов, подготовленных по тематике решения актуальных проблем образовательного процесса, общественных, технических (информационных), гуманитарных, экономических и юридических наук.

Эффективность генерации модели верифицируется с помощью предложенной методики экспертной оценки на основе балльной системы от 1 до 10 с двумя параметрами: логика и грамматика. Под грамматикой понимаются любые синтаксические, грамматические и иные ошибки, позволяющие идентифицировать аннотацию как сгенерированную; под логикой – смысловая корректность аннотации. Обработка результатов экспертной оценки показала, что распределение оценок сгенерированных и реальных аннотаций статистически неразличимы, что свидетельствует о высоком качестве генерации языковой модели.

Разработка внедрена в учебный процесс государственного вуза в виде программного продукта (веб-приложения), используемого в научном обеспечении и сопровождении образовательного процесса, оказывающего помощь в подготовке квалифицированных научных специалистов. Веб-приложение позволяет сформировать краткую характеристику научного материала с точки зрения его тематики, содержания, новизны и других особенностей. Работа основана на функционировании большой языковой модели архитектуры T5, дообученной на корпусе из тысячи размеченных научных публикаций, содержащих результаты научных и прикладных исследований в области экономики, юриспруденции, педагогики, а также технических наук в контексте правоохранительной деятельности

Дальнейшее исследование предполагает дообучение моделей и оценку сгенерированных аннотаций с точки зрения требований нормативных документов, а также рассмотрение дообучения мультиязычных больших языковых моделей для задачи генерации аннотаций к научным статьям на разных языках. Помимо этого, остаётся открытым вопрос исследования качества генерации модели при постепенном расширении датасета текстами из различных предметных областей.

В заключении целесообразно отметить, что концепция предлагаемой разработки позиционирует свое применение в качестве системы поддержки принятия решений. Крайне важно использовать результаты генерации в сочетании с собственными авторскими знаниями предметной области на базе персонального критического мышления, анализа и интерпретации данных.

Литература

1. Жмудь В.А. Методы научных исследований: учебное пособие. Москва: Ай Пи Ар Медиа. 2024. 344 с.
2. Мейлихов Е.З. Искусство писать научные статьи: научно-практическое руководство. Долгопрудный: Издательский Дом «Интеллект». 2020. 335 с.

3. ГОСТ 7.9-95 (ИСО 214-76). Система стандартов по информации, библиотечному и издательскому делу. Реферат и аннотация. Общие требования // М.: Госстандарт России. 1995.
4. ГОСТ Р 7.0.99-2018 (ИСО 214:1976). Система стандартов по информации, библиотечному и издательскому делу. Реферат и аннотация. Общие требования // М.: Госстандарт России. 2018.
5. ГОСТ 7.86-2003. Система стандартов по информации, библиотечному и издательскому делу. Издания. Общие требования к издательской аннотации // М.: Госстандарт России. 2003.
6. ГОСТ Р 7.0.7-2021. Система стандартов по информации, библиотечному и издательскому делу. Статьи в журналах и сборниках. Издательское оформление // М.: Госстандарт России. 2021.
7. Курицкая Е.В. Технология написания аннотации к техническому тексту // Актуальные вопросы современного языкознания и тенденции преподавания иностранных языков: теория и практика: Материалы III Всероссийской научно-практической конференции (Кострома, 20 октября 2022 г.). Кострома: Военная академия радиационной, химической и биологической защиты имени Маршала Советского Союза С.К. Тимошенко (г. Кострома) Министерства обороны Российской Федерации. 2023. С. 93–99.
8. Schmarzo B. The Economics of Data, Analytics, and Digital Transformation: The theorems, laws, and empowerments to guide your organization's digital transformation // Packt Publishing. 2020. 260 p.
9. Reinsel D., Gantz J., Rydning J. The Digitization of the World From Edge to Core // An IDC White Paper. 2018. 28 p.
10. Толстых М.Ю. К вопросу обеспечения процессов цифровой трансформации в системе обучения // Цифровая трансформация образования: современное состояние и перспективы: Сборник научных трудов по материалам II Международной научно-практической конференции (Курск, 17–18 ноября 2023 г.). Курск: Курский государственный медицинский университет, 2024. С. 439–442.
11. Хлыбова М.А. Цифровые технологии в обучении написанию аннотаций в магистратуре неязыкового вуза // Филологический аспект. 2023. № 05(22). С. 55–58.
12. Солдатенкова Ю.А. YandexGPT и ChatGPT: характеристика, сравнение и основные отличия нейросетей // Моя профессиональная карьера. 2023. Т. 3. № 55. С. 277–284.
13. Lal K., Sharma B. Research Integrity & Ethics Scientific Misconduct // National Seminar on Academic Integrity and Research Ethics. At: DIT University, Dehradun. 2023. pp. 129–143.
14. Zmitrovich D., Abramov A., Kalmykov A., Tikhonova M., Taktasheva E., Astafurov D., Baushenko M., Snegirev A., Kadulin V., Markov S., Shavrina T., Mikhailov V., Fenogenova A. A Family of Pretrained Transformer Language Models for Russian: arXiv:2309.10931. arXiv. 2023.
15. Touvron H. et al. Llama 2: Open Foundation and Fine-Tuned Chat Models: arXiv:2307.09288. arXiv. 2023.
16. Brown T.B. et al. Language Models are Few-Shot Learners: arXiv:2005.14165. arXiv. 2020.
17. Tunstall L., Werra L. von, Wolf T. Natural Language Processing with Transformers, Revised Edition. 1st edition. Sebastopol: O'Reilly Media, Inc. 2022. 406 p.
18. Lin C.-Y. ROUGE: A Package for Automatic Evaluation of Summaries // Text Summarization Branches Out. Barcelona. 2004. pp. 74–81.

19. Ravenscroft J., Oellrich A., Saha S., Liakata M. Multi-label Annotation in Scientific Articles – The Multi-label Cancer Risk Assessment Corpus // Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). 2016. pp. 4115–4123.
20. Sun J., Wang Y., Li Z. An Improved Template Representation-based Transformer for Abstractive Text Summarization // IEEE International Joint Conference on Neural Network. 2020. pp. 1–8.
21. Amusat O., Hegde H., Mungall C.J., Giannakou A., Byers N.P., Gunter D., Fagnan K., Ramakrishnan L. Automated Annotation of Scientific Texts for ML-based Keyphrase Extraction and Validation. arXiv.2311.05042. arXiv, 2023.
22. Гуцыкова С.В. Метод экспертных оценок: теория и практика. Москва: Издательство «Институт психологии РАН». 2011. 144 с.
23. Щеглов И.А. Роль студентоориентированного подхода в социализации экспертизы // Гуманитарный вестник. 2021. № 4(90). С. 1–15.
24. Уилке К. Основы визуализации данных. Пособие по эффективной и убедительной подаче информации. Москва: Бомбора, 2024. 352 с.
25. Иванов Б.Н. Теория вероятностей и математическая статистика: учебное пособие для вузов. Издание третье. Санкт-Петербург: Лань. 2024. 224 с.

Голубинский Андрей Николаевич — д-р техн. наук, доцент, и.о. заместителя директора по научной работе, Институт проблем передачи информации им. А.А. Харкевича Российской академии наук. Область научных интересов: машинное обучение, нейросетевое моделирование, автоматизированные системы управления с элементами искусственного интеллекта, обработка речевых сигналов. Число научных публикаций — 242. annikgol@mail.ru; Большой Каретный переулок, 19/1, 127051, Москва, Россия; р.т.: +7(495)650-2235.

Толстых Андрей Андреевич — канд. техн. наук, инженер-программист, ООО «РТК». Область научных интересов: искусственные нейронные сети, машинное обучение, обучение с подкреплением. Число научных публикаций — 63. tolstykh.aa@yandex.ru; проспект Высоковольтный, 1/49, 127566, Москва, Россия; р.т.: +7(910)242-7955.

Толстых Марина Юрьевна — канд. техн. наук, доцент, кафедра международной информационной безопасности, Московский государственный лингвистический университет; доцент кафедры, кафедры специальных информационных технологий учебно-научного комплекса информационных технологий, Московского университета МВД России им. В.Я. Кикотя. Область научных интересов: информационная безопасность, цифровая трансформация, машинное обучение. Число научных публикаций — 109. marina_lion@mail.ru; улица Коптевская, 63, 127299, Москва, Россия; р.т.: +7(920)440-3845.

DOI 10.15622/ia.24.1.10

A. GOLUBINSKIY, A. TOLSTYKH, M. TOLSTYKH

AUTOMATIC GENERATION OF SCIENTIFIC ARTICLES ABSTRACTS BASED ON LARGE LANGUAGE MODELS

Golubinskiy A., Tolstykh A., Tolstykh M. **Automatic Generation of Scientific Articles Abstracts Based on Large Language Models.**

Abstract. The concept of automation of the process of annotation of scientific materials (Russian-language scientific articles) is proposed and its practical implementation is carried out by means of machine learning technologies, and additional training of large language models. The relevance of correct and rational compilation of annotations is indicated, and the problems related to establishing a balance between the time-consuming process of annotation and ensuring compliance with key requirements for annotation are highlighted. The basics of annotation presented in the family of standards on information, librarianship, and publishing are analyzed, and the classification of annotations and requirements for their content and functionality is given. The essence and content of the annotation process, and the typical structure of the research object are presented schematically. The issue of integration of digital technologies into the annotation process is analyzed, and special attention is paid to the advantages of introducing machine learning and artificial intelligence technology. The digital toolkit used to generate text in natural language processing applications is briefly described. Its shortcomings for solving the problem posed in this scientific article are noted. The research part substantiates the choice of the machine learning model used to solve the problem of conditional text generation. The existing pre-trained large language models are analyzed and, considering the problem statement and existing limitations of computing resources, the ruT5-base model is selected. A description of the dataset is given, including scientific articles from journals included in the list of peer-reviewed scientific publications in which the main scientific results of dissertations for the degrees of candidate and doctor of science should be published. The data labeling technique based on the operation of the tokenizer of the pre-trained large language model is characterized, and the numerical characteristics of the dataset distributions and the parameters of the training pipeline are presented graphically and in tables. The ROUGE quality metric is used to evaluate the model, and the expert assessment method, including grammar and logic as basic criteria, is used to evaluate the results. The quality of automatic annotation generation is comparable to real texts and meets the requirements of information content, structure and compactness. The article may be of interest to an audience of scientists and researchers seeking to optimize their scientific activities in terms of integrating digitalization tools into the process of writing articles, as well as to specialists involved in training large language models.

Keywords: annotation, generation, large language models, digitalization, machine learning.

References

1. Zhmud' V.A. Metody nauchnyh issledovanij: uchebnoe posobie [Scientific research methods: textbook]. Moscow: Aj Pi Ar Media. 2024. 344 p. (In Russ.).
2. Mejlihov E.Z. Iskusstvo pisat' nauchnye stat'i: nauchno-prakticheskoe rukovodstvo [The art of writing scientific articles: a scientific and practical guide]. Dolgoprudnyj: Izdatel'skij Dom «Intellekt». 2020. 335 p. (In Russ.).
3. GOST R 7.9-95 (ISO 214-76). Sistema standartov po informacii, bibliotechnomu i izdatel'skomu delu. Referat i annotacija. Obshhie trebovanija [System of standards on

information, librarianship and publishing. Informative abstract and indicative abstract. General requirements]. M.: Gosstandart Rossii. 1995. (In Russ.).
4. GOST R 7.0.99-2018 (ISO 214:1976). Sistema standartov po informacii, bibliotechnomu i izdatel'skomu delu. Referat i annotacija. Obshhie trebovanija [System of standards on information, librarianship and publishing. Abstract and annotation. General requirements]. M.: Gosstandart Rossii. 2018. (In Russ.).
5. GOST 7.86-2003. Sistema standartov po informacii, bibliotechnomu i izdatel'skomu delu. Izdanija. Obshhie trebovanija k izdatel'skoj annotacii [System of standards on information, librarianship and publishing. Editions. General requirements for publishing annotations]. M.: Gosstandart Rossii. 2003. (In Russ.).
6. GOST R 7.0.7-2021. Sistema standartov po informacii, bibliotechnomu i izdatel'skomu delu. Stat'i v zhurnalah i sbornikah. Izdatel'skoe oformlenie [System of standards on information, librarianship and publishing. Articles in magazines and collections. Publishing design]. M.: Gosstandart Rossii. 2021. (In Russ.).
7. Kurickaja E.V. Tehnologija napisanija annotacii k tehnicheskomu tekstu [Technology for writing annotations for technical texts] Aktual'nye voprosy sovremennogo jazykoznanija i tendencii prepodavanija inostrannyh jazykov: teorija i praktika : Materialy III Vserossijskoj nauchno-prakticheskoj konferencii [Current issues of modern linguistics and trends in teaching foreign languages: theory and practice: Materials of the III All-Russian Scientific and Practical Conference]. Kostroma: Voennaja akademija radiacionnoj, himicheskoj i biologicheskoj zashhity imeni Marshala Sovetskogo Sojuza S.K. Timoshenko (g. Kostroma) Ministerstva oborony Rossijskoj Federacii. 2023. pp. 93–99. (In Russ.).
8. Schmarzo B. The Economics of Data, Analytics, and Digital Transformation: The theorems, laws, and empowerments to guide your organization's digital transformation. Packt Publishing. 2020. 260 p.
9. Reinsel D., Gantz J., Rydning J. The Digitization of the World from Edge to Core. An IDC White Paper. 2018. 28 p.
10. Tolstyh M.J. K voprosu obespechenija processov cifrovoj transformacii v sisteme obuchenija [On the issue of ensuring digital transformation processes in the education system]. Cifrovaja transformacija obrazovanija: sovremennoe sostojanie i perspektivy: Sbornik nauchnyh trudov po materialam II Mezhdunarodnoj nauchno-prakticheskoj konferencii [Digital transformation of education: current state and prospects: Collection of scientific papers based on the materials of the II International Scientific and Practical Conference]. Kursk: Kurskij gosudarstvennyj medicinskij universitet. 2024. pp. 439–442. (In Russ.).
11. Hlybova M.A. [Digital technologies in teaching annotation writing in a master's program at a non-linguistic university]. Filologicheskij aspekt – The philological aspect. 2023. no. 05(22). pp. 55–58. (In Russ.).
12. Soldatenkova J.A. [YandexGPT and ChatGPT: characteristics, comparison and main differences between neural networks]. Moja professional'naja kar'era – My professional career. 2023. vol. 3. no. 55. pp. 277–284. (In Russ.).
13. Lal K., Sharma B. Research Integrity & Ethics Scientific Misconduct. National Seminar on Academic Integrity and Research Ethics. At: DIT University, Dehradun. 2023. pp. 129–143.
14. Zmitrovich D., Abramov A., Kalmykov A., Tikhonova M., Taktasheva E., Astafurov D., Baushenko M., Snegirev A., Kadulin V., Markov S., Shavrina T., Mikhailov V., Fenogenova A. A Family of Pretrained Transformer Language Models for Russian: arXiv:2309.10931. arXiv. 2023.
15. Touvron H. et al. Llama 2: Open Foundation and Fine-Tuned Chat Models: arXiv:2307.09288. arXiv. 2023.

16. Brown T.B. et al. Language Models are Few-Shot Learners: arXiv:2005.14165. arXiv. 2020.
17. Tunstall L., Werra L. von, Wolf T. Natural Language Processing with Transformers, Revised Edition. 1st edition. Sebastopol: O'Reilly Media, Inc. 2022. 406 p.
18. Lin C.-Y. ROUGE: A Package for Automatic Evaluation of Summaries. Text Summarization Branches Out. Barcelona. 2004. pp. 74–81.
19. Ravenscroft J., Oellrich A., Saha S., Liakata M. Multi-label Annotation in Scientific Articles – The Multi-label Cancer Risk Assessment Corpus. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). 2016. pp. 4115–4123.
20. Sun J., Wang Y., Li Z. An Improved Template Representation-based Transformer for Abstractive Text Summarization. IEEE International Joint Conference on Neural Network. 2020. pp. 1–8.
21. Amusat O., Hegde H., Mungall C.J., Giannakou A., Byers N.P., Gunter D., Fagnan K., Ramakrishnan L. Automated Annotation of Scientific Texts for ML-based Keyphrase Extraction and Validation. arXiv.2311.05042. arXiv, 2023.
22. Gucykova S.V. [Expert assessment method: theory and practice] Metod jekspertnyh ocenok: teorija i praktika. Moscow: Institut psihologii RAN. 2019. 144 p. (In Russ.).
23. Shheglov I.A. [The role of the student-centered approach in the socialization of expertise]. Gumanitarnyj vestnik – Humanitarian Bulletin. 2021. no. 4(90). pp. 1–15. (In Russ.).
24. Uilke K. Osnovy vizualizacii dannyh. Posobie po jeffektivnoj i ubeditel'noj podache informacii [Basics of data visualization. A Guide to Effectively and Persuasively Presenting Information]. Moscow: Bombora, 2024. 352 p. (In Russ.).
25. Ivanov B.N. Teorija verojatnostej i matematicheskaja statistika: uchebnoe posobie dlja vuzov[Probability theory and mathematical statistics: a textbook for universities]. The third edition. Sankt-Peterburg: Lan'. 2024. 224 p. (In Russ.).

Golubinskiy Andrey — Ph.D., Dr.Sci., Associate Professor, Acting deputy director for research, Institute for Information Transmission Problems (Kharkevich Institute) Russian Academy of Sciences. Research interests: machine learning, neural network modeling, automated control systems with artificial intelligence elements, speech signal processing. The number of publications — 242. annikgol@mail.ru; 19/1, Bolshoy Karetny Lane, 127051, Moscow, Russia; office phone: +7(495)650-2235.

Tolstykh Andrey — Ph.D., Software engineer, OOO "RTK". Research interests: artificial neural networks, machine learning, reinforcement learning. The number of publications — 63. tolstykh.aa@yandex.ru; 1/49, Vysokovoltny Av., 127566, Moscow, Russia; office phone: +7(910)242-7955.

Tolstykh Marina — Ph.D., Associate professor, Department of international information security, Moscow State Linguistic University; Associate professor of the department, Department of special information technologies of the educational and scientific complex of information technologies, Moscow University of the Ministry of Internal Affairs of Russia. V.Ya. Kikotya. Research interests: information security, digital transformation, machine learning. The number of publications — 109. marina_lion@mail.ru; 63, Koptevskaya St., 127299, Moscow, Russia; office phone: +7(920)440-3845.

DOI 10.15622/ia.24.1.11

N.V. HUNG, P.T. DAT, N. TAN, N.A. QUAN, L.T.N. TRANG, L.M. NAM

HEVERL – VIEWPORT ESTIMATION USING REINFORCEMENT LEARNING FOR 360-DEGREE VIDEO STREAMING

Nguyen Viet Hung, Pham Tien Dat, Nguyen Tan, Nguyen Anh Quan, Le Thi Huyen Trang, Le Mai Nam. **HEVERL – Viewport Estimation Using Reinforcement Learning for 360-degree Video Streaming.**

Abstract. 360-degree video content has become a pivotal component in virtual reality environments, offering viewers an immersive and engaging experience. However, streaming such comprehensive video content presents significant challenges due to the substantial file sizes and varying network conditions. To address these challenges, view adaptive streaming has emerged as a promising solution, aimed at reducing the burden on network capacity. This technique involves streaming lower-quality video for peripheral views while delivering high-quality content for the specific viewport that the user is actively watching. Essentially, it necessitates accurately predicting the user's viewing direction and enhancing the quality of that particular segment, underscoring the significance of Viewport Adaptive Streaming (VAS). Our research delves into the application of incremental learning techniques to predict the scores required by the VAS system. By doing so, we aim to optimize the streaming process by ensuring that the most relevant portions of the video are rendered in high quality. Furthermore, our approach is augmented by a thorough analysis of human head and facial movement behaviors. By leveraging these insights, we have developed a reinforcement learning model specifically designed to anticipate user view directions and improve the experience quality in targeted regions. The effectiveness of our proposed method is evidenced by our experimental results, which show significant improvements over existing reference methods. Specifically, our approach enhances the Precision metric by values ranging from 0.011 to 0.022. Additionally, it reduces the Root Mean Square Error (RMSE) by 0.008 to 0.013, the Mean Absolute Error (MAE) by 0.012 to 0.018 and the F1-score by 0.017 to 0.028. Furthermore, we observe an increase in overall accuracy of 2.79 to 16.98. These improvements highlight the potential of our model to significantly enhance the viewing experience in virtual reality environments, making 360-degree video streaming more efficient and user-friendly.

Keywords: head-eye movement, reinforcement learning, deep learning, machine learning, video streaming, 360-degree video.

1. Introduction. In recent years, prediction models have gained significant attention in the research community, particularly in the field of 360-degree video streaming. Accurate prediction in this context is crucial as it enhances the viewer's immersion and understanding of the video content. However, achieving high prediction accuracy remains a challenging task, especially under varying network conditions.

Existing research has explored various methods to improve the performance of prediction models for 360-degree videos. For example, reinforcement learning has been used to control model predictions based on data-driven designs, significantly improving performance, as demonstrated in the work of the authors in paper [1].

In the context of 360-degree videos, accurate prediction is excellent since it increases the viewer's understanding and immersion of the video. Significantly, when network conditions change, adapting to meet viewers' needs is difficult. From these research issues [2-4], the prediction models are built based on head movements to adapt to different types of videos on the Viewport Adaptive Streaming (VAS) system. However,the adaptability and self-learning ability are not only low but also dependable on the initial data, so it is still difficult when the data changes continuously.

In the context of 360-degree videos, accurate viewport prediction is essential for adapting to viewers' needs, especially when network conditions fluctuate. Studies such as those by the authors in [2-4] have developed prediction models based on head movements to adapt to different types of videos on the Viewport Adaptive Streaming (VAS) system. However, these models often suffer from limited adaptability and self-learning capabilities, particularly when data changes continuously, making it difficult to maintain accuracy.

Virtual reality (VR) presents additional challenges in this domain. As VR technology becomes more widespread, ensuring users feel fully immersed and interactively engaged is critical. However, video streaming in VR is constrained by factors such as network bandwidth, video resolution, and content complexity. High-speed transmission in the viewer's viewport area, coupled with lower quality in other areas, is a fundamental requirement for VR video streaming. Many studies have attempted to address these challenges by analyzing network conditions and employing optimization methods, but achieving a balance between network optimization and user immersion remains a significant hurdle. Therefore, predicting the viewer's viewport area is valid and applicable, considering the user's perspective without downloading the entire content. This means that the video content will be offloaded, and the network will have more space, which will help to improve the user's viewing area. In fact, [5] research has also shown critical retinal areas of the viewer; these are considered core areas. Based on these areas, we can quickly fix minor problems, such as limiting the quality of areas that are not considered and thereby improving the quality of areas that are considered.

Predicting the audience's view is a real challenge. Because each person will have a completely different view angle when turning their head and moving eyes. One more challenge for this problem is that the heads may not be moved when the viewers move their eyes. Only the movements of the eyes do not provide enough basis for a prediction model since we need both head-eye movements to be analyzed. Figure 1 shows information and forecast areas

in recent times. However, in this paper, we build on the principles of head movements to determine a better viewport position.

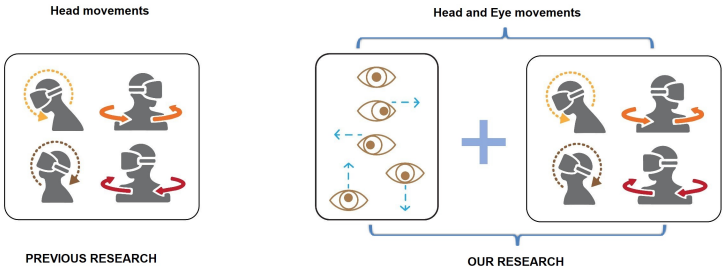


Fig. 1. Head-eye movements

Besides, psychology and perspective on movement are essential issues to analyze and make the right prediction [6-8]. First, the video contents partly affect viewers' psychology. For example, in emotional videos, viewers tend to change their head and eye movements when they have excess feelings. Second, many authors have been also researching perspective effects to evaluate the standard user's field of view. The factor of heads and eyes moving without following the rules also contributes significantly to incorrect prediction orders.

Furthermore, streaming 360-degree videos requires much more bandwidth compared to regular videos. The prediction method is necessary to achieve the user's perceived quality QoE because the user only sees a part of it. Thus, watching adaptive video streaming is an effective method to satisfy video quality [9-12]. However, this performance relies on the view adaptation scheme, view prediction and bandwidth. To overcome these problems, we propose a server-to-client streaming framework based on reinforcement learning, which optimizes 360-degree video streaming in viewport prediction to adapt to changing network conditions. We call this method HEVERL:

– To address these issues, we propose the HEVERL (Head-eye Movement Oriented Viewport Estimation Based on Reinforcement Learning) approach, which represents an advancement in viewport prediction for VR applications. Unlike traditional methods that rely solely on head orientation data, HEVERL incorporates both head-eye movement information to more accurately forecast the user's future viewport. This multi-modal sensing strategy provides a comprehensive understanding of the user's visual attention and behavior within the VR environment, leading to improved prediction accuracy.

– The HEVERL algorithm also introduces a novel content preparation and delivery mechanism that adaptively prefetches and updates the bitrate of previously viewed perspectives based on predicted viewport distribution. This proactive, viewport-aware content optimization enhances the user's perceived quality of experience by addressing network fluctuations and view prediction errors. By integrating prefetching and adaptive bitrate selection for previously visited viewports, HEVERL sets itself apart from traditional VR video streaming solutions, which generally rely on reactive strategies.

In summary, the HEVERL algorithm's dynamic adaptation to fluctuating network conditions and its ability to overcome potential view prediction errors represent a significant advancement in VR video streaming. The algorithm HEVERL may enhance the robustness and reliability of the VR experience, especially in latency-sensitive applications, and marks a step forward in achieving consistently high-quality VR viewing experiences. To provide a better understanding of our research, this report includes the following content: Section 2 discusses the related work. Section 3 describes the suggested viewport estimation technique. Section 4 evaluates the proposed method's performance compared to other methods. Section 5 concludes.

2. Related work

2.1. Streaming Video 360 Degrees. Recent research has focused on 360-degree video streaming, aiming to optimize bandwidth usage without compromising video quality. Studies [13-15] suggest that 360-degree video should be used as standard content to transmit the entire video, ensuring high viewing quality for users in all directions. However, streaming the full video consumes substantial bandwidth, allowing only a portion of the 360-degree video to be viewed at a time.

According to the research [16], there are two types of view-adaptive streaming: proposed tile-based streaming and asymmetric panorama image-based streaming. Panorama-based streaming generates multiple versions of a 360-degree video from different perspectives, necessitating video playback based on the user's orientation. While this approach reduces the apparent quality of the viewport and significantly lowers bandwidth usage, it also requires greater flexibility because limited versions result in poor display quality in viewer mode.

In tile-based streaming, the video is divided into multiple encrypted tiles, and different devices request tiles based on the user's perspective. Many algorithms for 360-degree video streaming [17, 18] transmit the Field of View (FoV) in this manner, effectively reducing bandwidth. However, this method is less flexible due to the dynamic changes in the user's perspective. As a result, recent viewport adaptation methods have relied on FoV [19, 20]. These

FoV-based prediction methods have improved significantly by reducing the performance impact caused by network distance to the predicted FoV and uneven bitrate assignment [21-24]. They dramatically reduce tile quality variation within the FoV. However, they still depend heavily on accurate bandwidth calculations, which can be influenced by network conditions, leading to estimation errors and performance degradation.

To overcome these limitations, we propose a reinforcement learning method combined with user behavior analysis to automatically adapt to network conditions and select tiles that optimize the predicted viewport area.

2.2. Synthetic prediction models. In this section, we will present some models built for prediction in recent years.

2.2.1. Head movements. In studies [3,4,25-27], the authors developed segment prediction models based on head movements. While many of these models are similar to our proposed model and aim to enhance the accuracy of predicting future user views in recommender systems, we identified some limitations. Notably, these methods primarily consider head movements while neglecting eye movements. The head can remain stationary while the eyes move. Therefore, experimental methods should account for head-eye movements to improve prediction accuracy.

Regarding head movements, most studies focus on changes in head position, acknowledging that head movements are generally slower compared to eye movements. However, addressing both types of movements presents a significant challenge. Many studies exclusively target head movements, overlooking the crucial role of eye movements. In reality, while the head may turn left or right, the eyes can independently look in different directions. This disparity underscores the importance of algorithmic adaptation to accommodate more complex movements for improved accuracy in prediction models.

2.2.2. Head-eye movements. In the study [28], the author developed cloud streaming for head-mounted displays, allowing viewers to experience the illusion of being in a virtual room by rotating their viewpoint. Additionally, in the study [29], the authors implemented a caching strategy that predicts user views based on cell resolution, aiming to forecast the viewing frequency of 360-degree video tiles. This method is particularly impactful under limited buffering conditions.

In another approach [30], the authors focused on predicting how different segments of a 360-degree video would be viewed on a head-mounted display. This method incorporated overlapping views and utilized techniques such as saliency detection, face detection, and object detection. However, the

algorithm primarily fine-tuned a fixed prediction network, leaving questions unanswered regarding the adaptability to changing movement dynamics.

While studies [28-30] have made significant strides in considering both head-eye movements, there is a pressing need for further research on the Viewport Adaptive Streaming (VAS) system's role in predicting user views. This gap in our understanding presents an exciting opportunity for future exploration and innovation in the field.

2.3. Reinforcement learning-based prediction. Viewport adaptation schemes for 360-degree video rely on estimated frequency width accuracy and are categorized based on throughput and buffering [31, 32]. However, this approach needs more flexibility and performs optimally only under specific network conditions. Therefore, adaptive algorithms are designed based on bitrate and user behavior to address these challenges and enhance adaptability and performance.

Approaches that rely on explicitly storing states and actions rather than using approximate functions are not scalable for real-world cyber environments. In response, D-DASH [33, 34] computes the action value Q using a neural network model (such as RNN or LSTM). D-DASH has shown superior performance and faster convergence compared to traditional Q-learning methods. However, its performance is still contingent on specific states and actions. To tackle this limitation, we propose an RL-based algorithm for decision-making that autonomously adapts to environmental changes.

Furthermore, the correlation between video perspective quality and video bitrate is non-linear. A neural network predicts video quality, while an RL algorithm selects the bitrate. This approach outperforms existing methods by delivering higher video quality and reducing latency.

While the authors have demonstrated that reinforcement learning optimizes adaptive bitrate for videos [35], this approach utilizes deep reinforcement learning (DRL) to train the curriculum autonomously. This enables bitrate decisions for 360-degree videos based on chunk selection and planning. This method has shown superior experimental results compared to state-of-the-art techniques. However, it primarily focuses on bitrate selection and chunk planning decisions, contrasting with our proposed method, which leverages user behavior to automatically adjust the bitrate and determine quality levels specifically for the viewport area.

On the contrary, in a recent study [36], researchers developed a system tailored for adaptation on Facebook's video platform using reinforcement learning (RL) in a live environment. They simulated the RL technique to train the agent effectively. Similarly, [37] introduced an advanced sequential reinforcement learning model to streamline decision-making and enhance the

Quality of Experience (QoE). These studies highlight the effectiveness of RL techniques in optimizing video streaming environments. However, these studies primarily concentrate on improving QoE by addressing factors like buffering, video quality, and timing without incorporating behavioral considerations.

Generally, reinforcement learning methods involve an agent making adaptive decisions in an interactive environment through trial and error [34]. Reinforcement learning empowers the agent to optimize its actions based on feedback, which is crucial for navigating dynamic and uncertain network conditions. However, these methods can be time-consuming, and their effectiveness hinges on the exploration strategy employed. Therefore, our proposed method aims to swiftly predict and make decisions that do not compromise the viewer's perception amidst fluctuating network conditions.

3. Proposed viewport estimation method – HEVERL. HEVERL is an acronym that stands for **H**ead **E**ye Movement Oriented **V**iewport **E**stimation Based on **R**einforcement **L**earning in Figure 2. Before discussing the HEVERL design, we will formulate the video streaming problem using the assumptions and constraints described in Section 3.

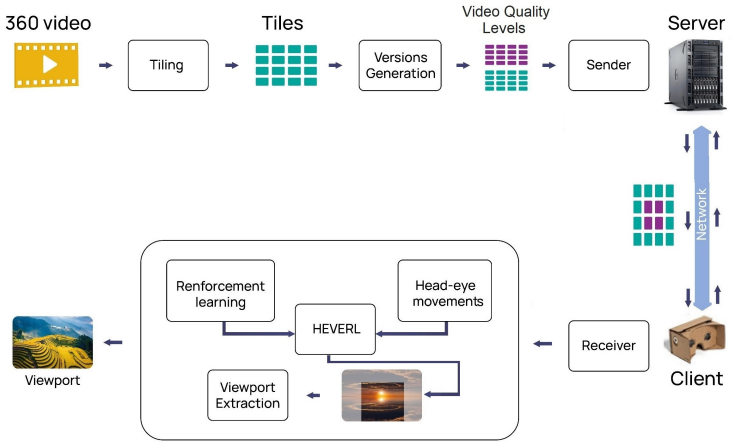


Fig. 2. HEVERL architecture

In this part, we present a problem that needs to be solved by predicting the viewport area that the human movement direction is using. Prediction is done when the direction of the user's movement does not change because it

is easier to predict and increase the quality of that area. However, in reality, prediction is very complicated because the more flexible the user is, the more significant the changes in prediction. Therefore, for each period t, it is necessary to predict the viewport area, and the next point will change, and so on, until the end of period t_n.

Furthermore, these changes will affect the user's perceived quality because when the user's movement direction is in any position, that area will increase in quality and reduce the near-quality area when not noticed. Therefore, this prediction increases the quality of user perception and limits bandwidth consumption in limited network conditions.

The core principle of tiling-based viewport adaptive streaming lies in the spatial partitioning of video content into distinct, granular sections known as tiles. This innovative architectural design deviates from the conventional view of the entire video frame as a single entity. By breaking down the video in this manner, the streaming system can handle the delivery of each tile independently, leading to more advanced adaptation strategies.

Expanding on the tiled structure, the tiling-based approach generates numerous encoded versions for each tile. This extensive range of tile variants empowers the system to enhance video quality based on the user's current viewport or field of view. Tiles that intersect with the user's viewport, called 'visible tiles,' are encoded at a higher quality to deliver an immersive viewing experience. In contrast, tiles outside the user's viewport, known as 'invisible tiles,' are encoded at a lower quality to conserve bandwidth and system resources in Figure 3.

The tiling-based viewport adaptive streaming approach is built on selectively assigning quality to visible and invisible tiles. By delivering the highest quality version of the tiles currently in the user's viewport, the system can provide an optimal visual experience without requiring high-quality data to be transmitted for the entire frame. This targeted quality allocation allows for efficient bandwidth utilization while reducing the risk of stalls or quality degradation during playback, as the system can dynamically adjust tile quality in response to user navigation and viewport changes.

The tiling-based viewport adaptive streaming model represents a significant step forward in video delivery, addressing the challenges of providing high-quality content while maximizing resource utilization. By spatially partitioning the video into tiles and selectively encoding multiple quality versions for each tile, the system can adaptively deliver the most appropriate content to the user based on their current viewport, resulting in a more immersive and bandwidth-efficient streaming experience.

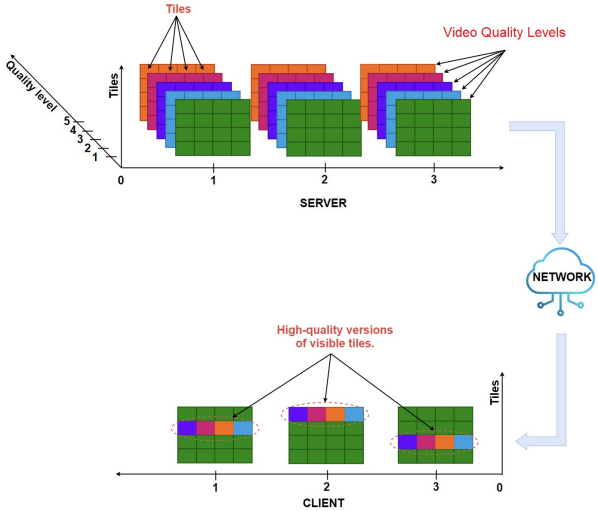


Fig. 3. Tiling-based Viewport Adaptive Streaming

3.1. Design of viewport prediction and selection. In this section, our focus is on designing a predictive model and devising methodologies for computing and categorizing viewport regions using reinforcement learning, illustrated in Figure 4.

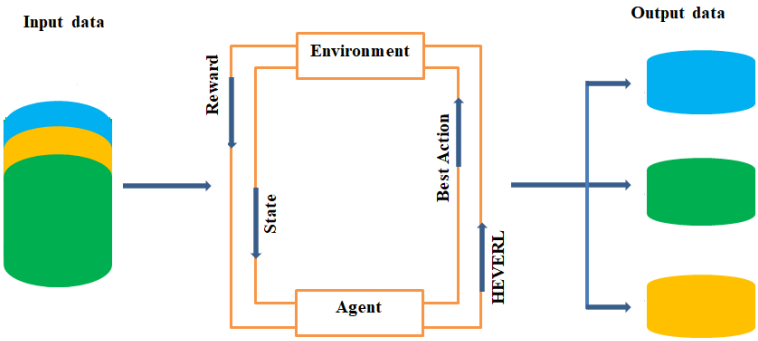


Fig. 4. HEVERL System

The system is structured as follows:

First, the data undergoes preprocessing before input. The data is represented through two states: t and t'. These states are stored as arrays and evolve spatially and temporally.

Second, we configure the environment settings and perform analysis based on these states. Subsequently, the algorithm calculates weights and dynamically predicts the user's viewing area throughout the video. The parameters are determined as follows:

– **Agent**. The Agent's objective is to locate the flag image, depicted in Figure 5. The Agent's path includes obstacles that influence the determination of the necessary route, impacting subsequent decision-making. Figure 5 illustrates how the Agent interacts with the Environment through actions such as left, right, up, and down.

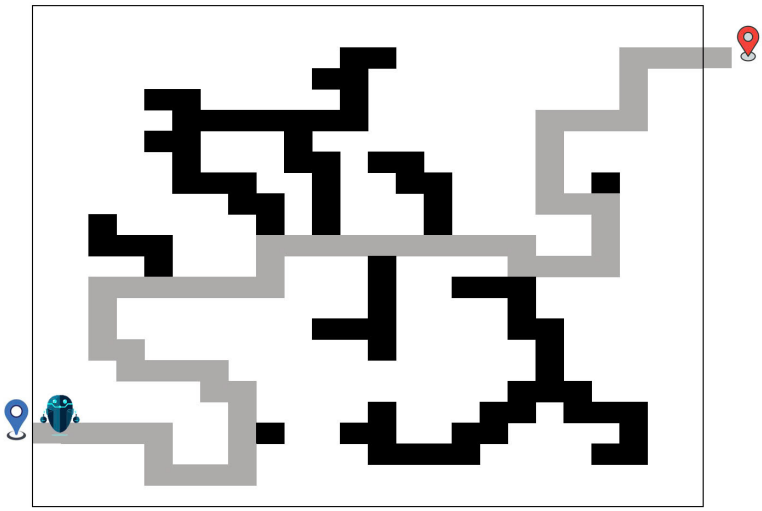


Fig. 5. Agent

– **State**. The state indicates the current position within the environment. Following each action, the environment provides the agent with a corresponding state.

– **Best Action**. The optimal action represents the transition process from the Agent to the environment. When the Agent reaches a forbidden box, the process terminates. The sequence of interactions between the Agent and the environment from start to finish is termed an Episode. Throughout the episode,

the Agent aims to select actions that maximize the Reward. The method by which the Agent selects these actions is known as the Policy.

– **HEVERL**. HEVERL will determine the final value to be saved and prepare for the next step based on the best action selections. Once identified and classified, the results will arrange the viewport sections sequentially and decide where to display information on the user's screen.

On the one hand, our approach utilizes the Markov Decision Process (MDP), a framework that aids agents in making decisions based on specific states. In applying this framework, we assume states possess the Markov property: the transition probability between two states is influenced solely by the preceding state.

Firstly, the concept of "probability of switching between two states" arises because, in reality, actions do not always yield deterministic outcomes. In an ideal scenario, repeating an action would consistently produce identical results. However, real-world processes are often stochastic. For instance, as depicted in Figure 6, if an agent decides to move upward and the environment's response is not deterministic, the outcomes can vary probabilistically. In this example, the agent might experience an 80% chance of returning to the "upper cell" state, with a 10% probability of transitioning to the "left cell" state and the "right cell" state each.

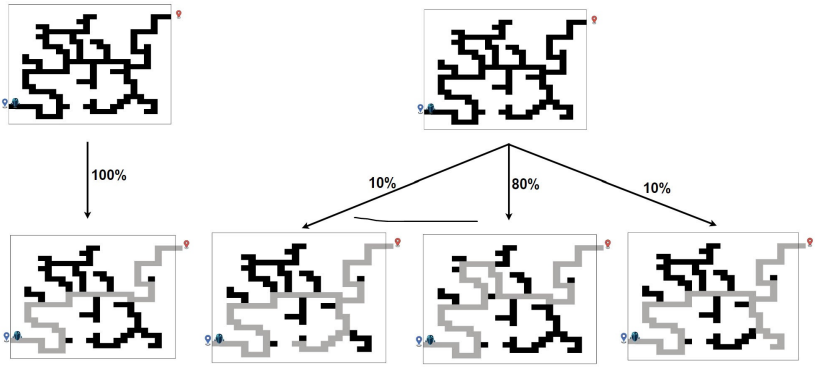


Fig. 6. Example process

3.2. Viewport Estimation Using Reinforcement Learning for 360-degree Video Streaming – HEVERL. HEVERL is the method we propose. It is based on the MDP model and is represented as follows. First, we calculate the Q_{Val} value when performing action h at state t by:

$$Q_{Val} = Q(t,h) = X(t,h) + \alpha \max_{h} Q(t',h). \quad (1)$$

Let X be the reward received when transferring state and $X(t,h)$ is the reward received υ'ith t' is the next state. Let α be the discount coefficient, ensuring that the farther "far away" from the Q_{Val} target, the smaller it is. Besides, let t be the state, and h be the action. This formula demonstrates that the Q_{Val} of action h at state t equals reward $X(t,h)$ plus the largest Q_{Val} of the following states t' when performing action h. As a result, we can create a state-action matrix as a lookup table using only that simple formula. As a result, for each state agent, the action with the highest Q_{Val} should be chosen. However, the Q_{Val} before and after acting will differ because RL is a stochastic process. This distinction is known as Temporal Difference:

$$f(h,t) = X(t,h) + \alpha \max_{h'} Q(t',h) - Q_{a-1}(t,h). \quad (2)$$

Therefore, the matrix Q_{Val} needs to be weighted based on TD by:

$$Q_a(t,h) = Q_{a-1}(t,h) + \sigma f_a(t,h), \quad (3)$$

where σ is the learning rate, through the times the agent performs actions, Q_{Val} will gradually converge. Thus, we aim to choose the appropriate action for a particular state. In other words, we use state as input and output as an action. During this stage, we realized that there is no constant solution using Neural Network (NN). All we need to do is remove the lookup table Q_{Val} and replace it with a simple NN in Figure 7. Besides, we employ a neural network structured with 4 layers. The configuration specifies the number of neurons per layer: 64, 128, 64, and 128 for layers 1, 2, 3, and 4, respectively. On the other hand, we use 3 neurons with x_1 as longitude, x_2 as latitude, and x_3 as the user's head-eye movement speed in Figure 7. In this part, we use x_3 represents the user's head-eye movements speed. It quantifies how quickly the user shifts their gaze. This variable can offer insights into user attention and focus, potentially indicating areas of interest or distraction. It could be measured in degrees per second if tracking angular movement per second for screen-based interactions. Understanding this speed, we can use adaptive content based on user engagement levels. All layers utilize ReLU activation functions, and regularization techniques include a Dropout set to 0.5 and an L2 regularization set to 0.01.

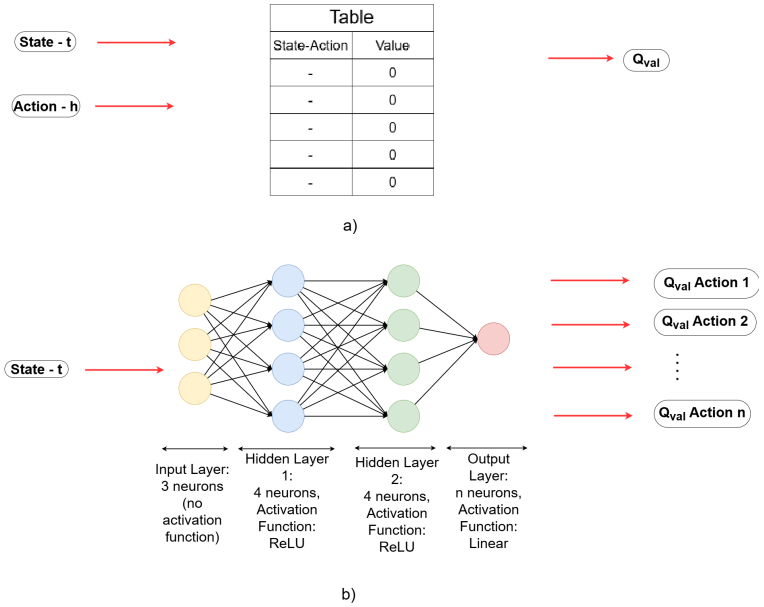


Fig. 7. State - Action

However, the most crucial part of NN is still missing. That is the Loss function. We aim to force the NN to learn how to accurately estimate the Q_{Val} for actions. Therefore, to determine the error between the actual and predicted Q_{Val}. The formula is determined and calculated as follows:

$$Loss\_function = (X + \alpha max_{h'} Q(t', h'; \varphi') - Q(t, h; \varphi))^2. \quad (4)$$

On the other hand, our HEVERL algorithm is proposed to perform as follows:

– **Step 01**: the setup environment injects a state into the network is t; The output is the Q_{Val} of the corresponding actions;

– **Step 02**: the agent chooses an action with a Policy and executes that action;

– **Step 03**: the environment returns state t' and reward x as the result of action h and saves the experience tuple $[t, h, x, t']$into memory;

– **Step 04**: sample the experiences into several batches and train the NN;

– **Step 05**: repeat until the end of M (M = 1000) episodes.

After performing the aforementioned steps, we calculate the predicted positions, which may fluctuate between different states. Experiments also indicate that our algorithm has shown improvement compared to conventional methods.

4. Performance Evaluation

4.1. Experimental Settings. To experiment, we use five videos: the Video Turtle describes People releasing baby turtles into the sea on the beach during the day. The Bar video describes the Bar as Light, with users moving and the bartender at work. The Video Ocean is described as follows: Under the ocean, people go underwater to see whales. Besides, there are two videos, Sofa and Po. Riverside is described as People sitting on sofas in the living room to talk, and Riverside videos outdoors during the day, with human activities. Each video contains traces of corresponding head-eye movements, and the information is also confirmed to change even when there is no head movement in Figure 8.

On the one hand, our dataset originates from the CSV file referenced [38]. We use two columns to indicate the viewer's position in latitude and longitude, normalized to a range of 0 to 1. Longitude values are scaled by multiplying by 2π, and latitude values by π to determine their on-screen positions. To display these positions accurately on an image, multiply longitude by the desired width and latitude by the desired height. Using these longitude and latitude coordinates, we can pinpoint the exact position of the observer. Besides, according to the authors in the article [38], head-eye movement data were collected from panoramic (360-degree) videos using head-eye tracking technologies. Head motion sensing technologies utilize accelerometers, gyroscope sensors, and kinematic trackers. Eye movement sensing technologies employ infrared eye trackers and eye-tracking glasses. 360-degree videos are recorded for users to view in virtual reality environments. Data from head-eye tracking sensors are recorded simultaneously with the video to provide information about how users interact with the content.

In this study, we utilized a dataset from [38] comprising head-eye movements data collected from 57 participants, including 25 women, whose ages ranged from 19 to 44 years (mean age: 25.7 years). Each participant viewed five distinct 360-degree videos for 20 seconds. The gaze data, sampled at 250Hz, yielded approximately 285,000 samples per video, totaling 1,425,000 samples across all videos. For model development, 80% of the data was

allocated for training and 20% for testing, ensuring comprehensive exposure during training and robust evaluation of model performance.

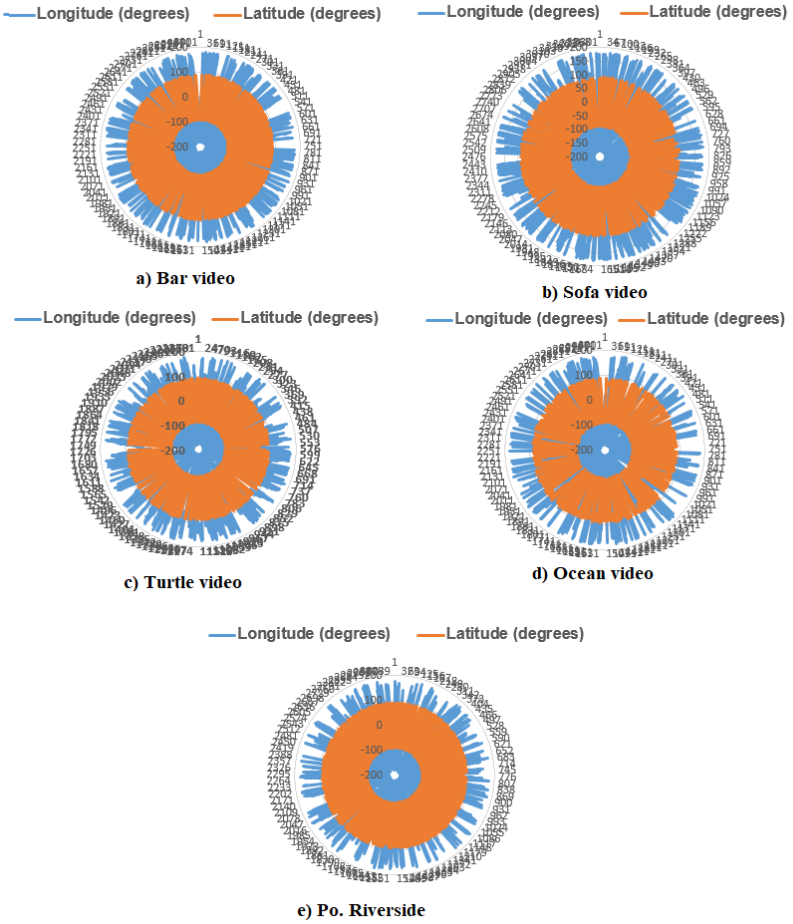


Fig. 8. Head-eye movements Dataset [38]

On the other hand, we experimented with the Windows 10 computer operating system, a Python-written experiment on a PC running 64-bit Windows 11, with 8192 MB RAM and an Intel® CoreTM i5-10400F Processor (6 Core, 12 Thread) CPU to measure the training time of different solutions. The proposed method HEVERL will be evaluated alongside other methods

by calculating the Root Mean Square Error (RMSE) based on precision calculation in the context of VAS.

The values defined in Table 1 should be replaced by the following abbreviations: TP for true positives, TN for true negatives, FP for false positives, and FN for false negatives.

– **Accuracy**. Accuracy is useful when the dataset's classes are well-balanced, with a similar number of instances in each class. However, accuracy can be misleading in imbalanced datasets, where one class significantly outnumbers the others.

$$Act = \frac{TN + TP}{TN + TP + FN + FP}. \tag{5}$$

– **Precision**. Indicates the precision with which Positive issues are detected.

$$Prec = \frac{TP}{TP + FP}. \tag{6}$$

– **Recall**. Recall measures the ability to find all the positive samples.

$$Recall = \frac{TP}{TP + FN}. \tag{7}$$

– **F1-Score**. F1-Score is the harmonic mean of precision and recall, providing a balance between the two.

$$F1 - Score = 2 * \frac{Prec * Recall}{Prec + Recall}. \tag{8}$$

Table 1. Definition of parameters

Values	Description
True Positives (TP)	True Positives are received True Positive;
False Positives (FP)	True Negatives are obtained False as Positive;
True negatives (TN)	True Negatives are received True Negatives;
False negatives (FN)	True Positives are received False as Negative.

Root Mean Square Error – RMSE. RMSE is one of the two leading performance indicators for a regression model. It computes the average difference between values predicted by a model and actual values. It estimates how well the model can predict the target value (accuracy):

$$RMSE = \sqrt{\frac{\sum_{a=1}^{H}(Prec_a - Act_a)^2}{H}}. \quad (9)$$

Mean Absolute Error – MAE. MAE is the average absolute magnitude of the errors between predicted and observed (true) viewport positions.

$$MAE = \frac{1}{H}\sum_{a=1}^{H}|Prec_a - Act_a|, \quad (10)$$

where:

– Let $Prec_a$ be the prediction rating,

– Let Act_a be the actual rating in the testing data set,

– H represents the number of rating prediction pairs between the testing data and the prediction result.

4.2. Viewport prediction performance. The viewport prediction performance of the HEVERL model is compared to the current reference models, including GLVP [3], AEVE [4], and GRU [39], in terms of Precision, RMSE, and MAE. This comparison aims to evaluate the viewport prediction capabilities of HEVERL against the benchmark models, intending to identify the advantages and effectiveness of the HEVERL model in applications that rely on accurate viewport prediction. Assessing these key performance metrics provides insights into the relative strengths and improvements offered by the HEVERL approach compared to the existing reference techniques.

The viewport prediction performance of HEVERL is compared with the current reference models such as GLVP [4], GRU [39], and AEVE [4] in terms of Precision, RMSE (Root Mean Square Error), MAE (Mean Absolute Error) in Table 2.

Table 2. HEVERL compared to the reference methods

Methods	Accuracy	Precision	Recall	F1-score	RMSE	MAE
GRU	71.23	0.865	0.860	0.861	0.248	0.147
GLVP	69.26	0.876	0.871	0.872	0.244	0.140
AEVE	83.45	0.869	0.862	0.864	0.249	0.144
HEVERL	**86.24**	**0.887**	**0.893**	**0.889**	**0.236**	**0.128**

The study compares the viewport prediction performance of HEVERL, a new proposed model, to existing reference models. Viewport prediction is essential in many applications, including adaptive streaming and

virtual/augmented reality, because it allows for efficient resource utilization and a better user experience.

In terms of precision, the study assesses each model's ability to predict the user's viewport. A higher Precision value indicates improved predictive performance. The results show that HEVERL outperforms the reference models in accurately predicting the user's viewport.

The study also examines the models' root mean square error (RMSE) and mean absolute error (MAE). These metrics are crucial in assessing the disparity between the predicted and actual viewport coordinates. Lower RMSE and MAE values indicate a higher level of predictive performance. The findings reveal that HEVERL exhibits lower RMSE and MAE than the reference models, suggesting that it delivers more accurate viewport predictions with fewer errors.

The study's results demonstrate that the HEVERL model is highly effective in viewport prediction. This model holds significant promise as a tool for optimizing resource allocation and enhancing the overall user experience in various applications that rely on accurate viewport prediction. Significantly, it surpasses the current reference models in terms of Precision, RMSE, and MAE.

4.3. Training time evaluation. Table 3 illustrates the performance of four algorithms (AEVE, GRU, GLVP, and HEVERL) across five datasets (Bar, Ocean, Po Riversides, Sofa, and Turtle). The performance metrics indicate that these algorithms yield favorable results, with average processing times below 100ms for the entire video. This demonstrates the algorithms' effectiveness in aiding decision-making processes.

Table 3. Training time overview

Methods	Bar	Ocean	Po. Riversides	Sofa	Turtle
AEVE	0.0953	0.0644	0.0722	0.0904	0.0933
GRU	0.1020	0.0766	0.0708	0.0921	0.0983
GLVP	0.1030	0.0951	0.0649	0.0954	0.0913
HEVERL	0.0885	0.0971	0.0863	0.1100	0.0782

However, it is critical to consider not only raw performance metrics but also the algorithm's consistency and stability. An algorithm that performs well on average but has a high degree of variability in results may be less desirable than one with slightly lower peak performance but is more stable and reliable.

Choosing the best algorithm is a nuanced decision based on the problem's requirements and constraints. If the goal is to maximize performance across all data sets, the HEVERL algorithm is the top choice. However, the algorithm's performance in specific data sets or use cases may be more relevant. A thorough understanding of the problem context and desired outcomes is

required before making a definitive recommendation on the best algorithm for training time evaluation.

5. Conclusions. In this paper, we tackle the difficult task of viewport prediction in the context of VR video streaming. The proposed solution outperformed four reference methods in several critical evaluation metrics, such as Precision, Root Mean Square Error (RMSE), and Mean Absolute Error. By accurately predicting the user's current and future viewport, the authors' approach has the potential to significantly improve VR content delivery, lowering latency and improving overall viewing quality. Accurate viewport prediction is a critical enabler for optimizing bandwidth utilization and selectively streaming high-quality content only for the regions of interest, ultimately increasing user satisfaction and engagement with more immersive and enjoyable VR services across various domains, such as gaming, education, and training.

Our strategy focuses on exploring and optimizing techniques to enhance the performance of Reinforcement Learning models within the VAS system, aiming to predict and improve user experience quality. Moving forward, we plan to conduct additional experiments to validate the effectiveness of this approach, while also investigating solutions for integrating and deploying these optimized models across real-world virtual reality platforms.

References

1. Pan X., Chen X., Zhang Q., Li N. Model predictive control: A rein-forcement learning-based approach. Journal of Physics: Conference Series. IOP Publishing. 2022. vol. 2203. no. 1. DOI: 10.1088/1742-6596/2203/1/012058.
2. Feng X., Swaminathan V., Wei S. Viewport prediction for live 360-degree mobile video streaming using user-content hybrid motion tracking. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies. 2019. vol. 3. no. 2. pp. 1–22. DOI: 10.1145/3328914.
3. Nguyen H., Dao T.N., Pham N.S., Dang T.L., Nguyen T.D., Truong T.H. An accurate viewport estimation method for 360 video streaming using deep learning. EAI Endorsed Transactions on Industrial Networks and Intelligent Systems. 2022. vol. 9. no. 4. DOI: 10.4108/eetinis.v9i4.2218.
4. Nguyen D. An evaluation of viewport estimation methods in 360-degree video streaming. 7th International Conference on Business and Industrial Research (ICBIR). IEEE, 2022. pp. 161–166. DOI: 10.1109/ICBIR54589.2022.9786513.
5. Nguyen V.H., Pham N.N., Truong C.T., Bui D.T., Nguyen H.T., Truong T.H. Retina-based quality assessment of tile-coded 360-degree videos. EAI Endorsed Transactions on Industrial Networks and Intelligent Systems. 2022. vol. 9. no. 32. DOI: 10.4108/eetinis.v9i32.1058.
6. Lee E.-J., Jang Y.J., Chung M. When and how user comments affect news readers' personal opinion: perceived public opinion and perceived news position as mediators. Digital Journalism. 2020. vol. 9. no. 1. pp. 42–63. DOI: 10.1080/21670811.2020.1837638.
7. Nguyen H.V., Tan N., Quan N.H., Huong T.T., Phat N.H. Building a chatbot system to analyze opinions of english comments. Informatics and Automation. 2023. vol. 22. no. 2.

pp. 289–315.
8. Raja U.S., Carrico A.R. A qualitative exploration of individual experiences of environmental virtual reality through the lens of psychological distance. Environmental Communication. 2021. vol. 15. no. 5. pp. 594–609. DOI: 10.1080/17524032.2020.1871052.
9. Jiang Z., Zhang X., Xu Y., Ma Z., Sun J., Zhang Y. Reinforcement learning based rate adaptation for 360-degree video streaming. IEEE Transactions on Broadcasting. 2021. vol. 67. no. 2. pp. 409–423. DOI: 10.1109/TBC.2020.3028286.
10. Nguyen V.H., Bui D.T., Tran T.L., Truong C.T., Truong T.H. Scalable and resilient 360-degree-video adaptive streaming over http/2against sudden network drops. Computer Communications. 2024. vol. 216. pp. 1–15. DOI: 10.1016/j.comcom.2024.01.001.
11. Kan N., Zou J., Li C., Dai W., Xiong H. Rapt360: Reinforcement learning-based rate adaptation for 360-degree video streaming with adaptive prediction and tiling. IEEE Transactions on Circuits and Systems for Video Technology. 2022. vol. 32. no. 3. pp. 1607–1623. DOI: 10.1109/TCSVT.2021.3076585.
12. Hung N.V., Chien T.D., Ngoc N.P., Truong T.H. Flexible http-based video adaptive streaming for good QoE during sudden bandwidth drops. EAI Endorsed Transactions on Industrial Networks and Intelligent Systems. 2023. vol. 10. no. 2. DOI: 10.4108/eetinis.v10i2.2994.
13. Wong E.S., Wahab N.H.A., Saeed F., Alharbi N. 360-degree video bandwidth reduction: Technique and approaches comprehensive review. Applied Sciences. 2022. vol. 12. no. 15. DOI: 10.3390/app12157581.
14. Lampropoulos G., Barkoukis V., Burden K., Anastasiadis T. 360-degree video in education: An overview and a comparative social media data analysis of the last decade. Smart Learning Environments. 2021. vol. 8. DOI: 10.1186/s40561-021-00165-8.
15. Ng K.-T., Chan S.-C., Shum H.-Y. Data compression and transmission aspects of panoramic videos. IEEE Transactions on Circuits and Systems for Video Technology. 2005. vol. 15. no. 1. pp. 82–95. DOI: 10.1109/TCSVT.2004.839989.
16. Xie L., Xu Z., Ban Y., Zhang X., Guo Z. 360ProbDASH: Improving QoE of 360 video streaming using tile-based http adaptive streaming. Proceedings of the 25th ACM international conference on Multimedia. 2017. pp. 315–323. DOI: 10.1145/3123266.3123291.
17. Hosseini M., Swaminathan V. Adaptive 360 VR video streaming: Divide and conquer. IEEE International Symposium on Multimedia (ISM). IEEE, 2016. pp. 107–110.
18. El-Ganainy T., Hefeeda M. Streaming virtual reality content. arXiv preprint arXiv:1612.08350. 2016. DOI: 10.48550/arXiv.1612.08350.
19. Xu M., Song Y., Wang J., Qiao M., Huo L., Wang Z. Predicting head movement in panoramic video: A deep reinforcement learning approach. IEEE transactions on pattern analysis and machine intelligence. 2019. vol. 41. no. 11. pp. 2693–2708. DOI: 10.1109/TPAMI.2018.2858783.
20. Hu H.-N., Lin Y.-C., Liu M.-Y., Cheng H.-T., Chang Y.-J., Sun M. Deep 360 pilot: Learning a deep agent for piloting through 360deg sports videos. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017. pp. 1396–1405.
21. Bao Y., Wu H., Zhang T., Ramli A.A., Liu X. Shooting a moving target: Motion-prediction-based transmission for 360-degree videos. IEEE International Conference on Big Data. IEEE. 2016. pp. 1161–1170. DOI: 10.1109/BigData.2016.7840720.
22. Petrangeli S., Swaminathan V., Hosseini M., De Turck F. An http/2-based adaptive streaming framework for 360 virtual reality videos. Proceedings of the 25th ACM international conference on Multimedia. 2017. pp. 306–314. DOI: 10.1145/3123266.3123453.

23. Hung N.V., Tien B.D., Anh T.T.T., Nam P.N., Huong T.T. An efficient approach to terminate 360-video stream on http/3. AIP Conference Proceedings. AIP Publishing. 2023. vol. 2909. no. 1.
24. Yu J., Liu Y. Field-of-view prediction in 360-degree videos with attention-based neural encoder-decoder networks. Proceedings of the 11th ACM Workshop on Immersive Mixed and Virtual Environment Systems. 2019. pp. 37–42. DOI: 10.1145/3304113.3326118.
25. Park S., Bhattacharya A., Yang Z., Das S.R., Samaras D. Mosaic: Advancing user quality of experience in 360-degree video streaming with machine learning. IEEE Transactions on Network and Service Management. 2021. vol. 18. no. 1. pp. 1000–1015. DOI: 10.1109/TNSM.2021.3053183.
26. Lee D., Choi M., Lee J. Prediction of head movement in 360-degree videos using attention model. Sensors. 2021. vol. 21. no. 11. DOI: 10.3390/s21113678.
27. Chen X., Kasgari A.T.Z., Saad W. Deep learning for content-based personalized viewport prediction of 360-degree VR videos. IEEE Networking Letters. 2020. vol. 2. no. 2. pp. 81–84. DOI: 10.1109/LNET.2020.2977124.
28. Vielhaben J., Camalan H., Samek W., Wenzel M. Viewport forecasting in 360 virtual reality videos with machine learning. IEEE international conference on artificial intelligence and virtual reality (AIVR). IEEE. 2019. pp. 74–747. DOI: 10.1109/AIVR46125.2019.00020.
29. Uddin M.M., Park J. Machine learning model evaluation for 360°video caching. IEEE World AI IoT Congress (AIIoT). IEEE. 2022. pp. 238–244. DOI: 10.1109/AIIoT54504.2022.9817292.
30. Fan C.-L., Yen S.-C., Huang C.-Y., Hsu C.-H. Optimizing fixation prediction using recurrent neural networks for 360° video streaming in head-mounted virtual reality. IEEE Transactions on Multimedia. 2020. vol. 22. no. 3. pp. 744–759. DOI: 10.1109/TMM.2019.2931807.
31. Yaqoob A., Bi T., Muntean G.-M. A survey on adaptive 360 video streaming: Solutions, challenges and opportunities. IEEE Communications Surveys and Tutorials. 2020. vol. 22. no. 4. pp. 2801–2838. DOI: 10.1109/COMST.2020.3006999.
32. Liu X. Deng Y. Learning-based prediction, rendering and association optimization for mec-enabled wireless virtual reality (VR) networks. IEEE Transactions on Wireless Communications. 2021. vol. 20. no. 10. pp. 6356–6370. DOI: 10.1109/TWC.2021.3073623.
33. Gadaleta M., Chiariotti F., Rossi M., Zanella A. D-DASH: A deep q-learning framework for dash video streaming. IEEE Transactions on Cognitive Communications and Networking. 2017. vol. 3. no. 4. pp. 703–718. DOI: 10.1109/TCCN.2017.2755007.
34. Souane N., Bourenane M., Douga Y. Deep reinforcement learning-based approach for video streaming: Dynamic adaptive video streaming over HTTP. Applied Sciences. 2023. vol. 13. no. 21. DOI: 10.3390/app132111697.
35. Xie Y., Zhang Y., Lin T. Deep curriculum reinforcement learning for adaptive 360 ○ video streaming with two-stage training. IEEE Transactions on Broadcasting. 2023. vol. 70. no. 2. pp. 441–452. DOI: 10.1109/tbc.2023.3334137.
36. Du L., Zhuo L., Li J., Zhang J., Li X., Zhang H. Video quality of experience metric for dynamic adaptive streaming services using dash standard and deep spatial-temporal representation of video. Applied Sciences. 2020. vol. 10. no. 5. DOI: 10.3390/app10051793.
37. Mao H., Chen S., Dimmery D., Singh S., Blaisdell D., Tian Y., Alizadeh M., Bakshy E. Real-world video adaptation with reinforcement learning. arXiv preprint arXiv:2008.12858. 2020. DOI: 10.48550/arXiv.2008.12858.

38. David E.J., Gutiérrez J., Coutrot A., Da Silva M.P., Callet P.L. A dataset of head and eye movements for 360 videos. Proceedings of the 9th ACM Multimedia Systems Conference. 2018. pp. 432–437.

39. Wu C., Zhang R., Wang Z., Sun L. A spherical convolution approach for learning long term viewport prediction in 360 immersive video. Proceedings of the AAAI Conference on Artificial Intelligence. 2020. vol. 34. no. 01. pp. 14003–14040. DOI: 10.1609/aaai.v34i01.7377.

Nguyen Viet Hung — Lecturer, Faculty of information technology, East Asia University of Technology. Research interests: multimedia communications, network security, artificial intelligence, traffic engineering in next-generation networks, QoE/QoS guarantee for network services, green networking, applications. The number of publications — 23. hungnv@eaut.edu.vn; Ky Phu - Ky Anh, Ha Tinh, Viet Nam; office phone: +84(098)911-2079.

Pham Tien Dat — Research assistant, East Asia University of Technology. Research interests: applications, networks. The number of publications — 1. 20212452@eaut.edu.vn; Vu Ninh – Kien Xuong, Thai Binh, Viet Nam; office phone: +84(036)239-6558.

Nguyen Tan — Research assistant, East Asia University of Technology. Research interests: applications, data analysis. The number of publications — 3. tan25102000@gmail.com; Trung Dung - Tien Lu, Hung Yen, Viet Nam; office phone: +84(035)919-0216.

Nguyen Anh Quan — Research assistant, East Asia University of Technology. Research interests: applications, networks. The number of publications — 1. anhq46724@gmail.com; Gia Lam, Hanoi, Viet Nam; office phone: +84(096)278-4293.

Le Thi Huyen Trang — Lecturer, Faculty of information technology, East Asia University of Technology. Research interests: multimedia communications, database management systems, artificial intelligence, applications. The number of publications — 2. tranglth@eaut.edu.vn; Phuong Tri - Thi Tran Phung - Dan Phuong, Hanoi, Viet Nam; office phone: +84(032)889-9334.

Le Mai Nam — Lecturer, Faculty of information technology, East Asia University of Technology. Research interests: software engineering, optimization mathematics, applications. The number of publications — 1. namlm@eaut.edu.vn; Phuong Trung - Thanh Oai, Hanoi, Viet Nam; office phone: +84(098)208-2117.

УДК 006.72 **DOI** 10.15622/ia.24.1.11

Н. Хунг, Ф.Т. Дат, Н. Тан, Н.А. Куан, Л. Транг, Л.М. Нам,

ОЦЕНКА ОБЛАСТИ ПРОСМОТРА С ИСПОЛЬЗОВАНИЕМ ОБУЧЕНИЯ С ПОДКРЕПЛЕНИЕМ ДЛЯ ПОТОКОВОЙ ПЕРЕДАЧИ ВИДЕО В ФОРМАТЕ 360 ГРАДУСОВ

Хунг Н., Дат Ф.Т., Тан Н., Куан Н.А., Транг Л., Нам Л.М. **Оценка области просмотра с использованием обучения с подкреплением для потоковой передачи видео в формате 360 градусов.**

Аннотация. Видеоконтент в формате 360 градусов стал ключевым компонентом в средах виртуальной реальности, предлагая зрителям захватывающий и увлекательный опыт. Однако потоковая передача такого комплексного видеоконтента сопряжена со значительными трудностями, обусловленными существенными размерами файлов и переменчивыми сетевыми условиями. Для решения этих проблем в качестве перспективного решения, направленного на снижение нагрузки на пропускную способность сети, появилась адаптивная потоковая передача просмотра. Эта технология предполагает передачу видео более низкого качества для периферийных зон просмотра, а высококачественный контент – для конкретной зоны просмотра, на которую активно смотрит пользователь. По сути, это требует точного прогнозирования направления просмотра пользователя и повышения качества этого конкретного сегмента, что подчеркивает значимость адаптивной потоковой передачи просмотра (VAS). Наше исследование углубляется в применение методов пошагового обучения для прогнозирования оценок, требуемых системой VAS. Таким образом, мы стремимся оптимизировать процесс потоковой передачи, обеспечивая высокое качество отображения наиболее важных фрагментов видео. Кроме того, наш подход дополняется тщательным анализом поведения движений головы и лица человека. Используя эти данные, мы разработали модель обучения с подкреплением, специально предназначенную для прогнозирования направлений взгляда пользователя и повышения качества изображения в целевых областях. Эффективность предлагаемого нами метода подтверждается нашими экспериментальными результатами, которые показывают значительные улучшения по сравнению с существующими эталонными методами. В частности, наш подход повышает метрику прецизионности на значения в диапазоне от 0,011 до 0,022. Кроме того, он снижает среднеквадратичную ошибку (RMSE) в диапазоне от 0,008 до 0,013, среднюю абсолютную ошибку (MAE) – от 0,012 до 0,018 и оценку F1 – от 0,017 до 0,028. Кроме того, мы наблюдаем увеличение общей точности с 2,79 до 16,98. Эти улучшения подчеркивают потенциал нашей модели для значительного улучшения качества просмотра в средах виртуальной реальности, делая потоковую передачу видео на 360 градусов более эффективной и удобной для пользователя.

Ключевые слова: движение головы и глаз, обучение с подкреплением, глубокое обучение, машинное обучение, потоковая передача видео, видео на 360 градусов.

Литература

1. Pan X., Chen X., Zhang Q., Li N. Model predictive control: A rein-forcement learning-based approach. Journal of Physics: Conference Series. IOP Publishing. 2022. vol. 2203. no. 1. DOI: 10.1088/1742-6596/2203/1/012058.
2. Feng X., Swaminathan V., Wei S. Viewport prediction for live 360-degree mobile video streaming using user-content hybrid motion tracking. Proceedings of the ACM on

Interactive, Mobile, Wearable and Ubiquitous Technologies. 2019. vol. 3. no. 2. pp. 1–22. DOI: 10.1145/3328914.

3. Nguyen H., Dao T.N., Pham N.S., Dang T.L., Nguyen T.D., Truong T.H. An accurate viewport estimation method for 360 video streaming using deep learning. EAI Endorsed Transactions on Industrial Networks and Intelligent Systems. 2022. vol. 9. no. 4. DOI: 10.4108/eetinis.v9i4.2218.
4. Nguyen D. An evaluation of viewport estimation methods in 360-degree video streaming. 7th International Conference on Business and Industrial Research (ICBIR). IEEE, 2022. pp. 161–166. DOI: 10.1109/ICBIR54589.2022.9786513.
5. Nguyen V.H., Pham N.N., Truong C.T., Bui D.T., Nguyen H.T., Truong T.H. Retina-based quality assessment of tile-coded 360-degree videos. EAI Endorsed Transactions on Industrial Networks and Intelligent Systems. 2022. vol. 9. no. 32. DOI: 10.4108/eetinis.v9i32.1058.
6. Lee E.-J., Jang Y.J., Chung M. When and how user comments affect news readers' personal opinion: perceived public opinion and perceived news position as mediators. Digital Journalism. 2020. vol. 9. no. 1. pp. 42–63. DOI: 10.1080/21670811.2020.1837638.
7. Nguyen H.V., Tan N., Quan N.H., Huong T.T., Phat N.H. Building a chatbot system to analyze opinions of english comments. Informatics and Automation. 2023. vol. 22. no. 2. pp. 289–315.
8. Raja U.S., Carrico A.R. A qualitative exploration of individual experiences of environmental virtual reality through the lens of psychological distance. Environmental Communication. 2021. vol. 15. no. 5. pp. 594–609. DOI: 10.1080/17524032.2020.1871052.
9. Jiang Z., Zhang X., Xu Y., Ma Z., Sun J., Zhang Y. Reinforcement learning based rate adaptation for 360-degree video streaming. IEEE Transactions on Broadcasting. 2021. vol. 67. no. 2. pp. 409–423. DOI: 10.1109/TBC.2020.3028286.
10. Nguyen V.H., Bui D.T., Tran T.L., Truong C.T., Truong T.H. Scalable and resilient 360-degree-video adaptive streaming over http/2against sudden network drops. Computer Communications. 2024. vol. 216. pp. 1–15. DOI: 10.1016/j.comcom.2024.01.001.
11. Kan N., Zou J., Li C., Dai W., Xiong H. Rapt360: Reinforcement learning-based rate adaptation for 360-degree video streaming with adaptive prediction and tiling. IEEE Transactions on Circuits and Systems for Video Technology. 2022. vol. 32. no. 3. pp. 1607–1623. DOI: 10.1109/TCSVT.2021.3076585.
12. Hung N.V., Chien T.D., Ngoc N.P., Truong T.H. Flexible http-based video adaptive streaming for good QoE during sudden bandwidth drops. EAI Endorsed Transactions on Industrial Networks and Intelligent Systems. 2023. vol. 10. no. 2. DOI: 10.4108/eetinis.v10i2.2994.
13. Wong E.S., Wahab N.H.A., Saeed F., Alharbi N. 360-degree video bandwidth reduction: Technique and approaches comprehensive review. Applied Sciences. 2022. vol. 12. no. 15. DOI: 10.3390/app12157581.
14. Lampropoulos G., Barkoukis V., Burden K., Anastasiadis T. 360-degree video in education: An overview and a comparative social media data analysis of the last decade. Smart Learning Environments. 2021. vol. 8. DOI: 10.1186/s40561-021-00165-8.
15. Ng K.-T., Chan S.-C., Shum H.-Y. Data compression and transmission aspects of panoramic videos. IEEE Transactions on Circuits and Systems for Video Technology. 2005. vol. 15. no. 1. pp. 82–95. DOI: 10.1109/TCSVT.2004.839989.
16. Xie L., Xu Z., Ban Y., Zhang X., Guo Z. 360ProbDASH: Improving QoE of 360 video streaming using tile-based http adaptive streaming. Proceedings

of the 25th ACM international conference on Multimedia. 2017. pp. 315–323. DOI: 10.1145/3123266.3123291.

17. Hosseini M., Swaminathan V. Adaptive 360 VR video streaming: Divide and conquer. IEEE International Symposium on Multimedia (ISM). IEEE, 2016. pp. 107–110.
18. El-Ganainy T., Hefeeda M. Streaming virtual reality content. arXiv preprint arXiv:1612.08350. 2016. DOI: 10.48550/arXiv.1612.08350.
19. Xu M., Song Y., Wang J., Qiao M., Huo L., Wang Z. Predicting head movement in panoramic video: A deep reinforcement learning approach. IEEE transactions on pattern analysis and machine intelligence. 2019. vol. 41. no. 11. pp. 2693–2708. DOI: 10.1109/TPAMI.2018.2858783.
20. Hu H.-N., Lin Y.-C., Liu M.-Y., Cheng H.-T., Chang Y.-J., Sun M. Deep 360 pilot: Learning a deep agent for piloting through 360deg sports videos. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017. pp. 1396–1405.
21. Bao Y., Wu H., Zhang T., Ramli A.A., Liu X. Shooting a moving target: Motion-prediction-based transmission for 360-degree videos. IEEE International Conference on Big Data. IEEE. 2016. pp. 1161–1170. DOI: 10.1109/BigData.2016.7840720.
22. Petrangeli S., Swaminathan V., Hosseini M., De Turck F. An http/2-based adaptive streaming framework for 360 virtual reality videos. Proceedings of the 25th ACM international conference on Multimedia. 2017. pp. 306–314. DOI: 10.1145/3123266.3123453.
23. Hung N.V., Tien B.D., Anh T.T.T., Nam P.N., Huong T.T. An efficient approach to terminate 360-video stream on http/3. AIP Conference Proceedings. AIP Publishing. 2023. vol. 2909. no. 1.
24. Yu J., Liu Y. Field-of-view prediction in 360-degree videos with attention-based neural encoder-decoder networks. Proceedings of the 11th ACM Workshop on Immersive Mixed and Virtual Environment Systems. 2019. pp. 37–42. DOI: 10.1145/3304113.3326118.
25. Park S., Bhattacharya A., Yang Z., Das S.R., Samaras D. Mosaic: Advancing user quality of experience in 360-degree video streaming with machine learning. IEEE Transactions on Network and Service Management. 2021. vol. 18. no. 1. pp. 1000–1015. DOI: 10.1109/TNSM.2021.3053183.
26. Lee D., Choi M., Lee J. Prediction of head movement in 360-degree videos using attention model. Sensors. 2021. vol. 21. no. 11. DOI: 10.3390/s21113678.
27. Chen X., Kasgari A.T.Z., Saad W. Deep learning for content-based personalized viewport prediction of 360-degree VR videos. IEEE Networking Letters. 2020. vol. 2. no. 2. pp. 81–84. DOI: 10.1109/LNET.2020.2977124.
28. Vielhaben J., Camalan H., Samek W., Wenzel M. Viewport forecasting in 360 virtual reality videos with machine learning. IEEE international conference on artificial intelligence and virtual reality (AIVR). IEEE. 2019. pp. 74–747. DOI: 10.1109/AIVR46125.2019.00020.
29. Uddin M.M., Park J. Machine learning model evaluation for 360°video caching. IEEE World AI IoT Congress (AIIoT). IEEE. 2022. pp. 238–244. DOI: 10.1109/AIIoT54504.2022.9817292.
30. Fan C.-L., Yen S.-C., Huang C.-Y., Hsu C.-H. Optimizing fixation prediction using recurrent neural networks for 360° video streaming in head-mounted virtual reality. IEEE Transactions on Multimedia. 2020. vol. 22. no. 3. pp. 744–759. DOI: 10.1109/TMM.2019.2931807.
31. Yaqoob A., Bi T., Muntean G.-M. A survey on adaptive 360 video streaming: Solutions, challenges and opportunities. IEEE Communications Surveys and Tutorials. 2020. vol. 22. no. 4. pp. 2801–2838. DOI: 10.1109/COMST.2020.3006999.

32. Liu X. Deng Y. Learning-based prediction, rendering and association optimization for mec-enabled wireless virtual reality (VR) networks. IEEE Transactions on Wireless Communications. 2021. vol. 20. no. 10. pp. 6356–6370. DOI: 10.1109/TWC.2021.3073623.
33. Gadaleta M., Chiariotti F., Rossi M., Zanella A. D-DASH: A deep q-learning framework for dash video streaming. IEEE Transactions on Cognitive Communications and Networking. 2017. vol. 3. no. 4. pp. 703–718. DOI: 10.1109/TCCN.2017.2755007.
34. Souane N., Bourenane M., Douga Y. Deep reinforcement learning-based approach for video streaming: Dynamic adaptive video streaming over HTTP. Applied Sciences. 2023. vol. 13. no. 21. DOI: 10.3390/app132111697.
35. Xie Y., Zhang Y., Lin T. Deep curriculum reinforcement learning for adaptive 360 ○ video streaming with two-stage training. IEEE Transactions on Broadcasting. 2023. vol. 70. no. 2. pp. 441–452. DOI: 10.1109/tbc.2023.3334137.
36. Du L., Zhuo L., Li J., Zhang J., Li X., Zhang H. Video quality of experience metric for dynamic adaptive streaming services using dash standard and deep spatial-temporal representation of video. Applied Sciences. 2020. vol. 10. no. 5. DOI: 10.3390/app10051793.
37. Mao H., Chen S., Dimmery D., Singh S., Blaisdell D., Tian Y., Alizadeh M., Bakshy E. Real-world video adaptation with reinforcement learning. arXiv preprint arXiv:2008.12858. 2020. DOI: 10.48550/arXiv.2008.12858.
38. David E.J., Gutiérrez J., Coutrot A., Da Silva M.P., Callet P.L. A dataset of head and eye movements for 360 videos. Proceedings of the 9th ACM Multimedia Systems Conference. 2018. pp. 432–437.
39. Wu C., Zhang R., Wang Z., Sun L. A spherical convolution approach for learning long term viewport prediction in 360 immersive video. Proceedings of the AAAI Conference on Artificial Intelligence. 2020. vol. 34. no. 01. pp. 14003–14040. DOI: 10.1609/aaai.v34i01.7377.

Хунг Нгуен Вьет — преподаватель, факультет информационных технологий, Восточноазиатский технологический университет. Область научных интересов: мультимедийные коммуникации, сетевая безопасность, искусственный интеллект, организация трафика в сетях нового поколения, гарантия качества сетевых услуг, экологичные сети, приложения. Число научных публикаций — 23. hungnv@eaut.edu.vn; Ки Фу - Ки Ань, Хатинь, Вьетнам; р.т.: +84(098)911-2079.

Дат Фам Тянь — научный сотрудник, Восточноазиатский технологический университет. Область научных интересов: приложения, сети. Число научных публикаций — 1. 20212452@eaut.edu.vn; Ву Нинь – Кьен Сюонг, Тхайбинь, Вьетнам; р.т.: +84(036)239-6558.

Тан Нгуен — научный сотрудник, Восточноазиатский технологический университет. Область научных интересов: приложения, анализ данных. Число научных публикаций — 3. tan25102000@gmail.com; Чынг Зунг - Тиен Лу, Хынгйен, Вьетнам; р.т.: +84(035)919-0216.

Куан Нгуен Ань — научный сотрудник, Восточноазиатский технологический университет. Область научных интересов: приложения, сети. Число научных публикаций — 1. anhq46724@gmail.com; Зялай, Ханой, Вьетнам; р.т.: +84(096)278-4293.

Транг Ле Тхи Хуэйен — преподаватель, факультет информационных технологий, Восточноазиатский технологический университет. Область научных интересов: мультимедийные коммуникации, системы управления базами данных, искусственный интеллект, приложения. Число научных публикаций — 2. tranglth@eaut.edu.vn; Фуонг Три - Тхи Тран Пхунг - Дан Фуонг, Ханой, Вьетнам; р.т.: +84(032)889-9334.

Нам Ле Май — преподаватель, факультет информационных технологий, Восточноазиатский технологический университет. Область научных интересов: разработка программного обеспечения, математика оптимизации, прикладные программы. Число научных публикаций — 1. namlm@eaut.edu.vn; Фуонг Чунг - Тхань Оай, Ханой, Вьетнам; р.т.: +84(098)208-2117.

DOI 10.15622/ia.24.1.12

A. Ageev, A. Konstantinov, L. Utkin

ADA-NAF: SEMI-SUPERVISED ANOMALY DETECTION BASED ON THE NEURAL ATTENTION FOREST

Ageev A., Konstantinov A., Utkin L. **ADA-NAF: Semi-Supervised Anomaly Detection Based on the Neural Attention Forest.**

Abstract. In this study, we present a novel model called ADA-NAF (Anomaly Detection Autoencoder with the Neural Attention Forest) for semi-supervised anomaly detection that uniquely integrates the Neural Attention Forest (NAF) architecture which has been developed to combine a random forest classifier with a neural network computing attention weights to aggregate decision tree predictions. The key idea behind ADA-NAF is the incorporation of NAF into an autoencoder structure, where it implements functions of a compressor as well as a reconstructor of input vectors. Our approach introduces several technical advances. First, a proposed end-to-end training methodology over normal data minimizes the reconstruction errors while learning and optimizing neural attention weights to focus on hidden features. Second, a novel encoding mechanism leverages NAF's hierarchical structure to capture complex data patterns. Third, an adaptive anomaly scoring framework combines the reconstruction errors with the attention-based feature importance. Through extensive experimentation across diverse datasets, ADA-NAF demonstrates superior performance compared to state-of-the-art methods. The model shows particular strength in handling high-dimensional data and capturing subtle anomalies that traditional methods often do not detect. Our results validate the ADA-NAF's effectiveness and versatility as a robust solution for real-world anomaly detection challenges with promising applications in cybersecurity, industrial monitoring, and healthcare diagnostics. This work advances the field by introducing a novel architecture that combines the interpretability of attention mechanisms with the powerful feature learning capabilities of autoencoders.

Keywords: anomaly detection, random forest, attention mechanism, neural attention forest.

1. Introduction.

Anomaly detection is a critical task in data analysis, focusing on identifying rare events or observations that significantly deviate from the norm in a given system or dataset [1]. Its importance spans numerous fields, including finance, healthcare, manufacturing, and network security, where anomalies often indicate critical issues or potential threats [2]. These deviations can arise from various sources such as measurement errors, deliberate attacks, equipment malfunctions, or rare natural phenomena. Traditional anomaly detection methods, primarily based on rule-based systems or statistical techniques, often struggle with complex, high-dimensional data [3]. This limitation has led to the development of more sophisticated approaches, particularly in the realm of unsupervised and semi-supervised learning, where the algorithms learn to identify anomalies with limited or no labeled examples [4–6].

Unsupervised anomaly detection is particularly valuable when datasets lack labeled anomalies or when the types of anomalies are not well-defined. These techniques aim to learn the underlying structure and distribution of

the data, enabling the identification of instances that diverge from learned patterns [7]. The applications of such methods are wide-ranging, encompassing areas like cybersecurity, fraud detection, network monitoring, manufacturing quality control, medical diagnostics, and environmental monitoring [8, 9]. In recent years, significant advancements have been made in deep learning-based approaches to anomaly detection. Autoencoders, a type of artificial neural network, have shown remarkable success in this domain [10]. By learning to reconstruct input data, autoencoders can effectively capture underlying patterns and dependencies, allowing for the identification of anomalies based on reconstruction errors.

In this article, we propose a novel semi-supervised anomaly detection approach called ADA-NAF (Anomaly Detection Autoencoder with Neural Attention Forest). This model uniquely adapts the Neural Attention Forest [11] as an autoencoder to learn representations of normal data. We benchmark ADA-NAF against prominent techniques such as Isolation Forest (IF) [12] and its Deep extension (DIF) [13].

The primary aim of this research is to address limitations in existing methods, particularly for semi-supervised scenarios and complex data distributions. ADA-NAF leverages the strengths of neural attention mechanisms and Random Forests within an autoencoder framework, offering potential improvements in accuracy and interpretability compared to current state-of-the-art methods.

Our key contributions are:

– Introduction of ADA-NAF, a novel anomaly detection model that implements a pretraining step using Random Forest to cluster feature vectors, enabling its use on unlabeled data.

– Development of a multi-head extension to ADA-NAF, enhancing robustness through the adaptation of multi-head attention mechanisms.

– Comprehensive evaluation of ADA-NAF's performance against other models on benchmark datasets.

The paper is structured as follows: Section 2 reviews related work, Section 3 provides background on autoencoders and the Neural Attention Forest model, Sections 4 and 5 detail our proposed approach, Sections 6 and 7 present the experimental setup and results, and Section 8 concludes the paper.

2. Related works. Anomaly detection techniques. The field of anomaly detection has evolved from traditional statistical methods to more advanced machine learning and deep learning approaches. The authors in [14] provide a comprehensive survey of network anomaly detection techniques, covering statistical, classification-based, and clustering-based methods. Recent

advancements in deep learning have led to promising results in anomaly detection tasks [15, 16].

Deep learning in anomaly detection. Deep learning-based approaches have gained significant traction in anomaly detection research. These include self-supervised learning [17], One-Class Classification (OCC) [18], and specialized techniques for time series anomaly detection [19]. Chalapathy and Chawla [20] offer a thorough survey of deep learning methods for anomaly detection, highlighting their effectiveness across various domains.

Specialized applications. The versatility of deep learning in anomaly detection is evident in its application to diverse fields. For instance, the authors in [21] focus on anomaly detection in log data, while the authors [22] explore GAN-based methods. Suarez and Naval [23] investigate deep learning techniques for video anomaly detection, and Tschuchnig and Gadermayr [24] review anomaly detection methods in medical imaging, specifically for brain MRI.

Attention mechanisms in anomaly detection. Attention mechanisms, which enable models to focus on the most relevant parts of the data, have been successfully applied to anomaly detection tasks [25, 26]. Notable examples include:

– Study [28] proposes a GAN-based approach with an attention mechanism for detecting anomalies in semiconductor production sensor data.

– The authors in study [29] introduce a deep learning model with an attention mechanism for anomaly detection in vector magnetic field data.

– Paper [30] presents a graph-based anomaly detection algorithm utilizing an attention mechanism.

These attention-based models differ in their underlying architectures and data representations. The GAN-based approach [28] focuses on generating normal data patterns, [29] and [30] present different approaches to anomaly detection using attention mechanisms. The model in [29] uses a multi-layer neural network for vector magnetic field data, with its depth determined by input complexity and desired feature abstraction. In contrast, [30] introduces a graph-based algorithm, where depth refers to the number of graph convolution layers. While both utilize attention mechanisms, they differ fundamentally in data representation and processing: [29] uses vector inputs and traditional neural networks, while [30] leverages graph structures to capture relational data information, a capability that traditional deep learning models may lack. The depth and complexity of these models are tailored to their specific application domains and data types.

Autoencoder-based approaches. Autoencoders have proven particularly effective for anomaly detection in high-dimensional and

unbalanced datasets [5]. The authors in [6] explore the use of autoencoder ensembles to enhance anomaly detection accuracy. These approaches leverage the autoencoder's ability to learn compact representations of normal data, facilitating the identification of anomalies through reconstruction errors.

Random Forest in anomaly detection. The integration of attention mechanisms with Random Forests has emerged as a promising direction in anomaly detection research. Utkin and Konstantinov [32, 33] introduced the Attention-based Random Forest, which assigns attention weights to data in tree leaves using neural networks. This approach, framed within Nadaraya-Watson kernel regression [34], offers a novel perspective on combining tree-based methods with neural attention mechanisms. Our work, ADA-NAF, builds upon these foundations, particularly the Neural Attention Forest framework, to create a unique autoencoder-based model for semi-supervised anomaly detection. By integrating the strengths of Random Forests, neural attention, and autoencoder architectures, ADA-NAF aims to address the challenges of detecting anomalies in complex, high-dimensional data with limited labeled examples.

3. Preliminaries

3.1. Autoencoders for Anomaly Detection. Autoencoders are neural networks designed to learn the internal representation of data by training it on input and reconstructing itself as output [17, 35, 36]. One of the important applications of autoencoders is anomaly detection, that is, the detection of unusual or anomalous patterns in data.

Let there be training data $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$, where $\mathbf{x}_i \in \mathbb{R}^d$ is a d-dimensional vector, N is a count of training data. The task of anomaly detection is to detect anomalous samples for this training dataset. Let $\mathbf{y} = (y_1, y_2, \ldots, y_N)$ be the label vector, where $y_i \in \{0, 1\}$ indicates whether $\mathbf{x}_i$ is an anomaly ($y_i = 1$) or a normal pattern ($y_i = 0$). The problem of anomaly detection can be formulated as a supervised learning problem, where it is required to build a model $f : \mathbb{R}^d \to \{0, 1\}$ that will classify new samples as anomalies or normal. Autoencoders use only normal data for training.

Autoencoders are a class of neural networks that allow you to model non-linear dependencies and extract important features from input data. They consist of two main components: an encoder and a decoder. The encoder transforms the input data into an internal representation called the encoding, and the decoder restores the data from the encoding back to its original space.

Mathematically, let $\mathbf{x} \in \mathbb{R}^d$ be an input vector of dimension d and $\mathbf{z} \in \mathbb{R}^h$ be an encoding vector of dimension h, where $h < d$. The encoder is modeled as a function $E : \mathbb{R}^d \to \mathbb{R}^h$ and the decoder is modeled as a function

$D : \mathbb{R}^h \to \mathbb{R}^d$. Then the process of encoding and decoding can be written as follows:

$$\mathbf{z} = E(\mathbf{x}), \quad \hat{\mathbf{x}} = D(\mathbf{z}), \tag{1}$$

where $\hat{\mathbf{x}} \in \mathbb{R}^d$ is the reconstruction of the input vector $\mathbf{x} \in \mathbb{R}^d$.

To train the autoencoder, we use a recovery method that minimizes the recovery error between the input data and its reconstruction. Denote the loss function as $L(\mathbf{x}, \hat{\mathbf{x}})$, which measures the discrepancy between the input vector and its reconstruction. Popular loss functions are root mean square error (MSE) and binary cross entropy (BCE). The goal is to minimize this loss function.

In the context of anomaly detection, autoencoders can be used to detect abnormal patterns based on differences between normal and abnormal data. Training the autoencoder on a set of only normal samples allows the model to learn the characteristics of the normal data and create a model that will have a high reconstruction error for the anomalies since the anomalies will differ from the expected normal distribution.

One of the popular approaches to anomaly detection using autoencoders is the threshold approach. After training the autoencoder on normal data, we can use it to reconstruct new samples and calculate the reconstruction error. We then set a threshold ϵ above which samples are considered anomalous. Samples for which the reconstruction error exceeds the threshold are considered anomalies.

Formally, let $\mathbf{x}_{\text{test}} \in \mathbb{R}^d$ be a new sample, and its reconstruction is denoted as $\hat{\mathbf{x}}_{\text{test}} \in \mathbb{R}^d$. Then the anomaly detection algorithm can be written as follows:

$$\text{Anomaly Score}(\mathbf{x}_{\text{test}}) = L(\mathbf{x}_{\text{test}}, \hat{\mathbf{x}}_{\text{test}}). \tag{2}$$

If Anomaly Score$(\mathbf{x}_{\text{test}}) > \tau$, where $\tau \in \mathbb{R}$ is the given threshold, then sample $\mathbf{x}_{\text{test}}$ is considered anomalous.

However, the threshold approach has its limitations, as determining the optimal threshold can be challenging. Several methods can be employed to address this issue:

– Statistical methods: using measures of the reconstruction error.

– Percentile-based approach: setting the threshold at a specific percentile (e.g., 95th or 99th) of the error distribution.

– ROC curve analysis: optimizing the threshold using labeled data to maximize a chosen metric.

– Cross-validation: determining a robust threshold that generalizes across data subsets.

– Adaptive thresholding: implementing dynamic thresholds adjusting to recent data patterns.

– Ensemble methods: combining multiple thresholds for a more robust decision boundary.

The choice of method depends on the application context, available data, and the relative costs of false positives versus false negatives.

3.2. The Neural Attention Forest. The Neural Attention Forest (NAF) is a novel approach that integrates the attention mechanism into the Random Forest [11]. The primary objective is to assign attention weights, computed by neural networks of a specific architecture, to data in the leaves of decision trees and to the Random Forest itself. This is achieved within the framework of the Nadaraya-Watson kernel regression.

The attention mechanism in NAF is implemented by two distinct parts of the neural network:

1. **Tree-specific Attention**. The first part consists of neural networks with shared weights, trained for all trees. This part computes the attention weights for data in the leaves. For each tree, the attention operation is implemented as:

$$\mathbf{A}_k(\mathbf{x}) = \sum_{j \in J_k(\mathbf{x})} \alpha(\mathbf{x}, \mathbf{x}_j, \theta)\mathbf{x}_j, \tag{3}$$

$$B_k(\mathbf{x}) = \sum_{j \in J_k(\mathbf{x})} \alpha(\mathbf{x}, \mathbf{x}_j, \theta)y_j, \tag{4}$$

where $\mathbf{A}_k(\mathbf{x}) \in \mathbb{R}^d$, $B_k(\mathbf{x}) \in \mathbb{R}$, $\mathbf{x} \in \mathbb{R}^d$ and $J_k(\mathbf{x})$ represents the set of indices for which the feature vectors $\mathbf{x}_j \in \mathbb{R}^d$ fall into the same leaf of the k-th tree as $\mathbf{x} \in \mathbb{R}^d$, $y_j \in \mathbb{R}$ is the output corresponding to the feature vector $\mathbf{x}_j$, θ is a parameter of neural network. The attention weight $\alpha(\mathbf{x}, \mathbf{x}_j, \theta) \in \mathbb{R}$ is calculated by a neural network with θ parameters.

2. **Global Attention**. The second part of the neural network aggregates all the keys $\mathbf{A}_k(\mathbf{x})$ and values $B_k(\mathbf{x})$ from the tree-specific attention. The global attention operation is:

$$\hat{y} = \sum_{k=1}^{T} \beta(\mathbf{x}, \mathbf{A}_k(\mathbf{x}), \psi)B_k(\mathbf{x}), \tag{5}$$

where $\mathbf{A}_k(\mathbf{x})$ and $B_k(\mathbf{x})$ are the keys and values, respectively, computed for each tree, $\beta(\mathbf{x}, \mathbf{A}_k(\mathbf{x}), \psi) \in \mathbb{R}$ is the global attention weight calculated by a neural network with ψ params, and T is the total number of trees in the Random Forest.

It should be noted that the neural network here has a specific architecture with scaled dot-product score output to implement the attention mechanism. The first part computes the trainable attention weights for each tree's data. The second part then aggregates the weighted outputs to produce the final prediction.

4. ADA-NAF for Anomaly Detection. ADA-NAF model uniquely integrates the Neural Attention Forest framework into an autoencoder architecture for anomaly detection. The key idea is to leverage ADA-NAF as an autoencoder that compresses input vectors into encoded feature representations, and then reconstructs the original input.

A schematic depiction of leveraging ADA-NAF for anomaly detection is presented in Figure 1.

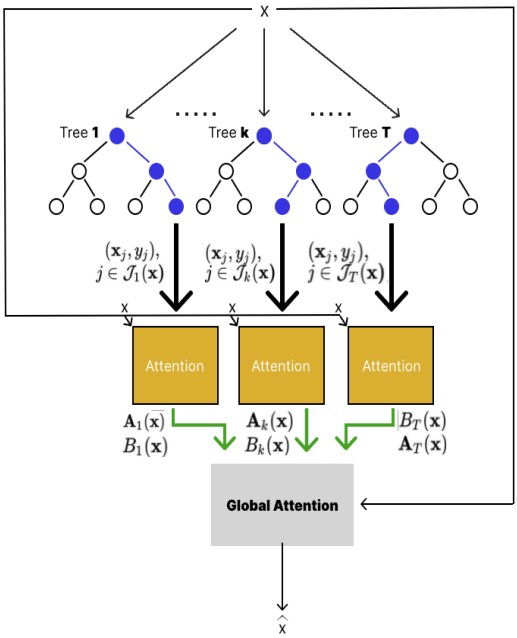


Fig. 1. The architecture of the proposed ADA-NAF model for anomaly detection

The input feature vector $\mathbf{x} \in \mathbb{R}^d$ passes through the Random Forest component to compute leaf node assignments $J_k(\mathbf{x})$ and attention-weighted aggregate representations $\mathbf{A}_k(\mathbf{x})$, $B_k(\mathbf{x})$ for each k tree. The global attention module uses these to produce a reconstructed vector $\hat{\mathbf{x}} \in \mathbb{R}^d$:

$$\hat{\mathbf{x}} = \sum_{k=1}^{T} \beta(\mathbf{x}, \mathbf{A}_k(\mathbf{x}), \psi) \cdot \mathbf{A}_k(\mathbf{x}). \tag{6}$$

The distance between the input $\mathbf{x}$ and reconstruction $\hat{\mathbf{x}}$ is computed, such as the Euclidean distance:

$$D(\mathbf{x}, \hat{\mathbf{x}}) = ||\mathbf{x} - \hat{\mathbf{x}}||_2. \tag{7}$$

If $D(\mathbf{x}, \hat{\mathbf{x}}) > \tau$, where τ is a threshold, then $\mathbf{x}$ is flagged as an anomaly. The threshold τ can be determined based on the distribution of distances for known non-anomalous data or through cross-validation.

The neural attention focuses on modeling normal data during training. At test time, anomalies result in larger reconstruction errors, allowing their detection. The key differentiating aspects from NAF are

– end-to-end training solely over normal data samples by minimizing reconstruction error;

– discovering a compressed feature encoding rather than performing supervised prediction;

– detecting anomalies based on deviation between inputs and reconstructions.

These structural modifications reshape the purpose of attention – instead of predictive performance, the focus is shifted towards the characterization of normality and subsequent identification of violations manifesting as anomalies at test time. This repurposing of the neural attention framework is the core innovation in ADA-NAF.

The general approach to training ADA-NAF for anomaly detection is shown in Figure 2:

1. Training the Random Forest component on a small labeled dataset:

– Dataset $\mathbf{X}_{labled} \in \mathbb{R}^{n \times d}$ contains n examples with dimension d with normal/anomalous class labels.

– Classification loss like cross-entropy is minimized:

$$\min_{\theta_{tree}} \mathcal{L}(\theta_{tree}) = \min_{\theta_{tree}} Loss_{CE}(\theta_{tree}, \mathbf{X}_{labled}, \mathbf{y}_{labled}), \tag{8}$$

where $\mathbf{y}_{labled} \in \mathbb{R}^n$ are labels from dataset $\mathbf{X}_{labled}$.

– The trained Random Forest is used to compute $J_k(\mathbf{x})$ - leaf indices for input vector $\mathbf{x}$.

2. Training ADA-NAF model parameters θ, ψ by minimizing reconstruction error on normal training set $\mathbf{X_{train}}$:

– X_{train} contains only normal class examples.
– Mean squared reconstruction error is minimized:

$$L(\theta, \psi) = \frac{1}{n} \sum_{\mathbf{x} \in \mathbf{X}_{train}} ||\mathbf{x} - \hat{\mathbf{x}}||^2. \tag{9}$$

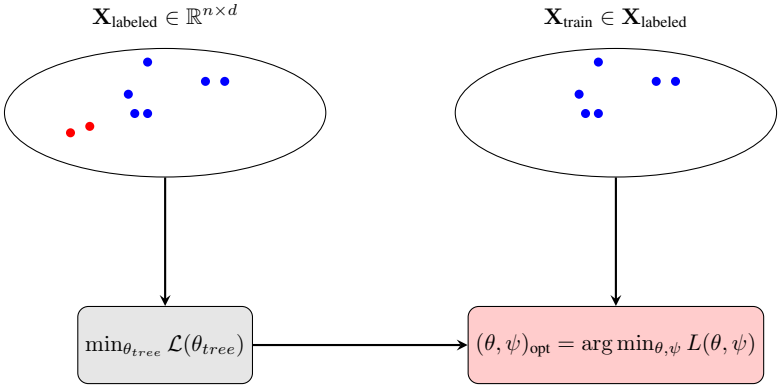


Fig. 2. ADA-NAF model training scheme for anomaly detection. Examples of normal data are shown in blue, and abnormal data in red. First, the Random Forest component is trained on a small labeled dataset. The entire ADA-NAF model is then trained to minimize the autoencoder reconstruction error using normal class examples only

The complete training and anomaly detection process for ADA-NAF is summarized in Listing 1. This algorithm outlines the two main stages of our approach: first, training the Random Forest component on a labeled subset of the data, and then training the whole ADA-NAF model on the normal data to minimize reconstruction error.

```
Input: X_train, X_test, number of trees T
Output: Anomaly scores

// Train the Random Forest
Train the Random Forest on a labeled subset of X_train
for each tree k in 1 to T:
```

```
    Compute leaf node assignments J_k(x) for all x in X_train

// Train ADA-NAF
Initialize neural attention weights θ, ψ
while not converged:
    for each x in X_train:
        Compute A_k(x) and B_k(x) using Eqs. (3) and (4)
        Compute reconstruction x̂ using Eq. (5)
        Update θ, ψ to minimize ||x - x̂||^2

// Detect anomalies
for each x in X_test:
    Compute reconstruction x̂
    Compute anomaly score as ||x - x̂||^2
```

Listing 1. Pseudocode of ADA-NAF training and anomaly detection algorithm

So the Random Forest component is first trained on labeled data, then the whole ADA-NAF model is fitted on unlabeled normal data by minimizing the autoencoder reconstruction error.

The neural network attention weights focus on normal examples during training. At test time, anomalies are then unlikely to be properly reconstructed, leading to larger errors that allow their detection.

Key advantages of the ADA-NAF anomaly detection approach include:

– Handles tabular data effectively through the Random Forest base model which stratifies the feature space.

– Learns a rich latent representation that captures boundaries between normal and anomalous data patterns.

– Leverages an initial labeled set to pre-train the Random Forest component for feature space stratification, then allows semi-supervised learning on unlabeled data for attention tuning.

– Provides local example-based explanations by selecting training points most influential for reconstructions based on learned attention weights. ADA-NAF's attention mechanisms assign importance weights to training examples. By analyzing the nearest neighbors of reconstructions according to these attention weights, we can extract influential examples that inform the model's predictions.

– Complementary to existing methods with competitive performance across various benchmark datasets.

We experimentally evaluate the proposed ADA-NAF anomaly detection method in Section 6.

5. Multi-Head ADA-NAF for Anomaly Detection. In the Multi-Head ADA-NAF (ADA-NAF-MH), we enhance the model by introducing multiple heads, as commonly done in transformer architectures [37]. Each attention head operates independently, with its own set of parameters, allowing for diverse attention patterns and reconstructions. As introduced in [37], multi-head attention projects the inputs into multiple subspaces and applies separate attention layers in parallel, before concatenating the outputs. This multi-head approach aims to capture complementary representations of the data, improving the model's robustness and accuracy.

Formally, for a given number of heads H, each head i is initialized with its own set of parameters $\theta^{(i)}$ and $\psi^{(i)}$.

Each head computes its own attention vectors as

$$\mathbf{A}_k^{(i)}(\mathbf{x}) = \sum_{j \in J_k(\mathbf{x})} \alpha^{(i)}(\mathbf{x}, \mathbf{x}_j, \theta^{(i)}) \mathbf{x}_j, \tag{10}$$

$$\mathbf{B}_k^{(i)}(\mathbf{x}) = \sum_{j \in J_k(\mathbf{x})} \alpha^{(i)}(\mathbf{x}, \mathbf{x}_j, \theta^{(i)}) y_j, \tag{11}$$

where $\alpha^{(i)}$ represents the attention weight computed by the i-th head.

The reconstruction for each head is given by:

$$\hat{\mathbf{x}}^{(i)} = \sum_{k=1}^{T} \beta^{(i)}(\mathbf{x}, \mathbf{A}_k^{(i)}(\mathbf{x}), \psi^{(i)}) \mathbf{B}_k^{(i)}(\mathbf{x}). \tag{12}$$

To aggregate the outputs of all heads, we compute an unweighted average:

$$\hat{\mathbf{x}} = \frac{1}{H} \sum_{i=1}^{H} \hat{\mathbf{x}}^{(i)}. \tag{13}$$

The model is trained to minimize the reconstruction loss on the normal training data:

$$L(\theta^{(1)}, ..., \theta^{(H)}, \psi^{(1)}, ..., \psi^{(H)}) = \sum_{\mathbf{x} \in X_{\text{train}}} ||\mathbf{x} - \hat{\mathbf{x}}||^2. \tag{14}$$

Specifically for anomaly detection, the multi-head extension enhances flexibility in capturing boundaries between normal and anomalous data points. Crucially, edge cases might be flagged by specialized heads finetuned through individual parameterizations. ADA-NAF-MHA architecture aims to fuse these signals into a unified detector of higher sensitivity.

6. ϵ-contamination attention regularization. We introduce an ϵ-contamination style regularization approach to impose robustness in the learned attention distributions while retaining sensitivity for anomaly detection. The global attention mechanism in ADA-NAF produces the reconstruction as:

$$\hat{\mathbf{x}} = \sum_{k=1}^{T} \beta(\mathbf{x}, \mathbf{A}_k(\mathbf{x}), \psi) \cdot \mathbf{A}_k(\mathbf{x}). \tag{15}$$

To enhance the robustness of this attention mechanism, we propose modifying the reconstruction process as follows:

$$\hat{\mathbf{x}}' = \sum_{k=1}^{T} ((1-\epsilon)\beta(\mathbf{x}, \mathbf{A}_k(\mathbf{x}), \psi) + \epsilon \cdot softmax(\mathbf{W})) \cdot \mathbf{A}_k(\mathbf{x}), \tag{16}$$

where $\mathbf{W} \in \mathbb{R}^T$ is a trainable parameter, randomly initialized. The contamination ratio ϵ controls the amount of mixing of $\mathbf{W}$ into the primary attention distribution β. This introduces a regularization effect that encourages the model to discover additional informative patterns beyond those identified by the original attention mechanism.

Experiments are conducted with ADA-NAF models trained using different fixed ϵ values.

We hypothesize that the ϵ-contamination attention regularization induces differing effects based on the ϵ level:

1. Small ϵ (0.1-0.2) introduces beneficial noise that improves stability and robustness to distortions without sacrificing sensitivity. Attention retains the focus on hidden features.
2. Large ϵ (>0.3) overwhelms useful signals, over-regularizing attention and reducing sensitivity to anomalies along with representations losing usefulness.
3. An optimal ϵ balances noise injection for desirable stability while avoiding dilution of attention selectivity. This optimal point is expected to be dataset and architecture-dependent.

7. Numerical experiments. The aim of this chapter is to provide a comprehensive evaluation of the proposed method using numerical experiments. The experiments are designed to demonstrate the effectiveness of the method in comparison to the other described in this article approaches, and to show the impact of various parameters on the performance of the method. In the experiments, we will compare the performance of the three models on a variety of datasets and use standard evaluation metrics such as AUC-ROC to assess the performance of each model. The results will be presented in the form of tables and graphs to allow for a clear and comprehensive comparison of the models.

Gradient descent is used for optimization with the following parameters: learning rate is 0.01, optimizer is AdamW, and the count epoch is 50.

ADA-NAF is implemented by means of software in Python. The software implementing ADA-NAF is available at https://github.com/AndreyAgeev/ada-naf.

7.1. Experimental Setup. In this section, we describe the dataset used in the experiments, the evaluation metrics, and the implementation details.

7.1.1. Datasets. We experiment over the following public anomaly detection benchmarks:

– **Arrhythmia** [38] – Collection of electrocardiogram (ECG) heartbeat segments from UCI Machine Learning Repository annotated with cardiac condition type.

– **Credit Card** [39] – Confidential credit card transactions dataset shared by a financial services company on Kaggle for analytics purposes and fraud detection.

– **Pima** [40] – Medical diagnostic measurements from Pima Indian diabetes patients released as part of open dataset collection by the National Institute for Diabetes and Digestive and Kidney Diseases.

– **Haberman** [41] – Historical dataset documenting breast cancer survival study featuring age of patients, year of surgery and number of detected axillary nodes. Hosted on UCI data repository.

– **Ionosphere** [42] – Radar data returns bounced off the ionosphere layer labeled as good or bad structures based on evidence of turbulence/stability patterns. Very common classification benchmark.

– **Seismic bumps** [43] – Recordings gathered from seismographic sensors in coal mines indicating warning signs of impending seismic bumps or just ambient tremors published as an open dataset.

– **Shuttle** [44] – System health data with sensor readings and component fault status during simulated space shuttle flights released by NASA for engineering challenges.

– **Annthyroid** [45] – Patient records tabulating biomarker readouts, test outcomes and diagnoses for differentiation of thyroid gland malfunction symptoms.

– **Bank Additional** [46] – Financial customer data on marketing campaign responses used for response modeling and fraud analysis hosted as a public dataset.

– **CelebA** [47] – Large-scale face image dataset with celebrity photos annotated for the presence/absence of multiple facial attributes like expressions, hair color, age, etc.

Table 1 shows a dataset information table.

For large real datasets, data slices are taken and instead of the full data. The slice size is indicated in the table with the description of the data. Before applying the methods, data preprocessing was carried out on some of the presented datasets, including data normalization and feature selection. The dataset preprocessing code is in https://github.com/AndreyAgeev/ada-naf.

Table 1. A brief introduction about the datasets

Dataset	normal	anomal	n feature
Arrhythmia	386	66	17
Credit	1500	400	30
Pima	500	268	8
Haberman	225	81	3
Ionosphere	225	126	33
Seismic bumps	2584	170	21
Shuttle	1000	13	9
Annthyroid	500	50	21
Bank additional	500	50	62
Celeba	500	50	39

7.1.2. Evaluation Metrics. In the experiments, we use the following evaluation metrics to assess the performance of the method:

– AUC-ROC.

To evaluate the AUC-ROC, 66.7% of the data were randomly selected for training and 33.3 % were randomly selected for testing. 33.3 % of the training dataset is also allocated to validation, which saves the best model.

7.1.3. Implementation Details. The proposed method was implemented using the programming language Python and the library PyTorch.

The following models were compared with each other:

– IF [12];

– DIF [13];

– autoencoder model with 2 hidden layers of size $d/2$ with ReLU activations;

– ADA-NAF (ADA-NAF-1) model with one hidden layer containing $d/2$ units;

– ADA-NAF (ADA-NAF-3) model with 3 hidden layers each again $d/2$ width with Tanh nonlinearities;

– multi-head ADA-NAF (ADA-MH-3-NAF-1) extending above using H=3 attention heads based on varied weight initializations (uniform, Xavier, normal distribution) that integrate both local tree-attention and global aggregation, with one hidden layer containing d/2 units;

where d is the input features for each dataset.

IF [12] is an unsupervised anomaly detection method based on the principle that anomalies are few and different, and thus should be easier to isolate in a dataset. The algorithm works by randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of that feature. This process is repeated recursively to create a tree structure. Anomalies are points that require fewer splits to be isolated from the rest of the data.

DIF [13] extends the concept of Isolation Forest by incorporating deep learning techniques. Instead of using raw features for splitting, DIF first transforms the input data using a neural network. This allows the model to learn complex, non-linear feature representations that can potentially lead to more effective isolation of anomalies. The depth of the model in DIF refers to both the depth of the neural network used for feature transformation and the depth of the isolation trees constructed on these transformed features. This combination of deep feature learning and isolation enables DIF to potentially capture more complex anomaly patterns than the original Isolation Forest.

8. Experimental Results. In this section, we present and discuss the experimental results.

8.1. Comparison between models. To measure the performance, we use the AUC-ROC score, which is a commonly used metric in anomaly detection. We compare the AUC-ROC score dependence on several datasets.

The results are shown in Table 2.

For these experiments, the number of trees was taken as 100 and the count epoch is 50. Mean squared error (MSE) was taken as a distance function for ADA-NAF-based models.

We use 3 different seeds when building trees, and 3 shuffle train/test datasets, and then average the results of the metrics. The results in Table 2 demonstrate that the proposed ADA-NAF approach is competitive with state-of-the-art anomaly detection techniques across the diverse benchmark datasets. On

the Credit Card, Ionosphere, Annthyroid, Shuttle, Celeba and Bank Additional datasets, ADA-NAF models achieve the highest or near highest AUC-ROC scores compared to the baseline methods. This underscores ADA-NAF's capability to effectively model the complex boundaries between normal and anomalous data patterns for these domains.

Table 2. Comparison of AUC-ROC for different models on different datasets

Dataset	**IF**	**DIF**	**Autoencoder**
Arrhytmia	$\mathbf{0.791 \pm 0.02}$	0.780 ± 0.01	0.771 ± 0.01
Credit	0.968 ± 0.01	0.935 ± 0.01	0.962 ± 0.01
Haberman	$\mathbf{0.658 \pm 0.08}$	0.602 ± 0.07	0.571 ± 0.07
Ionosphere	0.912 ± 0.03	0.909 ± 0.02	0.896 ± 0.03
Pima	$\mathbf{0.727 \pm 0.03}$	0.689 ± 0.03	0.689 ± 0.01
Seismic bumps	0.690 ± 0.02	$\mathbf{0.710 \pm 0.02}$	0.680 ± 0.02
Shuttle	0.842 ± 0.04	0.966 ± 0.04	0.967 ± 0.02
Annthyroid	0.848 ± 0.04	0.728 ± 0.03	0.697 ± 0.04
Celeba	0.732 ± 0.06	0.772 ± 0.05	0.740 ± 0.01
Bank additional	0.731 ± 0.03	0.717 ± 0.04	0.796 ± 0.06
Average	0.790	0.781	0.777
Dataset	**ADA-NAF-1**	**ADA-NAF-3**	**ADA-NAF-MH**
Arrhytmia	0.702 ± 0.06	0.674 ± 0.06	0.754 ± 0.04
Credit	0.972 ± 0.01	0.941 ± 0.02	$\mathbf{0.997 \pm 0.01}$
Haberman	0.580 ± 0.10	0.569 ± 0.09	0.567 ± 0.04
Ionosphere	0.960 ± 0.02	0.922 ± 0.02	$\mathbf{0.961 \pm 0.01}$
Pima	0.641 ± 0.03	0.617 ± 0.03	0.637 ± 0.03
Seismic bumps	0.695 ± 0.02	0.683 ± 0.02	0.704 ± 0.02
Shuttle	$\mathbf{0.990 \pm 0.01}$	0.790 ± 0.11	0.936 ± 0.09
Annthyroid	$\mathbf{0.891 \pm 0.02}$	0.748 ± 0.05	0.867 ± 0.03
celeba	0.697 ± 0.10	0.672 ± 0.09	$\mathbf{0.843 \pm 0.04}$
Bank additional	0.721 ± 0.03	0.700 ± 0.07	$\mathbf{0.797 \pm 0.03}$
Average	0.785	0.732	**0.806**

Notably, the DIF technique displays superior performance on the Arrhythmia, Haberman, and Pima datasets as evidenced by the higher AUC-ROC values. This indicates DIF's particular suitability for modeling anomalies in these medical-related tabular datasets. However, ADA-NAF variants still produce reasonable scores, elucidating the promise of the neural attention-based framework. An ablative analysis reveals that typically the multi-head ADA-NAF configuration demonstrates a slight edge over the single-headed version, corroborating the benefits of fusing diverse attention representations. Comparing shallow and deeper ADA-NAF models, gains from additional layers are dataset dependent – aligning with established knowledge that

optimal depth is contingent on data complexity. In summary, ADA-NAF puts forth a highly competitive semi-supervised technique for anomaly detection grounded in cutting-edge neural attention architectures. The experiments validate applicability to heterogeneous data domains with performance rivaling current state-of-the-art approaches. This underscores the potential of innovating tailored neural attention mechanisms for advancing anomaly detection.

8.2. Noise contamination of training set. Real-world datasets often contain some degree of inherent anomalies or mislabeled points overlapping with normal classes. It is vital to evaluate model robustness towards such contaminated training data. We simulate this by injecting anomalies masked as normal into the ADA-NAF training set in a controlled manner. To evaluate the impact of contaminated training data, we construct noisy variants of the normal set X_{normal} as follows:

1. We start with the completely normal training examples X_{normal} with size as $|X_{anomalous}|$, where $X_{anomalous}$ is the anomalous samples.

2. Noisy normal sets are prepared by combining normal and anomalous points:

$X_{normal}^{(A)}$: Take A% of the anomalous samples from $X_{anomalous}$ and combine them with (100 - A)% of the normal instances in X_{normal}.

3. Train ADA-NAF separately on X_{normal}, $X_{normal}^{(12.5)}$, $X_{normal}^{(25)}$, $X_{normal}^{(37.5)}$ and $X_{normal}^{(50)}$ while RF trains on balanced data.

The experiment evaluates how introducing different levels of noise into the training set affects the performance of ADA-NAF models in detecting anomalies. Understanding the impact of noise on ADA-NAF training is critical to improving model robustness and reliability, especially in real-world scenarios where the data often contains some level of noise or anomalies.

We use 3 seeds while building RF with 3 shuffle dataset cross-validation, the count epoch is 50, number of trees is 100.

Results in Figure 3 – 5 showcase the impact on AUC scores for the Ionosphere, Annthyroid and Celeba datasets. We compare 3 model variants: single hidden layer ADA-NAF, 3 hidden layer ADA-NAF and 3 headed attention with 1 hidden layer. Overall we observe performance degradation as anomalous points increasingly pollute the normal class data. The Annthyroid dataset displays the most graceful lowering of AUC compared to sudden drops for Celeba, indicating robustness. The multi-head architecture appears significantly more stable than standard ADA-NAF, retaining higher accuracy despite up to 50% contamination. This underscores the flaw of diversified attention heads in establishing reliable boundaries between normality and anomaly. The findings highlight the variability in the impact of label noise on

different models and datasets. For real-world anomaly detection, pre-filtering training data to minimize contamination would enhance ADA-NAF's detection capability. Online learning schemes to continually adapt to new normal and anomalous data are another strategy to offset declining performance over time. The analysis provides vital perspectives on the reliability of semi-supervised approaches in practice.

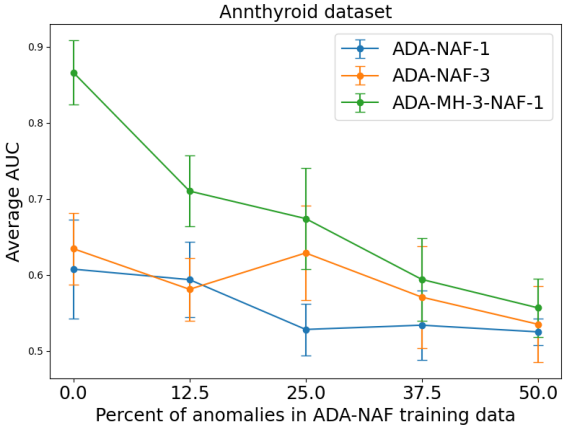


Fig. 3. AUC graphs for the Annyhyroid dataset with different noise injection

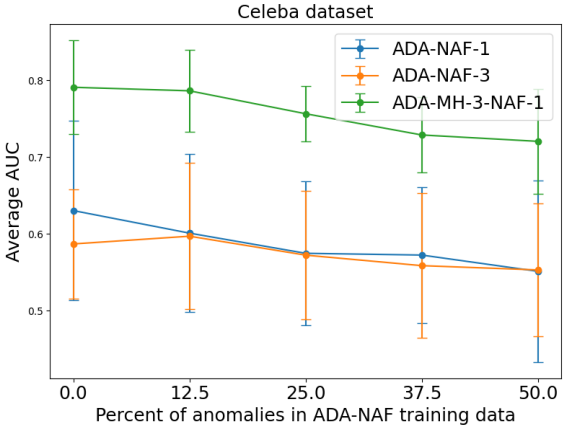


Fig. 4. AUC graphs for the Celeba dataset with different noise injection

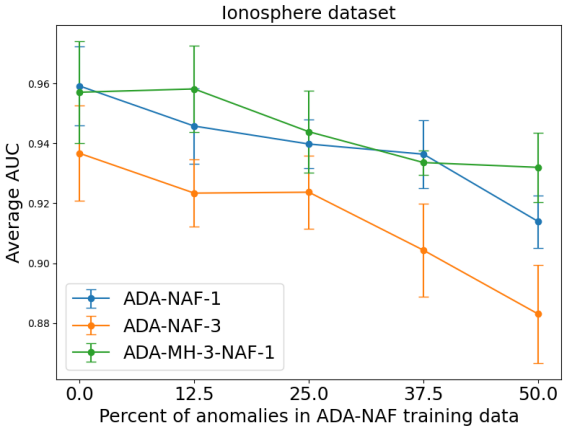


Fig. 5. AUC graphs for the Ionosphere dataset with different noise injection

8.3. ϵ-contamination. To evaluate the impact of the proposed regularization method, we conduct experiments with ADA-NAF models trained using different fixed ϵ values. We train multiple models while keeping all settings identical except for the ϵ hyperparameter controlling the admixture amount of the random attention matrix W. Models are trained end-to-end on the same normal training set for a fixed number of epochs. Specifically, we select ϵ values spread over the [0, 0.5] range. For each value, we train an ADA-NAF variant regularized with the corresponding ϵ contamination ratio. We plot the metrics vs ϵ to analyze the regularization impact. We hypothesize a non-linear influence, with small ϵ likely improving robustness hence detection accuracy by balancing noise, while large values can overwhelm true signals. The optimal ϵ is expected to be dataset-dependent.

The key findings from the experiments are:

1. In the Celeba Dataset (Figure 6), we observe that the AUC score for both ADA-NAF-1 and ADA-NAF-3 gradually decreases as the regularization parameter ϵ increases. ADA-NAF-1 demonstrates more stability, maintaining a relatively consistent AUC score. On the other hand, ADA-NAF-3 experiences a significant drop in performance, followed by a slight improvement. This pattern suggests that ADA-NAF-1, with its potentially simpler architecture, is less affected by increasing regularization, thereby indicating a steadier performance against the variations in ϵ. Conversely, the initial decline in ADA-NAF-3's performance up to a critical point $\epsilon = 0.3$ before it begins to recover slightly, underscores its vulnerability to stronger regularization effects but also hints at a possible resilience mechanism that kicks in beyond that point.

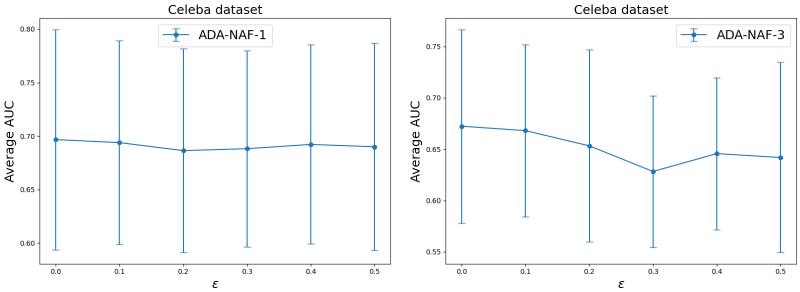


Fig. 6. AUC graphs for the Celeba dataset with different ϵ

2. Bank Additional Dataset (Figure 7). Both shallow and deeper ADA-NAF variants exhibit a peak AUC at the 0.4 contamination level. Performance is maximized with mild attention noise injection. The improvement suggests that low ϵ ratios serve more as useful perturbations rather than obstruction of meaningful attention patterns. This accords with literature on carefully tuned noise amplification improving generalization.

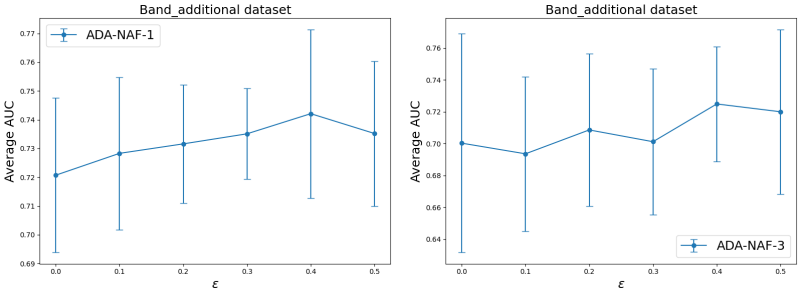


Fig. 7. AUC graphs for the Bank Additional dataset with different ϵ

3. Annthyroid Dataset (Figure 8). The model with one layer shows deterioration with increasing contamination, while the 3-layer model appears to be more stable, one can note the influence of the architecture and the choice of pollution level for different datasets.

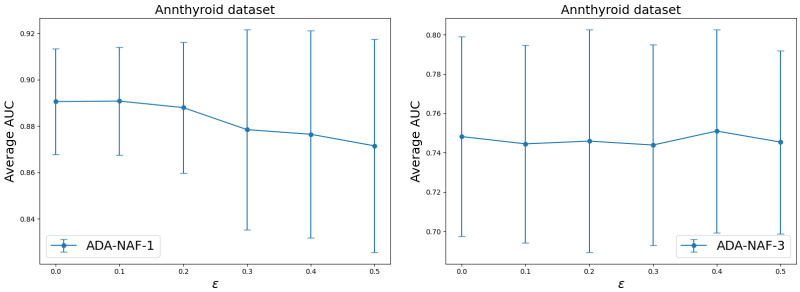


Fig. 8. AUC graphs for the Annthyroid dataset with different ϵ

9. Conclusion. This research introduced ADA-NAF - an innovative neural attention model for semi-supervised anomaly detection adapted from the Neural Attention Forest framework. We structurally modify NAF to operate as a reconstructive autoencoder powered by neural attention. Experimental results demonstrate that ADA-NAF provides a competitive approach rivaling state-of-the-art techniques over diverse anomaly detection benchmarks. The integrated architecture allows ADA-NAF to overcome the limitations of standard neural networks concerning tabular data while benefiting from the representational richness afforded by deep learning. By tuning neural attention, the model focuses intrinsically on normal data characteristics. At test time anomalies result in greater reconstruction errors enabling their detection without explicit labels. An interesting opportunity for further research includes replacing the neural network with a full-fledged transformer architecture. By tokenizing training samples into input sequences, transformers can capture complex data relationships through self-attention. This can potentially enhance the anomaly detection accuracy and interpretability of reconstructions provided by the ADA-NAF model. While our work focuses on adapting NAF for anomaly detection, the flexibility and power of this architecture suggest potential applications in various other semi-supervised learning tasks. The unique combination of Random Forests with neural attention mechanisms in NAF makes it particularly suitable for scenarios with limited labeled data. In many real-world applications across different domains, obtaining large amounts of labeled data can be expensive, time-consuming, or sometimes impossible. NAF's ability to leverage both labeled and unlabeled data effectively could prove valuable in such contexts. The model's capacity to capture complex feature interactions and its interpretability through attention weights could be beneficial in fields such as healthcare, finance, or industrial monitoring, where understanding the model's decision-making process is crucial. Furthermore,

the hierarchical structure of NAF could be advantageous in handling high-dimensional data or in tasks requiring multi-level feature extraction. While these potential applications remain to be explored, they highlight the versatility of NAF architecture and open up exciting avenues for future research beyond anomaly detection, particularly in semi-supervised learning scenarios. In conclusion, ADA-NAF contributes an interpretable semi-supervised technique to complement existing methodologies. The work highlights the importance of constructing innovative neural attention architectures tailored for anomaly detection challenges.

References

1. Chandola V., Banerjee A., Kumar V. Anomaly detection: A survey. ACM Computing Surveys. 2009. vol. 41. no. 3. pp. 1–58. DOI: 10.1145/1541880.1541882.
2. Barnett V., Lewis T. Outliers in statistical data. 3rd Edition. New York: Wiley, 1994. 608 p.
3. Grubbs F.E. Procedures for detecting outlying observations in samples. Technometrics. 1969. vol. 11. pp. 1–21. DOI: 10.1080/00401706.1969.10490657.
4. Goldstein M. Special Issue on Unsupervised Anomaly Detection. Applied Sciences. 2023. vol. 13(10). DOI: 10.3390/app13105916
5. Zhang C., Liu J., Chen W., Shi J., Yao M., Yan X., Xu N., Chen D. [Retracted] Unsupervised Anomaly Detection Based on Deep Autoencoding and Clustering. Security and Communication Networks. 2021. vol. 2021. DOI: 10.1155/2021/7389943.
6. Sarvari H., Domeniconi C., Prenkaj B., Stilo G. Unsupervised boosting-based autoencoder ensembles for outlier detection. Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining. 2021. pp. 91–103. DOI: 10.1007/978-3-030-75762-5_8.
7. Yoshihara K., Takahashi K. A simple method for unsupervised anomaly detection: An application to Web time series data. Plos one. 2022. vol. 17. no. 1. DOI: 10.1371/journal.pone.0262463.
8. Kiran B.R., Thomas D.M., Parakkal R. An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos. Journal of Imaging. 2018. vol. 4. no. 2. DOI: 10.3390/jimaging4020036.
9. Al-amri R., Murugesan R.K., Man M., Abdulateef A.F., Al-Sharafi M.A., Alkahtani A.A. A review of machine learning and deep learning techniques for anomaly detection in IoT data. Applied Sciences. 2021. vol. 11. no. 12. DOI: 10.3390/app11125320.
10. Finke T., Kramer M., Morandini A., Muck A., Oleksiyuk I. Autoencoders for unsupervised anomaly detection in high energy physics. Journal of High Energy Physics. 2021. vol. 2021. no. 6. DOI: 10.1007/JHEP06(2021)161.
11. Konstantinov A.V., Utkin L.V., Lukashin A.A., Muliukha V.A. Neural attention forests: Transformer-based forest improvement. Proceedings of International Conference on Intelligent Information Technologies for Industry. 2023. pp. 158–167.
12. Liu F.T., Kai M.T., Zhou Z.H. Isolation forest. Proceedings of 8th IEEE International Conference on Data Mining. 2008. pp. 413–422. DOI: 10.1109/ICDM.2008.17.
13. Xu H., Pang G., Wang Y., Wang Y. Deep Isolation Forest for Anomaly Detection. IEEE Transactions on Knowledge and Data Engineering. 2023. vol. 35. no. 12. pp. 12591–12604. DOI: 10.1109/TKDE.2023.3270293.
14. Ahmed M., Mahmood A.N., Hu J. A survey of network anomaly detection techniques. Journal of Network and Computer Applications. 2016. vol. 60. pp. 19–31. DOI: 10.1016/j.jnca.2015.11.016.

15. Liao Y., Bartler A., Yang B. Anomaly detection based on selection and weighting in latent space. Proceedings of 2021 IEEE 17th International Conference on Automation Science and Engineering (CASE). 2021. pp. 409–415. DOI: 10.1109/CASE49439.2021.9551267.
16. Xu J., Wu H., Wang J., Long M. Anomaly Transformer: Time Series Anomaly Detection with Association Discrepancy. Proceedings of Tenth International Conference on Learning Representations. 2022.
17. Hojjati H., Ho T.K.K., Armanfard N. Self-supervised anomaly detection: A survey and outlook. arXiv preprint arXiv:2205.05173. 2022.
18. Perera P., Oza P., Patel V.M. One-class classification: A survey. arXiv preprint arXiv:2101.03064. 2021.
19. Darban Z.Z., Webb G.I., Pan S., Aggarwal C.C., Salehi M. Deep learning for time series anomaly detection: A survey. ACM Computing Surveys. 2024. vol. 57. no. 1. DOI: 10.1145/369133.
20. Chalapathy R., Chawla S. Deep learning for anomaly detection: A survey. arXiv preprint arXiv:1901.03407. 2019.
21. Landauer M., Onder S., Skopik F., Wurzenberger M. Deep learning for anomaly detection in log data: A survey. Machine Learning with Applications. 2023. vol. 12. DOI: 10.1016/j.mlwa.2023.100470.
22. Di Mattia F., Galeone P., De Simoni M., Ghelfi E. A survey on gans for anomaly detection. arXiv preprint arXiv:1906.11632. 2019.
23. Suarez J.J.P., Naval Jr P.C. A survey on deep learning techniques for video anomaly detection. arXiv preprint arXiv:2009.14146. 2020.
24. Tschuchnig M.E., Gadermayr M. Anomaly detection in medical imaging-a mini review. Proceedings of the 4th International Data Science Conference–iDSC 2021. 2022. pp. 33–38.
25. Niu Z., Zhong G., Yu H. A review on the attention mechanism of deep learning. Neurocomputing. 2021. vol. 452. pp. 48–62. DOI: 10.1016/j.neucom.2021.03.091.
26. Bahdanau D., Cho K., Bengio Y. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473. 2014.
27. Zhu Y., Newsam S. Motion-aware feature for improved video anomaly detection. arXiv preprint arXiv:1907.10211. 2019.
28. Hashimoto M., Ide Y., Aritsugi M. Anomaly detection for sensor data of semiconductor manufacturing equipment using a GAN. Procedia Computer Science. 2021. vol. 192. pp. 873–882. DOI: 10.1016/j.procs.2021.08.090.
29. Wu X., Huang S., Li M., Deng Y. Vector magnetic anomaly detection via an attention mechanism deep-learning model. Applied Sciences. 2021. vol. 11. no. 23. DOI: 10.3390/app112311533.
30. Yu Y., Zha Z., Jin B., Wu G., Dong C. Graph-Based Anomaly Detection via Attention Mechanism. Proceedings of on: 18th International Conference on Intelligent Computing Theories and Application. 2022. pp. 401–411. DOI: 10.1007/978-3-031-13870-6_33.
31. Tang T.W., Hsu H., Huang W.R., Li K.M. Industrial Anomaly Detection with Skip Autoencoder and Deep Feature Extractor. Sensors. 2022. vol. 22. no. 23. DOI: 10.3390/s22239327.
32. Utkin L.V., Konstantinov A.V. Attention-based random forest and contamination model. Neural Networks: the official journal of the International Neural Network Society. 2022. vol. 154. pp. 346–359.
33. Utkin L., Ageev A., Konstantinov A., Muliukha V. Improved Anomaly Detection by Using the Attention-Based Isolation Forest. Algorithms. 2023. vol. 16. no. 1. DOI: 10.3390/a16010019.

34. Cai Z. Weighted nadaraya–watson regression estimation. Statistics and probability letters. 2001. vol. 51. no. 3. pp. 307–318. DOI: 10.1016/S0167-7152(00)00172-3.
35. Rumelhart D.E., Hinton G.E., Williams R.J. Learning internal representations by error propagation. Parallel Distributed Processing: Explorations in the Microstructure of Cognition. 1986. vol. 1. pp. 318–362.
36. Hawkins S., He H., Williams G., Baxter R. Outlier detection using replicator neural networks. International Conference on Data Warehousing and Knowledge Discovery. 2002. pp. 170–180. DOI: 10.1007/3-540-46145-0_17.
37. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin I. Attention is all you need. Advances in neural information processing systems. 2017. vol. 30.
38. Arrhythmia Dataset. Available at: https://www.kaggle.com/code/mtavares51/binary-classification-on-arrhythmia-dataset. (accessed 30.05.2024).
39. Credit Card Fraud Detection Dataset. Available at: https://www.kaggle.com/code/shivamsekra/credit-card-fraud-detection-eda-isolation-forest. (accessed 30.05.2024).
40. Pima Indians Diabetes Dataset. Available at: https://www.kaggle.com/code/hafizramadan/data-science-project-iii. (accessed 30.05.2024).
41. Haberman's Survival Dataset. Available at: https://www.kaggle.com/datasets/gilsousa/habermans-survival-data-set. (accessed 30.05.2024).
42. Ionosphere Dataset. Available at: https://www.kaggle.com/code/zymzym/classification-of-the-ionosphere-dataset-by-knn. (accessed 30.05.2024).
43. Seismic Bumps Dataset. Available at: https://www.kaggle.com/datasets/pranabroy94/seismic-bumps-data-set. (accessed 30.05.2024).
44. Shuttle Dataset. Available at: https://github.com/xuhongzuo/deep-iforest/tree/main. (accessed 30.05.2024).
45. Annthyroid Dataset. Available at: https://github.com/GuansongPang/deviation-network. (accessed 30.05.2024).
46. Bank Additional Dataset. Available at: https://github.com/GuansongPang/deviation-network. (accessed 30.05.2024).
47. CelebA Dataset. Available at: https://github.com/GuansongPang/deviation-network. (accessed 30.05.2024).

Ageev Andrey — Ph.D. student, Institute of computer science and technology, Peter the Great St. Petersburg Polytechnic University. Research interests: machine learning, bioinformatics, large language models, computer vision. The number of publications — 5. andreyageev1@mail.ru; 29, Polytechnicheskaya St., 195251, St. Petersburg, Russia; office phone: +7(812)775-0510.

Konstantinov Andrei — Ph.D. student, Institute of computer science and technology, Peter the Great St. Petersburg Polytechnic University; Assistant of the laboratory, Research laboratory of neural network technologies and artificial intelligence, Peter the Great St. Petersburg Polytechnic University. Research interests: machine learning, computer vision, image processing. The number of publications — 37. andrue.konst@gmail.com; 29, Polytechnicheskaya St., 195251, St. Petersburg, Russia; office phone: +7(911)954-5565.

Utkin Lev — Ph.D., Dr.Sci., Professor, Head of the institute, Institute of computer science and technology, Peter the Great St. Petersburg Polytechnic University; Head of the laboratory, Research laboratory of neural network technologies and artificial intelligence, Peter the Great St. Petersburg Polytechnic University. Research interests: machine learning, imprecise probability theory, decision making. The number of publications — 300. lev.utkin@gmail.com; 29, Polytechnicheskaya St., 195251, St. Petersburg, Russia; office phone: +7(921)344-6390.

Acknowledgements. The research is partially funded by the Ministry of Science and Higher Education of the Russian Federation as part of state assignments "Development and research of machine learning models for solving fundamental problems of artificial intelligence for the fuel and energy complex"(topic FSEG-2024-0027).

УДК 004.85 **DOI** 10.15622/ia.24.1.12

А.Ю. Агеев, А.В. Константинов, Л.В. Уткин

ADA-NAF: ПОЛУКОНТРОЛИРУЕМОЕ ОБНАРУЖЕНИЕ АНОМАЛИЙ НА ОСНОВЕ НЕЙРОННОГО ЛЕСА ВНИМАНИЯ

Агеев А.Ю., Константинов А.В., Уткин Л.В. **ADA-NAF: Полуконтролируемое обнаружение аномалий на основе нейронного леса внимания.**

Аннотация. В этом исследовании мы представляем новую модель под названием ADA-NAF (автоэнкодер обнаружения аномалий с нейронным лесом внимания) для полуконтролируемого обнаружения аномалий, которая уникальным образом интегрирует архитектуру нейронного леса внимания (NAF), которая была разработана для объединения случайного классификатора леса с нейронной сетью, вычисляющей веса внимания для агрегации прогнозов дерева решений. Ключевая идея ADA-NAF заключается в включении NAF в структуру автоэнкодера, где он реализует функции компрессора, а также реконструктора входных векторов. Наш подход представляет несколько технических достижений. Во-первых, предлагаемая сквозная методология обучения по обычным данным, которая минимизирует ошибки реконструкции при обучении и оптимизации нейронных весов внимания для фокусировки на скрытых признаках. Во-вторых, новый механизм кодирования, который использует иерархическую структуру NAF для захвата сложных шаблонов данных. В-третьих, адаптивная структура оценки аномалий, которая объединяет ошибки реконструкции с важностью признаков на основе внимания. Благодаря обширным экспериментам с различными наборами данных ADA-NAF демонстрирует превосходную производительность по сравнению с современными методами. Модель демонстрирует особую силу в обработке многомерных данных и выявлении тонких аномалий, которые традиционные методы часто не обнаруживают. Наши результаты подтверждают эффективность и универсальность ADA-NAF как надежного решения для реальных задач обнаружения аномалий с перспективными приложениями в кибербезопасности, промышленном мониторинге и диагностике здравоохранения. Эта работа продвигает эту область, представляя новую архитектуру, которая сочетает в себе интерпретируемость механизмов внимания с мощными возможностями обучения признакам автоэнкодеров.

Ключевые слова: обнаружение аномалий, случайный лес, механизм внимания, нейронный лес внимания.

Литература

1. Chandola V., Banerjee A., Kumar V. Anomaly detection: A survey. ACM Computing Surveys. 2009. vol. 41. no. 3. pp. 1–58. DOI: 10.1145/1541880.1541882.
2. Barnett V., Lewis T. Outliers in statistical data. 3rd Edition. New York: Wiley, 1994. 608 p.
3. Grubbs F.E. Procedures for detecting outlying observations in samples. Technometrics. 1969. vol. 11. pp. 1–21. DOI: 10.1080/00401706.1969.10490657.
4. Goldstein M. Special Issue on Unsupervised Anomaly Detection. Applied Sciences. 2023. vol. 13(10). DOI: 10.3390/app13105916
5. Zhang C., Liu J., Chen W., Shi J., Yao M., Yan X., Xu N., Chen D. [Retracted] Unsupervised Anomaly Detection Based on Deep Autoencoding and Clustering. Security and Communication Networks. 2021. vol. 2021. DOI: 10.1155/2021/7389943.
6. Sarvari H., Domeniconi C., Prenkaj B., Stilo G. Unsupervised boosting-based autoencoder ensembles for outlier detection. Proceedings of Pacific-Asia Conference on

Knowledge Discovery and Data Mining. 2021. pp. 91–103. DOI: 10.1007/978-3-030-75762-5_8.

7. Yoshihara K., Takahashi K. A simple method for unsupervised anomaly detection: An application to Web time series data. Plos one. 2022. vol. 17. no. 1. DOI: 10.1371/journal.pone.0262463.
8. Kiran B.R., Thomas D.M., Parakkal R. An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos. Journal of Imaging. 2018. vol. 4. no. 2. DOI: 10.3390/jimaging4020036.
9. Al-amri R., Murugesan R.K., Man M., Abdulateef A.F., Al-Sharafi M.A., Alkahtani A.A. A review of machine learning and deep learning techniques for anomaly detection in IoT data. Applied Sciences. 2021. vol. 11. no. 12. DOI: 10.3390/app11125320.
10. Finke T., Kramer M., Morandini A., Muck A., Oleksiyuk I. Autoencoders for unsupervised anomaly detection in high energy physics. Journal of High Energy Physics. 2021. vol. 2021. no. 6. DOI: 10.1007/JHEP06(2021)161.
11. Konstantinov A.V., Utkin L.V., Lukashin A.A., Muliukha V.A. Neural attention forests: Transformer-based forest improvement. Proceedings of International Conference on Intelligent Information Technologies for Industry. 2023. pp. 158–167.
12. Liu F.T., Kai M.T., Zhou Z.H. Isolation forest. Proceedings of 8th IEEE International Conference on Data Mining. 2008. pp. 413–422. DOI: 10.1109/ICDM.2008.17.
13. Xu H., Pang G., Wang Y., Wang Y. Deep Isolation Forest for Anomaly Detection. IEEE Transactions on Knowledge and Data Engineering. 2023. vol. 35. no. 12. pp. 12591–12604. DOI: 10.1109/TKDE.2023.3270293.
14. Ahmed M., Mahmood A.N., Hu J. A survey of network anomaly detection techniques. Journal of Network and Computer Applications. 2016. vol. 60. pp. 19–31. DOI: 10.1016/j.jnca.2015.11.016.
15. Liao Y., Bartler A., Yang B. Anomaly detection based on selection and weighting in latent space. Proceedings of 2021 IEEE 17th International Conference on Automation Science and Engineering (CASE). 2021. pp. 409–415. DOI: 10.1109/CASE49439.2021.9551267.
16. Xu J., Wu H., Wang J., Long M. Anomaly Transformer: Time Series Anomaly Detection with Association Discrepancy. Proceedings of Tenth International Conference on Learning Representations. 2022.
17. Hojjati H., Ho T.K.K., Armanfard N. Self-supervised anomaly detection: A survey and outlook. arXiv preprint arXiv:2205.05173. 2022.
18. Perera P., Oza P., Patel V.M. One-class classification: A survey. arXiv preprint arXiv:2101.03064. 2021.
19. Darban Z.Z., Webb G.I., Pan S., Aggarwal C.C., Salehi M. Deep learning for time series anomaly detection: A survey. ACM Computing Surveys. 2024. vol. 57. no. 1. DOI: 10.1145/369133.
20. Chalapathy R., Chawla S. Deep learning for anomaly detection: A survey. arXiv preprint arXiv:1901.03407. 2019.
21. Landauer M., Onder S., Skopik F., Wurzenberger M. Deep learning for anomaly detection in log data: A survey. Machine Learning with Applications. 2023. vol. 12. DOI: 10.1016/j.mlwa.2023.100470.
22. Di Mattia F., Galeone P., De Simoni M., Ghelfi E. A survey on gans for anomaly detection. arXiv preprint arXiv:1906.11632. 2019.
23. Suarez J.J.P., Naval Jr P.C. A survey on deep learning techniques for video anomaly detection. arXiv preprint arXiv:2009.14146. 2020.
24. Tschuchnig M.E., Gadermayr M. Anomaly detection in medical imaging-a mini review. Proceedings of the 4th International Data Science Conference–iDSC 2021. 2022. pp. 33–38.

25. Niu Z., Zhong G., Yu H. A review on the attention mechanism of deep learning. Neurocomputing. 2021. vol. 452. pp. 48–62. DOI: 10.1016/j.neucom.2021.03.091.
26. Bahdanau D., Cho K., Bengio Y. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473. 2014.
27. Zhu Y., Newsam S. Motion-aware feature for improved video anomaly detection. arXiv preprint arXiv:1907.10211. 2019.
28. Hashimoto M., Ide Y., Aritsugi M. Anomaly detection for sensor data of semiconductor manufacturing equipment using a GAN. Procedia Computer Science. 2021. vol. 192. pp. 873–882. DOI: 10.1016/j.procs.2021.08.090.
29. Wu X., Huang S., Li M., Deng Y. Vector magnetic anomaly detection via an attention mechanism deep-learning model. Applied Sciences. 2021. vol. 11. no. 23. DOI: 10.3390/app112311533.
30. Yu Y., Zha Z., Jin B., Wu G., Dong C. Graph-Based Anomaly Detection via Attention Mechanism. Proceedings of on: 18th International Conference on Intelligent Computing Theories and Application. 2022. pp. 401–411. DOI: 10.1007/978-3-031-13870-6_33.
31. Tang T.W., Hsu H., Huang W.R., Li K.M. Industrial Anomaly Detection with Skip Autoencoder and Deep Feature Extractor. Sensors. 2022. vol. 22. no. 23. DOI: 10.3390/s22239327.
32. Utkin L.V., Konstantinov A.V. Attention-based random forest and contamination model. Neural Networks: the official journal of the International Neural Network Society. 2022. vol. 154. pp. 346–359.
33. Utkin L., Ageev A., Konstantinov A., Muliukha V. Improved Anomaly Detection by Using the Attention-Based Isolation Forest. Algorithms. 2023. vol. 16. no. 1. DOI: 10.3390/a16010019.
34. Cai Z. Weighted nadaraya–watson regression estimation. Statistics and probability letters. 2001. vol. 51. no. 3. pp. 307–318. DOI: 10.1016/S0167-7152(00)00172-3.
35. Rumelhart D.E., Hinton G.E., Williams R.J. Learning internal representations by error propagation. Parallel Distributed Processing: Explorations in the Microstructure of Cognition. 1986. vol. 1. pp. 318–362.
36. Hawkins S., He H., Williams G., Baxter R. Outlier detection using replicator neural networks. International Conference on Data Warehousing and Knowledge Discovery. 2002. pp. 170–180. DOI: 10.1007/3-540-46145-0_17.
37. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin I. Attention is all you need. Advances in neural information processing systems. 2017. vol. 30.
38. Arrhythmia Dataset. Available at: https://www.kaggle.com/code/mtavares51/binary-classification-on-arrhythmia-dataset. (accessed 30.05.2024).
39. Credit Card Fraud Detection Dataset. Available at: https://www.kaggle.com/code/shivamsekra/credit-card-fraud-detection-eda-isolation-forest. (accessed 30.05.2024).
40. Pima Indians Diabetes Dataset. Available at: https://www.kaggle.com/code/hafizramadan/data-science-project-iii. (accessed 30.05.2024).
41. Haberman's Survival Dataset. Available at: https://www.kaggle.com/datasets/gilsousa/habermans-survival-data-set. (accessed 30.05.2024).
42. Ionosphere Dataset. Available at: https://www.kaggle.com/code/zymzym/classification-of-the-ionosphere-dataset-by-knn. (accessed 30.05.2024).
43. Seismic Bumps Dataset. Available at: https://www.kaggle.com/datasets/pranabroy94/seismic-bumps-data-set. (accessed 30.05.2024).

44. Shuttle Dataset. Available at: https://github.com/xuhongzuo/deep-iforest/tree/main. (accessed 30.05.2024).
45. Annthyroid Dataset. Available at: https://github.com/GuansongPang/deviation-network. (accessed 30.05.2024).
46. Bank Additional Dataset. Available at: https://github.com/GuansongPang/deviation-network. (accessed 30.05.2024).
47. CelebA Dataset. Available at: https://github.com/GuansongPang/deviation-network. (accessed 30.05.2024).

Агеев Андрей Юрьевич — аспирант, институт компьютерных наук и технологий, Санкт-Петербургский политехнический университет Петра Великого (СПбПУ). Область научных интересов: машинное обучение, биоинформатика, большие языковые модели и компьютерное зрение. Число научных публикаций — 5. andreyageev1@mail.ru; улица Политехническая, 29, 195251, Санкт-Петербург, Россия; р.т.: +7(812)775-0510.

Константинов Андрей Владимирович — аспирант, институт компьютерных наук и технологий, Санкт-Петербургский политехнический университет Петра Великого (СПбПУ); ассистент лаборатории, научно-исследовательская лаборатория нейросетевых технологий и искусственного интеллекта, Санкт-Петербургский политехнический университет Петра Великого (СПбПУ). Область научных интересов: машинное обучение, компьютерное зрение, обработка изображений. Число научных публикаций — 37. andrue.konst@gmail.com; улица Политехническая, 29, 195251, Санкт-Петербург, Россия; р.т.: +7(911)954-5565.

Уткин Лев Владимирович — д-р техн. наук, профессор, директор института, институт компьютерных наук и технологий, Санкт-Петербургский политехнический университет Петра Великого (СПбПУ); руководитель лаборатории, научно-исследовательская лаборатория нейросетевых технологий и искусственного интеллекта, Санкт-Петербургский политехнический университет Петра Великого (СПбПУ). Область научных интересов: машинное обучение, теория неточных вероятностей, принятие решений. Число научных публикаций — 300. lev.utkin@gmail.com; улица Политехническая, 29, 195251, Санкт-Петербург, Россия; р.т.: +7(921)344-6390.

Поддержка исследований. Исследования частично финансируются Министерством науки и высшего образования РФ в рамках государственного задания «Разработка и исследование моделей машинного обучения для решения фундаментальных задач искусственного интеллекта для ТЭК» (тема ФСЭГ-2024-0027).

Руководство для авторов

Взаимодействие автора с редакцией осуществляется через личный кабинет на сайте журнала «Информатика и автоматизация» http://ia.spcras.ru/. При р егистрации авторам рекомендуется заполнить все предложенные поля данных. Подготовка статьи ведется с помощью текстовых редакторов MS Word 2007 и выше или LaTex. Объем основного текста (до раздела Литература) - от 20 до 30 страниц включительно. Переносы запрещены. Номера страниц не проставляются. Основная часть текста статьи разбивается на разделы, среди которых являются обязательными: введение, хотя бы один «содержательный» раздел и заключение. Допускается также мотивированное содержанием и структурой материал а выделение подразделов. В основную часть опускается помещать рисунки, таблицы, листинги и формулы. Правила их оформления подробно рассмотрены на нашем сайте в разделе «Руководство для авторов».

Author guidelines

Interaction between each potential author and the Editorial board is realized through the personal account on the website of the journal "Informatics and Automation" http://ia.spcras.ru/. At the registration the authors are requested to fill out all data fields in the proposed form. The submissions should be prepared using MS Word 2007, LaTex. The text of the paper in the main part should not exceed 30 pages. Pages are not numbered; hyphenations are not allowed. Certain figures, tables, listings and formulas are allowed in the main section, and their typography is considered in more detail at the journal web.

Signed to print 15.01.2025. Passed for print 20.01.2025.
Printed in Publishing center GUAP.
Address: 67 litera A, B. Morskaya, St. Petersburg, 190000, Russia

Founder and Publisher: SPC RAS.
Address: 39 litera A, 14th Line V.O., St.Peterburg, 199178, Russia.
The journal is registered in the Federal Service for Supervision of Communications, Information Technology, and Mass Media,
Registration Certificate (registration number) ПИ № ФС77-79228 dated September 25, 2020
Subscription Index П5513, Russian Post Catalog

Подписано к печати 15.01.2025. Дата выхода в свет 20.01.2025.
Формат 60×90 1/16. Усл. печ. л. 20,8. Заказ № 5. Тираж 300 экз., цена свободная.
Отпечатано в Редакционно-издательском центре ГУАП.
Адрес типографии: Б. Морская, д. 67, лит. А, г. Санкт-Петербург, 190000, Россия

Учредитель и издатель: СПб ФИЦ РАН.
Адрес учредителя и издателя: 14-я линия В.О., д. 39, лит. А, г. Санкт-Петербург, 199178, Россия
Журнал зарегистрирован Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций,
свидетельство о регистрации (регистрационный номер) ПИ № ФС77-79228 от 25 сентября 2020 г.
Подписной индекс П5513 по каталогу «Почта России»