ISSN 2713-3192
DOI 10.15622/ia.2024.23.2
http://ia.spcras.ru

ТОМ 23 № 2

# ИНФОРМАТИКА И АВТОМАТИЗАЦИЯ

# INFORMATICS AND AUTOMATION

СПб ФИЦ РАН

Санкт-Петербург
2024

# INFORMATICS AND AUTOMATION

**Volume 23 № 2, 2024**

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. Lopotova **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 23 № 2, 2024**

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. Лопотова **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** 519.872.[1+3+8] **DOI** 10.15622/ia.23.2.1

О.И. Заяц, М.М. Кореневская, А.С. Ильяшенко, В.А. Мулюха

# СИСТЕМА МАССОВОГО ОБСЛУЖИВАНИЯ С АБСОЛЮТНЫМ ПРИОРИТЕТОМ, ВЕРОЯТНОСТНЫМ ВЫТАЛКИВАЮЩИМ МЕХАНИЗМОМ И ПОВТОРНЫМИ ЗАЯВКАМИ

*Заяц О.И., Кореневская М.М., Ильяшенко А.С., Мулюха В.А.* **Система массового обслуживания с абсолютным приоритетом, вероятностным выталкивающим механизмом и повторными заявками.**

**Аннотация.** Статья посвящена исследованию одноканальной системы массового обслуживания. На вход системы подаются два стационарных пуассоновских потока заявок. Первый из них обладает абсолютным приоритетом по отношению ко второму. Емкость системы ограничена $k$ заявками. В системе присутствует вероятностный выталкивающий механизм: если подошедшая высокоприоритетная заявка застает все места в накопителе занятыми, то она с заданной вероятностью выталкивания $a$ может вытеснить из накопителя одну низкоприоритетную заявку, если таковые в нем имеются. Все заявки обслуживаются по одному и тому же показательному закону. Заявки, не сумевшие попасть в систему из-за ограниченности объема накопителя, а также вытесненные из накопителя при срабатывании выталкивающего механизма, не теряются сразу безвозвратно, а направляются в особую часть системы, называемую орбитой и предназначенную для сохранения повторных заявок. На орбите формируются две отдельные неограниченные очереди, состоящие, соответственно, из низкоприоритетных и высокоприоритетных повторных заявок. При отсутствии свободного места в накопителе вновь подошедшие заявки с заданной вероятностью настойчивости $q$ присоединяются к соответствующей орбитальной очереди. Время пребывания повторных заявок на орбите распределено по показательному закону, параметр этого закона различается для разных типов требований. После ожидания на орбите вторичные заявки вновь направляются в систему. Вероятностные характеристики описанной системы рассчитываются методом производящих функций, ранее предложенным авторами для расчета аналогичных систем без повторных требований. Детально исследуется зависимость вероятностей потери обоих типов заявок от параметров системы, прежде всего от вероятности выталкивания $a$, емкости системы $k$ и вероятности повторного обращения (вероятности настойчивости) $q$. Показано, что ранее выявленные в аналогичных задачах без повторных обращений эффект запирания системы и эффект линейности закона потерь сохраняют свою силу и при наличии вторичных заявок. Теоретические результаты подкрепляются численными расчетами. Построены области запирания системы и области действия линейного закона потерь. Исследуется влияние вероятности повторного обращения $q$ на форму этих областей, а также на кривые зависимости вероятностей потери обоих типов заявок от вероятности выталкивания $a$.

**Ключевые слова:** приоритетные системы массового обслуживания, теория массового обслуживания, абсолютный приоритет, повторные заявки, вероятностный выталкивающий механизм, линейный закон потерь, эффект запирания системы.

**1. Введение.** Системы массового обслуживания (СМО) широко используются при изучении и моделировании реальных процессов, таких, например, как передача данных с использованием компьютерных сетей, при математическом анализе разнообразных

экономических и социальных явлений, а также при моделировании различного рода производственных систем.

Основными элементами любой СМО являются входящий поток заявок, накопитель очереди и исполнительная часть, включающая некоторое число каналов обслуживания. При этом в модель системы часто приходится добавлять специфические усложнения, позволяющие описать реальные особенности ее функционирования. Так, например, аппроксимируя входящий поток в рамках модели потока Пальма, можно выбрать конкретный вид плотности вероятности интервала времени между моментами поступления заявок, а также задать надлежащий закон распределения длительности обслуживания. В многопотоковых системах часто целесообразно установить соответствующий приоритет заявок того или иного типа, а также усилить этот приоритет добавлением выталкивающего механизма, который при полном заполнении накопителя дает право высокоприоритетным заявкам выталкивать из него низкоприоритетные и занимать их место [1]. Разработан также целый ряд более специфических и тонких усложнений модели СМО, например, учитывающих фактор разогрева или охлаждения системы [2, 3], динамически изменяющиеся приоритеты [4], и, наконец, возможность повторной подачи заявок в случае их первоначального отклонения системой [5, 6, 7, 8]. В контексте рассмотрения приоритетных СМО необходимо выделить работы, в которых разработан подход к численному расчету вероятностных характеристик многоканальных приоритетных СМО с абсолютными и относительными приоритетами, при котором для преодоления проблемы размерности предложено использование распределения периода полной занятости системы обслуживанием заявок с высшим приоритетом [9, 10].

Модели систем обслуживания с повторными заявками относятся к числу наиболее важных, востребованных и практически значимых моделей современной теории массового обслуживания. Впервые возможность повторной подачи заявок была введена в науку более полувека тому назад [11]. Последние десятилетия характеризуются новой волной интереса исследователей к этим задачам. В настоящее время СМО с повторными заявками, совместно с имитационными методами используются для моделирования работы многочисленных технических систем, включая компьютерные сети, телекоммуникационные сети и другие IT-приложения. Поэтому вполне естественно, что в последние десятилетия на эту тему было опубликовано большое число статей, появившихся, в частности, в журналах по прикладной теории вероятностей, стохастическим

моделям, исследованию операций, компьютерным наукам и разнообразным инженерным приложениям.

Регулярно проводятся семинары и конференции по проблемам теории очередей с повторными заявками. Одним из наиболее известных и авторитетных таких форумов является «International Workshop on Retrial Queues and Related Topics» (WRQ). Первый семинар из этой серии состоялся в 1998 году в Мадриде, а последний на данный момент, тринадцатый – в 2021 году в онлайн формате [12].

Имеется ряд монографических исследований по теме повторных заявок. Поведение классических СМО при наличии повторных заявок с акцентом на численные методы анализа детально разобрано в книге С.Н. Степанова [13]. Необходимо также упомянуть содержательное и подробное руководство Г. Фалина и Дж. Темплтона [14], а также книгу Х. Арталехо и А. Гомес-Коррала [15], содержащую, в частности, обширную библиографию по этой тематике, включающую более семисот работ, а также изложение ряда новых современных техник численного анализа.

Новые работы по СМО с повторными заявками продолжают выходить постоянно, причем по сравнению с классическими результатами, известными ранее, значительно усложняются постановки соответствующих задач. Авторы работ стремятся максимально приблизить постановки решаемых ими задач к моделям, актуальным для современной телематики и программной инженерии. Поясним этот тезис на примере ряда последних публикаций.

Так, например, статья [16] содержит краткий обзор совместных результатов двух групп исследователей. Первая группа, работающая в Томске и возглавляемая А.А. Назаровым, занимается исследованием систем с повторными вызовами и конфликтом заявок. Под конфликтом понимается такая ситуация, когда вновь поступившая заявка может с некоторой заданной вероятностью «захватить» заявку, находящуюся на обслуживании, после чего обе они вместе покидают систему и уходят на орбиту. Вторая группа работает под руководством Я. Штрика в Венгрии и специализируется на изучении СМО с отказами каналов обслуживания. В работе [16] разбирается модель СМО с повторными заявками, способными вступать в конфликт, причем учитывая отказы канала обслуживания. Такие системы моделируют многие реальные ситуации, в частности, телекоммуникационные системы с протоколами множественного доступа при наличии коллизий (так называемый CSMA/CD протокол [17]). Похожие системы в несколько усложненном их варианте изучаются в недавно опубликованной работе [18].

В классической теории массового обслуживания не учитывается расход ресурсов, необходимых для обработки требований, что эквивалентно допущению о неограниченном изобилии ресурсов в обслуживающем приборе. Если же объем ресурса ограничен, то заявкам придется либо ждать его пополнения, либо досрочно покидать систему необслуженными. Такая ситуация типична для систем бытового обслуживания, производственных и транспортных систем, медицинских учреждений, автосервисов, торговых предприятий и других реальных систем, подобных перечисленным. Данная модель весьма интересна и для сферы IT-технологий, если под «ресурсом» понимать, например, объем памяти, требуемой для запоминания информации.

Модель СМО с повторными заявками, ограниченной емкостью и ограниченными ресурсами на обслуживание предложена в работе [19]. Модель является одноканальной и двухпотоковой, с дополнительным потоком пополнения ресурса. Оба потока считаются марковскими, первый имеет абсолютный приоритет над вторым. Приоритет проявляется во внеочередном обслуживании высокоприоритетных требований, и в первоочередном выделении им ресурса. Если заявка не может попасть в систему или ей не хватает ресурса, она направляется на орбиту.

В работе [20] содержится обзор последних работ по СМО с повторными обращениями при наличии так называемых «отрицательных» заявок. Под отрицательными заявками (G-заявками) понимаются особые заявки, которые поступают в систему не для того, чтобы обслужиться, а для того, чтобы захватить обычные («положительные») заявки и удалить их из системы. Повторные очереди в присутствии отрицательных заявок представляют собой идеальную модель для описания многих реальных ситуаций, встречающихся в программной инженерии, например, проникновения вируса в телематическую систему, сбоев в работе колл-центров, функционирования простого протокола передачи почты (SMPT).

В статье [21] рассматривается одноканальная СМО с ограниченным буфером и несколькими простейшими входящими потоками заявок, обслуживаемых по одному и тому же показательному закону. В настоящей статье исследуется аналогичная система с двумя входящими потоками, но зато при наличии приоритета первого потока, а также вероятностного (рандомизированного) выталкивающего механизма. Считалось, что время пребывания на орбите распределено по показательному закону. Это допущение является наиболее распространенным в литературе,

хотя в последнее время стали рассматривать и произвольный закон распределения [22]. Наша работа нацелена на то, чтобы прежде всего учесть наличие вероятностного выталкивания, поэтому указанное обобщение пока оставлено в стороне.

Использование вероятностного выталкивающего механизма было продиктовано особенностями космического эксперимента «Контур-2» [23]. В этом эксперименте реально использовалась модель с абсолютным приоритетом. Изменяя параметр выталкивающего механизма $\alpha$ (он равен вероятности выталкивания низкоприоритетной заявки высокоприоритетной) удавалось очень эффективно управлять вероятностями потери. Например, для типичных вариантов задания исходных данных, представленных в работах [24, 25], при увеличении вероятности выталкивания от нуля до единицы вероятность потери высокоприоритетных заявок уменьшалась в $10^{22}$ раз.

Основной задачей космического эксперимента «Контур-2» в части, касающейся обработки информационных потоков, была организация эффективной передачи информации по каналам связи ограниченной пропускной способности с целью управления непосредственно с поверхности Земли объектами, расположенными в космосе на борту МКС. Для решения данной задачи была реализована приоритетная модель СМО, в которой сетевые пакеты с более высоким приоритетом занимали в накопителе системы место ближе к каналу обслуживания, чем пакеты низкоприоритетных заявок, а также имели преимущество по постановке в очередь за счет своего права выталкивать низкоприоритетные пакеты.

Использование такой модели, однако, все равно не учитывало полностью все нюансы проведения космического эксперимента. Для повышения точности моделирования было бы весьма целесообразным учесть также возможность повторной подачи заявок. Соответствующие методы детально разработаны [13, 14, 15] и в комбинации с вероятностным выталкиванием позволяют существенно повысить надежность и отказоустойчивость системы [23].

В имеющейся литературе по СМО с повторными требованиями приоритетные модели ранее рассматривались, однако, насколько нам известно, они не охватывали выталкивающий механизм. Некоторое представление о состоянии исследований по теории приоритетных СМО с вероятностным выталкивающим механизмом дают работы авторов настоящей статьи [4, 5, 24, 25] и приведенная там библиография. Повторные обращения в них фигурировали только в работе [5], но в ней разобран другой тип приоритета –

относительный, а не абсолютный, как того требует физическая постановка задачи [23].

Согласно системе обозначений приоритетных СМО, предложенной Г.П. Башариным [1], которая расширяет классическую нотацию Д. Кендалла [26, 27], наша система имеет обозначение:

$$\overrightarrow{M_2}/M/1/k/f_2^1. \quad (1)$$

Здесь первый символ $\overrightarrow{M_2}$ означает, что на вход поступают два простейших потока требований, второй символ M говорит, что оба они обслуживаются по одному и тому же показательному закону, единица в третьей позиции указывает на наличие одного канала обслуживания, а символ приоритета $f_2^1$ соответствует абсолютному приоритету и вероятностному выталкивающему механизму. Отметим, что в оригинальной работе Башарина [1] для верхнего индекса в символе приоритета предусматривались только два значения 0 (без выталкивающего механизма) и 2 (детерминированный выталкивающий механизм). Использовать 1 в случае вероятностного выталкивающего механизма предложили авторы настоящей статьи [4, 5, 24, 25]. Система вида (1) уже изучалась ранее авторами [24, 25], но без возможности повторной подачи заявок. Между тем, хорошо известно, что учет повторных обращений способен кардинально изменить свойства системы [13, 14, 15].

**2. Сведение модели к системе без повторных заявок.** Рассматриваемая нами СМО класса $\overrightarrow{M_2}/M/1/k/f_2^1$, но без повторных заявок ранее была детально рассмотрена авторами в работах [24, 25]. Учесть возможность повторного возвращения отклоненных заявок в систему удается сравнительно просто, поскольку без учета фактора повторных обращений задача уже была решена ранее.

Рассмотрим заявку, которая после первичного поступления в СМО покинула систему необслуженной (вообще не попала в нее или была вытеснена из накопителя). Такая заявка должна встать в особую очередь на повторное попадание в СМО. Данная очередь формируется независимо от системы вне ее границ и называется орбитой системы массового обслуживания [13].

Заявка, которая попала на орбиту, находится на орбите случайный промежуток времени, а потом вновь пытается вернуться в СМО. Первым приближением при описании орбиты является представление ее в рамках марковской модели, в которой каждая заявка независимо от других занимает орбиту случайное время,

распределенное по показательному закону.

Описанная СМО представляет собой двухпотоковую приоритетную систему класса $\overrightarrow{M_2}/M/1/k/f_2^1$. В ней интервалы между первичными поступлениями в систему высокоприоритетных и низкоприоритетных заявок распределены по показательному закону. Оба потока являются простейшими с параметрами $\lambda_{0,i}$, где $i = \overline{1,2}$ определяет номер потока. Для i-го потока интервал между требованиями распределен по показательному закону:

$$a_i(\tau) = \lambda_{0,i} e^{-\lambda_{0,i}\tau}, (\ i=\overline{1,2}), \tag{2}$$

причем время обслуживания не зависит от типа потока и для всех требований имеет показательное распределение с параметром $\mu$:

$$b_1(x) = b_2(x) = \mu e^{-\mu x}. \tag{3}$$

Допустим, что время пребывания заявки i-го типа на орбите имеет показательное распределение с параметром $\gamma_i$ $(i = \overline{1,2})$, так что:

$$c_i(\tau) = \gamma_i e^{-\gamma_i \tau}, (i = \overline{1,2}). \tag{4}$$

В общем случае заявка типа i, пробыв на орбите время, распределенное по закону (4), с заданной вероятностью $q_i$ пытается вновь попасть в систему и встать в очередь на обслуживание. Соответственно, с вероятностью $(1 - q_i)$ эта заявка теряется окончательно, покидая не только СМО, но и орбиту.

Если обозначить длину очереди заявок i-го типа в момент времени $t$ на орбите как $N_{orb,i}(t)$, то очевидно, что поток повторных заявок i-го типа на выходе орбитальной очереди будет простейшим с интенсивностью $\gamma_i$, что обусловлено показательным распределением (4). При этом суммарный поток всех заявок с орбиты, направляемых в СМО, будет суперпозицией простейших потоков всех типов заявок. Как показано в [13], поток повторных заявок i-го типа будет являться простейшим с интенсивностью:

$$\lambda_{rep,i} = \gamma_i \overline{n_{orb,i}}, (i = \overline{1,2}), \tag{5}$$

где $\overline{n_{orb,i}}$ является средней длиной очереди повторных заявок i-го типа. При этом необходимо отметить, что поток повторных заявок статистически не зависит от исходного входящего потока первичных заявок.

В работе [13] также показано, что средняя длина очереди заявок i-го типа на орбите определяется следующим выражением:

$$\overline{n_{orb,i}} = \frac{\lambda_{0,i} P_{loss}^{(i)} q_i}{(1 - P_{loss}^{(i)} q_i)\gamma_i}, (i = \overline{1,2}), \quad (6)$$

где $P_{loss}^{(i)}$ обозначает полную вероятность потери заявки i-го типа. Подставляя выражение (6) в уравнение для интенсивности потока (5), получим:

$$\lambda_{rep,i} = \frac{\lambda_{0,i} P_{loss}^{(i)} q_i}{\left(1 - P_{loss}^{(i)} q_i\right)}, \left(i = \overline{1,2}\right). \quad (7)$$

Фактическая интенсивность i-го потока заявок на входе СМО, учитывающая как первичные, так и повторные заявки, может быть вычислена в виде суммы:

$$\lambda_i = \lambda_{0,i} + \lambda_{rep,i}, (i = \overline{1,2}), \quad (8)$$

с помощью (7) ее можно преобразовать к виду:

$$\lambda_i = \frac{\lambda_{0,i}}{1 - P_{loss}^{(i)} q_i}, \left(i = \overline{1,2}\right). \quad (9)$$

Это выражение показывает, что из-за наличия орбиты, суммарная интенсивность полного потока на входе системы может заметно превзойти интенсивность соответствующего первичного потока заявок.

В работах [24, 25] показано, что вероятность потери для заявок обоих типов в системах класса $\overrightarrow{M_2}/M/1/k/f_2^1$ представляет собой функцию от следующих параметров модели:

$$P_{loss}^{(i)} = \phi_i(\rho_1, \rho_2, k, \alpha), (i = \overline{1,2}), \quad (10)$$

где $\rho_1, \rho_2$ – полные коэффициенты загрузки по каждому из входящих потоков заявок, $k$ – емкость системы, $\alpha$ – параметр выталкивающего механизма ($\rho_i > 0, i = \overline{1,2}; 0 \le \alpha \le 1;\ k = \overline{0,\infty}$). Важно заметить, что функция $\varphi_i$ монотонно возрастает относительно аргумента $\rho_i$, так как при увеличении интенсивности потока частота потерь требований из этого потока не может уменьшиться.

Вводя коэффициенты загрузки по первичным потокам $\rho_{i,0}$, получим выражения для полных коэффициентов загрузки $\rho_1, \rho_2$ с учетом повторных поступлений заявок с орбиты в следующем виде:

$$\rho_i = \frac{\lambda_i}{\mu} = \frac{\rho_{i,0}}{1 - \phi_i(\rho_1, \rho_2, k, \alpha) q_i}, \left(i = \overline{1,2}\right). \quad (11)$$

Решая систему уравнений (11) относительно неизвестных $\rho_1$ и $\rho_2$, легко найти значения вероятностей потери для заявок обоих типов.

Необходимо отметить, что система (11) имеет единственное решение. Для исследования разрешимости данной системы преобразуем ее уравнения к виду:

$$1 - \phi_i(\rho_1, \rho_2, k, \alpha) q_i = \frac{\rho_{i,0}}{\rho_i}, (i = \overline{1,2}), \quad (12)$$

далее можно графически представить на одном рисунке функции зависимостей правой и левой частей уравнения (12) от аргумента $\rho_i$ при неизменных значениях остальных параметров.

Ниже на рисунке 1 график, обозначенный цифрой «1», представляет функцию из левой части равенства (12), а гипербола, обозначенная цифрой «2», выражает правую часть (12). Уравнение (12) имеет единственное решение, лежащее на интервале $(\rho_{i,0}; \frac{\rho_{i,0}}{1-q_i})$:

$$\rho_{i,0} < \rho_i < \frac{\rho_{i,0}}{1 - q_i}, (i = \overline{1,2}).$$

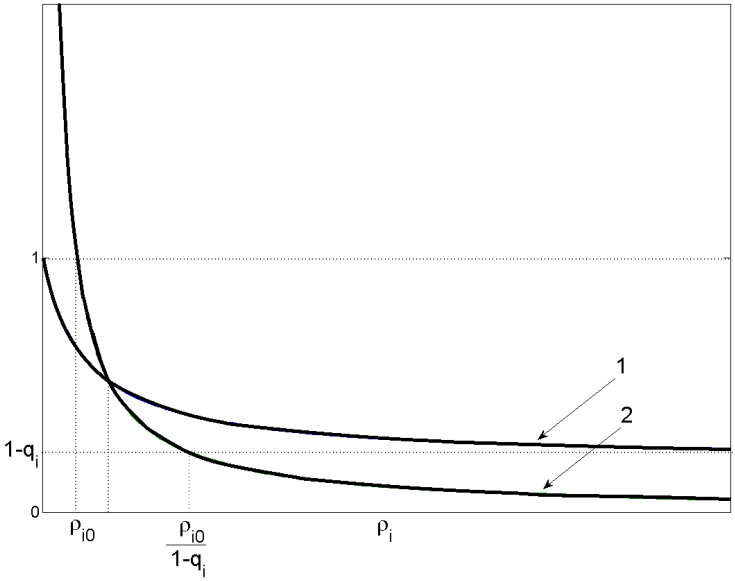



Рис. 1. Левая (1) и правая (2) части уравнения (12)

**3. Традиционная двухпотоковая система с повторными заявками.** Входящий поток нашей СМО включает два вида потоков: потоки первичных и повторных заявок, причем все эти маргинальные потоки независимы и являются простейшими. Суммарный входящий поток также будет простейшим, поэтому система с повторными вызовами будет принадлежать к тому же классу $\overrightarrow{M_2}$/M/1/k/$f_2^1$, что и аналогичная система без повторных заявок. Различие между двумя указанными системами состоит только в числовых значениях интенсивностей входящих потоков: при повторении запросов они увеличиваются.

Обозначим интенсивность первичного поступления заявок в систему для высокоприоритетного и низкоприоритетного трафика, соответственно, через $\lambda_1$ и $\lambda_2$, вероятность выталкивания – через α, интенсивность обслуживания любой заявки – через μ. Вытесненные из очереди высокоприоритетные и низкоприоритетные заявки с вероятностями настойчивости $q_1$ и $q_2$, соответственно, направляются на свои участки орбиты, емкость последней не ограничивается. Оттуда заявки вновь пытаются попасть в систему. Схема такой СМО представлена на рисунке 2.

Интенсивности появления вторичных высокоприоритетных и низкоприоритетных заявок обозначим через $\lambda_{1,rep} = q_1\gamma_1$ и $\lambda_{2,rep} = q_2\gamma_2$, соответственно, где $\gamma_1$ и $\gamma_2$ – интенсивности выхода заявок с орбиты, а $\lambda_{1,rep}$ и $\lambda_{2,rep}$ задаются выражениями (7).

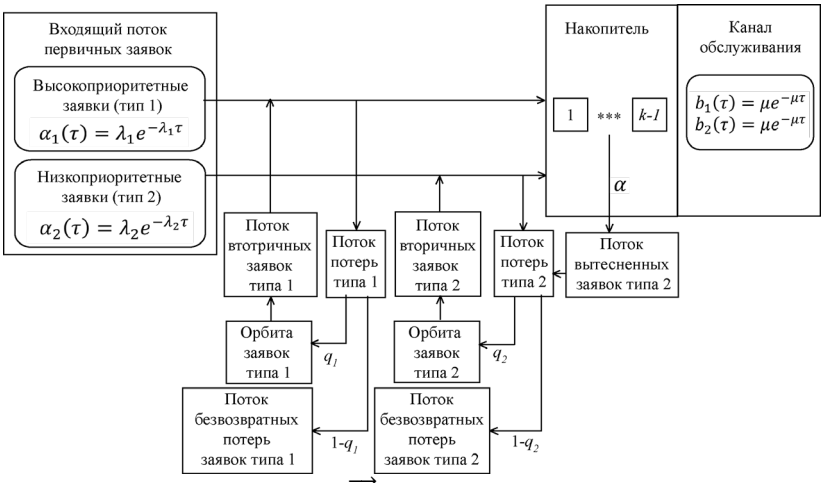



Рис. 2. Схема СМО класса $\overrightarrow{M_2}/M/1/k/f_2^1$ с повторными заявками

Работа системы, представленной на рисунке 2, полностью повторяет работу СМО без повторных заявок, детально разобранную в статьях [24, 25]. Если пересчитать эффективные интенсивности входящих потоков так, как это было описано в разделе 2 настоящей статьи, то дальнейшие вычисления полностью повторяют алгоритм решения задачи, изложенный в [24, 25].

Финальные вероятности системы вычисляются методом производящих функций. Сущность этого метода состоит в следующем. Вначале записывается система уравнений Колмогорова для финальных вероятностей состояния системы $p_{i,j}$, где i обозначает число высокоприоритетных, а j – низкоприоритетных требований в системе, причем $0 \le i + j \le k$. Обозначим через G(u,v) производящую функцию вероятностей $p_{i,j}$. Для ее вычисления уравнение Колмогорова с номером (i,j) домножается на $u^i v^j$, после чего левые части всех таких уравнений суммируются по множеству допустимых значений i, j.

В результате функция G(u,v) представляется в виде отношения полинома степени k+2 относительно аргументов u и v к полиному

второй степени от тех же аргументов. Полином в знаменателе имеет относительно u два корня $u_1(v)$ и $u_2(v)$, которые являются полюсами функции G. Но по смыслу задачи G(u,v) сама является полиномом степени k и не должна иметь никаких полюсов. Приравнивая к нулю вычеты G при u=$u_1$ и u=$u_2$, получаем выражение для G, в котором сохраняются только так называемые «опорные» вероятности $p_i = p_{i,k-i}(i=\overline{0,k})$.

Достоинство этого метода состоит в том, что он позволяет перейти от решения полной системы уравнений Колмогорова, имеющей порядок $\frac{k(k+1)}{2}$, к решению «укороченной» системы порядка (k+1) для «опорных» вероятностей, все остальные финальные вероятности линейно выражаются через $p_i$ . Указанным методом получены все приводимые ниже числовые результаты.

**4. Однопотоковая система с повторными заявками.** В этом разделе будет рассмотрен модифицированный практически значимый вариант системы п.3. Речь идет фактически об однопотоковой СМО, на вход которой поступают заявки лишь одного типа, имеющие право повторного обращения. При этом заявки из первичного потока рассматриваются как высокоприоритетные. Первичные заявки, которые были потеряны из-за отсутствия мест в очереди, с вероятностью $q_1 = 1$, попадают на орбиту системы. Если заявка была вытеснены из системы не впервые, то есть, если она уже успела посетить орбиту, то такая заявка вновь направляется на орбиту с заданной вероятностью настойчивости $q$ , и, соответственно, с вероятностью $1 - q$ теряется безвозвратно. Схема описанной однопотоковой системы приведена на рисунке 3.

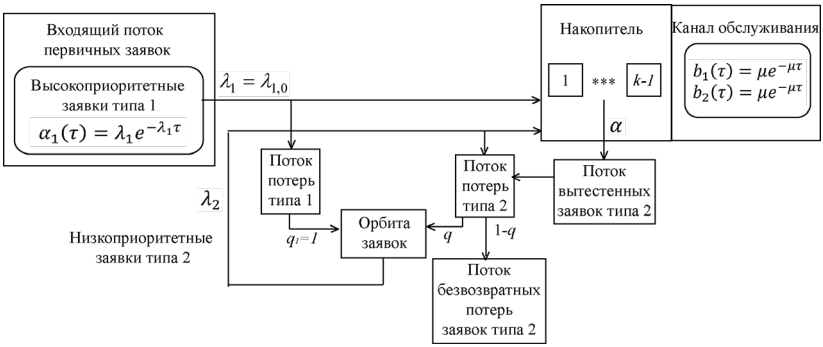



Рис. 3. Схема однопотоковой СМО с орбитой и повторными заявками

Для данной СМО интенсивности входящих потоков заявок вычисляются очевидным образом:

$$\begin{cases} \lambda_1 = \lambda_{1,0}, \\ \lambda_2 = \lambda_{1,rep} + \lambda_{2,rep} = \dfrac{\lambda_{1,0} P_{loss}^{(1)}}{1 - P_{loss}^{(1)}} + \dfrac{\lambda_2 q P_{loss}^{(2)}}{1 - q P_{loss}^{(2)}}. \end{cases} \quad (13)$$

Тогда эффективные коэффициенты загрузки по каждому типу заявок можно получить, решая систему уравнений:

$$\begin{cases} \rho_1 = \rho_{1,0}, \\ \rho_2 = \dfrac{\rho_{1,0}\phi_1(\rho_1, \rho_2, k, \alpha)}{1 - \phi_1(\rho_1, \rho_2, k, \alpha)} \cdot \dfrac{1 - q\phi_2(\rho_1, \rho_2, k, \alpha)}{1 - 2q\phi_2(\rho_1, \rho_2, k, \alpha)}. \end{cases} \quad (14)$$

Далее, используя найденные значения $\rho_1$ и $\rho_2$, можно исследовать и эту СМО методами, изложенными в работах [24, 25]. В результате были численно построены зависимости вероятностей потерь каждого из типов заявок от параметров системы, в том числе от вероятности настойчивости $q$.

**5. Числовые результаты и их анализ.** В СМО с ограниченным накопителем наиболее интересными для исследования являются вероятности потерь. В процессе их вычисления для обоих типов заявок в работах [24, 25] вначале определялись «опорные» вероятности $p_i (i = \overline{0, k})$, то есть вероятности состояний, в которых система целиком заполнена требованиями, причем из них i являются низкоприоритетными, а $(k - i)$ высокоприоритетными. Вероятности потери для заявок обоих типов выражаются через опорные вероятности следующим образом:

$$P_{loss}^{(1)} = p_0 + (1 - \alpha)\sum_{i=1}^{k-1} p_i \,, P_{loss}^{(2)} = \sum_{i=0}^{k} p_i + \frac{\rho_1}{\rho_2}\alpha \sum_{i=1}^{k-1} p_i + \frac{\rho_1}{\rho_2} p_k, \quad (15)$$

где:

1. $\rho_1, \rho_2$ – коэффициенты загрузки по каждому типу заявок;
2. $\alpha$ – вероятность срабатывания выталкивающего механизма;

3. $q_1, q_2$ – вероятности настойчивости каждого типа отклоненных заявок;

4. $k$ – емкость системы (включая $k$-1 место ожидания и одно место обслуживания).

Для исследования влияния различных факторов на вероятности потерь рассмотрим два тестовых варианта загрузки, в которых полная загрузка близка к единице, но с преобладанием разных типов заявок. Значения маргинальных коэффициентов загрузки зададим так: в первом варианте превалируют низкоприоритетные заявки ($\rho_1 = 0.2, \rho_2 = 0.9$), а во втором – наоборот ($\rho_1 = 1.2, \rho_2 = 0.2$). Ряд качественно важных результатов, касающихся зависимости вероятности потери от параметров ( $\alpha, q_1, q_2$ ), были получены для высокоприоритетного трафика в случае слабой загрузки (рисунок 4) и для низкоприоритетного в случае сильной загрузки (рисунок 7).

Графики рисунка 4 демонстрируют близкую к линейной зависимость вероятности потери высокоприоритетных заявок при слабой загрузке системы. Этот же факт численно подтверждается для низкоприоритетных заявок (рисунок 5). Данный эффект в работах [24, 25] был назван «линейным законом потерь», и он наблюдается, как видим, и для модели с повторными заявками. Важно заметить, что учет повторных заявок позволяет существенно (в 3-4 раза) снизить уровень потерь, что хорошо согласуется с данными натурных наблюдений.

Аналогичные зависимости были построены для случая сильной загрузки системы (рисунки 6, 7). При сильной загрузке линейность кривой потерь с повторными заявками отсутствует. При этом их повторная подача снижает уровень потерь в 2-3 раза.

В работах [24, 25] введено понятие области запирания (это совокупность значений коэффициентов загрузки$(\rho_{1,}, \rho_2)$, при которых, увеличивая вероятность выталкивания $\alpha$, можно сделать вероятность потери низкоприоритетных требований близкой к единице). Рисунок 7 иллюстрирует этот эффект для системы с повторными требованиями. Если вероятность настойчивости $q_1$ близка к единице, то запирание происходит уже для α порядка 10-20%.

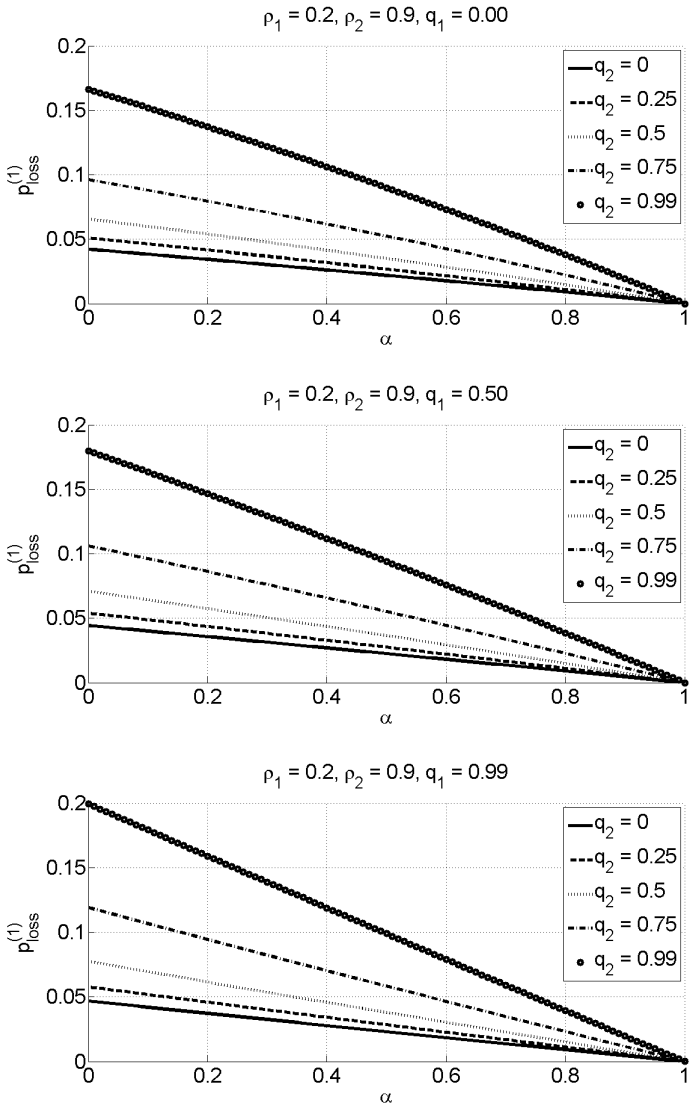



Рис. 4. График зависимости вероятности потери высокоприоритетных заявок от параметра $\alpha$ при различных $q_1$ и $q_2$ для слабой загрузки системы ($\rho_1 = 0.2, \rho_2 = 0.9$).

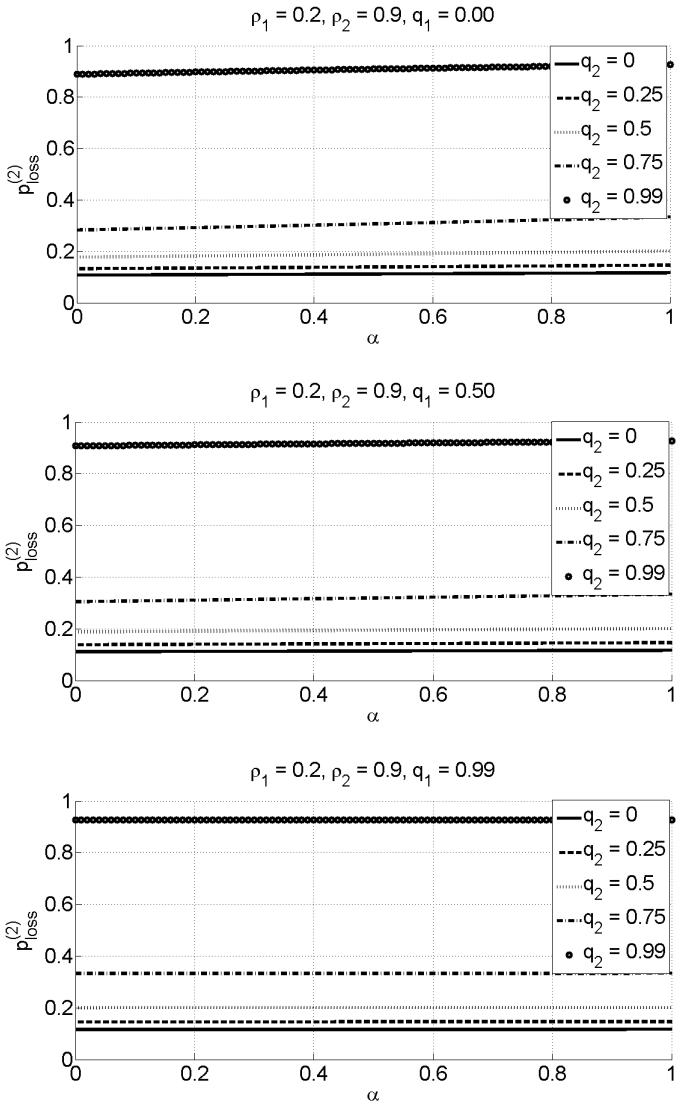



Рис. 5. График зависимости вероятности потери низкоприоритетных заявок от параметра $\alpha$ для различных $q_1$ и $q_2$ в случае слабой загрузки системы ($\rho_1 = 0.2, \rho_2 = 0.9$).

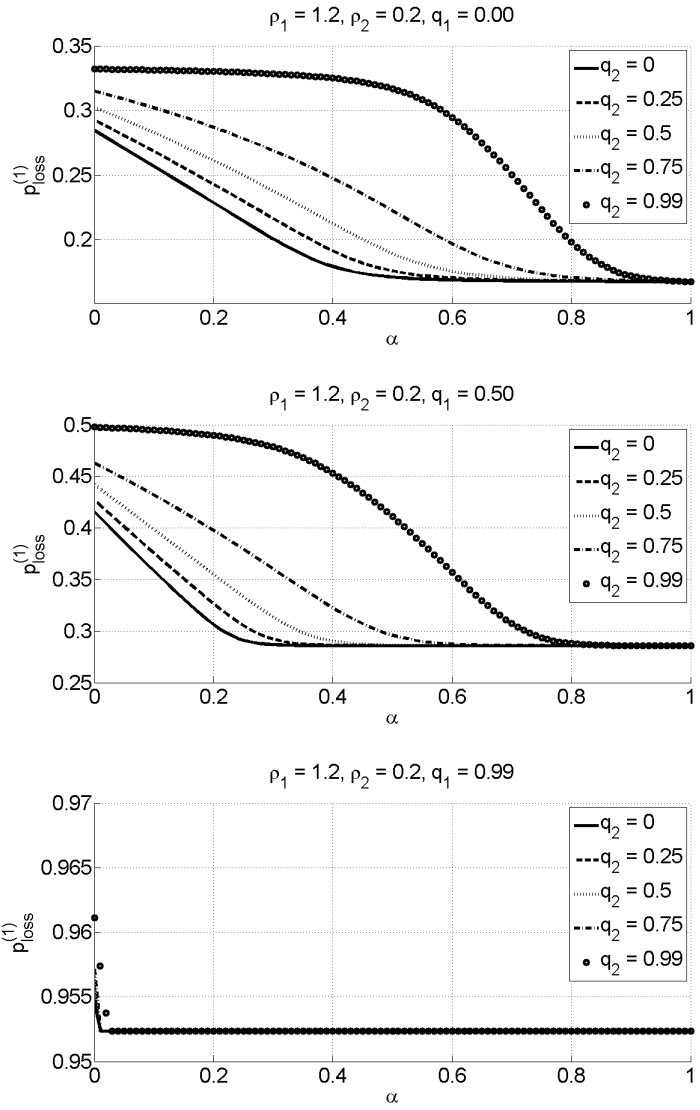



Рис. 6. График зависимости вероятности потерь высокоприоритетных заявок от параметров $\alpha, q_1, q_2$в случае сильной загрузки $\rho_1 = 1.2, \rho_2 = 0.2$

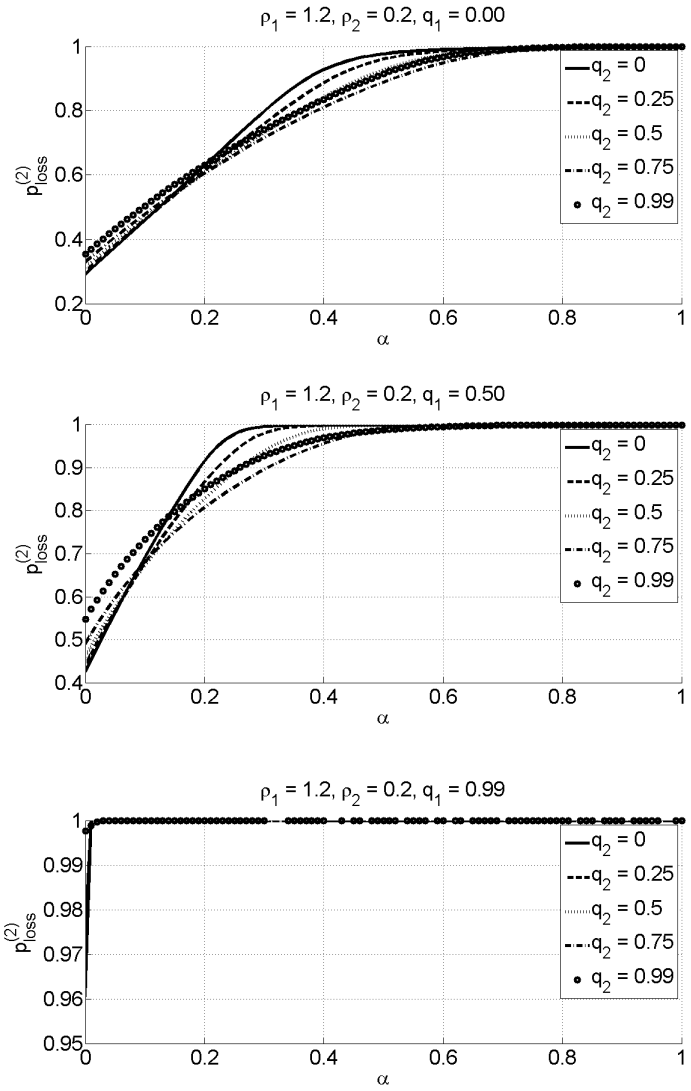



Рис. 7. График зависимости вероятности потерь низкоприоритетных заявок от параметров $\alpha, q_1, q_2$в случае сильной загрузки $\rho_1 = 1.2, \rho_2 = 0.2$

Последним из остававшихся открытым вопросов является вопрос о влиянии на потери суммарной емкости k системы.

На рисунке 8 приведена соответствующая зависимость в диапазоне от 5 до 100 единиц. Дальнейшее увеличение k оказалось лишенным смысла, так как незначительно влияло на вероятность потерь.

Численные расчеты были проведены также и для модифицированной однопотоковой системы, описанной в п. 4 настоящей работы. На рисунке 9 представлены графики зависимости вероятностей потери первичных заявок, а также повторных заявок от параметра $q$. Из графиков видно, что увеличение данного параметра позволяет существенно уменьшить вероятность потери (как минимум, в 5-6 раз).

**Заключение.** В рамках данной работы вначале была исследована традиционная марковская двухпотоковая СМО, сочетающая в себе комбинацию следующих трех элементов: 1) наличие повторных заявок; 2) абсолютный приоритет одного из типов заявок; 3) вероятностный выталкивающий механизм. Далее приводится исследование еще одной новой модели СМО, на вход которой подается всего лишь один первичный поток требований, заявки из которого трактуются как высокоприоритетные. Появляющиеся повторные заявки формируют дополнительный низкоприоритетный входящий поток. Доказано, что анализ такой, фактически однопотоковой модели сводится к анализу традиционной двухпотоковой модели, если надлежащим образом задать параметры последней. Для обеих упомянутых выше моделей СМО с повторными заявками методом производящих функций получено финальное распределение вероятностей состояния. Детально изучена зависимость значений вероятностей потери заявок от основных параметров системы.

Одним из практически значимых результатов работы является изучение эффекта запирания системы в условиях подачи повторных требований. Этот результат можно применить на практике для снижения вычислительной нагрузки на телематические устройства, работающие в режиме реального времени. Знание областей запирания позволяет заранее рационально выбрать параметры модели и избежать громоздких трудоемких вычислений в реальном времени. Это способно существенно повысить производительность телекоммуникационных устройств и оперативность управления ими. Об этом свидетельствует, в частности, опыт применения изложенной в статье теории в задачах управления роботами в серии космических экспериментов на борту МКС.

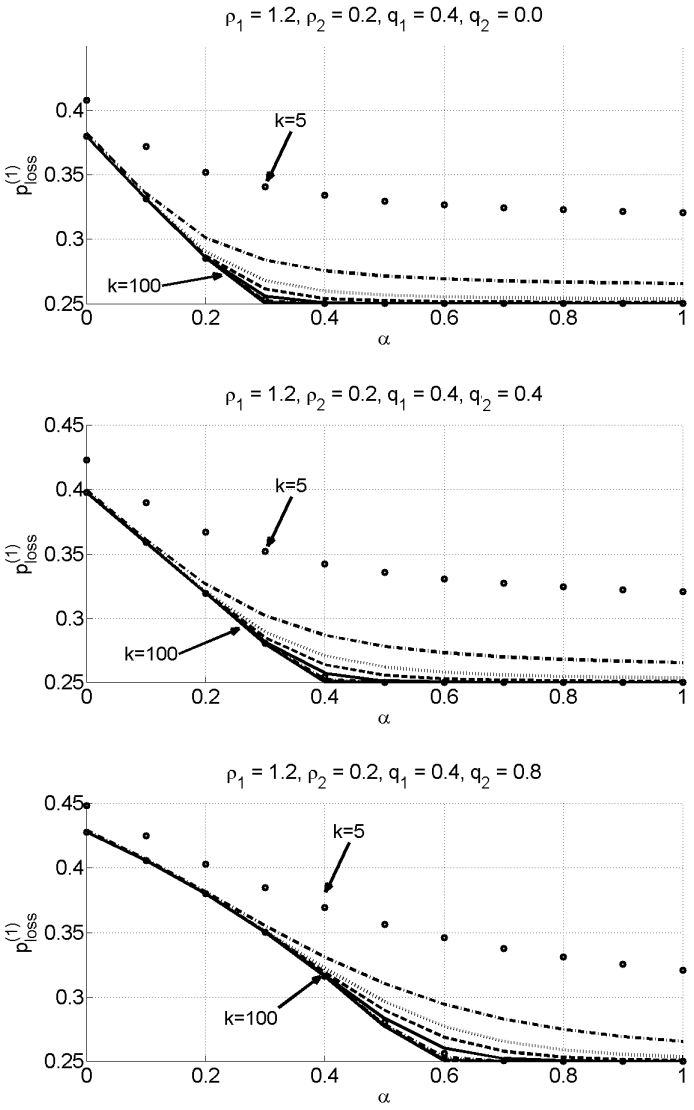



Рис. 8. График зависимости вероятности потерь от параметра $\alpha$ в случае сильной загрузки системы

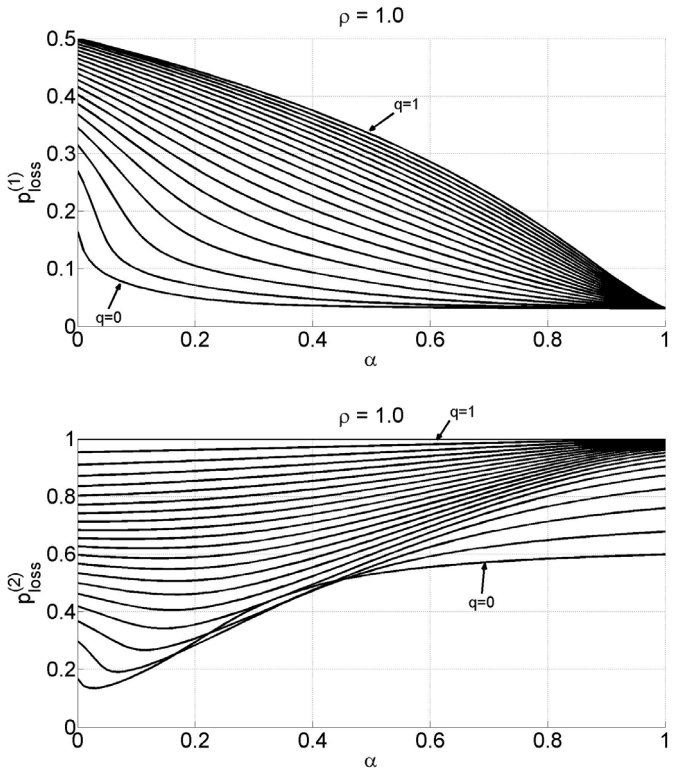



Рис. 9. График зависимости вероятности потери от параметра $\alpha$ в модифицированной системе для различных значений вероятности $q$

Другим важным прикладным результатом работы является теоретическое обоснование возможности построения предложенным методом областей действия линейного закона потерь. Эта техника ранее была подробно описана авторами в работах [24, 25]. Статья содержит разъяснения, касающиеся особенностей ее применения в случае наличия повторных обращений.

## Литература

1. Башарин Г.П. Некоторые результаты для систем с приоритетом // Массовое обслуживание в системах передачи информации. 1969. С. 39–53.
2. Хабаров Р.С., Лохвицкий В.А., Корчагин П.В. Расчет временных характеристик системы массового обслуживания с процессами расщепления и слияния заявок и разогревом // Вестник российского нового университета. Серия: сложные системы: модели, анализ и управление. 2021. № 2. С. 10–19.

3. Лохвицкий В.А., Гончаренко В.А., Левчик Э.С. Модель масштабируемого микросервиса на основе системы массового обслуживания с «охлаждением» // Интеллектуальные технологии на транспорте. 2022. № 1(29). С. 39–44.
4. Ilyashenko A., Zayats O., Muliukha V. and Lukashin A. Alternating priorities queueing system with randomized push-out mechanism // Internet of Things, Smart Spaces, and Next Generation Networks and Systems: 15th International Conference, NEW2AN, and 8th Conference, ruSMART. 2015. pp. 436–445.
5. Korenevskaya M., Zayats O., Ilyashenko A., Muliukha V. Retrial queuing system with randomized push-out mechanism and non-preemptive priority // Procedia Computer Science. 2019. vol. 150. pp. 716–725.
6. Keerthiga S., Indhira K. Two phase of service in M/G/1 queueing system with retrial customers // The Journal of Analysis. 2023. pp. 1–27.
7. Saravanan V., Poongothai V., Godhandaraman P., Performance analysis of a multi server retrial queueing system with unreliable server, discouragement and vacation model // Mathematics and Computers in Simulation. 2023. vol. 214. pp. 204–226.
8. Danilyuk E.Yu., Moiseeva S.P., Sztrik J. Asymptotic analysis of retrial queueing system M/M/1 with impatient customers, collisions and unreliable server // Journal of Siberian Federal University. Mathematics and Physics. 2020. vol. 13. no. 2. pp. 218–230.
9. Хабаров Я.С., Хомоненко А.Д. Расчет многоканальной системы массового обслуживания с прерываниями и гиперэкспоненциальными распределениями времен обработки заявок и периода непрерывной занятости // Наукоемкие технологии в космических исследованиях Земли. 2019. Т. 11. № 5. С. 48–56.
10. Краснов С.А., Лохвицкий В.А., Хабаров Р.С. Численный анализ многоканальных систем массового обслуживания с абсолютным приоритетом на основе фазовой аппроксимации периода непрерывной занятости // Труды Военно-космической академии имени А.Ф. Можайского. 2022. № 682. С. 7–20.
11. Cohen J.W. Basic problems of telephone traffic theory and the influence of repeated calls // Philips telecommunications review. 1957. vol. 18. no. 2. pp. 49–100.
12. Malayil S.K.C., Varghese C.J., Krishnamoorthy K. On A Queueing Inventory System with Marked Compound Poisson Input and Exponentially distributed Batch Service. 13th International Workshop on Retrial Queues and Related Topics (WRQ-2021). 2021.
13. Степанов С.Н. Численные методы расчета систем с повторными вызовами. М.: Наука, 1983. 230 с.
14. Falin G.I., Templeton J.G.C. Retrial Queues. London: Chapman and Hall. 1997. 320 p.
15. Artalejo J.R., Gomes-Corral A. Retrial Queueing Systems. A Computational Approach. Berlin: Springer. 2008. 318 p.
16. Nazarov A., Strik J., Kvach A. A survey of recent results in finite-source retrial queues with collisions // Information technologies and mathematical modelling. Queueing Theory and Applications: 17th International Conference and 12th Workshop on Retrial Queues and Related Topics. 2018. pp. 1–15.
17. Choi B.D., Shin Y.W., Ahn W.C. Retrial queues with collision arising from unslotted CMSA/CD protocols // Queueing systems. 1992. vol. 11. no. 4. pp. 335–356.
18. Полховская А.В., Данилюк Е.Ю., Моисеева С.П., Бобкова О.С. Вероятностная модель совместного доступа с коллизиями, Н-настойчивостью и отказами // Вестник ТГУ. Управление, вычислительная техника и информатика. 2022. № 58. С. 35–46.
19. Shajin D., Dudin A.N., Dudina O., Krishnamoorthy A. A two-priority single server retrial queue with additional items // Journal of industrial and management optimization. 2020. vol. 16. no. 6. pp. 2891–2912.

20. Malik G., Upadhyaya S., Sharma R. A study of retrial G-queues under different scenarios: a review // Proceedings of international conference on scientific and natural computing (SNC 2021). 2021. pp. 211–220.
21. Morozov E., Rumyantsev A., Dey S., Deepack T.G. Performance analysis and stability of multiclass orbit queue with constant retrial rates and buckling // Performance evaluation. 2019. vol. 134(1). no. 102005.
22. Meziani S., Kernane T. Extended generator and associated martingales for M/G/1 retrial queue with classical retrial policy and general retrial times // Probability in the Engineering and Informational Sciences. 2022. vol. 37. no. 1. pp. 206–213.
23. Zaborovsky V., Muliukha V., Ilyashenko A. Cyber-Physical Approach in a Series of Space Experiments «Kontur» // Lecture Notes in Computer Science. 2015. vol. 9247. pp. 745–758.
24. Ilyashenko A., Zayats O., Muliukha V., Laboshin L. Further Investigations of the Priority Queuing System with Preemptive Priority and Randomized Push-Out Mechanism // Internet of Things, Smart Spaces, and Next Generation Networks and Systems: 14th International Conference and 7th Conference, ruSMART. 2014. pp. 433–443.
25. Muliukha V., Ilyashenko A., Zayats O., Zaborovsky V. Preemptive Queuing System with Randomized Push-Out Mechanism // Communications in Nonlinear Science and Numerical Simulation. 2015. vol. 21. no. 1-3. pp. 147–158.
26. Джейсуол Н. Очереди с приоритетами. М.: Мир, 1973. 280 с.
27. Клейнрок Л. Теория массового обслуживания. М.: Машиностроение, 1979. 430 с.

**Заяц Олег Иванович** — канд. физ.-мат. наук, доцент, высшая школа прикладной математики и вычислительной физики физико-механического института, Санкт-Петербургский политехнический университет Петра Великого. Область научных интересов: прикладные задачи теории случайных процессов, прикладные задачи теории вероятностей, методы решения уравнения Фоккера-Планка-Колмогорова, методы решения уравнения Пугачева, теория массового обслуживания, стохастическая механика. Число научных публикаций — 112. zay.oleg@gmail.com; улица Политехническая, 29, 195251, Санкт-Петербург, Россия; р.т.: +7(812)775-0530.

**Кореневская Мария Максимовна** — выпускница, высшая школа прикладной математики и вычислительной физики физико-механического института, Санкт-Петербургский политехнический университет Петра Великого. Область научных интересов: прикладные задачи теории вероятностей, теория массового обслуживания. Число научных публикаций — 6. korenevskayamasha@gmail.com; улица Политехническая, 29, 195251, Санкт-Петербург, Россия; р.т.: +7(812)775-0530.

**Ильяшенко Александр Сергеевич** — канд. физ.-мат. наук, старший научный сотрудник, высшая школа технологий искусственного интеллекта института компьютерных наук и кибербезопасности, Санкт-Петербургский политехнический университет Петра Великого. Область научных интересов: теория массового обслуживания, численное моделирование, методологии разработки программного обеспечения, компьютерные науки. Число научных публикаций — 41. Ilyashenko.alex@gmail.com; улица Политехническая, 29, 195251, Санкт-Петербург, Россия; р.т.: +7(812)775-0530.

**Мулюха Владимир Александрович** — канд. техн. наук, директор, высшая школа технологий искусственного интеллекта института компьютерных наук и кибербезопасности, Санкт-Петербургский политехнический университет Петра

Великого. Область научных интересов: искусственный интеллект, компьютерные сети, суперкомпьютерные вычисления, кибербезопасность, управление роботами. Число научных публикаций — 75. vladimir.muliukha@spbstu.ru; улица Политехническая, 29, 195251, Санкт-Петербург, Россия; р.т.: +7(911)937-8207.

**Поддержка исследований.** Работа выполнена в рамках гос.задания ФГАОУ ВО СПбПУ (тема № FSEG-2024-0027).

**DOI** 10.15622/ia.23.2.1

O. ZAYATS, M. KORENEVSKAYA, A. ILYASHENKO, V. MULIUKHA

# PRIORITIZED RETRIAL QUEUEING SYSTEMS WITH RANDOMIZED PUSH-OUT MECHANISM

*Zayats O., Korenevskaya M., Ilyashenko A., Muliukha V.* **Prioritized Retrial Queueing Systems with Randomized Push-Out Mechanism.**

**Abstract.** The article is focused on a single-channel preemptive queuing system. Two stationary Poisson flows of customers are incoming to the system. The first flow has an absolute priority over the second one: a new high-priority customer from the first flow displaces a low-priority one from the service channel and takes its place. The capacity of the system is limited to $k$ customers. There is a probabilistic push-out mechanism in the system: if a new high-priority customer finds that all the places in the queue are occupied, then it has the right to displace one low-priority customer from the queue with probability $a$. Both types of customers have the same exponentially distributed service times. Customers who failed to enter the system due to the limited size of the queue, as well as those expelled from the queue or service channel when the push-out mechanism is triggered, are not lost immediately, but they are sent to a special part of the system called the orbit and designed to store repeated customers. In orbit, there are two separate unlimited queues, consisting of low-priority and high-priority repeated customers, respectively. If there are no free places in the system, new customers with a probability $q$ are added to the corresponding orbital queue. The waiting time of repeated customers in orbit is distributed according to an exponential law. The parameter of this law may differ for different types of customers. After waiting in orbit, secondary customers try to re-enter the system. The probabilistic characteristics of the described queuing system are calculated by the method of generating functions, previously proposed by the authors for calculating a similar system without repeated customers. This method allows finding the main probabilistic characteristics of distributions for both types of customers. Particular attention is paid to the study of the dependence of the loss probabilities for both types of customers on the parameters of the system, primarily on the push-out probability $a$, the capacity of the system $k$, and the probability of repeated circulation (probability of persistence) $q$. It is shown that the effect of blocking the system and the effect of the linear law of customers' losses, previously identified in similar problems without repeated customers, remain valid even in the presence of secondary repeated customers. The theoretical results are proved by numerical calculations. The blocking area for the second type of customers was calculated along with the area of linear loss law for both types of customers. We studied the influence of the probability of repeated circulation $q$ on the shape of these areas and on the dependence of the loss probabilities for both types of customers on the push-out probability $a$.

**Keywords:** priority queueing systems, queueing theory, absolute priority, retrial customers, randomized push-out mechanism, linear loss law, system locking effect.

## References

1. Basharin G.P. [Some results for priority systems]. Massovoe obslujivanie v sistemah peredachi infomacii – Queueing theory in data transfer systems. 1969. pp. 39–53. (In Russ.).
2. Khabarov R.S., Lokhvitskij V.A., Korchagin P.V. [Calculation of a split-merge queueing system with warm up]. Vestnik rossiyskogo novogo universiteta. Seria: Slojnie sistemi: modeli, analiz i upravlenie – Bulletin of the Russian New University. Series: complex systems: models, analysis and management. 2021. no. 2. pp. 10–19. (In Russ.).

3. Lokhvitsky V.A., Goncharenko V.A., Levchik E.S. A Scalable Microservice Model Based on a Queuing System with «Cooling». Intellectual Technologies on Transport. 2022. no. 4. pp. 46–51.
4. Ilyashenko A., Zayats O., Muliukha V. and Lukashin A. Alternating priorities queueing system with randomized push-out mechanism. Internet of Things, Smart Spaces, and Next Generation Networks and Systems: 15th International Conference, NEW2AN, and 8th Conference, ruSMART. 2015. pp. 436–445.
5. Korenevskaya M., Zayats O., Ilyashenko A., Muliukha V. Retrial queuing system with randomized push-out mechanism and non-preemptive priority. Procedia Computer Science. 2019. vol. 150. pp. 716–725.
6. Keerthiga S., Indhira K. Two phase of service in M/G/1 queueing system with retrial customers. The Journal of Analysis. 2023. pp. 1–27.
7. Saravanan V., Poongothai V., Godhandaraman P., Performance analysis of a multi server retrial queueing system with unreliable server, discouragement and vacation model. Mathematics and Computers in Simulation. 2023. vol. 214. pp. 204–226.
8. Danilyuk E.Yu., Moiseeva S.P., Sztrik J. Asymptotic analysis of retrial queueing system M/M/1 with impatient customers, collisions and unreliable server. Journal of Siberian Federal University. Mathematics and Physics. 2020. vol. 13. no. 2. pp. 218–230.
9. Khabarov Ya.S., Khomonenko A.D. [Calculation of a multi-channel queuing system with interruptions and hyper-exponential distributions of application processing times and periods of continuous employment]. Naukoyemkiye tekhnologii v kosmicheskikh issledo-vaniyakh Zemli – High-tech technologies in space exploration of the Earth. 2019. vol. 11. no. 5. pp. 48–56. (In Russ.).
10. Krasnov S.A., Lokhvitskiy V.A., Khabarov R.S. [Numerical analysis of multi-channel queuing systems with absolute priority based on phase approximation of the period of continuous employment]. Trudy Voyenno-kosmicheskoy akademii imeni A.F. Mozhayskogo – Proceedings of the Military Space Academy named after A.F. Mozhaisky. 2022. no. 682. pp. 7–20. (In Russ.).
11. Cohen J.W. Basic problems of telephone traffic theory and the influence of repeated calls. Philips telecommunications review. 1957. vol. 18. no. 2. pp. 49–100.
12. Malayil S.K.C., Varghese C.J., Krishnamoorthy K. On A Queueing Inventory System with Marked Compound Poisson Input and Exponentially distributed Batch Service. 13th International Workshop on Retrial Queues and Related Topics (WRQ-2021). 2021.
13. Stepanov S.N. Chislennie metody rascheta sistem s povtornimi vizovami [Numerical methods in retrial systems]. M.:Nauka, 1983. 230 p. (In Russ.).
14. Falin G.I., Templeton J.G.C. Retrial Queues. London: Chapman and Hall. 1997. 320 p.
15. Artalejo J.R., Gomes-Corral A. Retrial Queueing Systems. A Computational Approach. Berlin: Springer. 2008. 318 p.
16. Nazarov A., Strik J., Kvach A. A survey of recent results in finite-source retrial queues with collisions. Information technologies and mathematical modelling. Queueing Theory and Applications: 17th International Conference and 12th Workshop on Retrial Queues and Related Topics. 2018. pp. 1–15.
17. Choi B.D., Shin Y.W., Ahn W.C. Retrial queues with collision arising from unslotted CMSA/CD protocols. Queueing systems. 1992. vol. 11. no. 4. pp. 335–356.
18. Polhovskaya A.V., Daniluk E.Yu., Moiseeva S.P., Bobkova O.S. [Probabilistic sharing model with collisions, H-persistence and failures]. Bulletin of TSU. Management, Computer Engineering and Informatics – Vestnik TGU. Upravlenie, vichislitelnaya tehnika i informatica. 2022. no. 58. pp. 35–46. (In Russ).
19. Shajin D., Dudin A.N., Dudina O., Krishnamoorthy A. A two-priority single server retrial queue with additional items. Journal of industrial and management optimization. 2020. vol. 16. no. 6. pp. 2891–2912.

20. Malik G., Upadhyaya S., Sharma R. A study of retrial G-queues under different scenarios: a review. Proceedings of international conference on scientific and natural computing (SNC 2021). 2021. pp. 211–220.
21. Morozov E., Rumyantsev A., Dey S., Deepack T.G. Performance analysis and stability of multiclass orbit queue with constant retrial rates and buckling. Performance evaluation. 2019. vol. 134(1). no. 102005.
22. Meziani S., Kernane T. Extended generator and associated martingales for M/G/1 retrial queue with classical retrial policy and general retrial times. Probability in the Engineering and Informational Sciences. 2022. vol. 37. no. 1. pp. 206–213.
23. Zaborovsky V., Muliukha V., Ilyashenko A. Cyber-Physical Approach in a Series of Space Experiments «Kontur». Lecture Notes in Computer Science. 2015. vol. 9247. pp. 745–758.
24. Ilyashenko A., Zayats O., Muliukha V., Laboshin L. Further Investigations of the Priority Queuing System with Preemptive Priority and Randomized Push-Out Mechanism. Internet of Things, Smart Spaces, and Next Generation Networks and Systems: 14th International Conference and 7th Conference, ruSMART. 2014. pp. 433–443.
25. Muliukha V., Ilyashenko A., Zayats O., Zaborovsky V. Preemptive Queuing System with Randomized Push-Out Mechanism. Communications in Nonlinear Science and Numerical Simulation. 2015. vol. 21. no. 1-3. pp. 147–158.
26. Jayswal N. Ocheredi s prioritetami [Priority queues]. M.: Mir, 1973. 280 p. (In Russ.).
27. Kleinrock L. Teoria massovogo obslujivaniya [Queueing theory]. M.: Mashinostroenie, 1979. 430 p. (In Russ.).

**Zayats Oleg** — Ph.D., Associate professor, Higher school of applied mathematics and computational physics in institute of physics and mechanics, Peter the Great St. Petersburg Polytechnic University. Research interests: applied problems of the theory of random processes, applied problems of probability theory, methods for solving the Fokker-Planck-Kolmogorov equation, methods for solving the Pugachev equation, queuing theory, stochastic mechanics. The number of publications — 112. zay.oleg@gmail.com; 29, Polytechnicheskaya St., 195251, St. Petersburg, Russia; office phone: +7(812)775-0530.

**Korenevskaya Mariia** — Graduate, Higher school of applied mathematics and computational physics in institute of physics and mechanics, Peter the Great St. Petersburg Polytechnic University. Research interests: applied problems of probability theory, queuing theory. The number of publications — 6. korenevskayamasha@gmail.com; 29, Polytechnicheskaya St., 195251, St. Petersburg, Russia; office phone: +7(812)775-0530.

**Ilyashenko Alexander** — Ph.D., Senior scientific researcher, Higher school of artificial intelligence technologies in institute of computer science and cybersecurity, Peter the Great St. Petersburg Polytechnic University. Research interests: computer science, queuing theory, prioritized systems, remote control. The number of publications — 41. Ilyashenko.alex@gmail.com; 29, Polytechnicheskaya St., 195251, St. Petersburg, Russia; office phone: +7(812)775-0530.

**Muliukha Vladimir** — Ph.D., Director, Higher school of artificial intelligence technologies in institute of computer science and cybersecurity, Peter the Great St. Petersburg Polytechnic University. Research interests: AI, computer networks, HPC, cyber security, robot control. The number of publications — 75. vladimir.muliukha@spbstu.ru; 29, Polytechnicheskaya St., 195251, St. Petersburg, Russia; office phone: +7(911)937-8207.

**Acknowledgements.** The research was done with the support of the state assignment of SPbPU (Theme No. FSEG-2024-0027).

**УДК** 519.25 **DOI** 10.15622/ia.23.2.2

Е.Д. Карепова, В.С. Петракова

# СТАТИСТИЧЕСКИ ОБОСНОВАННАЯ КОРРЕКТИРОВКА ПОКАЗАНИЙ ДАТЧИКОВ СТАНЦИЙ CITYAIR УРОВНЯ КОНЦЕНТРАЦИИ ВЗВЕШЕННЫХ ЧАСТИЦ PM2.5 В ПРИЗЕМНОМ СЛОЕ АТМОСФЕРЫ ГОРОДА

*Карепова Е.Д., Петракова В.С.* **Статистически обоснованная корректировка показаний датчиков станций CityAir уровня концентрации взвешенных частиц PM2.5 в приземном слое атмосферы города.**

**Аннотация.** В качестве маркера, характеризующего загрязнение воздуха в приземном слое атмосферы современных городов, часто используется уровень концентрации твердых частиц диаметром 2.5 микрона и меньше (Particulate Matter, PM2.5). В работе обсуждается практика применения для измерения концентрации PM2.5 в условиях городской среды относительно дешевого оптического датчика, входящего в состав станции CityAir. В статье предложена статистически обоснованная корректировка получаемых станциями CityAir первичных данных о значениях концентрации взвешенных частиц PM2.5 в приземном слое атмосферы г. Красноярска. Для построения регрессионных моделей эталонными считались измерения, получаемые от анализаторов Е-BAM, расположенных на тех же постах наблюдения, что и корректируемые датчики. Для анализа использовались первичные данные 1) с 9 автоматизированных постов наблюдения краевой ведомственной информационно-аналитической системы данных о состоянии окружающей среды Красноярского края (КВИАС); 2) с 21-й станции CityAir системы мониторинга Красноярского научного центра СО РАН. В работе продемонстрировано, что при корректировке показаний датчиков необходимо учитывать метеорологические показатели. Кроме того, показано, что коэффициенты регрессии существенно зависят от сезона. Проведено сравнение методов обучения с учителем для решения задачи корректировки показаний недорогих датчиков. Дополнительная информация по результатам анализа данных, не вошедшая в текст статьи, размещена на электронном ресурсе https://asm.krasn.ru/.

**Ключевые слова:** уровень концентрации PM2.5, обучение с учителем, регрессионные модели, корректировка системы датчиков.

**1. Введение.** По данным Министерства природных ресурсов и экологии РФ город Красноярск является одним из нескольких городов России с самым грязным воздухом, концентрация вредных веществ в атмосфере города часто превышает допустимые нормы. Только за февраль 2023 года Красноярск дважды (5 и 13 февраля) попал на первое место в рейтинге крупных городов мира с высоким уровнем загрязнения атмосферы по версии сервиса IQAir, отслеживающего качество воздуха в реальном времени (https://www.iqair.com/ru/world-air-quality-ranking).

Общепринятым маркером и одновременно одним из самых вредных загрязнителей воздуха в приземном слое атмосферы современных городов являются твердые частицы диаметром 2.5 микрона и меньше (Particulate Matter, PM2.5). Взвешенные частицы PM2.5 имеют как

естественное происхождение (частицы почвы, пыль, сажа, споры растений, цветочная пыльца, а также дым от лесных пожаров) так и антропогенное (выхлопные газы двигателей автомобилей, выбросы промышленных предприятий, продукты сгорания угля или дров при отоплении). Концентрация взвешенных частиц PM2.5 является базовым показателем загрязнения городов и широко обсуждается в научной литературе [1 – 6]. Модели многофакторной линейной регрессии широко описаны в силу легкости их получения и возможности применения в практических задачах оперативного прогноза загрязнения [7 – 10]. Представляет также интерес развивающийся подход к моделированию согласованности измерений [11, 12].

В последнее время количество и качество собираемых данных, а также их детализация имеют тенденцию к росту. Поэтому помимо непосредственной работы с данными о загрязнениях и моделями их распространения уделяется большое внимание проблемам сбора и накопления информации о загрязнениях. В связи с этим представляет особый интерес оценка эффективности использования недорогих сенсоров [13 – 18].

Красноярск является одним из городов России в котором ведется мониторинг качества атмосферного воздуха на стационарных постах наблюдения. Во-первых, Министерство экологии и рационального природопользования Красноярского края поддерживает краевую ведомственную информационно-аналитическую систему данных о состоянии окружающей среды Красноярского края (КВИАС). Девять автоматизированных постов наблюдений (АПН) КВИАС расположены в г. Красноярске. Раз в 20 минут выполняется автоматическое измерение метеорологических параметров и концентрации загрязнений в приземном слое атмосферы. В КВИАС для мониторинга концентрации PM2.5 используются анализаторы пыли модели E-BAM (Met One Instruments Inc., США) [19], принцип действия которых основан на измерении поглощения $\beta$-излучения частицами пыли, осажденными на фильтрующую ленту. Эта методтка сертифицирована U.S. EPA (United States Environmental Protection Agency) [20]. Анализаторы этого класса рекомендованы для измерения содержания фракций PM10 и PM2.5 в атмосфере, сертифицированы и аккредитованы во многих странах мира, в том числе и в России (№ 57884-14 в Госреестре средств измерений).

Во-вторых, в г. Красноярск действует система мониторинга качества воздуха Красноярского научного центра СО РАН (КНЦ СО РАН) [21]. Каждый пост оснащен станцией мониторинга воздуха CityAir [22], разработанной группой компаний из новосибирского

технопарка и инновационного центра Сколково. Станция раз в 20 минут выдает основные метеорологические параметры и концентрацию аэрозольных частиц PM2.5 и PM10. Для мониторинга концентрации PM2.5 используются оптические датчики (№ 75984-19 в Госреестре средств измерений), в которых проходящий через поток загрязненного воздуха фокусированный лазерный луч рассеивается на твердых частицах, что регистрируется фотодиодом и позволяет количественно оценить загрязнение. Практика применения таких датчиков показала, что они уступают по точности анализаторам Е-BAM, однако пригодны для оценки уровня концентрации PM2.5 [18]. Система мониторинга КНЦ СО РАН имеет около 30 постов, расположенных в разных районах г. Красноярск, что обеспечивает хорошую детализацию информации о загрязнениях.

В статье анализируется согласованность показаний датчиков, принадлежащих разным системам мониторинга и предлагается статистически обоснованная корректировка первичных данных о концентрации PM2.5, получаемых с постов системы мониторинга КНЦ СО РАН.

Статистический анализ выполнен на языке Python с использованием библиотек `numpy`, `pandas`, `sklearn`, `statsmodels`. Отметим, что дополнительная информация по результатам анализа данных, не вошедшая в текст статьи, размещена на электронном ресурсе [23].

**2. Используемые для анализа данные и их обозначения.** Для анализа использовались первичные данные 1) с 9-ти автоматизированных постов наблюдения сети КВИАС; 2) с 21-й станции CityAir системы мониторинга КНЦ СО РАН. Далее в зависимости от принадлежности поста АПН КВИАС или системе мониторинга КНЦ СО РАН показатели будут иметь префикс “m_” или “s_”, соответственно. Кроме того, для краткости датчики концентрации PM2.5 АПН КВИАС и станции CityAir будем упоминать как “анализаторы Е-BAM” и “оптические датчики”, соответственно.

По каждому посту данные представлены временными рядами измерений в приземном слое атмосферы температуры ($t$) в °C, давления ($p$) в мм рт. ст., относительной влажности воздуха ($h$) в % и концентрации взвешенных частиц PM2.5 ($PM$) в мкг/м$^3$. В скобках указаны используемые далее обозначения факторов. Каждый ряд содержит до 105192 измерений (с 01.01.2019 00:00 по 31.12.2022 23:40, 3 измерения в час).

Поскольку целью работы является построение регрессионной модели для корректировки данных о концентрациях PM2.5 станций CityAir, мы будем рассматривать данные не как временные ряды,

а как связанные выборки случайных величин. Описательная статистика показывает, что распределение температуры имеет ярко выраженную трехмодальность (зимний, летний и демисезонный периоды), причем зимний сезон имеет «тяжелый хвост» в сторону низких отрицательных температур. Распределение давления близко к нормальному, распределение влажности имеет большую асимметрию с «тяжелым хвостом» влево. Гистограммы распределения для концентрации PM2.5, температуры, давления и влажности для 4-х дублирующих датчиков (карусель изображений), а также описательная статистика данных представлены в разделе «Описательная статистика» ресурса [23].

Распределение концентрации PM2.5 близко к логнормальному (таблица 1), однако имеет две моды, первая из которых соответствует фоновым значениям концентрации, а вторая, менее выраженная, – периодам высоких концентраций. Отметим, что значения, обычно определяемые в описательной статистике как выбросы (т.е. превышающие в полтора межквартильного расстояния значение третьего квартиля [24]), для нас таковыми не являются, поскольку отражают ситуацию значительного превышения предельно допустимых концентраций (ПДК) PM2.5 в атмосфере.

Таблица 1. Уровни концентрации твердых взвешенных частиц PM2.5. Описательная статистика. Данные поста «Ветлужанка»

| Статистика | Первичные данные (мкг/м$^3$) | | Первичные данные после очистки (мкг/м$^3$) | | Логарифм от первичных данных | | Логарифм от первичных данных после очистки | |
|---|---|---|---|---|---|---|---|---|
| | s | m | s | m | s | m | s | m |
| Максимум | 797.00 | 403.00 | 797.00 | 403.00 | 6.68 | 6.00 | 6.68 | 6.00 |
| Среднее | 45.90 | 24.78 | 48.19 | 25.99 | 2.85 | 2.63 | 2.94 | 2.70 |
| Ошибка среднего | 0.34 | 0.14 | 0.37 | 0.16 | 0.01 | 0.00 | 0.01 | 0.00 |
| Среднеквадратичное отклонение | 79.12 | 33.04 | 81.33 | 34.13 | 1.37 | 1.09 | 1.33 | 1.04 |
| 25% ($Q_1$) | 6.50 | 7.00 | 7.17 | 8.00 | 1.87 | 1.95 | 1.97 | 2.08 |
| Медиана ($Q_2$) | 14.50 | 14.00 | 15.50 | 14.00 | 2.67 | 2.64 | 2.74 | 2.64 |
| 75% ($Q_3$) | 42.50 | 27.00 | 45.26 | 28.00 | 3.75 | 3.30 | 3.81 | 3.33 |
| $IQR = Q_3 - Q_1$ | 36.00 | 20.00 | 38.00 | 20.00 | 1.88 | 1.35 | 1.84 | 1.25 |
| Асимметрия | 3.16 | 3.33 | 3.06 | 3.24 | 0.35 | -0.03 | 0.42 | 0.11 |
| Эксцесс | 14.64 | 17.71 | 13.83 | 16.72 | 2.56 | 3.08 | 2.54 | 3.02 |

В данных присутствуют пропуски, соответствующие периодам поломки аппаратуры. Для нашего исследования заполнение пропусков нецелесообразно, такие данные были просто удалены из выборок. Кроме того, данные о концентрации PM2.5 содержали небольшое количество случаев, которые были расценены как сбои аппаратуры. Например, если в период фоновых концентраций PM2.5 в трех идущих подряд измерениях между двумя измерениями, которые соответствуют небольшим значениям концентраций PM2.5, происходит резкий скачок показаний датчика, мы считали его сбоем аппаратуры. Такие ситуации редки ($< 10$ случаев), хаотично разбросаны по 30 выборкам и, следовательно, не отражают какую-либо тенденцию. Эти данные тоже были удалены из выборок. Следует иметь в виду, что не все ошибки измерений и сбои аппаратуры мы могли выявить, и они остались в выборке.

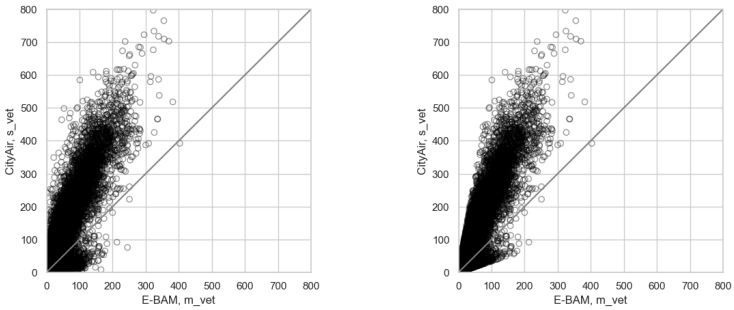
Для сравнения показаний датчиков, принадлежащих разным системам мониторинга существует 4 поста, на которых установлена измерительная аппаратура обоих типов. Это посты “Ветлужанка”, “Покровка”, “Свердовский”, “Кировский”[1].

Уже на этапе описательной статистики становится ясно, что имеются существенные различия в данных, полученных с помощью измерительной аппаратуры разного типа. Например, среднее значение концентрации PM2.5 по выборке измерений оптического датчика станции CityAir, почти в 2 раза превышают среднее значение в выборке анализатора Е-BAM.

Поскольку методика измерения концентрации PM2.5, используемая анализаторами Е-BAM, тщательно верифицирована [20], а оптические датчики принято использовать, как сравнительно дешевую альтернативу [13 – 18], то актуально построение статистически обоснованного правила корректировки первичных данных о концентрации PM2.5 оптических датчиков по данным, полученным от Е-BAM.

Диаграмма рассеяния рисунка 1(а) показывает систематическое завышение концентрации оптических датчиков. Мы приняли решение для каждой пары датчиков исключить из анализа $\sim$5 % пар значений, которые соответствуют максимальным расхождениям в показаниях уровней концентраций PM2.5 в паре. В результате из последующего анализа были исключены пары значений, в которых показания оптического датчика более чем в 6 раз отличались от показаний анализатора Е-BAM. Диаграмма рассеяния после корректировки отображена на рисунке 1(б). В таблице 1 приведена описательная статистика для скорректированных выборок концентраций PM2.5.

[1] раздел «Карты» ресурса [23]




а) б)

Рис. 1. Диаграмма рассеяния показаний о концентрации взвешенных частиц PM2.5 (мкг/м$^3$) на анализаторе E-BAM (`m_vet`) и оптическом датчике станции CityAir (`s_vet`), расположенных в Ветлужанке а) до; б) после очистки данных

Далее в тексте будут приведены результаты анализа для пары дублирующих датчиков с поста “Ветлужанка”. Для краткости будем помечать “s_vet” и “m_vet” измерения станции CityAir и датчиков АПН КВИАС, соответственно. Анализ для оставшихся трех пар датчиков, расположенных в Покровке, Свердловском и Кировском районах г. Красноярска, проводился аналогично, необходимые ссылки на электронный ресурс [23], содержащий результаты анализа по этим постам, будут даны в тексте статьи.

**3. Описание используемых для анализа методов.** Основными хорошо формализованными средствами статистики, которые используются при поиске взаимосвязей между выборками, являются 1) корреляционный анализ, выявляющий наличие линейных связей между двумя выборками, и 2) аппарат парной или множественной регрессии, устанавливающий более общую взаимосвязь между наборами данных [25 – 27]. В обоих случаях, выборки должны быть достаточного объема и хорошо описывать генеральную совокупность.

Пусть известно множество $X_{\mathcal{T}} \in R^{D\times n}$, состоящее из $n$ *связанных* выборок (факторов) мощности $D$. Каждая выборка представляется вектор-столбцом $x^j = (x_1^j, \ldots, x_D^j)^T$, $j = 1, \ldots, n$. Отметим, что $X_{\mathcal{T}}$ также можно рассматривать как множество строк-наблюдений $x_i = (x_i^1, \ldots, x_i^n)$, $i = 1, \ldots, D$. Пусть каждому наблюдению $x_i$ соответствует скалярный отклик $y_i^{\mathcal{T}}$, $i = 1, \ldots, D$. На основе этой информации в регрессионном анализе необходимо построить алгоритм оценки значения отклика $\hat{y}$ по входному набору значений факторов $\hat{x} = (\hat{x}^1, \ldots, \hat{x}^n) \notin X_{\mathcal{T}}$. Алгоритм называют *регрессионной моделью*, а множество пар $(y_i^{\mathcal{T}}, x_i)$ – *обучающим*.

В настоящее время в регрессионном анализе наряду с классическими статистическими методами используются методы машинного обучения с учителем. Кратко опишем методы, используемые в данной работе.

В большинстве случаев, мы рассматривали линейные регрессионные модели, для которых прогнозное значение отклика $\hat{y} \in R$ ищется в виде:

$$\hat{y} = f(\hat{x}) := \omega_0 + \omega_1 \hat{x}^1 + ... + \omega_n \hat{x}^n, \tag{1}$$

линейной комбинации факторов $\hat{x} = \left(\hat{x}^1, ..., \hat{x}^n\right) \in R^n$. В этом случае построение регрессионной модели сводится к определению набора параметров $\omega = (\omega_0, ..., \omega_n)$ на основе обучающего множества пар $(y_i^{\mathcal{T}}, x_i)$, где $x_i \in X_{\mathcal{T}}$, $i = 1, \ldots, D$.

Классический метод определения параметров $\omega$, разработанный еще К.Ф. Гауссом и А.А. Марковым – это метод наименьших квадратов (МНК), в котором минимизируется сумма квадратов невязок:

$$\min_{\omega} \sum_{i=1}^{D} \left(y_i^{\mathcal{T}} - f(x_i)\right)^2. \tag{2}$$

Если обучающее множество содержит выборки почти коллинеарных факторов (нечеткая мультиколлинеарность набора факторов), то МНК становится неустойчив, в том числе вычислительно. В таких случаях используют некоторую регуляризацию (2), не позволяющую параметрам $\omega$ сильно возрастать. В методе *гребневой регрессии* (Ridge) используется регуляризация Тихонова в норме пространства $L_2$ [28], в методе *LASSO* (Least Absolute Shrinkage and Selection Operator) – в норме пространства $L_1$ [29], а модель *эластичной сети* (Elastic Net) использует обе эти регуляризации [30]. При этом в моделях появляются дополнительные параметры, которые необходимо подбирать на основе специального исследования набора данных, например, методом кросс-валидации. Метод LASSO часто приводит к разреженным регрессионным моделям, поскольку некоторые компоненты в $\omega$ могут стать нулевыми и, следовательно, будут исключены из модели. Метод гребневой регрессии может делать часть компонент в $\omega$ малыми, но редко зануляет их. Эластичная сеть дает более гладкую зависимость $\omega$ от параметра регуляризации, чем метод LASSO. Таким образом, эти методы также

можно использовать для обоснования уменьшения размерности пространства факторов.

*Метод опорных векторов* (SVM, Support Vector Method) был предложен в [31], как метод классификации (метод обобщенных портретов) полезный для нечетко разделенных классов. В [32, 33] метод был расширен на решение задач регрессии, аппроксимации и оценивания функций. В методе опорных векторов наряду с минимизацией функционала (2) и его регуляризацией в пространстве $L_2$ оптимизируется и множество наблюдений (опорных векторов), по которым вычисляется сумма в (2). Метод игнорирует ошибки, меньшие заданного $\varepsilon$.

Предсказание отклика по линейной модели регрессии, обученной по обучающему множеству без выбросов, довольно устойчиво. Однако, поскольку накладывается сильное ограничение на структуру модели (линейный вид функции $f(x)$ в (1)), предсказание отклика может быть неточным [27]. Существует широкий набор методов, в которых априорные допущения о структуре модели очень слабые. В результате модель, обычно, перестает быть линейной и гибко подстраивается под обучающее множество $X_{\mathcal{T}}$. При отсутствии переобучения (о чем необходимо заботиться специально) предсказания таких методов могут быть весьма точны. Однако, обратной стороной адаптивности модели является ее неустойчивость. Структура модели сильно зависит от обучающего множества; и предсказания откликов могут сильно отличаться, если они получены по моделям, обученным даже на мало отличающихся обучающих множествах.

В работе мы сравнили результаты, полученные по описанным выше линейным моделям и двум методам, в которых структура модели заранее не фиксируется. Оба метода разрабатывались для задач классификации, но были адаптированы к задачам регрессии. В этих методах пространство объектов рассматривается, как $n$-мерное метрическое пространство с определенным расстоянием $\rho(\cdot,\cdot)$, а наблюдения $x$ – как точки в нем.

Метод $k$ *ближайших соседей* предсказывает значение отклика $\hat{y}$ по входному набору значений факторов $\hat{x} = (\hat{x}^1, \ldots, \hat{x}^n)$ на основе осреднения откликов от $k$ наблюдений из обучающего множества $X_{\mathcal{T}}$, которые являются ближайшими к значению входной переменной $\hat{x}$ по метрике $\rho$.

В регрессионной модели *дерева решений* [34, 35] пространство объектов представляется дизъюнктным объединением непересекающихся областей $R_m$, $m = 1, \ldots, M$, в каждом из которых настраивается простая

регрессионная модель (например, константа) для предсказания отклика:

$$\hat{y} = f(\hat{x}) := \sum_{m=1}^{M} c_m(\hat{x}) I(\hat{x} \in R_m).$$

Здесь $I(x \in R)$ – идентификационная функция области $R$, а $c_m(x)$ – соответствующая $R_m$ регрессионная модель. Для идентификации регрессионной модели необходимо определить правило разбиения пространства объектов на области и соответствующую каждой области $c_m$. В большинстве пакетов в настоящее время для обучения деревьев решений используется итерационный алгоритм CART [35], в котором разбиение на подобласти происходит гиперплоскостями, параллельными одной из координатных осей, а $c_m = \mathtt{avg}(y_i^{\mathcal{T}} | x_i \in R_m)$ усредняет отклики из обучающей выборки, соответствующие всем $x_i \in R_m$.

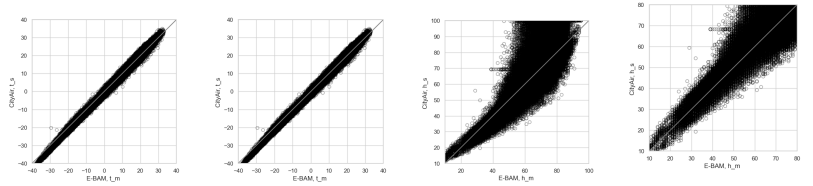
Поскольку модели, полученные по алгоритму CART являются неустойчивыми относительно обучающего множества и имеют тенденцию к переобучению, то обычно используют специальную процедуру баггинга [36], повышающую устойчивость прогноза. Мы использовали алгоритм *случайного леса* (Random Forest) [37], который генерирует ансамбль решающих деревьев и усредняет по нему предсказание.

**4. Корреляционный анализ. Сезонность.** Построенные на рисунке 2 диаграммы рассеяния для а) температуры; б) давления; в) влажности, полученных с АПН КВИАС (`m_vet`) и станции CityAir (`s_vet`) иллюстрируют[2] согласованность показаний температуры и давления (коэффициент детерминации показаний одного типа датчика относительно другого в этом случае $R^2 > 0.98$). Однако измерения влажности разными приборами существенно отличаются (рисунок 2(в)). Регрессия вида:

$$h'_{\mathtt{m\_vet}} = 0.000028 h^3_{\mathtt{s\_vet}} - 0.01305 h^2_{\mathtt{s\_vet}} + 1.9667 h_{\mathtt{s\_vet}} - 14.62964,$$

с коэффициентом детерминации $R^2 = 0.78$ улучшает согласованность показаний разных типов датчиков (рисунок 2(г)), но все равно при повышенной влажности разброс показаний станции CityAir относительно АПН КВИАС остается более 40 %.

[2]Аналогичный анализ для оставшихся пар дублируемых датчиков доступен в разделе Диаграммы рассеяния ресурса [23]




а) температура б) давление в) влажность г) скорректированная влажность

Рис. 2. Диаграммы рассеяния метеорологических параметров, полученных со станции CityAir и АПН КВИАС на посту Ветлужанка

Климат г. Красноярск континентальный с относительно морозной малоснежной зимой и жарким летом с малым количеством осадков. Более того, в холодный сезон большой вклад в концентрацию загрязнителя в атмосфере вносят работающие на полную мощность ТЭЦ и печное отопление. Поэтому естественна гипотеза о разном проявлении связи значений концентрации загрязнителя в атмосфере и метеорологических параметров в разные сезоны, что неизбежно будет влиять на модель корректировки датчиков. Для подсети дублирующих датчиков сгруппированные по месяцам коэффициенты корреляции между концентрацией PM2.5 и температурой, давлением и влажностью приведены в таблицах 2 (а–в), соответственно[3]. Заметим, что взятые по месяцам значения метеорологических параметров распределены нормально.

Корреляционный анализ позволяет быстро оценить возможность линейной связи между значениями концентраций PM2.5 и метеопараметрами. На его основе можно сделать следующие предварительные выводы. Во-первых, в холодное время года существенны отрицательная корреляция концентрации PM2.5 с температурой и положительная с давлением и влажностью, а поздней весной и летом эти корреляции практически отсутствуют. Это легко объясняется следующими причинами. Низкие температуры, отсутствие ветра и высокая влажность являются причиной температурных инверсий в нижних слоях атмосферы, что затрудняет рассеяние загрязняющих веществ [38] и приводит к периодам устойчивой повышенной концентрации их маркера PM2.5. Более того, с понижением температуры повышается интенсивность печного отопления и работы ТЭЦ, что тоже увеличивает загрязнение нижних слоев атмосферы в холодный сезон.

[3] Данные для всех 30 датчиков доступны в разделе «Корреляционный анализ» ресурса [23].

Таблица 2. Коэффициент корреляции между показаниями подсети дублирующих датчиков о концентрации частиц PM2.5 и измерениями а) температуры; б) влажности; в) давления

а)

| | Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sep | Oct | Nov | Dec |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **m_vet** | -0.48 | -0.58 | -0.49 | -0.05 | -0.13 | -0.04 | 0.03 | 0.08 | -0.06 | -0.36 | -0.10 | -0.64 |
| **m_pok** | -0.45 | -0.47 | -0.36 | 0.08 | 0.10 | 0.12 | 0.12 | 0.21 | 0.02 | -0.12 | -0.28 | -0.42 |
| **m_svr** | -0.35 | -0.42 | -0.32 | 0.21 | 0.13 | 0.27 | 0.12 | 0.21 | 0.20 | -0.05 | -0.22 | -0.52 |
| **m_kir** | -0.32 | -0.44 | -0.37 | 0.08 | 0.10 | 0.17 | -0.05 | 0.14 | -0.01 | -0.11 | -0.22 | -0.38 |
| **s_vet** | -0.55 | -0.57 | -0.52 | -0.15 | -0.22 | -0.07 | -0.16 | 0.07 | -0.25 | -0.33 | -0.33 | -0.67 |
| **s_pok** | -0.48 | -0.55 | -0.40 | -0.09 | -0.09 | -0.01 | -0.12 | 0.09 | -0.13 | -0.17 | -0.26 | -0.55 |
| **s_svr** | -0.46 | -0.36 | -0.38 | 0.16 | 0.02 | 0.07 | 0.12 | 0.22 | 0.04 | -0.13 | -0.22 | -0.45 |
| **s_kir** | -0.45 | -0.38 | -0.42 | -0.02 | -0.06 | -0.05 | -0.06 | 0.08 | -0.15 | -0.24 | -0.24 | -0.41 |

б)

| | Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sep | Oct | Nov | Dec |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **m_vet** | 0.49 | 0.46 | 0.40 | 0.12 | 0.04 | 0.10 | 0.18 | 0.21 | 0.28 | 0.27 | 0.37 | 0.49 |
| **m_pok** | 0.39 | 0.38 | 0.22 | -0.01 | -0.15 | -0.04 | -0.00 | 0.08 | 0.06 | 0.04 | 0.36 | 0.32 |
| **m_svr** | 0.48 | 0.38 | 0.28 | -0.05 | -0.07 | -0.18 | -0.04 | 0.02 | 0.01 | 0.04 | 0.36 | 0.52 |
| **m_kir** | 0.54 | 0.47 | 0.35 | 0.03 | -0.04 | 0.01 | 0.11 | 0.11 | 0.23 | 0.18 | 0.40 | 0.41 |
| **s_vet** | 0.35 | 0.19 | 0.34 | 0.18 | 0.14 | 0.18 | 0.03 | 0.12 | 0.30 | 0.22 | 0.25 | 0.05 |
| **s_pok** | 0.31 | 0.12 | 0.18 | 0.07 | 0.01 | 0.11 | 0.05 | 0.12 | 0.14 | 0.14 | 0.20 | 0.19 |
| **s_svr** | 0.35 | 0.21 | 0.18 | 0.02 | 0.04 | 0.21 | 0.13 | 0.01 | 0.16 | 0.12 | 0.15 | 0.15 |
| **s_kir** | 0.35 | 0.34 | 0.23 | 0.15 | 0.05 | 0.24 | 0.27 | 0.13 | 0.27 | 0.14 | 0.34 | 0.32 |

в)

| | Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sep | Oct | Nov | Dec |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **m_vet** | 0.13 | 0.21 | 0.07 | 0.05 | 0.15 | -0.12 | -0.28 | -0.21 | 0.02 | 0.19 | 0.03 | 0.26 |
| **m_pok** | 0.31 | 0.34 | 0.05 | 0.00 | 0.04 | -0.02 | -0.02 | -0.11 | 0.06 | 0.16 | 0.31 | 0.14 |
| **m_svr** | 0.14 | 0.26 | 0.02 | -0.05 | -0.03 | -0.04 | -0.08 | -0.14 | -0.07 | 0.09 | 0.27 | 0.40 |
| **m_kir** | 0.20 | 0.34 | 0.02 | -0.00 | -0.01 | -0.15 | -0.06 | -0.09 | -0.05 | 0.16 | 0.20 | 0.31 |
| **s_vet** | 0.35 | -0.07 | 0.08 | 0.09 | 0.19 | -0.02 | -0.07 | -0.09 | 0.09 | 0.23 | 0.29 | -0.32 |
| **s_pok** | 0.33 | 0.43 | 0.07 | 0.03 | 0.02 | 0.04 | 0.03 | 0.00 | 0.10 | 0.10 | 0.32 | 0.37 |
| **s_svr** | 0.30 | 0.26 | 0.08 | -0.05 | 0.02 | -0.23 | -0.14 | -0.12 | -0.06 | 0.17 | 0.31 | 0.34 |
| **s_kir** | 0.28 | 0.33 | 0.09 | 0.05 | 0.06 | -0.16 | -0.10 | -0.10 | 0.05 | 0.17 | 0.27 | 0.30 |

Есть и более общее замечание. Летом в целом концентрации PM2.5 ниже, чем в холодный период, а точность измерений анализатора E-BAM [19] и оптического датчика при низких концентрациях взвешенных частиц ниже, чем при высоких. Таким образом, дисперсия обоих измерений растет, что плохо отражается на их согласованности. Во-вторых, в холодный период положительная корреляция между показаниями концентрации PM2.5 и влажностью для датчиков эталонной подсети КВИАС выше, чем для датчиков станций CityAir, что, по всей видимости, объясняется невысокой точностью измерений станции CityAir.

Таким образом, из результатов корреляционного анализа следует, что параметры регрессионной модели согласования показаний парных

датчиков должны зависеть от сезона. Для последующего анализа нами были условно выделены три периода: 1) зимний (октябрь – март) с сильной корреляцией между температурой и концентрацией PM2.5); 2) летний (июнь – август); 3) демисезонный (апрель, май, сентябрь). Следует отметить, что в августе наблюдается слабая положительная корреляция загрязнений с температурой, тогда как в другие летние месяцы эта связь отсутствует. Такое поведение косвенно подтверждает, что измерения при низких концентрациях плохо согласованы, поскольку в августе в Красноярске часто возникают периоды повышения концентрации PM2.5, связанные с лесными пожарами.

**5. Модели регрессии для корректировки значений концентрации PM2.5 от датчиков станций CityAir.** Сначала мы выполним регрессионный анализ для выявления зависимости значений концентрации взвешенных частиц PM2.5, полученных с помощью оптического датчика станции CityAir, от эталонных значений концентраций PM2.5, полученных от анализатора Е-BAM, с учетом значений метеорологических параметров. Кроме того, мы сравним методы обучения регрессионной модели, описанные в разделе 3.

Имеющиеся данные были разбиты на обучающую (80% объема) и тестовую (20% объема) выборки. Качество модели оценивалось коэффициентом детерминации $R^2$, который показывает, какую долю дисперсии тестовой выборки концентраций PM2.5 объясняет модель. Поскольку значение коэффициента детерминации зависит от разбиения данных на обучающую и тестовую выборки, то процедура повторялась для 100 случайных разбиений. Далее в таблицах приведен средний коэффициент детерминации по всем попыткам, при этом среднеквадратичное отклонение не превышает процента. Ниже приведены результаты для дублирующих датчиков с поста Ветлужанка[4].

В регрессионном анализе рассматривались различные комбинации следующих факторов: концентрации PM2.5, полученные анализатором Е-BAM ($PM_m$); температура ($t_s$), давление ($p_s$) и влажность ($h_s$), полученные с помощью датчиков станции CityAir. В качестве отклика рассматривались значения концентрации PM2.5, полученные с оптического датчика CityAir ($PM_s$). Сделаем несколько замечаний относительно всех построенных в статье линейных регрессионных моделей. Во-первых, все остатки имеют нулевое среднее, медиану в районе 0,8 мкг/м$^3$, слабую отрицательную симметрию ($\sim$-0,2) и умеренный эксцесс ($\sim$ 15). По критерию Дарбина-Уотсона автокорреляции первого

[4]Результаты регрессионного анализа по остальным парам датчиков представлены в разделе «Регрессионный анализ» ресурса [23].

порядка у остатков отсутствуют. Тем не менее, статистическую проверку на нормальность остатки не проходят. Во-вторых, $t$-статистика с 5% уровнем значимости показывает, что во всех случаях коэффициенты значимо отличны от нуля. $F$-статистики с 5% уровнем значимости показывает, что во всех случаях существуют коэффициенты отличные от нуля, т.е. в этом смысле линейная модель приемлема. В-третьих, с помощью информационного критерия Акаике AIC и байесовского критерия BIC [40, 41] проведено сравнение качества линейных регрессий с различным набором факторов, построенных на одном и том же обучающем множестве. Анализ показал, что наименьшие AIC и BIC имеет регрессия, учитывающая $PM_m$, $t_s$ и $p_s$.

В таблице 3 представлены коэффициенты детерминации, вычисленные для каждой из рассмотренных регрессионных моделей, обученных на всем объеме обучающей выборки. $R^2$ оценивался на основе полного объема данных тестовой выборки с учетом множественности факторов.

Таблица 3. Коэффициент детерминации $R^2$ регрессионных моделей, обученных на полном объеме данных обучающей выборки

| Модель | Факторы | | | | |
|---|---|---|---|---|---|
| | $PM_m$ | $PM_m$, $t_s$ | $PM_m$, $t_s, h_s$ | $PM_m$, $t_s, p_s$ | $PM_m, t_s$, $p_s, h_s$ |
| Линейная регрессия (МНК) | 0.844 | 0.856 | 0.857 | 0.858 | 0.859 |
| LASSO | 0.844 | 0.856 | 0.856 | 0.858 | 0.859 |
| Эластичная сеть | 0.844 | 0.856 | 0.856 | 0.858 | 0.859 |
| Метод опорных векторов | 0.831 | 0.850 | 0.853 | 0.760 | 0.748 |
| k ближайших соседей | 0.830 | 0.870 | 0.882 | 0.883 | 0.889 |
| Дерево решений | 0.847 | 0.789 | 0.803 | 0.815 | 0.825 |
| Случайный лес | 0.848 | 0.864 | 0.883 | 0.893 | *0.902* |

На основе результатов регрессионного анализа можно сделать следующие выводы. Во-первых, множественная линейная регрессия с помощью наименьших квадратов (МНК) даёт хорошее приближение, сравнимое по точности с более сложными и вычислительноемкими методами машинного обучения. При этом линейная регрессия позволяет в явном виде получать коэффициенты, отражающие зависимость значения отклика от значений факторов. Во-вторых, для всех пар дублирующих датчиков лучшую точность предсказания отклика дают непараметрические методы “случайный лес” и “$k$ ближайших соседей”. В-третьих, добавление в анализ зависимости факторов влажности и давления не дает значительного улучшения точности моделей.

Это можно объяснить двумя причинами: 1) всесезонность выборки гасит разнонаправленное влияние этих факторов в разные сезоны; 2) обсуждаемая ранее некорректность измерений влажности.

Множественная линейная регрессия, учитывающая максимальное количество факторов, для парных датчиков поста Ветлужанка имеет следующий вид:

$$PM_s = a_0 + a_1 \cdot PM_m + a_2 \cdot t_s + a_3 \cdot p_s + a_4 \cdot h_s = \\ = 88.068 + 2.100\, PM_m - 0.781\, t_s - 0.127\, p_s + 0.054\, h_s, \quad (3)$$

где коэффициенты определены со следующими доверительными интервалами: $a_0 \in [86.584; 89.209]$, $a_1 \in [2.097; 2.103]$, $a_2 \in [-0.785; -0.777]$, $a_3 \in [-0.129; -0.125]$, $a_4 \in [0.054; 0.056]$. Отметим, что учет логнормальности распределений концентраций PM2.5 не дает существенного улучшения в прогнозе[5].

Построенные по полной обучающей выборке регрессионные модели мы оценили для отдельных групп значений тестовой выборки (таблица 4). Во-первых, коэффициент детерминации $R^2$ был вычислен по группам значений тестовой выборки, относящимся к одному сезону (строки Зима, Лето и Демисезон). Во-вторых, $R^2$ был вычислен для наблюдений из тестовой выборки, соответствующих моментам, когда скользящее среднее за сутки значение концентрации PM2.5 не превышало принятого в России [39] среднесуточного значения ПДК PM2.5 (35 мкг/м$^3$) и наблюдениям, в которых среднесуточная концентрация превышала ПДК. В таблице 4 соответствующие строки помечены “Не превышает ПДК” и “Превышает ПДК”, соответственно. В этом случае для того, чтобы избежать запаздывания периодов роста и спада концентрации PM2.5 текущее среднее значение вычислялось по наблюдениям, взятым за период 12 часов до текущего значения и 12 часов, начиная с текущего значения. Наконец, $R^2$ был вычислен для усредненных скользящим средним данных с окном 1 час, 6 часов, сутки.

Из данных таблицы 4 следует, что линейная регрессия хорошо приближает скользящее среднее, и тем лучше, чем больше окно. В то же время ожидаемо предсказания отклика в период высоких концентраций точнее, чем в период низких. Это подтверждается и очень низкой точностью модели для демисезонного, и, особенно, летнего периодов. Кроме того, учет в модели температуры повышает ее точность

[5]раздел «Учет логнормальности» ресурса [23]

($R^2$ увеличивается на проценты), добавление в модель давления еще незначительно улучшает точность прогноза отклика ($R^2$ увеличивается на десятые доли процента). Введение в модель влажности нецелесообразно.

Таблица 4. Коэффициент детерминации $R^2$, рассчитанный для групп значений уровня концентрации PM2.5 датчика "`s_vet`" тестовой выборки, для линейной регрессии (МНК), построенной на всём объеме данных обучающей выборки

| Обучающая выборка | Факторы | | | | |
|---|---|---|---|---|---|
| | $PM_m$ | $PM_m$, $t_s$ | $PM_m$, $t_s, h_s$ | $PM_m$, $t_s, p_s$ | $PM_m, t_s$, $p_s, h_s$ |
| Скользящее среднее за сутки | 0.939 | 0.944 | 0.952 | *0.954* | 0.953 |
| Скользящее среднее за 6 ч. | 0.936 | 0.941 | 0.947 | 0.947 | 0.947 |
| Скользящее среднее за час | 0.895 | 0.901 | 0.907 | 0.907 | 0.907 |
| Не превышает ПДК | 0.324 | 0.304 | 0.357 | 0.356 | 0.357 |
| Превышает ПДК | 0.739 | 0.770 | 0.764 | 0.771 | 0.771 |
| Зима | 0.862 | 0.863 | 0.866 | 0.867 | 0.868 |
| Лето | 0.220 | 0.314 | 0.338 | 0.367 | 0.363 |
| Демисезон | 0.582 | 0.600 | 0.603 | 0.603 | 0.602 |

Из проведенного исследования следует, что для повышения точности корректировки датчика необходимо обучать модели не на всей совокупности обучающей выборки, а предварительно выделять из всего множества данных целевую группу. В таблице 5 представлены значения коэффициента детерминации $R^2$ после обучения трех моделей: линейной регрессии на основе метода наименьших квадратов (LR), случайного леса (RF) и «$k$ ближайших соседей» ($k$-N). В названии строки указана группа данных, на которой проходило обучение. В частности, из таблицы 5 следует, что обучение на осредненных данных с учетом всех факторов дает $R^2$ близкий к единице. Кроме того, мы видим значительное улучшение моделей на данных с небольшими значениями концентрации PM2.5. Таким образом, результаты анализа, представленные в таблицах 4, 5, показывают, что корректировка концентраций PM2.5 для оптического датчика CityAir относительно анализатора E-BAM должна выполнятся, по крайней мере, по сезонным данным.

Для оперативной корректировки $PM_s$ удобно использовать параметрическую модель множественной линейной регрессии методом наименьших квадратов, учитывающую показания температуры и давления. В этом случае фактором является концентрация PM2.5, измеренная оптическим датчиком CityAir, а для обучения модели в качестве отклика используются концентрации PM2.5, измеренные анализатором E-BAM.

Таблица 5. Коэффициент детерминации $R^2$, рассчитанный для значений уровня концентрации PM2.5 оптического датчика "s_vet" тестовой выборки для нескольких моделей регрессии, построенных на указанной в строке группе данных обучающей выборки

| Обучающая выборка | $PM_m$ | | | $PM_m, t_s$ | | | $PM_m, t_s, p_s$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | LR | RF | $k$-N | LR | RF | $k$-N | LR | RF | $k$-N |
| Скользящее среднее | | | | | | | | | |
| за сутки | 0.945 | 0.945 | 0.949 | 0.959 | 0.991 | 0.988 | 0.959 | 0.997 | 0.997 |
| за 6 часов | 0.944 | 0.944 | 0.945 | 0.955 | 0.978 | 0.977 | 0.955 | 0.985 | 0.984 |
| за час | 0.904 | 0.904 | 0.898 | 0.914 | 0.930 | 0.931 | 0.915 | 0.943 | 0.938 |
| Не превышает ПДК | 0.616 | 0.621 | 0.577 | 0.631 | 0.654 | 0.683 | 0.631 | 0.725 | 0.712 |
| Превышает ПДК | 0.772 | 0.769 | 0.752 | 0.795 | 0.796 | 0.799 | 0.807 | 0.840 | 0.820 |
| Зима | 0.889 | 0.893 | 0.883 | 0.890 | 0.879 | 0.889 | 0.890 | 0.898 | 0.893 |
| Лето | 0.795 | 0.782 | 0.788 | 0.797 | 0.787 | 0.799 | 0.799 | 0.874 | 0.823 |
| Демисезон | 0.693 | 0.709 | 0.667 | 0.715 | 0.712 | 0.723 | 0.716 | 0.773 | 0.737 |

Ниже приведены формулы пересчета показаний оптического датчика для поста "Ветлужанка"[6]:

$$PM^{corr}_{\texttt{s\_vet}} = 36.109 + 0.367\, PM_{\texttt{s\_vet}} - 0.143\, t_{\texttt{s\_vet}} - 0.041\, p_{\texttt{s\_vet}}, \quad (4)$$

$$PM^{corr}_{\texttt{s\_vet}} = 286.693 + 0.535\, PM_{\texttt{s\_vet}} + 0.300\, t_{\texttt{s\_vet}} - 0.385\, p_{\texttt{s\_vet}}, \quad (5)$$

$$PM^{corr}_{\texttt{s\_vet}} = 28.500 + 0.361\, PM_{\texttt{s\_vet}} + 0.166\, t_{\texttt{s\_vet}} - 0.0310\, p_{\texttt{s\_vet}}, \quad (6)$$

для зимнего ($R^2 \sim 0.88$), летнего ($R^2 \sim 0.81$) и демисезонного ($R^2 \sim 0.67$) периодов, соответственно.

Более того, однопараметрическая регрессия $PM^{corr}_s = a \cdot PM_s$, построенная по тем же данным дает коэффициенты пересчета $a$ равные $0.43$, $0.7$ и $0.54$ для зимнего, летнего и демисезонного периодов, соответственно. В этом случае, коэффициенты детерминации равны $0.85$, $0.57$ и $0.40$.

[6]Аналогичные формулы для трех других постов можно найти в разделе «Формулы для корректировки» ресурса [23].

**6. Корректировка значений концентрации PM2.5 по постам системы мониторинга КНЦ СО РАН.** В результате описанного статистического анализа мы имеем подсистему из четырех оптических датчиков (`s_vet`, `s_pok`, `s_svr`, `s_kir`), откалиброванных по показаниям эталонных анализаторов Е-BAM.

Будем использовать эти датчики для корректировки оптических датчиков CityAir всех других постов системы мониторинга КНЦ СО РАН.

На рисунке 3 представлены коэффициенты корреляции между показаниями о концентрации PM2.5 каждого откалиброванного оптического датчика CityAir и всеми другими оптическими датчиками CityAir постов системы мониторинга КНЦ СО РАН. Для каждого датчика найден откалиброванный, у которого с ним максимальный коэффициент корреляции. На рисунке 4 все датчики нанесены на карту и одним цветом закрашены топографические области, объединяющие датчики, коэффициенты корреляции которых максимальны с одним из четырех откалиброванных датчиков.

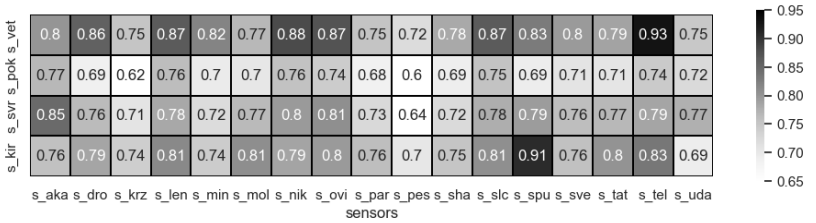



Рис. 3. Корреляция между показаниями об уровнях концентрации частиц PM2.5 подсети дублирующих датчиков и датчиков, оставшейся подсети КНЦ СО РАН

Полученные топографические области хорошо объясняются следующими географическими фактами. Во-первых, датчик `s_pok` расположен на хорошо проветриваемом высоком холме вдали от транспортных развязок. Ожидается, что этот датчик будет наименее коррелирован с другими. Поэтому в зону его действия не попал ни один датчик. Во-вторых, самая большая зона влияния (Зона 1) оказалась у датчика `s_vet`. Это можно объяснить преобладанием северо-западных ветров и расположением датчика `s_vet` на северо-западе в жилом районе Красноярска. Все датчики Зоны 1 относятся к левому берегу Енисея. Эти датчики будут аналогичным образом реагировать на суточные колебания уровней концентрации PM2.5, связанные, например, с пробками на дорогах, и длительные периоды высокой концентрации PM2.5, связанные с неблагоприятными метеорологическими условиями.


Рис. 4. Области максимальной корреляции показаний датчиков сети КНЦ СО РАН с показаниями скорректированных датчиков постов "vet "pok "svr "kir"

Может показаться удивительным, что в зону влияния (Зона 2) датчика `s_svr`, расположенного на правом берегу Енисея, попали датчики с левого берега. Однако все эти датчики расположены в хорошо проветриваемой относительно чистой части города, расположенной в долине реки Енисей. Поэтому понятна и их хорошая корреляция с датчиком `s_svr`, находящимся в аналогичных условиях. Зона влияния (Зона 3) датчика `s_kir` объединяет небольшую группу датчиков, расположенных в наиболее загрязненной части долины реки Енисей, а также на островах, через которые проходит мост с плотным транспортным потоком.

**7. Благодарности.** Авторы благодарят Алексея Токорева за предоставленные данные.

**8. Заключение.** В статье предложена статистически обоснованная корректировка первичных данных об уровне концентрации взвешенных частиц PM2.5 в приземном слое атмосферы г. Красноярска, получаемых оптическими датчиками станции CityAir. Для построения регрессионных моделей эталонными считались измерения, получаемые от анализаторов E-BAM, расположенных на тех же постах наблюдения, что и корректируемые датчики.

На основе проведенного анализа можно сделать следующие выводы.

1. При корректировке показаний датчиков необходимо учитывать метеорологические параметры. В нашем случае в линейной регрессионной модели лучше учитывать зависимость от температуры и давления. Точность измерения влажности на станциях CityAir не позволяет

учитывать этот показатель при корректировке, однако не исключает его влияния на показания датчика.

2. Параметры корректировки уровня концентрации PM2.5 с помощью регрессионной модели существенно зависят от сезона.

3. Показания оптических датчиков станций CityAir даже после корректировки не могут быть использованы в качестве эталонного оборудования для определения уровня концентрации PM2.5. В то же время их показания полностью отражают тренды показателей загрязнения, поэтому эти датчики могут использоваться для описания сценариев и прогнозов периодов повышенных концентраций взвешенных частиц в атмосфере городов.

4. Для оперативной корректировки значений концентрации PM2.5 датчиков станций CityAir достаточно использовать линейную многофакторную регрессионную модель на основе метода наименьших квадратов. Для научных ретроспективных исследований временных рядов концентраций PM2.5 рекомендуется использовать многофакторную регрессию, основанную на методе случайного леса.

## Литература


1. Chae S., Shin J., Kwon S., Lee S., Kang S., Lee D. PM10 and PM2.5 real-time prediction models using an interpolated convolutional neural network // Science Report. 2021. vol. 11(1). no. 11952.
2. Kim B., Lim Y., Wan Cha J. Short-term prediction of particulate matter (PM10 and PM2.5) in Seoul, South Korea using tree-based maching learning algorithms // Atmospheric Pollution Research. 2022. vol. 13(10). no. 101547.
3. Perrino C., Catrambone M., Pietrodangelo A. Influence of atmospheric stability on the mass concentration and chemical composition of atmospheric particles: A case study in Rome, Italy // Environment International. 2008. vol. 34. pp. 621–628.
4. Perez P., Menares C., Ramirez C. PM2.5 forecasting in Coyhaique, the most polluted city in the Americas // Urban Climate. 2020. vol. 32. no. 100608.
5. Zhang Zh., Wu L., Chen Y. Forecasting PM2.5 and PM10 concentrations using GMCN(1,N) model with the similar meteorological condition: Case of Shijiazhuang in China // Ecological Indicators. 2020. vol. 119. no. 106871.
6. Yang J., Yan R., Nong M., Liao J., Li F., Sun W. PM2.5 concentrations forecasting in Beijing through deep learning with different inputs model structures and forecast time // Atmospheric Pollution Research. 2021. vol. 12(9). no. 101168.
7. Лыченко Н.М., Великанова Л.И., Верзунов С.Н., Сороковая А.В. Модели прогноза уровня загрязнения атмосферного воздуха г. Бишкек // Вестник Кыргызско-Российского Славянского университета. 2021. Т. 21. № 4. С. 87–95.
8. Vlachogianni A., Kassomenos P., Karppinen A., Karakitsios S., Kukkonen J. Evaluation of a multiple regression model for the forecasting of the concentrations of NOx and PM10 in Athens and Helsinki // Science of the total environment. 2011. vol. 409. pp. 1559–1571.
9. Iglesias-Gonzalez S., Huertas-Bolanos M.E., Hernandez-Paniagua I.Y., Mendoza A. Explicit Modeling of Meteorological Explanatory Variables in Short-Term Forecasting

of Maximum Ozone Concentrations via a Multiple Regression Time Series Framework // Atmosphere. 2020. vol. 11(12). no. 1304.

10. Zhou Q., Jiang H., Wang J., Zhou J. A hybrid model for PM 2.5 forecasting based on ensemble empirical mode decomposition and a general regression neural network // Science of the Total Environment. 2014. vol. 496. pp. 264–274.
11. Аронов П.М. Оценка согласованного значения результатов межлабораторных измерений с минимальным увеличением их неопределённости // Эталоны. Стандартные образцы. 2019. Т. 15. № 4. С. 49–52.
12. Носков С.И. Метод максимальной согласованности в регрессионном анализе // Известия ТулГУ. Технические науки. 2021. № 10. С. 380–385.
13. Badura M., Batog P., Drzeniecka-Osiadacz A., Modzel P. Evaluation of Low-Cost Sensors for Ambient PM 2.5 Monitoring // Journal of Sensors. 2018. vol. 1. no. 5096540.
14. Shen H., Hou W., Zhu Y., Zheng S., Ainiwaer S., Shen G., Chen Y., Cheng H., Hu J., Wan Y., Tao S. Temporal and spatial variation of PM2.5 in indoor air monitored by low-cost sensors // Science of The Total Environment. 2021. vol. 770. no. 145304.
15. Jayaratne R., Liu X., Ahn K.H., Asumadu-Sakyi A., Fisher G., Gao J., Mabon A., Mazaheri M., Mullins B., Nyaku M., Ristovski Z., Scorgie Y., Thai P., Dunbabin M., Morawska L. Low-cost PM2.5 sensors: An assessment of their suitability for variousapplications // Aerosol and Air Quality Research. 2020. vol. 20. no. 3. pp. 520–532.
16. Gao M., Cao J., Seto E. A distributed network of low-cost continuous reading sensors to measure spatiotemporal variations of PM2.5 in Xi'an, China // Environmental Pollution. 2015. vol. 199. pp. 56–65.
17. Wang W., Lung S., Liu Ch. Application of Machine Learning for the in-Field Correction of a PM2.5 Low-Cost Sensor Network // Sensors. 2020. vol. 20(17). no. 5002.
18. Bi J., Stowell J., et al. Contribution of low-cost sensor measurements to the prediction of PM2.5 levels: A case study in Imperial County, California, USA // Environmental research. 2020. vol. 180. no. 108810.
19. E-BAM particulate monitor operation manual. Available at: https://metone.com/wp-content/uploads/2022/06/E-BAM-9805-Manual-Rev-G.pdf (accessed: 08.05.2023).
20. Environmental Technology Verification Report. Available at: https://archive.epa.gov/nrmrl/archive-etv/web/pdf/01_vr_metone_bam1020.pdf (accessed: 18.08.2023).
21. Заворуев В.В., Якубайлик О.Э., Кадочников А.А., Токарев А.В. Система мониторинга воздуха Красноярского научного центра СО РАН // Региональные проблемы дистанционного зондирования Земли: Материалы VII Международной научной конференции (г. Красноярск, 29 сентября – 2 октября 2020 г.). Красноярск: СФУ, 2020. С. 70–73.
22. Станция мониторинга воздуха CityAir. Электронный ресурс. URL: https://cityair.ru/ru/equipment/ (дата обращения: 08.05.2023).
23. Мониторинг состояния воздуха. Электронный ресурс. URL: https://asm.krasn.ru/ (дата обращения: 08.05.2023).
24. Тьюки Дж. Анализ результатов наблюдений. Разведочный анализ. М. 1981. 696 с.
25. Себер Дж. Линейный регрессионный анализ. М. 1980. 456 с.
26. Демиденко Е.З. Линейная и нелинейная регрессии. М.: Финансы и статистика. 1981. 302 с.
27. Хасти Т., Тибришани Р., Фридман Д. Основы статистического обучения. Интеллектуальный анализ данных, логический вывод и прогнозирование. М: Вильямс, 2020. 768 с.

28. Hoerl A.E., Kennard R.W. Ridge regression: biased estimation for nonorthogonal problems // Technometrics. 1970. vol. 12. no. 1. pp. 55–67.
29. Tibshirani R. Regression shrinkage and selection via the lasso // Journal of the Royal Statistical Society, Series B: Statistical Methodology. 1996. vol. 58. pp. 267–288.
30. Zou H., Hastie T. Regularization and variable selection via the elastic net // Journal of the Royal Statistical Society, Series B: Statistical Methodology. 2005. vol. 67. no. 2. pp. 301–320.
31. Вапник В.Н., Червоненкис А.Я. Теория распознавания образов (статистические проблемы обучения). М. 1974. 416 с.
32. Vapnik V. The Nature of Statistical Learning Theory. New- York: Springer Verlag N.Y. 1995. 188 p. DOI: 10.1007/978-1-4757-2440-0.
33. Vapnik V., Golowich S., Smola A. Support Vector Method for Function Approximation Regression Estimation and Signal // Advances in Neural Information Processing Systems. 1996. p. 281–287.
34. Morgan J.N. Sonquist J.A. Problems in the analysis of survey data, and a proposal // Journal of the American Statistical Association. 1963. vol. 58. pp. 415–434.
35. Breiman L., Friedman J., Olshen R., Stone C. Classification and regression trees. CA: Wadsworth and Brooks/Cole Advanced Books and Software. 1984. 368 p.
36. Breiman L. Bagging Predictors // Machine Learning. 1996. vol. 24. pp. 123–140.
37. Breiman L. Random Forests // Machine Learning. 2001. vol. 45. pp. 5–32.
38. Wallace J., Kanaroglou P. The effect of temperature inversions on ground-level nitrogen dioxide (NO2) and fine particulate matter (PM2.5) using temperature profiles from the Atmospheric Infrared Sounder (AIRS) // Science of The Total Environment. 2009. vol. 407. no. 18. pp. 5085–5095.
39. Санитарные правила и нормы СанПиН 1.2.3685-21 «Гигиенические нормативы и требования к обеспечению безопасности и (или) безвредности для человека факторов среды обитания». 2021 г. URL: https://docs.cntd.ru/document/573500115 (дата обращения: 26.01.2024).
40. Akaike H. A new look at statistical model identification // IEEE Transactions on Automatic Control. 1974. vol. 19. pp. 716–723.
41. Stoica P., Selen Y. Model-order selection: a review of information criterion rules // IEEE Signal Processing Magazine. 2004. vol. 21. no. 4. pp. 36–47.


**Карепова Евгения Дмитриевна** — канд. физ.-мат. наук, доцент, ведущий научный сотрудник, Отдел "Вычислительная математика", Институт вычислительного моделирования СО РАН. Область научных интересов: вычислительная математика, математическое моделирование природных и технических процессов, анализ данных, высокопроизводительное программирование. Число научных публикаций — 175. e.d.karepova@icm.krasn.ru; Академгородок, 50/44, 660036, Красноярск, Россия; р.т.: +7(391)290-7476.

**Петракова Виктория Сергеевна** — канд. физ.-мат. наук, научный сотрудник, Отдел "Вычислительная математика", Институт вычислительного моделирования СО РАН. Область научных интересов: математическое моделирование, вычислительная математика, анализ данных. Число научных публикаций — 22. rikka@icm.krasn.ru; Академгородок, 50/44, 660036, Красноярск, Россия; р.т.: +7(923)267-3748.

**Поддержка исследований.** Исследование выполнено при финансовой поддержке «Красноярского краевого фонда поддержки научной и научно-технической деятельности» в рамках реализации научного проекта № 2022110809055 «Оценка эффективности использования сети недорогих сенсорных датчиков для сбора данных о загрязнениях в пограничных слоях атмосферы на основе анализа наблюдений за динамикой концентрации взвешенных частиц PM2.5».

**DOI** 10.15622/ia.23.2.2

E. KAREPOVA, V. PETRAKOVA

# STATISTICAL SUBSTANTIATION OF THE REVISING OF READINGS BY THE CITYAIR STATION OF PM2.5 CONCENTRATION LEVELS IN THE ATMOSPHERIC BOUNDARY LAYER OF THE CITY

*Karepova E., Petrakova V.* **Statistical Substantiation of the Revising of Readings by the CityAir Station of PM2.5 Concentration Levels in the Atmospheric Boundary Layer of the City.**

**Abstract.** As a marker characterizing air pollution in the surface layer of the atmosphere of modern cities, the concentration level of particulate matter with a diameter of 2.5 microns or less (Particulate Matter, PM2.5) is often used. The paper discusses the practice of using a relatively cheap optical sensor, which is part of the CityAir station, to measure the concentration of PM2.5 in an urban environment. The article proposes a statistically justified correction of the primary data obtained by CityAir stations on the values of the concentration of suspended particles PM2.5 in the surface layer of the atmosphere of Krasnoyarsk. For the construction of regression models, measurements obtained from E-BAM analyzers located at the same observation posts as the corrected sensors were considered as a reference. For the analysis, primary data was used 1) from 9 automated observation posts of the regional departmental information and analytical system of data on the state of the environment of the Krasnoyarsk Territory (KVIAS); 2) from the 21st CityAir station of the monitoring system of the Krasnoyarsk Scientific Center of the Siberian Branch of the Russian Academy of Sciences. The paper demonstrates that when correcting sensor readings, it is necessary to take into account meteorological indicators. In addition, it is shown that the regression coefficients significantly depend on the season. Supervised learning methods are compared for solving the problem of correcting the readings of inexpensive sensors. Additional information on the results of data analysis, which was not included in the text of the article, is available on the electronic resource https://asm.krasn.ru/.

**Keywords:** particulate Matter, PM2.5 concentration level, supervised learning, regression models, sensor system revising.

## References

1. Chae S., Shin J., Kwon S., Lee S., Kang S., Lee D. PM10 and PM2.5 real-time prediction models using an interpolated convolutional neural network. Science Report. 2021. vol. 11(1). no. 11952.
2. Kim B., Lim Y., Wan Cha J. Short-term prediction of particulate matter (PM10 and PM2.5) in Seoul, South Korea using tree-based maching learning algorithms. Atmospheric Pollution Research. 2022. vol. 13(10). no. 101547.
3. Perrino C., Catrambone M., Pietrodangelo A. Influence of atmospheric stability on the mass concentration and chemical composition of atmospheric particles: A case study in Rome, Italy. Environment International. 2008. vol. 34. pp. 621–628.
4. Perez P., Menares C., Ramírez C. Perez P., Menares C., Ramirez C. PM2.5 forecasting in Coyhaique, the most polluted city in the Americas. Urban Climate. 2020. vol. 32. no. 100608.
5. Zhang Zh., Wu L., Chen Y. Forecasting PM2.5 and PM10 concentrations using GMCN(1,N) model with the similar meteorological condition: Case of Shijiazhuang in China. Ecological Indicators. 2020. vol. 119. no. 106871.

6. Yang J., Yan R., Nong M., Liao J., Li F., Sun W. PM2.5 concentrations forecasting in Beijing through deep learning with different inputs model structures and forecast time. Atmospheric Pollution Research. 2021. vol. 12(9). no. 101168.
7. Lychenko N.M., Velikanova L.I., Verzunov S.N., Sorokovaya A.V. [Models for predicting the level of atmospheric air pollution in Bishkek]. Vestnik Kyrgyzsko-Rossiyskogo Slavyanskogo universiteta – Bulletin of the Kyrgyz-Russian Slavic University. 2021. vol. 21. no. 4. pp. 87–95. (In Russ.).
8. Vlachogianni A., Kassomenos P., Karppinen A., Karakitsios S., Kukkonen J. Evaluation of a multiple regression model for the forecasting of the concentrations of NOx and PM10 in Athens and Helsinki. Science of the total environment. 2011. vol. 409. pp. 1559–1571.
9. Iglesias-Gonzalez S., Huertas-Bolanos M.E., Hernandez-Paniagua I.Y., Mendoza A. Explicit Modeling of Meteorological Explanatory Variables in Short-Term Forecasting of Maximum Ozone Concentrations via a Multiple Regression Time Series Framework. Atmosphere. 2020. vol. 11(12). no. 1304.
10. Zhou Q., Jiang H., Wang J., Zhou J. A hybrid model for PM 2.5 forecasting based on ensemble empirical mode decomposition and a general regression neural network. Science of the Total Environment. 2014. vol. 496. pp. 264–274.
11. Aronov P.M. [Estimation of consensus value of interlaboratory measurement results accompanied by a minimum increase in associated uncertainty]. Etalony. Standartnye obrazcy – Standards. Reference materials. 2019. vol. 15. no. 4. pp. 49–52. (In Russ.).
12. Noskov S.I. [Maximum Consistency Method in Regression Analysis]. Izvestija TulGU. Tehnicheskie nauki – Proceedings of TulGU. Technical science. 2021. № 10. С. 380–385. (In Russ.).
13. Badura M., Batog P., Drzeniecka-Osiadacz A., Modzel P. Evaluation of Low-Cost Sensors for Ambient PM 2.5 Monitoring. Journal of Sensors. 2018. vol. 1. no. 5096540.
14. Shen H., Hou W., Zhu Y., Zheng S., Ainiwaer S., Shen G., Chen Y., Cheng H., Hu J., Wan Y., Tao S. Temporal and spatial variation of PM2.5 in indoor air monitored by low-cost sensors. Science of The Total Environment. 2021. vol. 770. no. 145304.
15. Jayaratne R., Liu X., Ahn K.H., Asumadu-Sakyi A., Fisher G., Gao J., Mabon A., Mazaheri M., Mullins B., Nyaku M., Ristovski Z., Scorgie Y., Thai P., Dunbabin M., Morawska L. Low-cost PM2.5 sensors: An assessment of their suitability for variousapplications. Aerosol and Air Quality Research. 2020. vol. 20. no. 3. pp. 520–532.
16. Gao M., Cao J., Seto E. A distributed network of low-cost continuous reading sensors to measure spatiotemporal variations of PM2.5 in Xi'an, China. Environmental Pollution. 2015. vol. 199. pp. 56–65.
17. Wang W., Lung S., Liu Ch. Application of Machine Learning for the in-Field Correction of a PM2.5 Low-Cost Sensor Network. Sensors. 2020. vol. 20(17). no. 5002.
18. Bi J., Stowell J., et al. Contribution of low-cost sensor measurements to the prediction of PM2.5 levels: A case study in Imperial County, California, USA. Environmental research. 2020. vol. 180. no. 108810.
19. E-BAM particulate monitor operation manual. Available at: https://metone.com/wp-content/uploads/2022/06/E-BAM-9805-Manual-Rev-G.pdf (accessed: 08.05.2023).
20. Environmental Technology Verification Report. Available at: https://archive.epa.gov/nrmrl/archive-etv/web/pdf/01_vr_metone_bam1020.pdf (accessed: 18.08.2023).
21. Zavoruev V.V., Yakubailik O.E., Kadochnikov A.A., Tokarev A.V. [Air monitoring system of the Krasnoyarsk Scientific Center of the SB RAS]. Regionalnyye problem distantsionnogo zondirovaniya Zemli: Materialy VII Mezhdunarodnoy nauchnoy konferentsii [Regional problems of remote sensing of the Earth: Proceedings of the VII International Scientific Conference]. Krasnoyarsk: SFU. 2020. pp. 70-73. (In Russ.).

22. Stancija monitoringa vozduha CityAir. Jelektronnyj resurs [ Air Monitoring Station CityAir. Electronic resource]. Available at: https://cityair.ru/ru/equipment/ (accessed: 08.05.2023). (In Russ.).
23. Monitoring sostojanija vozduha. Jelektronnyj resurs [Air state monitoring. Electronic resource.] Available at: https://asm.krasn.ru/ (accessed: 08.05.2023). (In Russ.).
24. Tyuki J. Analiz rezul'tatov nabljudenij. Razvedochnyj analiz [Analysis of observation results. Exploratory analysis]. Moscow. 1981. 696 p. (In Russ.).
25. Seber G. Linejnyj regressionnyj analiz [Linear regression analysis]. Moscow. 1980. 456 p. (In Russ.).
26. Demidenko E.Z. Linejnaja i nelinejnaja regressii. [Linear and nonlinear regression]. Moscow: Finansy i Statistika. 1981. 302 p. (In Russ.).
27. Hastie Tr., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd Edition. Springer. 2017. 768 p.
28. Hoerl A.E., Kennard R.W. Ridge regression: biased estimation for nonorthogonal problems. Technometrics. 1970. vol. 12. no. 1. pp. 55–67.
29. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B: Statistical Methodology. 1996. vol. 58. pp. 267–288.
30. Zou H., Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society, Series B: Statistical Methodology. 2005. vol. 67. no. 2. pp. 301–320.
31. Vapnik V., Chervonenkis A. Teorija raspoznavanija obrazov (statisticheskie problem obuchenija) [Pattern recognition theory (statistical learning problems)]. M.: Nauka. 1974. 416 p. (In Russ.).
32. Vapnik V. The Nature of Statistical Learning Theory. New- York: Springer Verlag N.Y. 1995. 188 p. DOI: 10.1007/978-1-4757-2440-0.
33. Vapnik V., Golowich S., Smola A. Support Vector Method for Function Approximation Regression Estimation and Signal. Advances in Neural Information Processing Systems. 1996. pp. 281–287.
34. Morgan J.N. Sonquist J.A. Problems in the analysis of survey data, and a proposal. Journal of the American Statistical Association. 1963. vol. 58. pp. 415–434.
35. Breiman L., Friedman J., Olshen R., Stone C. Classification and regression trees. CA: Wadsworth and Brooks/Cole Advanced Books and Software. 1984. 368 p.
36. Breiman L. Bagging Predictors. Machine Learning. 1996. vol. 24. pp. 123–140.
37. Breiman L. Random Forests. Machine Learning. 2001. vol. 45. pp. 5–32.
38. Wallace J., Kanaroglou P. The effect of temperature inversions on ground-level nitrogen dioxide (NO2) and fine particulate matter (PM2.5) using temperature profiles from the Atmospheric Infrared Sounder (AIRS). Science of The Total Environment. 2009. vol. 407. no. 18. pp. 5085–5095.
39. anitary rules and norms SanPiN 1.2.3685-21 «Hygienic standards and requirements for ensuring the safety and (or) harmlessness of environmental factors for humans». 2021. Available at: https://docs.cntd.ru/document/573500115 (accessed: 26.01.2024). (In Russ.).
40. Akaike H. A new look at statistical model identification. IEEE Transactions on Automatic Control. 1974. vol. 19. pp. 716–723.
41. Stoica P., Selen Y. Model-order selection: a review of information criterion rules. IEEE Signal Processing Magazine. 2004. vol. 21. no. 4. pp. 36–47.

**Karepova Eugeniya** — Ph.D., Associate Professor, Leading researcher, Department "Computational mathematics", Institute of Computational Modeling of the Siberian Branch of the Russian Academy of Sciences. Research interests: computational mathematics, mathematical modeling of natural and technical processes, data analysis, high-performance programming. The number of publications — 175. e.d.karepova@icm.krasn.ru; 50/44, Akademgorodok, 660036, Krasnoyarsk, Russia; office phone: +7(391)290-7476.

**Petrakova Viktoriya** — Ph.D., Researcher, Department "Computational mathematics", Institute of Computational Modeling of the Siberian Branch of the Russian Academy of Sciences. Research interests: mathematical modeling, computational mathematics, data analysis. The number of publications — 22. rikka@icm.krasn.ru; 50/44, Akademgorodok, 660036, Krasnoyarsk, Russia; office phone: +7(923)267-3748.

**Acknowledgements.** The study was financially supported by the Krasnoyarsk Regional Fund of Science and Technology Support, project No. 2022110809055 «Evaluation of the effectiveness of using a network of the low-cost sensors to collect data on pollution in the atmospheric boundary layers based on an analysis of observations of the dynamics of the concentration of suspended particles PM2.5».

**DOI** 10.15622/ia.23.2.3

A. EBRAHEEM, I. IVANOV

# TOWARDS AUTOMATED AND OPTIMAL IIOT DESIGN

*Ebraheem A., Ivanov I.* **Towards Automated and Optimal IIoT Design.**

**Abstract.** In today's world, the Internet of Things has become an integral part of our lives. The increasing number of intelligent devices and their pervasiveness has made it challenging for developers and system architects to plan and implement systems of the Internet of Things and Industrial Internet of Things effectively. The primary objective of this work is to automate the design process of Industrial Internet of Things systems while optimizing the quality of service parameters, battery life, and cost. To achieve this goal, a general four-layer fog-computing model based on mathematical sets, constraints, and objective functions is introduced. This model takes into consideration the various parameters that affect the performance of the system, such as network latency, bandwidth, and power consumption. The Non-dominated Sorting Genetic Algorithm II is employed to find Pareto optimal solutions, while the Technique for Order of Preference by Similarity to Ideal Solution is used to identify compromise solutions on the Pareto front. The optimal solutions generated by this approach represent servers, communication links, and gateways whose information is stored in a database. These resources are chosen based on their ability to enhance the overall performance of the system. The proposed strategy follows a three-stage approach to minimize the dimensionality and reduce dependencies while exploring the search space. Additionally, the convergence of optimization algorithms is improved by using a biased initial population that exploits existing knowledge about how the solution should look. The algorithms used to generate this initial biased population are described in detail. To illustrate the effectiveness of this automated design strategy, an example of its application is presented.

**Keywords:** IIoT, IoT, NGSA-II, TOPSIS, cloud, fog computing, multiobjective optimization, gateway, edge devices.

**1. Introduction.** One of the drivers of Industry 4.0 is the Industrial Internet of Things (IIoT), the development of which is a consequence of the widespread use of computer technology. Cloud computing is one of the factors driving the success of the Internet of Things (IoT) and Industrial Internet of Things. It allows users to solve computing problems using the resources of connected servers and data centers scattered around the world and working as a single ecosystem [1]. In some cases, the long network distance between edge devices and remote cloud data centers reduces the quality of service, resulting in high latency, high bandwidth usage, and unreliable connections. The concept of fog computing helps solve these problems by bringing computing and storage closer to end users. It can also help reduce unplanned downtime, improve efficiency, and keep the Internet from being flooded with data from a myriad of sources. Thus, fog computing provides services in the same way as the cloud, with better quality parameters that meet the critical requirements of IIoT [2]. Therefore, it can be used as a basis for IIoT systems and models. For fog networks to reach their full potential, good and careful planning is required.

The structure of the paper is as follows: A review of the revised literature is presented in the following section. The architecture of the IIoT system and the suggested techniques for choosing and allocating its components are subsequently specified. After that, the mathematical model is extensively described. The methods related to optimization and decision-making follow. The next steps before the conclusion are the simulation process and outcomes.

**2. Related work.** The problem of optimizing and synthesizing IIoT systems belongs to a much wider class of problems related to the synthesis of the structure of complex systems, examples of which can be found in [3 – 6]. In this context, special attention was paid to IIoT and cyber-physical systems in [7], where models, methods, and algorithms for synthesizing complex management plans in cyber-physical systems and industrial internet are proposed.

This topic was also addressed by different organizations, alliances and consortiums. As a result, different frameworks and architectures were suggested like RAMI 4.0 (*R*eference *A*rchitectural *M*odel *I*ndustrie 4.0) [8], IIRA (*I*ndustrial *I*nternet *R*eference *A*rchitecture) [9], and IVRA (*I*ndustrial *V*alue *C*hain *R*eference *A*rchitecture) [10]. However, these frameworks do not propose any kind of optimization or automation methods. They rather provide definitions, guidelines and suggestions on how to construct and organize IIoT systems.

Study [11] reviewed the various methods used to optimize IoT networks, and classified them into 8 groups based on the network aspect being optimized (network routing, power consumption, congestion control, heterogeneity, scalability, reliability, quality of service , security). Optimizing and synthesizing IIoT and IoT systems has also been addressed as a problem of network planning and optimization for different technologies like mobile networks [12, 13], LoRaWAN (*Lo*ng *Ra*nge *W*ide *A*rea *N*etwork) [14], SigFox [15], NB-IoT [16].

Paper [17] proposed a model based on a three-layer fog computing architecture that takes into account transmission, propagation, and processing delays, as well as network traffic while keeping the overall deployment cost within the desired budget. Paper [18] uses the same model with modified objective functions, introduces a new optimization method, and compares the results of several optimization algorithms.

The revised literature can be loosely divided into the following categories:

– technology-specific literature addressing certain types of networks or certain layers in the technology stack.

– literature that is more general in nature that addresses the structure of the system and the territorial distribution of its components.

This work falls into the second category and is a continuation and an expansion of the approach presented in [17, 18] by adding the necessary elements to enable an automated buttom-up design approach for IIoT, which, to the best of the authors' knowledge, is not encountered in previous literature.

**3. Architecture and methodology.** To develop the mathematical model in a simple yet realistic and feasible manner, a four-layer architecture based on fog computing is used (Figure 1).

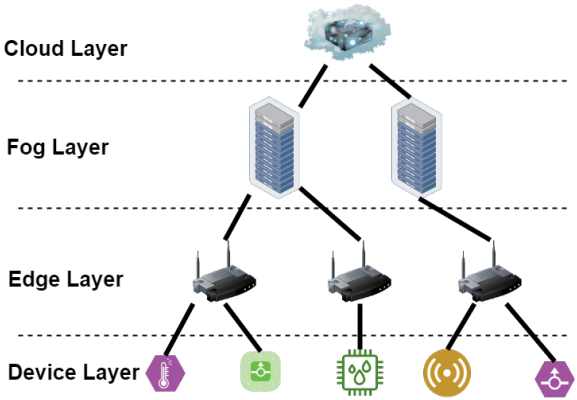



Fig. 1. Fog architecture for IIoT [19]

The layers from bottom to top are:

– layer 1 (the device layer): it includes sensors and actuators used in the system.

– layer 2 (the edge layer): it contains edge gateways that provide edge analysis, data flow multiplexing and throttling, and local data storage.

– layer 3 (the fog layer): it contains fog network nodes, which can be servers or any specialized computing devices.

– layer 4 (the cloud): it is seen as a large network of connected servers operating as a single ecosystem that provides a set of services such as data storage and management, and application hosting [1].

As stated earlier, improving the quality characteristics of IIoT systems is critical and should start early in the design process. To accomplish this purpose, resource allocation at various levels of the aforementioned architecture is used.

The starting point is a set of edge devices, each of which supports one communication technology. These devices have to be connected to edge gateways, the type and location of which are to be selected from a set of possible choices. The gateways must then be connected to a set of fog network nodes, the type and location of which are also to be selected from a set of possible

choices, or directly to the cloud. In this work, we consider that the servers of the cloud are located in one data center to which gateways are either connected directly through the internet (the fog layer is bypassed) or indirectly through the fog layer. The fog layer is connected to the cloud using special links.

The approach used is to divide the solution process into two main optimization stages and one intermediate stage (Figure 2).

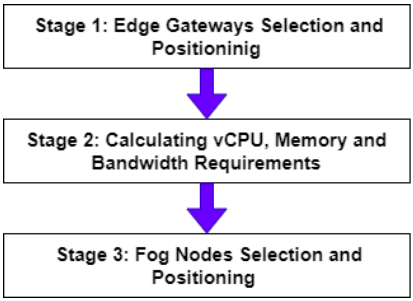



Fig. 2. Solution stages

The types and location of edge gateways are established in the first stage to reduce equipment cost and device-to-gateway distance, which also reduces deployment overhead, power consumption, and enhances connection quality by reducing link errors and packet loss [20]. In preparation for the third stage, the vCPU (virtual Central Processing Unit), memory, and bandwidth requirements are computed in the second stage. The third stage specifies the type and location of the fog network nodes, as well as whether or not to connect to the fog network node or to the cloud directly. The main goal in this stage is to reduce the cost, latency, and traffic traveling over the network.This split simplifies the search for optimal solutions and avoids expanding the dimensionality of the solution space.

**4. Mathematical Model.** The major purpose of the suggested methodology is to give users the ability to select technological solutions (gateways, servers, communication channels) that best match the needs of their IIoT system at every level. While the majority of the revised literature focuses on either the highest three levels of the used model or the lowest two levels, the contribution of this study is the offering of a holistic approach that addresses the system as a whole. The incorporation of technology-specific algorithms or procedures (routing, automatic configuration, etc.) is beyond the scope of this study. The main components of the model (Figure 3) are:

– available resources represented by mathematical sets that could possibly reflect information stored in database tables for example,

– binary decision variables, whose values determine whether a specific resource is allocated or not and where it is installed,

– constraints that guarantee the solution to be feasible and to meet the technical requirements,

– cost functions that have to be either minimized or maximized. In this work, all the cost functions are to be minimized.

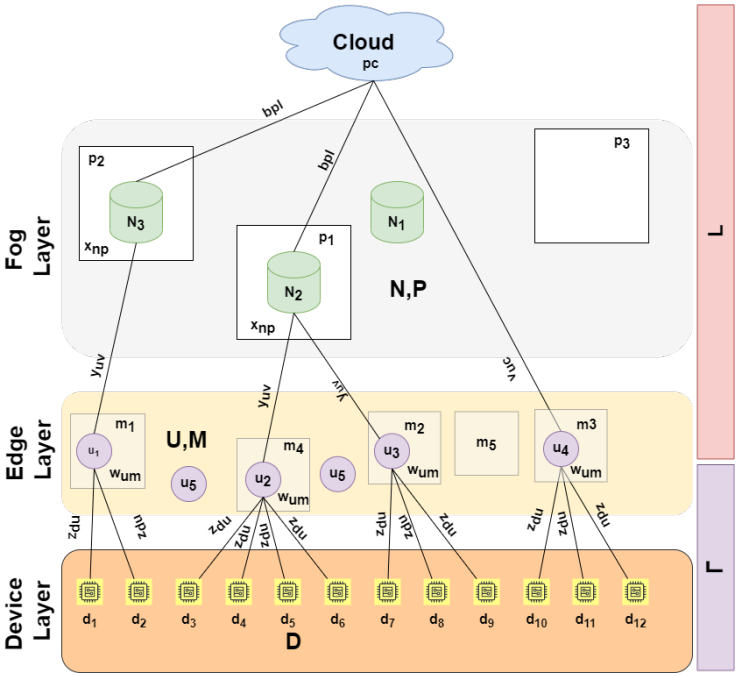



Fig. 3. Architecture and mathematical model of the fog network

The model expands the model proposed in [17, 18] by adding an additional low-level layer, which is the device layer mentioned earlier. It is worth noting that the formulation of the sets $U, N, P, L$, the decision variables $x_{np}, y_{up}, v_{uc}, b_{pl}$ , the constraints from 8-14, and objectives from 3 to 5 described next is based on these two works. The following are the seven mathematical sets employed in this model:

– $\Gamma$ is the set of communication technologies $\{\gamma_1, \gamma_2, \dots\}$ that can be used. Commnication technology can be something like ZigBee , wireless, and so on. Each technology $\gamma_i$ is characterized by the maximum number of supported devices $d^{\gamma}_{max}$ and the carrier frequency $f^{\gamma}$. Some technologies

support adaptive data rates and some do not, so we assume that devices and gateways can be configured to operate at the same data rate as long as they support the same technology.

– $D$ is the set of edge devices $\{d_1, d_2, \dots\}$. Each edge device $d_i$ is characterized by its communication technology $\gamma^d$, location $p^d$, the number of bytes per second transmitted by the device $\theta^d$, receiver antenna gain $G^d_{RX}$ (dBi), receiver sensitivity $S^d_{RX}$ (dBm), transmitter power output $P^d_{TX}$ (dBm). If $\theta^d$ is not precisely known, because the device sends packets on the occurrence of some events and not on a timely basis, worst-case scenario can be used to estimate $\theta^d$. Sometimes $\theta^d$ might be implied by the used technology $\gamma^d$.

– $U$ is the set of edge gateways that can be installed in the network $\{u_1, u_2, \dots\}$. Each gateway is characterized by the speed of the communication channel $\kappa^u$, the cost $\xi^u$ and the set of communication technologies supported by the gateway $\Gamma^u$. For each $\gamma^u \in \Gamma^u$ there is a receiver antenna gain $G^{\gamma u}_{RX}$ (dBi), a receiver sensitivity $S^{\gamma u}_{RX}$ (dBm), and a transmitter power output $P^{\gamma u}_{TX}$ (dBm). In our model, these values are known from the very beginning. However, some of the values need to be calculated during stage 2. These values include the total amount of memory $\lambda^u$ and the number of virtual processors (vCPU) $\alpha^u$ required to process the data sent by the gateway to the fog node or the cloud $\theta^u$ which also requires calculation.

– $N$ is the set of fog nodes $\{n_1, n_2, \dots\}$. Each node is characterized by the total available memory $\lambda^n$, the number of available vCPUs $\alpha^n$, the network interface bandwidth $\theta^n$ in bytes per second, and the cost $\xi^n$ in currency units.

– $L$ is the set of possible types of links that can be used to connect the nodes of the fog network and the cloud $\{l_1, l_2, \dots\}$ . Each channel is characterized by its bandwidth in bytes per second $\theta^l$, and a cost in currency unit per meter $\xi^l$ .

– $M$ is the set of possible places to install gateways $\{m_1, m_2, \dots\}$.

– $P$ is the set of possible places to install the fog nodes $\{p_1, p_2, \dots\}$.

In this model, six decision variables are utilized to allocate the aforementioned resources. This allocation involves the installation of network hardware at specific locations and establishing links between the installed hardware.

– $w_{um}$ is a binary decision variable such that $w_{um} = 1$ if and only if the gateway $u \in U$ is installed at the location $m \in M$.

– $z_{du}$ is a binary decision variable such that $z_{du} = 1$ if and only if the edge device $d \in D$ is connected to the gateway $u \in U$.

– $x_{np}$ is a binary decision variable such that $x_{np} = 1$ if and only if the fog net node $n \in N$ is set to the location $p \in P$.

– $y_{up}$ is a binary decision variable such that $y_{up} = 1$ if and only if gateway $u \in U$ is connected to the location $p \in P$.

– $v_{uc}$ is a binary decision variable such that $v_{uc} = 1$ if and only if the gateway $u \in U$ is connected to the cloud.

– $b_{pl}$ is a binary decision variable such that $b_{pl} = 1$ if and only if the node of the fog network, set at the location $p \in P$, is connected to the cloud by a channel of type $l \in L$.

The first stage in the adopted approach employs the following constraints:

– Constraint 1: the maximum number of devices connected to a gateway via a certain technology must not be exceeded:

$$\sum_{d \in D} z_{du} \delta(d, \gamma) \leqslant q(u, \gamma) d_{max}^{\gamma}, \tag{1}$$

where $\delta(d, \gamma)$ is a binary function that returns 1 if device $d$ supports technology $\gamma$, and $q(u, \gamma)$ is a function that returns the number of physical interfaces of type $\gamma$ in gateway U.

– Constraint 2: this constraint is related to the placement of the gateways. A good approach is to use a link budget constraint, which can be defined as the difference between the gains and losses of the system and must be greater than zero:

$$\left.\begin{array}{l} P_{TX}^{d} + G_{RX}^{\gamma u} - S_{RX}^{\gamma u} - P_{L}^{\gamma u d} > 0 \\ P_{TX}^{u} + G_{RX}^{d} - S_{RX}^{d} - P_{L}^{\gamma u d} > 0 \end{array}\right\} \gamma^{\mu} \in \Gamma^{u}, u \in U, d \in D, \tag{2}$$

where $P_{L}^{\gamma u d}$ is the path loss from the device to the gateway for the frequency used by communication technology $\gamma$. Computing this value requires knowing the precise locations of the transmitter and receiver and the environment in which the signal is traveling. There are various approaches to get close to this value. One approach is to analyze site-specific radio wave propagation using ray tracing techniques, which are used by software like Wireless Insite, that include a collection of RF propagation models (3D ray-tracing, quick ray-based methods, and empirical models) [21]. Matlab communications toolbox provides such capabilities by using RayTracing objects which are propagation models that compute propagation paths using 3-D environment geometry [22]. Empirical models can also be used to estimate the value of the path loss like Cost-231 Walflsch-Ikegami model [23, 24], Hata model, close-in model,

floating-intercept model, Longly-Rice model, communications research centre predict model [25]. In cases where there is a line of sight, the free space path loss model can be utilized:

$$P_L = 32.45 + 20log(d) + 20log(f), \tag{3}$$

where $d$ is the distance between the device and the edge gateway in $km$ and f is the frequency in $MHz$.

– Constraint 3: each device is connected to exactly one gateway:

$$\sum_{u \in U} z_{du} = 1, d \in D. \tag{4}$$

– Constraint 4: each location has a maximum of one gateway:

$$\sum_{u \in U} w_{um} \leqslant 1, m \in M. \tag{5}$$

– Constraint 5 : a gateway can only be installed if at least one device is actually connected to it:

$$\vartheta(u) = 1, u \in U, \tag{6}$$

$$\vartheta(u) = \begin{cases} 1: & (\sum_{m \in M} w_{um} > 0 \text{ and } \sum_{d \in D} z_{du} > 0) \\ & \text{or } (\sum_{m \in M} w_{um} = 0 \text{ and } \sum_{d \in D} z_{du} = 0) \\ .0: & \text{otherwise} \end{cases}, u \in U. \tag{7}$$

– Constraint 6: a device and a gateway can be linked if they support the same communication technology:

$$z_{du} \leqslant \sum_{\gamma \in \Gamma} \gamma(d, \gamma) q(u, \gamma), d \in D, u \in U. \tag{8}$$

– Constraint 7: the traffic passing through a gateway cannot exceed its bandwidth capability:

$$\sum_{d \in D} z_{du} \theta^d \leqslant \kappa^u. \tag{9}$$

The following set of constraints is used in stage 3 described earlier:

– Constraint 8: no more than one fog node can be installed at any given location:

$$\sum_{n \in N} x_{np} \leqslant 1, p \in P. \tag{10}$$

– Constraint 9: each gateway used is connected to only one fog node or to the cloud:

$$\sum_{p \in P} y_{up} + v_{uc} = 1, u \in U. \tag{11}$$

– Constraint 10: each fog node used is connected to the cloud:

$$\sum_{n \in N} x_{np} = \sum_{l \in L} b_{pl}, p \in P. \tag{12}$$

– Constraint 11: the number of vCPUs required by edge gateways does not exceed the number of vCPUs that can be provided by the fog node:

$$\sum_{u \in U} y_{up} \alpha^u \leqslant \sum_{n \in N} x_{np} \alpha^n, p \in P. \tag{13}$$

– Constraint 12: the amount of memory required by the gateways does not exceed the amount of memory that can be provided by the fog node to which they are connected:

$$\sum_{u \in U} y_{up} \lambda^u \leqslant \sum_{n \in N} x_{np} \lambda^n, p \in P. \tag{14}$$

– Constraint 13: the total bandwidth required to connect edge gateways to the fog node does not exceed the bandwidth of the fog node:

$$\sum_{u \in U} y_{up} \theta^u \leqslant \sum_{n \in N} x_{np} \theta^n, p \in P. \tag{15}$$

– Constraint 14: the bandwidth required to send data from the fog node to the cloud does not exceed the bandwidth of the link used:

$$\sum_{n\in N}\sum_{u\in U} y_{up}x_{np}\theta^u r^n \leqslant \sum_{l\in L} b_{pl}\theta^l, p \in P. \tag{16}$$

where $r^n$ is the average percentage of data sent to the cloud from the incoming data at the fog network node.

During the first stage the goal is to:

– Objective 1: minimize the total distance from devices to gateways:

$$\Theta = \min\left(\sum_{d\in D}\sum_{u\in U} z_{du}\mathscr{E}\left(d, \sum_{m\in M} w_{um}m\right)\right), \tag{17}$$

where $\mathscr{E}(a,b)$ is a function that returns the distance between two points $a$ and $b$.

– Objective 2: minimize the cost of deployment, which is primarily the total cost of the gateways:

$$\Lambda = \min \sum_{m\in M}\sum_{u\in U} w_{um}\xi^u. \tag{18}$$

During stage 3, the goal is to:

– Objective 3: minimize network latency:

$$D_T = \min(D_t + D_n + D_p), \tag{19}$$

where:

– $D_t$ is the transmission delay, which can be calculated using the following formula:

$$D_t = \sum_{u\in U} \theta^u/\kappa^u \left((h_1+1)\left[\sum_{p\in P} y_{up}\right] + (h_1+h_2+1)v_{uc}\right) + (h_2+1)\sum_{n\in N}\sum_{p\in P}\sum_{u\in U} y_{up}x_{np}\theta^u r^n\left(\sum_{l\in L}\frac{b_{pl}}{\theta^l}\right), \tag{20}$$

where $h_1$ is the average number of hops from edge gateways to fog network nodes, and $h_2$ is the average number of hops from fog network nodes to the cloud. It is assumed that the average number of transitions from edge gateways to the cloud is $(h_1+h_2)$.

– $D_n$ is the propagation delay, which can be calculated using the following formula:

$$D_n = \sum_{u \in U} \frac{\mathscr{E}(m_u, p_c)}{v} v_{uc} + \sum_{u \in U} \left( \frac{\mathscr{E}(m^u, p^u) + \mathscr{E}(p^u, p^c)}{v} \sum_{p \in P} y_{up} \right), \quad (21)$$

where $v$ is the signal propagation speed, $m^u = \sum_{m \in M} w_{um} m$ is the location of the gateway $u$, $p^u = \sum_{p \in P} \sum_{n \in N} x_{np} y_{up} p$ is the location at which is installed the fog node to which the gateway $u$ is connected, and $p^c$ is the location at which the data center representing the cloud system (or at least part of it) is installed.

– $D_p$ is the processing delay, which can be calculated using the following formula:

$$D_p = \sum_{u \in U} (h_1 + h_2) k, \quad (22)$$

where $k$ is the average processing delay in seconds per transition.

– Objective 4: minimize the total traffic going to the cloud, which is the sum of the traffic from all the edge gateways connected to the cloud, plus the traffic coming from the various fog network nodes:

$$\mathscr{T} = \min \left( \sum_{u \in U} \left[ v_{uc} + \sum_{n \in N} \sum_{p \in P} x_{np} y_{up} r^n \right] \theta^u \right). \quad (23)$$

– Objective 5: minimize costs:

$$\mathscr{C} = \min \left( \sum_{p \in P} \left[ \sum_{n \in N} x_{np} \xi^n + \sum_{l \in L} b_{pl} \mathscr{E}(p, p^c) \xi^l \right] \right). \quad (24)$$

At the intermediate stage 2, the following values are calculated for the used gateways:

– the number of bytes per second sent by the gateway $\theta^u$:

$$\theta^u = \sum_{d \in D} z_{du} \theta^d, u \in U, \quad (25)$$

– the number of vCPUs $\alpha^u$ and the amount of RAM $\lambda^u$ needed to process the data sent by the gateway to the fog nodes. These quantities are

difficult to calculate accurately. When visiting the websites of popular cloud providers and technology companies [26 – 28], it turned out that the maximum bandwidth is related to the number of virtual processors and RAM in addition to the type of workload and the required performance in addition to other factors. To determine these values in our model, we assume a linear relationship between the outbound bandwidth of the edge gateway and the number of required virtual processors and RAM.

$$\begin{aligned} \alpha^u &= \kappa_\alpha \theta^u + \alpha_{min} + \alpha_{margin}, \\ \lambda^u &= \kappa_\lambda \theta^u + \lambda_{min} + \lambda_{margin}, \end{aligned} \tag{26}$$

where:

– $\kappa_\alpha$ is a coefficient that relates throughput to the number of vCPUs.

– $\alpha_{min}$ is the minimum number of vCPUs required for a virtual machine to function properly.

– $\alpha_{margin}$ is the number of vCPUs dedicated to handle unexpected load changes or ensure proper system operation.

– $\kappa_\lambda$ is a coefficient that relates bandwidth to the amount of RAM.

– $\lambda_{min}$ is the minimum amount of RAM required for a virtual machine to function correctly.

– $\lambda_{margin}$ is the amount of RAM, reserved to handle unexpected changes in load or ensure proper system operation.

The proposed model does not claim to be complete and has a number of limitations:

– Devices are treated as stationary and do not move from one gateway to another. In such cases, a network should be planned to ensure and maximize coverage in areas where assets are moving. However, these cases are not considered by the current model.

– Energy consumption is not taken into account in an explicit way. This model aims to minimize the distance between IoT devices and gateways, which can lead to situations where some technology-specific settings that allow for less energy consumption can be applied.

It is also worth mentioning that the system structure and the software operation requirements are two important correlated factors in the context of territorially distributed systems. On the one hand, the system architecture can affect how efficiently data is processed. On the other hand, the structure of the system may be determined by the requirements for software operation. While the proposed method does not take into consideration the technicalities of the used software, it aims to enhance the quality of service factors, which usually positively impact the software performance.

The environmental factors also affect the quality of the obtained solutions, by affecting the link budget of the communication systems. Rain, terrain, and interference from other sources can all affect the quality of the communication link. While this might sometimes be accounted for by choosing an appropriate propagation model and good quality equipment (antennas, cables,....), it is hard to take into consideration all the factors. The authors recommend using equation 2 with a positive margin or safety factor $P_{margine}$:

$$\left.\begin{array}{l} P_{TX}^{d} + G_{RX}^{\gamma u} - S_{RX}^{\gamma u} - P_{L}^{\gamma ud} > P_{margine} \\ P_{TX}^{u} + G_{RX}^{d} - S_{RX}^{d} - P_{L}^{\gamma ud} > P_{margine} \end{array}\right\} \gamma^{u} \in \Gamma^{u}, u \in U, d \in D. \qquad (27)$$

**5. Searching for optimal solutions.** The problem is formulated as a multiobjective optimization problem. Optimality is understood as a situation where no objective function can be improved without worsening at least one other (Pareto optimality). In the absence of any additional information, it cannot be said that one of the Pareto optimal solutions is better than the other. This means that a slight increase in costs may lead to a decrease in traffic or latency, or an increase in delay may be accompanied by a decrease in traffic or costs. In such situations, it is important to find as many solutions as possible on the Pareto front. By isolating one specific optimal solution at a time or by scalarizing the problem, certain common optimization approaches offer to reduce a multi-objective optimization problem to a single-objective optimization problem. When such techniques are employed, they must be used repeatedly in order to acquire the greatest number of solutions, which increases the time costs. However, there exist techniques that aim to guarantee convergence to a Pareto-optimal set while maintaining diversity in solutions [29]. They simultaneously consider all objective functions reserving the nature of the original problem. Evolutionary algorithms are one such method. This article uses the Non-dominated Sorting Genetic Algorithm II (NSGA II), which is a very famous and widely used variant of the genetic algorithm. According to [30], by solving several test problems using the NSGA-II, it was found that this method outperforms other algorithms under testing, such as PAES (Pareto Achieved Evolution Strategy) and SPEA (Strength Pareto evolutionary algorithm), in terms of finding a diverse set of solutions. The algorithm follows the general scheme of the genetic algorithm with modified crossover and selection. By definition, A solution $p_1$ is said to dominate the other solution $p_2$ if $p_1$ is no worse than $p_2$ for all objectives and $p_1$ is strictly better than $p_2$ in at least one objective.

The crowding distance measures the density of solutions around an individual in the objective space based on the distances between neighboring solutions in each dimension. It is used in NSGA-II to select diverse solutions for the next generation. For each objective function, the population is initially sorted in ascending order of magnitude. Intermediate solutions are given a distance value equal to the absolute normalized difference in the function values of two adjacent solutions, while the boundary solutions are given an infinite distance value. The overall crowding-distance value is calculated as the sum of individual distance values corresponding to each objective [29].

The selection of individuals is carried out using a binary tournament selection, which involves several tournaments between two individuals chosen at random from the population each time. The winner of each tournament is selected for crossover. The steps of NSGA-II can be summarized as follows [31]:

1. initialize the NSGA-II parameters like the population size $N$, the generation counter $t = 0$, and the offspring population obtained by applying crossover and mutation $Q_t = \emptyset$.

2. create the initial parent population $P_t$ according to the procedures described later on.

3. merge the parent and offspring populations for maintaining elitism $R_t = P_t \cup Q_t$.

4. sort the population of candidate solutions $R_t$ into different fronts $F_1, F_2, \ldots, F_r$ using non-dominated sorting.

5. calculate the crowding distance of each candidate solution in each front $F_i$.

6. using binary tournament selection, choose the parent population for the next generation $P_{t+1}$ based on the non-domination status and crowding distance .

7. apply crossover and mutation on $P_{t+1}$, generate offspring population $Q_{t+1}$ for next generation.

8. increment $t$.

9. check for termination criteria which might be a maximum number of generations or a satisfactory level of convergence. If the termination criteria is not met repeat steps from 3 to 9.

Figure 4 illustrates the main steps of the algorithms as described in the original article [29].

In step 1 during the generation of the initial population, individuals that represent *pseudo-solutions* are generated. A *pseudo-solution* might not satisfy all the constraints but its *structure* is similar to a feasible solution.

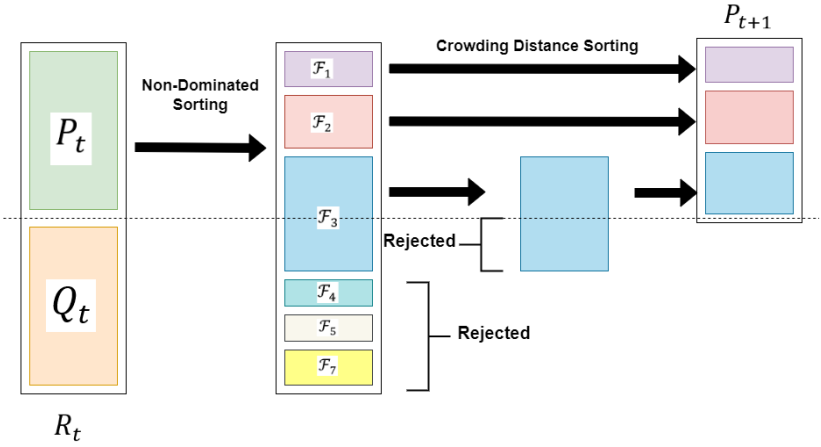



Fig. 4. NSGA II procedure [29]

This helps reduce the time required to reach feasible and optimal solutions. The following procedure is used:

1. determine the set of communication technologies $\Gamma_{used}$ used by all the devices in $D$.

2. filter out any gateway in $U$ that does not support any $\gamma \in \Gamma_{used}$. As a result, $U$ can be written as ($U = U_{usefull} \cup U_{useless}$).

3. determine the maximum number of gateways $i_{u,max}$ that can be installed, which is the minimum between the size of $U_{usefull}$ and the size of $M$.

4. choose a random number $i_u \in \{1, ..., i_{u,max}\}$ of gateways from $U_{usefull}$ and install them at random locations from $M$. As a result, we obtain $w_{um}$.

5. assign each device to an installed gateway that supports its communication technology. As a result, we obtain $z_{du}$.

6. repeat 3, 4 and 5 until the required number of individuals in the initial population is reached.

A similar procedure is used to generate the initial population for the second step:

1. choose a subset $U_{cloud} \subset U_{installed} = \{u \in U : \exists m \in M, w_{um} \neq 0\}$. For each $u \in U_{cloud}$ set $v_{uc} = 1$.

2. if the size of $U_{cloud}$ is equal to the size of $U_{installed}$ set $x_{np} = 0, \forall n \in N, p \in P$ and $b_{pl} = 0, \forall p \in P, l \in L$ then go the previous step if the number of individuals is not reached.

3. if the size of $U_{cloud}$ is less than to the size of $U_{installed}$ determine the maximum number of fog nodes $j_{n,max}$ that can be installed, which is the minimum between the size of $P$, the size of $N$ and the size of $U_{installed} \setminus U_{cloud}$.

4. choose a random number $j_n \in \{1, ..., j_{n,max}\}$ of fog nodes $N_{installed} \subset N$ and install them at random locations $P_{occupied} \subset P$. As a result, we obtain $x_{np}$.

5. assign to each location $p \in P_{occupied}$ its closest gateway first to make sure all fog nodes are used then assign the rest of the gateways in the same way. As a result we obtain $y_{up}$.

6. for every $p \in P_{occupied}$ assign a random $l \in L$. As a result we obtain $b_{pl}$.

7. repeat the previous steps until the required number of individuals in the initial population is reached.

In this work, the algorithm implementation within the pymoo framework [32] is used. Pymoo is a python framework that supports modern single-objective and multi-objective optimization algorithms.

To automatically select the best compromise solution among those obtained, the Technique for Order of Preference by Similarity to the Ideal Solution (TOPSIS) method is used. It is a multi-criteria decision making (MCDM) technique that compares a set of alternatives based on predefined criteria. According to this method, the chosen optimal solution must have the shortest Euclidean distance from the positive-ideal solution and the longest Euclidean distance from the negative-ideal solution [33]. In this work, the TOPSIS method is implemented in python based on [34]:

– Construct a normalized objective matrix with $m$ rows and $n$ columns:

$$t_{ij} = \frac{f_{ij}}{\sqrt{\sum_{i=1}^{m} f_{ij}^2}}. \tag{28}$$

– Construct a weighted normalized objective matrix by multiplying each column by a weight $w_j$ corresponding to the importance of the objective function:

$$v_{ij} = t_{ij} w_j. \tag{29}$$

– Determine the positive ideal solution, $A^+$, and negative-ideal solution, $A^-$ by finding the best value of each objective function. Where maximization is required, the best value is the largest value within the column of the objective matrix and where minimization is required (which is the case for all the

objective functions used in our case), the best value is the smallest value in the column. Mathematically, the positive ideal solution is given by:

$$A^+ = \left\{ \left( \max_i(v_{ij}) \;\middle|\; j \in J \right), \left( \min_i(v_{ij}) \;\middle|\; j \in J' \right) \;\middle|\; i \in 1,2,\ldots,m \right\} = \left\{ v_1^+, v_2^+, \ldots, v_j^+, \ldots, v_n^+ \right\}, \quad (30)$$

where $J$ is the set of indexes of maximization objectives and $J'$ is the set of indexes of minimization objectives in the overall set $\{1,2,3,4,...,n\}$. Next, find the worst value of each objective, which is the smallest and largest value within the column of the objective matrix for maximization and minimization objectives respectively. These values constitute the negative-ideal solution is given by:

$$A^- = \left\{ \left( \min_i(v_{ij}) \;\middle|\; j \in J \right), \left( \max_i(v_{ij}) \;\middle|\; j \in J' \right) \;\middle|\; i \in 1,2,\ldots,m \right\} = \left\{ v_1^-, v_2^-, \ldots, v_j^-, \ldots, v_n^- \right\}. \quad (31)$$

– Calculate the Euclidean distance between each solution and the positive-ideal and negative-ideal solutions:

$$S_{i+} = \sqrt{(\sum_{j=1}^{n} \left( v_{ij} - v_j^+ \right)^2}, i = 1,2,3,\ldots,m. \quad (32)$$

$$S_{i-} = \sqrt{\sum_{j=1}^{n} \left( v_{ij} - v_j^- \right)^2}, i = 1,2,3,\ldots,m. \quad (33)$$

– Calculate the closeness of each optimal solution:

$$C_i = \frac{S_{i-}}{S_{i-} + S_{i+}}, \quad (34)$$

when $S_{i-} = 0$, $C_i = 0$ and solution $i$ is the closest to the negative ideal. When $S_{i+} = 0$,$C_i = 1$ and solution $i$ is the closest to the positive ideal. The solution having the largest $C_i$ is the recommended solution.

The order in which the previously mentioned algorithms and procedures are applied is shown in Figure 5.

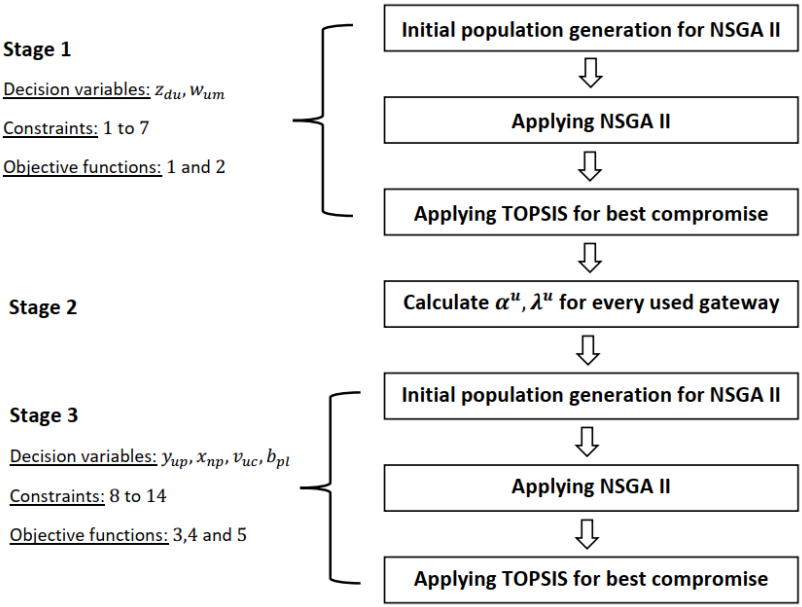



Fig. 5. The sequence of applying NSGA II and TOPSIS

**6. Simulation and Results.** A simple example is provided to demonstrate the results of applying the proposed method. However, the method was successfully applied to much larger and more complex cases. The success of the application refers to the convergence of the method and the quality of obtained solutions. Some of the numerical values chosen for the simulation are similar to what might be found in a real scenario, but some values, such as the position of the fog nodes and the cloud data center, are chosen to make it easier to visually demonstrate the concept. In real life, they would exist at much greater distances from the edge of the system. The example uses a set of abstract communication technologies that are characterized by the maximum number of supported devices $d_{max}^{\gamma}$ and the used frequency $f^{\gamma}$ (Table 1). These communication technologies are supported by both the set of edge devices and the set of gateways. The edge devices in this example are listed in Table 2 and are described by the supported interface $\gamma^d$, the number of bytes sent per second $\theta^d$, and their location $p^d = (x^d, y^d)$. It is worth noting that for simplicity and convenience of demonstration, it

is assumed that all components of the IIoT system can be distributed over a geographical area whose dimensions are $4000m \times 4000m$.

Table 1. Parameters of the used communication technologies

| | $d^{\gamma}_{max}$ | $f^{\gamma}$ MHz |
|---|---|---|
| $\gamma_0$ | 32 | 915 |
| $\gamma_1$ | 128 | 433.92 |
| $\gamma_2$ | 128 | 17.12 |

Table 2. Parameters of the edge devices

| [h!] | $\gamma^d$ | $\theta^d$ (bytes per second) | $x^d$ (m) | $y^d$ (m) | $G^d_{RX}$ (dBi) | $S^d_{RX}$ (dBm) | $P^d_{TX}$ (dBm) |
|---|---|---|---|---|---|---|---|
| $d_0$ | $\gamma_0$ | 512 | 1391 | 1945 | 2 | -70 | 14 |
| $d_1$ | $\gamma_0$ | 25600 | 1408 | 1461 | 2 | -70 | 14 |
| $d_2$ | $\gamma_0$ | 6400 | 488 | 1578 | 2 | -72 | 14 |
| $d_3$ | $\gamma_0$ | 12800 | 834 | 518 | 2 | -71 | 14 |
| $d_4$ | $\gamma_0$ | 12800 | 594 | 387 | 2 | -75 | 14 |
| $d_5$ | $\gamma_0$ | 3200 | 949 | 1628 | 2 | -73 | 14 |
| $d_6$ | $\gamma_0$ | 25600 | 904 | 235 | 2 | -72 | 14 |
| $d_7$ | $\gamma_1$ | 25600 | 401 | 3323 | 2 | -82 | 22 |
| $d_8$ | $\gamma_1$ | 3200 | 200 | 3306 | 2 | -81 | 20 |
| $d_9$ | $\gamma_2$ | 51200 | 3348 | 2960 | 2 | -90 | 20 |
| $d_10$ | $\gamma_2$ | 12800 | 159 | 2917 | 2 | -86 | 22 |
| $d_11$ | $\gamma_0$ | 12800 | 3267 | 3273 | 2 | -89 | 22 |
| $d_12$ | $\gamma_2$ | 25600 | 3406 | 3512 | 2 | -91 | 22 |
| $d_13$ | $\gamma_2$ | 12800 | 3800 | 2899 | 2 | -88 | 22 |
| $d_14$ | $\gamma_1$ | 6400 | 3748 | 3546 | 2 | -89 | 22 |

The gateways to choose from are listed in Table 3. In this case, there are 5 gateways described by their link speed $\kappa^u$, cost $\xi^u$, and supported communication technologies $\Gamma^u$. A gateway can support multiple technologies and multiple instances of the same technology.

Gateways can be physically installed in 6 possible locations, the coordinates of which are indicated in Table 4. The proposed method selects the best gateway and the best location given the given constraints at the first stage of optimization. It also determines which devices should be connected to the selected gateway.

The fog nodes to choose from are listed in Table 5. In this case, there are 4 nodes that can be described using available memory $\lambda^n$, number of available virtual processors $\alpha^n$, network bandwidth interface $\theta^n$ and cost $\xi^n$.

Table 3. Parameters of the gateways

| | $\kappa^u$ (Mbs) | $\xi^u$ ($) | $\Gamma^u$ | | | | |
|---|---|---|---|---|---|---|---|
| | | | $\gamma$ | $q(u,\gamma)$ | $G_{RX}^{\gamma u}$ dBi | $S_{RX}^{\gamma u}$ dBm | $P_{TX}^{\gamma u}$ dBm |
| $u_0$ | 95 | 20000 | $\gamma_0$ | 1 | 3 | −120 | 26 |
| | | | $\gamma_1$ | 1 | 4 | −110 | 24 |
| $u_1$ | 92 | 25000 | $\gamma_1$ | 1 | 4 | 110 | 24 |
| | | | $\gamma_2$ | 1 | 25 | 3 | -100 |
| $u_2$ | 32 | 18000 | $\gamma_1$ | 1 | 4 | 110 | 24 |
| | | | $\gamma_2$ | 1 | 25 | 3 | -100 |
| $u_3$ | 67 | 15000 | $\gamma_1$ | 1 | 4 | 110 | 24 |
| $u_4$ | 85 | 12000 | $\gamma_0$ | 1 | 3 | -120 | 26 |

Table 4. Coordinates of the possible locations for installing gateways

| | $x^m$ (mm) | $y^m$ (mm) |
|---|---|---|
| $m_0$ | 921 | 1254 |
| $m_1$ | 339 | 3093 |
| $m_2$ | 3592 | 3239 |
| $m_3$ | 489 | 2830 |
| $m_4$ | 3847 | 3225 |
| $m_5$ | 1026 | 956 |

Table 5. Parameters of fog nodes

| | $\lambda^n$ (Gigabyte) | $\alpha^n$ | $\theta^n$ (megabit per second) | $\xi^n$ ($) |
|---|---|---|---|---|
| $n_0$ | 256 | 96 | 3598 | 1000000 |
| $n_1$ | 128 | 48 | 2768 | 520000 |
| $n_2$ | 256 | 24 | 3486 | 600000 |
| $n_3$ | 128 | 96 | 2265 | 4800000 |

Nodes can be physically installed in 4 possible locations, the coordinates of which are indicated in Table 6. The proposed method will select the best variant and the best location, taking into account the given constraints, at the second stage of optimization. It will also determine which gateways should be connected to the selected node. The data center or cloud system is usually far from the edges of the IIoT network, but for demonstration purposes, it is considered to be set to $(x^c, y^c) = (2000, 3800)$.

Table 6. Coordinates of the possible locations for installing fog network nodes

| | $x^p$ (mm) | $y^p$ (mm) |
|---|---|---|
| $p_0$ | 2569 | 3084 |
| $p_1$ | 1679 | 2589 |
| $p_2$ | 1209 | 2896 |

Four possible communication channels or links can be used to connect the fog node to the data center. For each communication channel, the transmission rate $\theta^l$ and the price $\xi^l$ are listed in Table 7. The price represents the cost of deployment in dollars per meter.

Table 7. Parameters of the communication links

| | $\theta^l$ (megabits per second) | $\xi^l$ ($/m) |
|---|---|---|
| $l_0$ | 10 | 6 |
| $l_1$ | 100 | 14 |
| $l_2$ | 1000 | 24 |
| $l_3$ | 10000 | 50 |

The initial data on the system are depicted in Figure 6.

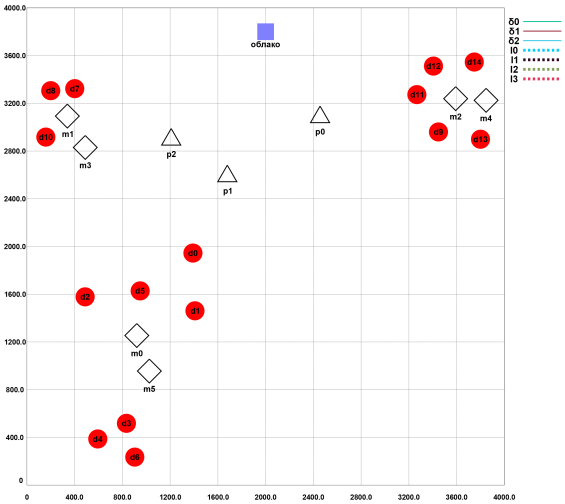



Fig. 6. Initial data about the IIoT system

In the first optimization step, NSGA II is applied first. As a result, we get 3 optimal solutions after 200 generations. As for the path loss, we consider that there is a direct line of sight between the devices and gateways for simplicity. Hence, the free space path loss model is considered.

In this example, we have no preference as to which solution to use, so TOPSIS is applied, resulting in the following solution:

$$w_{um} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

$$z_{du} = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

These results are shown in Figure 7.

Note that, for example, $w_{22} = 1$ means that $u_2$ is installed at $m_2$, and $z_{04} = 1$ means that $d_0$ is connected to $u_4$. At the intermediate stage, additional parameters of the installed gateways are determined, as in Table 8.

As a result of applying NSGA II at the second stage of optimization, 7 optimal solutions were obtained. In this example, we have no preference as to which solution to use, so TOPSIS is applied again, resulting in the following solution:

$$y_{up} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}, x_{np} = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0, \end{bmatrix}, b_{pl} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0, \end{bmatrix}, v_{uc} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}. \quad (35)$$

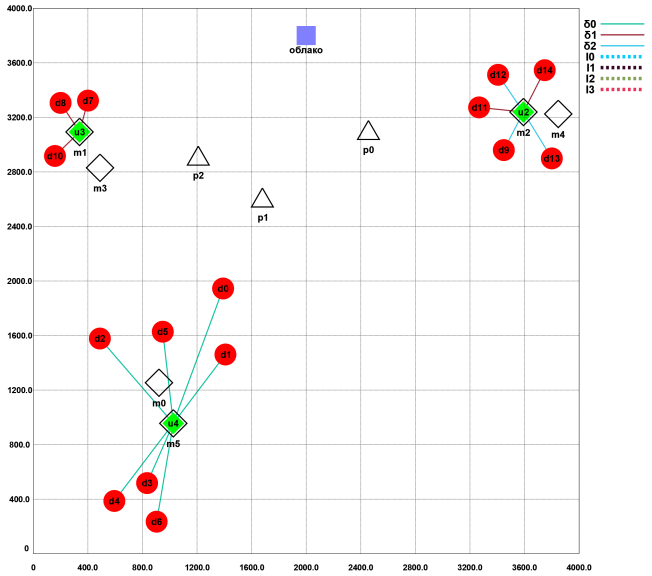



Fig. 7. Visualization of the results of the first stage

Table 8. Additional parameters for the installed gateways

| | $\lambda^u$ (Gigabytes) | $\alpha^u$ | $\theta^u$ (bytes per second) |
|---|---|---|---|
| $u_0$ | - | - | - |
| $u_1$ | - | - | - |
| $u_2$ | 2 | 2 | 3720 |
| $u_3$ | 2 | 2 | 1520 |
| $u_4$ | 2 | 2 | 5756 |

The final results are shown in Figure 8.

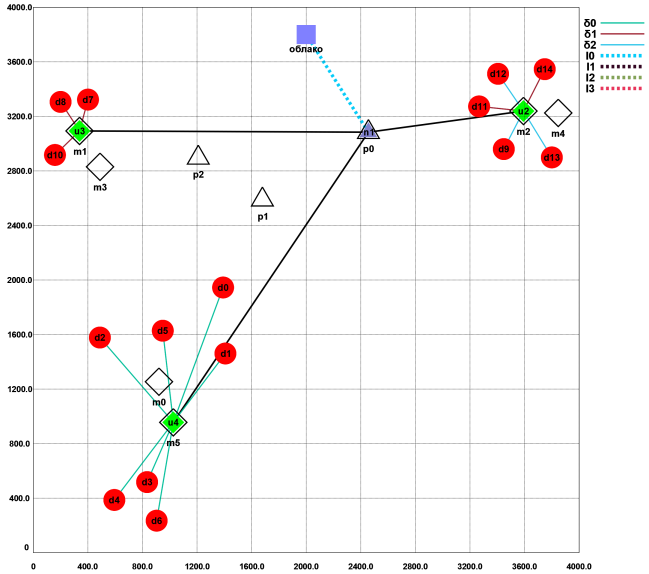



Fig. 8. Visualization of the results of the third stage

Note that, for example, $y_{20} = 1$ means that $u_2$ is connected to a node that is installed at $p_0$, and $x_{10} = 1$ means that $n_1$ is assigned to $p_0$. $b_{00} = 1$ means that a channel of type $l_0$ is used to connect the fog node at $p_0$ to the cloud.

Thus, the proposed method automatically allocates network resources using TOPSIS and NSGA II. The user is only required to provide a database of available resources and possible places to install these resources. The algorithm then decides what to use and where to install it.

**7. Conclusion.** Although the proposed technique does not consider mobility or technology-specific factors, it has the potential to ensure IIoT performance and effectiveness and decrease the dependency on the knowledge of system architects by looking for the best non-dominated Pareto solutions. The importance of this work lies in its holistic approach and generality. It not only handles specific architectural layers as most methods presented in the revised literature do but also includes all layers in the system hierarchy. The generality comes from the technique being agnostic to the type of IoT technology by using only general properties that characterize each level component. While the presented example demonstrates the applicability of the method on small-scale problems, applying the method to larger-scale problems has also yielded

similar results. The next step in this research is to introduce more hardware acceleration by exploiting the potentials of graphics cards, collect statistical data about the results of applying the proposed technique in different projects, which might help further verify its usefulness and practicality, and comparing the results and performance indicators of using the proposed model with optimization algorithms other that NSGA-II.

## References


1. Microsoft Azure official web site. Available at: https://azure.microsoft.com/en-us/resources/cloud-computing-dictionary/what-is-the-cloud. (accessed 02.01.2023).
2. Basir R., Qaisar S., Ali M., Aldwairi M., Ashraf M.I., Mahmood A., Gidlund M. Fog Computing Enabling Industrial Internet of Things: State-of-the-Art and Research Challenges. Sensors. 2019. vol. 19(21). no. 4807.
3. Tsvirkun A.D. Osnovy sinteza struktury slozhnyh system [Fundamentals of synthesis of the structure of complex systems]. M.: Nauka, 1982. 200 p. (In Russ.).
4. Tsvirkun A.D., Akinfiev V.K., Solov'ev M.M. Modelirovanie razvitija krupnomasshtabnyh sistem: (Na primere toplivno-jenergeticheskih otraslej i kompleksov) [Modeling the development of large-scale systems: (On the example of fuel and energy industries and complexes)]. M.: Ekonomika, 1983. 176 p. (In Russ.).
5. Akinfiev V.K., Cvirkun A.D. Metody i instrumental'nye sredstva upravlenija razvitiem kompanij so slozhnoj strukturoj aktivov [Methods and tools for managing the development of companies with a complex asset structure.]. M.: IPU RAN, 2020. 307 p. (In Russ.).
6. Tsvirkun A.D., Akinfiev V.K., Filippov V.A. Imitacionnoe modelirovanie v zadachah sinteza struktury slozhnyh system [Simulation modeling in problems of synthesis of the structure of complex systems]. M.: Nauka, 1985. 173 p. (In Russ.).
7. Potryasayev S.A. Sintez tehnologij i kompleksnyh planov upravlenija informacionnymi processami v promyshlennom internete. dis. d-r teh. nauk. [Synthesis of technologies and compex plans for managing information processes in the industrial Internet. doct. diss. in technical sciences.]. St. Petersburg, 2020. (In Russ.).
8. International society of automation official website. Available at: https://www.isa.org/intech-home/2019/march-april/features/rami-4-0-reference-architectural-model-for-industr. (accessed 13.09.2023).
9. Industry IoT consortium official website. Available at: https://www.iiconsortium.org/pdf/IIRA-v1.9.pdf. (accessed 12.09.2023).
10. Industrial value chain initiative official website. Available at: https://docs.iv-i.org/doc_161208_Industrial_Value_Chain_Reference_Architecture.pdf. (accessed 14.09.2023).
11. Srinidhi N.N., Kumar S.D., Venugopal K.R. Network optimizations in the Internet of Things: A review. Engineering Science and Technology, an International Journal. 2019. vol. 22. no. 1. pp. 1–21.
12. Ceselli A., Premoli M., Secci S. Mobile Edge Cloud Network Design Optimization. IEEE/ACM Transactions on Networking. 2017. vol. 25. no. 3. pp. 1818–1831.
13. Chimmanee K., Jantavongso S. Practical mobile network planning and optimization for Thai smart cities: Towards a more inclusive globalization. Research in Globalization. 2021. vol. 3. no. 100062.
14. Gava M.A., Rocha H.R.O., Faber M.J., Segatto M.E.V., Wortche H., Silva J.A.L. Optimizing Resources and Increasing the Coverage of Internet-of-Things (IoT) Networks: An Approach Based on LoRaWAN. Sensors. 2023. vol. 23(3). no. 1239.

15. Purnama A.A.F., Nashiruddin M.I. SigFox-based Internet of Things Network Planning for Advanced Metering Infrastructure Services in Urban Scenario. IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT). 2020. pp. 15–20.
16. Nashiruddin M.I., Purnama A.A.F. NB-IOT network planning for advanced metering infrastructure in Surabaya, Sidoarjo, and gresik. 8th International Conference on Information and Communication Technology (ICoICT). 2020. pp. 1–6.
17. Haider F., Zhang D., St-Hilaire M., Makaya C. On the Planning and Design Problem of Fog Computing Networks. IEEE Transactions on Cloud Computing. 2018. vol. 9. no. 2. pp. 724–736.
18. Zhang D., Haider F., St-Hilaire M., Makay C. Model and algorithms for the planning of Fog Computing Networks. IEEE Internet of Things Journal. 2019. vol. 6. no. 2. pp. 3873–3884.
19. Ebraheem A., Ivanov I.A. Internet of Things: Analysis of Parameters and Requirements. International Conference on Smart Applications, Communications and Networking (SmartNets). 2022. pp. 01–04.
20. Kaur S., Mir R.N. Base station positioning in Wireless Sensor Networks. International Conference on Internet of Things and Applications (IOTA). 2016. pp. 116–120.
21. REMCOM official web site. Available at: https://www.remcom.com/wireless-insite-em-propagation-software. (accessed 04.07.2023).
22. Mathworks official web site. Available at: https://mathworks.com/help/comm/ref/rfprop.raytracing.html. (accessed 04.07.2023).
23. Alqudah Y.A. On the performance of Cost 231 Walfisch Ikegami model in deployed 3.5 GHz network. The International Conference on Technological Advances in Electrical, Electronics and Computer Engineering (TAEECE). 2013. pp. 524–527.
24. Correia L.M. A view of the COST 231-Bertoni-Ikegami model. 3rd European Conference on Antennas and Propagation. 2009. pp. 1681–1685.
25. Zhang J., Gentile C., Garey W. On the Cross-Application of Calibrated Pathloss Models Using Area Features: Finding a way to determine similarity between areas. IEEE Antennas and Propagation Magazine. 2019. vol. 62. no. 1. pp. 40–50.
26. Rackspace technology official web site. Available at: https://docs.rackspace.com/blog/different-types-of-oci-servers-in-the-cloud. (accessed 12.05.2023).
27. Google cloud official web site. Available at: https://cloud.google.com/compute/docs/machine-resource. (accessed 12.05.2023).
28. Amazon web services official web site. Available at: https://aws.amazon.com/ec2/instance-types. (accessed 12.05.2023).
29. Deb K., Pratap A., Agarwal S., Meyarivan T. A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Transactions on Evolutionary Computation. 2002. vol. 6. no. 2. pp. 182–197.
30. Yusoff Y., Ngadiman M., Zain A. Overview of NSGA-II for optimizing machining process parameters. Procedia Engineering. 2011. vol. 15. pp. 3978–3983.
31. Palaparthi A., Riede T., Titze I.R. Combining Multiobjective Optimization and Cluster Analysis to Study Vocal Fold Functional Morphology. IEEE Transactions on Biomedical Engineering. 2014. vol. 61. no. 7. pp. 2199–2208.
32. Blank J., Kalyanmoy D. Pymoo: Multi-objective optimization in python. IEEE Access. 2020. vol. 8. pp. 89497–89509.
33. Halicka K. Technology Selection Using the TOPSIS Method. Foresight and STI Governance. 2020. vol. 14. no. 1. pp. 85–96.

34. Sarraf A., Mohaghar A., Bazargani H. Developing TOPSIS method using statistical normalization for Selecting Knowledge Management Strategies. Journal of Industrial Engineering and Management. 2013. vol. 6. no. 4. pp. 860–875.

**Ebraheem Ali** — Ph.D. student, HSE University. Research interests: industrial internet of things, control theory, software technologies and development of information systems. The number of publications — 4. aebrakhim@hse.ru; 34, Tallinskaya St., 123592, Moscow, Russia; office phone: +7(495)772-9590 [15166].

**Ivanov Ilya** — Ph.D., Associate professor, academic supervisor of the programme (internet of things and cyber-physical systems), HSE University. Research interests: internet of things, cyber-physical systems, control and diagnostics of electronic devices. The number of publications — 105. i.ivanov@hse.ru; 34, Tallinskaya St., 123592, Moscow, Russia; office phone: +7(495)772-9590 [15166].

**УДК** 004.021 **DOI** 10.15622/ia.23.2.3

А. Эбрахим, И.А. Иванов

# НА ПУТИ К АВТОМАТИЗИРОВАННОМУ И ОПТИМАЛЬНОМУ ПРОЕКТИРОВАНИЮ СИСТЕМ IIOT

*Эбрахим А., Иванов И.А.* **На пути к автоматизированному и оптимальному проектированию систем IIoT.**

**Аннотация.** В современном мире Интернет вещей стал неотъемлемой частью нашей жизни. Растущее число умных устройств и их повсеместное распространение усложняют разработчикам и системным архитекторам эффективное планирование и внедрение систем Интернета вещей и промышленного Интернета вещей. Основная цель данной работы – автоматизировать процесс проектирования промышленных систем Интернета вещей при оптимизации параметров качества обслуживания, срока службы батареи и стоимости. Для достижения этой цели вводится общая четырехуровневая модель туманных вычислений, основанная на математических множествах, ограничениях и целевых функциях. Эта модель учитывает различные параметры, влияющие на производительность системы, такие как задержка сети, пропускная способность и энергопотребление. Для нахождения Парето-оптимальных решений используется генетический недоминируемый алгоритм сортировки II, а для определения компромиссных решений на Парето-фронте – метод определения порядка предпочтения по сходству с идеальным решением. Оптимальные решения, сгенерированные этим подходом, представляют собой серверы, коммуникационные каналы и шлюзы, информация о которых хранится в базе данных. Эти ресурсы выбираются на основе их способности улучшить общую производительность системы. Предлагаемая стратегия следует трехэтапному подходу для минимизации размерности и уменьшения зависимостей при исследовании пространства поиска. Кроме того, сходимость оптимизационных алгоритмов улучшается за счет использования предварительно настроенной начальной популяции, которая использует существующие знания о том, как должно выглядеть решение. Алгоритмы, используемые для генерации этой начальной популяции, описываются подробно. Для иллюстрации эффективности автоматизированной стратегии приводится пример ее применения.

**Ключевые слова:** IoT, IIoT, NGSA-II, TOPSIS, облако, туманные вычисления, многокритериальная оптимизация, шлюз, пограничные устройства.

## Литература

1. Официальный сайт Microsoft Azure. URL: https://azure.microsoft.com/en-us/resources/cloud-computing-dictionary/what-is-the-cloud (дата обращения: 02.01.2023).
2. Basir R., Qaisar S., Ali M., Aldwairi M., Ashraf M.I., Mahmood A., Gidlund M. Fog Computing Enabling Industrial Internet of Things: State-of-the-Art and Research Challenges. Sensors. 2019. vol. 19(21). no. 4807.
3. Цвиркун А.Д. Основы синтеза структуры сложных систем. М.: Наука, 1982. 200 с.
4. Цвиркун А.Д., Акинфиев В.К., Соловьев М.М. Моделирование развития крупномасштабных систем: (На примере топливно-энергетических отраслей и комплексов). М.: Экономика, 1983. 176 с.
5. Акинфиев В.К., Цвиркун А.Д. Методы и инструментальные средства управления развитием компаний со сложной структурой активов. М.: ИПУ РАН, 2020. 307 с.

6. Цвиркун А.Д., Акинфиев В.К., Филиппов В.А. Имитационное моделирование в задачах синтеза структуры сложных систем. М.: Наука, 1985. 173 с.
7. Потрясаев С.А. Синтез технологий и комплексных планов управления информационными процессами в промышленном интернете: дис. д-р тех. наук. СПБ., 2020.
8. Официальный сайт Международного общества автоматизации. URL: https://www.isa.org/intech-home/2019/march-april/features/rami-4-0-reference-architectural-model-for-industr (дата обращения: 13.09.2023).
9. Официальный сайт промышленного IoT-консорциума. URL: https://www.iiconsortium.org/pdf/IIRA-v1.9.pdf (дата обращения: 12.09.2023).
10. Официальный сайт инициативы в области промышленной цепочки создания стоимости. URL: https://docs.iv-i.org/doc_161208_Industrial_Value_Chain_Reference_Architecture.pdf (дата обращения: 14.09.2023).
11. Srinidhi N.N., Kumar S.D., Venugopal K.R. Network optimizations in the Internet of Things: A review. Engineering Science and Technology, an International Journal. 2019. vol. 22. no. 1. pp. 1–21.
12. Ceselli A., Premoli M., Secci S. Mobile Edge Cloud Network Design Optimization. IEEE/ACM Transactions on Networking. 2017. vol. 25. no. 3. pp. 1818–1831.
13. Chimmanee K., Jantavongso S. Practical mobile network planning and optimization for Thai smart cities: Towards a more inclusive globalization. Research in Globalization. 2021. vol. 3. no. 100062.
14. Gava M.A., Rocha H.R.O., Faber M.J., Segatto M.E.V., Wortche H., Silva J.A.L. Optimizing Resources and Increasing the Coverage of Internet-of-Things (IoT) Networks: An Approach Based on LoRaWAN. Sensors. 2023. vol. 23(3). no. 1239.
15. Purnama A.A.F., Nashiruddin M.I. SigFox-based Internet of Things Network Planning for Advanced Metering Infrastructure Services in Urban Scenario. IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT). 2020. pp. 15–20.
16. Nashiruddin M.I., Purnama A.A.F. NB-IOT network planning for advanced metering infrastructure in Surabaya, Sidoarjo, and gresik. 8th International Conference on Information and Communication Technology (ICoICT). 2020. pp. 1–6.
17. Haider F., Zhang D., St-Hilaire M., Makaya C. On the Planning and Design Problem of Fog Computing Networks. IEEE Transactions on Cloud Computing. 2018. vol. 9. no. 2. pp. 724–736.
18. Zhang D., Haider F., St-Hilaire M., Makay C. Model and algorithms for the planning of Fog Computing Networks. IEEE Internet of Things Journal. 2019. vol. 6. no. 2. pp. 3873–3884.
19. Ebraheem A., Ivanov I.A. Internet of Things: Analysis of Parameters and Requirements. International Conference on Smart Applications, Communications and Networking (SmartNets). 2022. pp. 01–04.
20. Kaur S., Mir R.N. Base station positioning in Wireless Sensor Networks. International Conference on Internet of Things and Applications (IOTA). 2016. pp. 116–120.
21. Официальный сайт REMCOM. URL: https://www.remcom.com/wireless-insite-em-propagation-software (дата обращения: 04.07.2023).
22. Официальный сайт Mathworks. URL: https://mathworks.com/help/comm/ref/rfprop.raytracing.html (дата обращения: 04.07.2023).
23. Alqudah Y.A. On the performance of Cost 231 Walfisch Ikegami model in deployed 3.5 GHz network. The International Conference on Technological Advances in Electrical, Electronics and Computer Engineering (TAEECE). 2013. pp. 524–527.

24. Correia L.M. A view of the COST 231-Bertoni-Ikegami model. 3rd European Conference on Antennas and Propagation. 2009. pp. 1681–1685.
25. Zhang J., Gentile C., Garey W. On the Cross-Application of Calibrated Pathloss Models Using Area Features: Finding a way to determine similarity between areas. IEEE Antennas and Propagation Magazine. 2019. vol. 62. no. 1. pp. 40–50.
26. Официальный сайт Rackspace. URL: https://docs.rackspace.com/blog/different-types-of-oci-servers-in-the-cloud (дата обращения: 12.05.2023).
27. Официальный сайт Google Cloud. URL: https://cloud.google.com/compute/docs/machine-resource (дата обращения: 12.05.2023).
28. Официальный сайт Amazon Web Services. URL: https://aws.amazon.com/ec2/instance-types (дата обращения: 12.05.2023).
29. Deb K., Pratap A., Agarwal S., Meyarivan T. A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Transactions on Evolutionary Computation. 2002. vol. 6. no. 2. pp. 182–197.
30. Yusoff Y., Ngadiman M., Zain A. Overview of NSGA-II for optimizing machining process parameters. Procedia Engineering. 2011. vol. 15. pp. 3978–3983.
31. Palaparthi A., Riede T., Titze I.R. Combining Multiobjective Optimization and Cluster Analysis to Study Vocal Fold Functional Morphology. IEEE Transactions on Biomedical Engineering. 2014. vol. 61. no. 7. pp. 2199–2208.
32. Blank J., Kalyanmoy D. Pymoo: Multi-objective optimization in python. IEEE Access. 2020. vol. 8. pp. 89497–89509.
33. Halicka K. Technology Selection Using the TOPSIS Method. Foresight and STI Governance. 2020. vol. 14. no. 1. pp. 85–96.
34. Sarraf A., Mohaghar A., Bazargani H. Developing TOPSIS method using statistical normalization for Selecting Knowledge Management Strategies. Journal of Industrial Engineering and Management. 2013. vol. 6. no. 4. pp. 860–875.

**Эбрахим Али** — аспирант, Национальный исследовательский университет «Высшая школа экономики». Область научных интересов: промышленный интернет вещей, теория управления, технология разработки программных комплексов. Число научных публикаций — 4. aebrakhim@hse.ru; улица Таллинская, 34, 123592, Москва, Россия; р.т.: +7(495)772-9590 [15166].

**Иванов Илья Александрович** — канд. техн. наук, доцент, научный руководитель программы (интернет вещей и киберфизические системы), Национальный исследовательский университет «Высшая школа экономики». Область научных интересов: интернет вещей, киберфизические системы, контроль и диагностика электронных устройств. Число научных публикаций — 105. i.ivanov@hse.ru; улица Таллинская, 34, 123592, Москва, Россия; р.т.: +7(495)772-9590 [15166].

**УДК** 519.85 **DOI** 10.15622/ia.23.2.4

А.А. Сирота, А.В. Акимов, Р.Р. Отырба

# ДЕФОРМИРУЮЩИЕ ПРЕОБРАЗОВАНИЯ ИЗОБРАЖЕНИЙ И ИХ ПРИМЕНЕНИЕ ПРИ АУГМЕНТАЦИИ ДАННЫХ ДЛЯ ОБУЧЕНИЯ ГЛУБОКИХ НЕЙРОННЫХ СЕТЕЙ

*Сирота А.А., Акимов А.В., Отырба Р.Р.* **Деформирующие преобразования изображений и их применение при аугментации данных для обучения глубоких нейронных сетей.**

**Аннотация.** Проведены исследования возможностей аугментации (искусственного размножения) обучающих данных в задаче классификации с использованием деформирующих преобразований обрабатываемых изображений. Представлены математическая модель и быстродействующий алгоритм выполнения деформирующего преобразования изображения, при использовании которых исходное изображение преобразуется с сохранением своей структурной основы и отсутствием краевых эффектов. Предложенный алгоритм используется для аугментации наборов изображений в задаче классификации, содержащих относительно небольшое количество обучающих примеров. Аугментация исходной выборки осуществляется в два этапа, включающих зеркальное отображение и деформирующее преобразование каждого исходного изображения. Для проверки эффективности подобной техники аугментации в статье проводится обучение нейронных сетей – классификаторов различного вида: сверточных сетей стандартной архитектуры (convolutional neural network, CNN) и сетей с остаточными связями (deep residual network, DRN). Особенностью реализуемого подхода при решении рассматриваемой задачи является также отказ от использования предобученных нейронных сетей с большим количеством слоев и дальнейшим переносом обучения, поскольку их применение несет за собой затраты с точки зрения используемого вычислительного ресурса. Показано, что эффективность классификации изображений при реализации предложенного метода аугментации обучающих данных на выборках малого и среднего объема повышается до статистически значимых значений используемой метрики.

**Ключевые слова:** глубокие нейронные сети, аугментация обучающих данных, деформирующие искажения изображений, эффективность глубоких нейронных сетей.

**1. Введение.** В современных обучающихся системах зачастую возникает проблема недостатка данных, используемых для обучения, связанная как с ресурсными ограничениями при их подготовке, так и с объективными факторами, мешающими получить достаточный набор примеров. Для разрешения этой проблемы в системах классификации на основе глубоких нейронных сетей (ГНС) применяются разнообразные технологии аугментации данных, используемых для обучения. Под аугментацией далее будем понимать искусственное размножение данных (ИРД), т.е. использование некоторых «опорных» образов для увеличения объема обучающих выборок на основе стохастических или детерминистских моделей представления данных. Применение аугментации в условиях малой и средней по объему выборки во многих случаях позволяет получить более высокую

эффективность алгоритмов классификации и сегментации изображений при их тестировании на новых данных. Такой подход к подготовке обучающих данных позволяет снизить затраты времени, а также, в определенной степени, преодолеть проблему несбалансированности обучающих данных различных классов.

Существует и альтернативный подход в случаях, когда сбор необходимого числа обучающих образов оказывается сложен из-за специфического характера предметной области. Он состоит в использовании заранее предобученных моделей классификаторов. Так, часто реализуемым подходом применительно к практике применения ГНС является использование готовых (предобученных) нейронных сетей с большим количеством слоев и последующей реализации техники переноса обучения в контексте решаемой задачи на основе имеющейся малой выборки, которая применяется для дообучения. Однако такой подход видится не всегда оправданным, так как использование архитектуры предообученных сетей требует существенных вычислительных ресурсов. Поэтому вариант с проведением обучения сетей собственной архитектуры, удовлетворяющей требованиям по ресурсоемкости, остается востребованным.

Вопросы повышения качества классификации образов, основанные на ИРД, продемонстрированы в работах [1 –15]. В большинстве из них для генерации новых изображений в обучающей выборке при обучении ГНС используются различные преобразования: поворот, сжатие и растяжение, наклон, зеркальное отражение, обрезка, смещение и другие. Возможности современных программных сред, например, Keras (Tensorflow) предусматривают включение опций по использованию подобных стандартных средств аугментации. В работах отечественных авторов [7 –15] также представлено применение различных сочетаний алгоритмов аугментации в задачах обработки изображений.

В [7] новые данные в обучающей выборке генерируются с помощью морфинг-преобразований путем «скрещивания» исходных данных между собой. В [8] предложены алгоритмы внесения реалистической деформации изображений лиц, с помощью которых была размножена стандартная обучающая выборка изображений для алгоритма Виолы-Джонса. Показано, что подобным образом объем обучающей выборки может быть уменьшен в примерно в 10 раз при снижении вероятности распознавания не более чем на 2–4%. В [10] описывается алгоритм ИРД для задач машинного обучения классификаторов биологических объектов по спектрам, основанный

на использовании ядерных оценок функций правдоподобия классов образов.

Существуют также реализации алгоритмов деформирующих преобразований, используемых при аугментации изображений, включенные в современные библиотеки и фреймворки для Python, такие, например, как библиотека Albumentations. Наиболее популярный среди данной категории алгоритм, известный под названием ElasticTransform [16], в Albumentations ориентирован на аугментацию при решении задач классификации, сегментации и поиска объектов на изображениях и основан на использовании разнообразных методов интерполяции данных. Примеры применения этого алгоритма представлены в [15, 17, 18].

Помимо задач аугментации данных [4, 5, 8, 9] в известных работах описано использование деформирующих преобразований и их моделей внутри самих алгоритмов распознавания [19 –21], для генерации и модификации изображений при построении сверхразрешения [22], 3D-моделировании [9, 23], для защиты биометрических систем от атак с использованием этих деформаций [24, 25]. В [26, 27] представлены реализации данных подходов на основе ГНС.

В целом по результатам ранее выполненных исследований следует отметить высокую эффективность применения деформирующих преобразований изображений, показанную при обучении алгоритмов распознавания.

В то же время, следует отметить, что применение любых алгоритмов аугментации сопряжено с увеличением времени обучения, вследствие естественного его расхода на преобразование входных изображений. В особенности это касается алгоритмов, реализующих деформирующие преобразования, как весьма затратных.

В связи с изложенным целью настоящей работы является исследование вопросов построения быстродействующих алгоритмов формирования плавных деформирующих преобразований (ДП), обеспечивающих сохранение структурной основы исходного изображения, а также возможностей их применения для аугментации данных при обучении ГНС различной архитектуры в задачах классификации и оценки получаемого при этом выигрыша.

**2. Математическая модель и алгоритм деформирующих преобразований.** Пусть $\Omega_x \subset \mathrm{R}^n$ – некоторое множество значений аргумента непрерывной функции. Определим деформирующее преобразование (ДП) скалярной функции $n$ переменных

$f:\Omega_x \to G \subset \mathrm{R}^1,\ \mathbf{g}=f(\mathbf{x}),\ \mathbf{x}\in\Omega_x,\ \mathbf{g}\in G$ как преобразование, модифицирующее значение ее аргументов в области определения по заданному закону. Очевидным способом такой модификации является добавление значения функции деформации известного вида к исходным значениям этих переменных по следующей формуле:

$$\begin{gathered} f[x_1+r_1(x_1,...,x_n),...,x_n+r_n(x_1,...,x_n)]= \\ =f[u_1(x_1,...,x_n),...,u_n(x_1,...,x_n)]=g(x_1,...,x_n), \end{gathered} \tag{1}$$

где $g(x_1,...,x_n)$ – результирующая деформированная функция; $r_i(x_1,...,x_n)$, $i=\overline{1,n}$ – непрерывные функции ДП по каждой координате, которые могут быть детерминированными функциями фиксированной формы или представлять реализации многомерного случайного поля.

Перейдем к использованию векторных обозначений $\mathbf{x}=(x_1,...,x_n)^{\mathrm{T}}$, $\mathbf{r}(\mathbf{x})=(r_1(\mathbf{x}),...,r_n(\mathbf{x}))^{\mathrm{T}}$ и $\mathbf{u}(\mathbf{x})=(u_1(\mathbf{x}),...,u_n(\mathbf{x}))^{\mathrm{T}}$. Тогда можно записать выражение для выполняемого ДП в векторном виде следующем образом:

$$f[\mathbf{x}+\mathbf{r}(\mathbf{x})]=f[\mathbf{u}(\mathbf{x})]=g(\mathbf{x}).$$

При реализации алгоритмов ДП на основе (1) следует учесть ряд особенностей реализуемых преобразований.

Первая из них заключается в том, что для корректного формирования значений деформированной функции $f(\mathbf{u}(\mathbf{x}))=g(\mathbf{x})$, значения ее аргументов не должны выходить за пределы области определения исходной функции $f(\mathbf{x})$, имеющей в общем случае вид многомерной регулярной дискретной сетки. Другими словами, если для $f(\mathbf{x})$ вектор $\mathbf{x}\in\Omega_x$, то необходимо, чтобы вектор $\mathbf{u}(\mathbf{x})\in\Omega_x$.

Подобное условие может быть выполнено [28], если использовать специальные масочные функции $\mathbf{h}(\mathbf{x})=(h(x_1),...,h(x_n))^{\mathrm{T}}$, ограничивающие краевые значения $\mathbf{u}(\mathbf{x})$:

$$\mathbf{u}(\mathbf{x})=\mathbf{x}+\mathbf{u}'(\mathbf{x})=\mathbf{x}+\mathbf{h}(\mathbf{x})\otimes\mathbf{r}(\mathbf{x}),$$

где $\otimes$ – обозначение операции поэлементного умножения векторов. В качестве масочных функций могут быть использованы оконные

функции, имеющие трапециевидную форму или форму многомерного гауссовского закона распределения. В дальнейшем в реализованных алгоритмах обработки изображений будут использоваться трапециевидные функции, которые, в отличие от гауссиан, практически не модифицируют функцию деформирующего преобразования в центральной области определения.

Вторая особенность алгоритмов внесения ДП состоит (по крайней мере, для задач обработки изображений) в необходимости их применения по отношению к решетчатым функциям, заданным на многомерных дискретных сетках. В этом случае, как показано в [8] для обеспечения возможности модификации аргументов решетчатой функции $\hat{f}(\mathbf{x})$ при внесении в них непрерывных деформаций произвольного характера необходимо предварительно выполнить ее интерполяцию и представление в виде функции $\tilde{f}(\mathbf{x})$ вещественных переменных. Затем для интерполированной непрерывной функции необходимо выполнить обратный переход к дискретному представлению деформированной решетчатой функции $\hat{g}(\mathbf{x})$, путем дискретизации с исходным шагом.

Анализ применения возможных способов интерполяции, например, с использованием радиально-базисных функций или сплайна тонкой пластины показывает, что при реализации подобных подходов, оптимизированные алгоритмы имеют сложность порядка $O(N_{xy} \cdot \log(N_{xy}))$ [29 –31], где $N_{xy}$ – общее число элементов интерполируемого массива данных произвольной размерности (в данном случае общее число пикселей изображения). Также при использовании интерполяции дополнительные накладные расходы возникают при выполнении сдвиговых деформаций, сложность которых при заранее вычисленной функции деформации можно оценить как $O(N_{xy})$. Все это делает подобные алгоритмы ДП затратными в вычислительном отношении, особенно, если требуется преобразовать большое число изображений при аугментации обучающих примеров. Поэтому предлагается реализовать упрощенный подход, основанный на использовании ограничивающей маски и реализации перестановки элементов решетчатой функции в соответствие с используемым механизмом внесения ДП, сложность которого составляет $O(N_{xy})$.

Рассмотрим применительно к задаче обработки изображений следующую постановку. Пусть дифференцируемая функция

$f(x_1, x_2)$, $\mathbf{x}=(x_1, x_2) \in \Omega_x$ определена на прямоугольной области $\Omega_x$ размера $R_1 \times R_2$. Требуется преобразовать ее путем внесения ДП:

$$g(x_1, x_2) = f[x_1 + A_m r_1(x_1, x_2),\ x_2 + A_m r_2(x_1, x_2)],$$

где $\mathbf{r}(\mathbf{x}) = (r(x_1, x_2), r(x_1, x_2))^{\mathrm{T}}$ – непрерывные функции, описывающие деформирующие искажения и являющиеся либо функциями известной формы, либо, в общем случае, реализациями случайного поля, причем $\left|r_j(\mathbf{x})\right| \le 1$, $j$=1,2 ; $A_m$ – амплитуда ДП, обеспечивающая с учетом ограничений, вносимых наложением масочной функции, выполнение условия $\mathbf{x} + \mathbf{u}'(\mathbf{x}) \in \Omega_x$.

При цифровой обработке изображений в результате дискретизации по пространственным координатам с интервалом $\delta x = \delta x_1 = \delta x_2$ формируется цифровой эквивалент исходной функции путем фиксации ее значений в точках $\hat{f}(i,k)$ и цифровой эквивалент $\hat{g}(i,k)$ деформированной функции, заданных на прямоугольной дискретной сетке $(i,k) \in \Psi = \{i = \overline{1,N}, k = \overline{1,M}\}$:

$$\begin{gathered} \hat{f}(i,k) = f(t_i, s_k), \\ t_i = (i-1)\delta x,\ i = \overline{1,N},\ s_k = (k-1)\delta x,\ k = \overline{1,M}, \\ N = R_1 / \delta x,\ t_1 = 0,\ t_N = R_1,\ M = R_2 / \delta x,\ s_1 = 0,\ s_M = R_2, \\ \hat{g}(i,k) = g(t_i, s_k) = f[t_i + A_m r_1(t_i, s_k), s_k + A_m r_2(t_i, s_k)]. \end{gathered} \tag{2}$$

Введем вектор целочисленных индексов $\mathbf{v}(i,k) = (v_1(i,k), v_2(i,k))^{\mathrm{T}}$, $v_1(i,k) = \mathrm{round}[A_m r_1(t_i, s_k) / \delta x]$, $v_2(i,k) = \mathrm{round}[A_m r_2(t_i, s_k) / \delta x]$, где round[...] обозначает операцию округления до ближайшего значения. Тогда компоненты аргумента в (2) можно представить как:

$$\begin{gathered} t_i + A_m r_1(t_i, s_k) = \delta x(i-1) + \delta x v_1(i,k) + \varepsilon_1(i,k), \\ s_k + A_m r_2(t_i, s_k) = \delta x(k-1) + \delta x v_2(i,k) + \varepsilon_2(i,k), \\ -\delta x / 2 \le \varepsilon_1(i,k) < \delta x / 2,\ -\delta x / 2 \le \varepsilon_2(i,k) < \delta x / 2. \end{gathered} \tag{3}$$

Вектор $\boldsymbol{\varepsilon}(i,k) = (\varepsilon_1(i,k),\ \varepsilon_2(i,k))^{\mathrm{T}}$ в (3) также является случайным. Он характеризует погрешность квантования по уровню

заданной в непрерывном времени функции деформации, которая сопутствует процессу дискретизации исходного изображения по пространству:

$$\begin{gathered}\hat{g}(i,k) = g(t_i, s_k) = \\ = f[t_i + \delta x v_1(i,k) + \varepsilon_1(i,k), s_k + \delta x v_2(i,k) + \varepsilon_2(i,k)] .\end{gathered} \quad (4)$$

На основе представления ДП в (4) используем для цифрового деформированного изображения первое приближение:

$$\begin{gathered}\hat{g}(i,k) = f[\delta x(i-1) + \delta x v_1(i,k),\ \delta x(k-1) + \delta x v_2(i,k)] + \\ + f_t'[t_i + \delta x v_1(i,k), s_k + \delta x v_2(i,k)]\varepsilon_1(i,k) + \\ + f_s'[t_i + \delta x v_1(i,k), s_k + \delta x v_2(i,k)]\varepsilon_2(i,k) ,\end{gathered}$$

где $f_t', f_s'$ – частные производные исходной функции $f$. Тогда для (4) можно записать в скалярной и операторной форме:

$$\begin{gathered}\hat{g}(i,k) = \hat{f}(i + v_1(i,k), k + v_2(i,k)) + O(\varepsilon_1) + O(\varepsilon_2) , \\ \hat{\mathbf{g}} = G\left[\hat{\mathbf{f}}, \mathbf{v}\right] + G\left[\hat{\mathbf{f}}_t', \mathbf{v}\right] \otimes \boldsymbol{\varepsilon}_1 + G\left[\hat{\mathbf{f}}_s', \mathbf{v}\right] \otimes \boldsymbol{\varepsilon}_2 ,\end{gathered} \quad (5)$$

где $\hat{\mathbf{g}}$ – обозначение полученной после перестановки решетчатой функции; $G\left[\hat{\mathbf{f}}, \mathbf{v}\right]$ – оператор перестановки позиций размещения исходной решетчатой функции $\hat{f}(i,k)$ в соответствии со значениями $\mathbf{v}(i,k)$; $G\left[\hat{\mathbf{f}}_{t,s}', \mathbf{v}\right]$ – обозначение синхронно выполняемых операторов перестановки производных функции $f$, взятых в точках дискретизации по пространству.

Оператор $G(\hat{\mathbf{f}}, \mathbf{v})$ является случайным оператором, определяющим возможные размещения элементов $\hat{\mathbf{f}}$ и $\hat{\mathbf{f}}'$ в элементы $\hat{\mathbf{g}}_1 = G\left[\hat{\mathbf{f}}, \mathbf{v}\right]$ и $\hat{\mathbf{g}}_2 = G\left[\hat{\mathbf{f}}', \mathbf{v}\right]$, соответственно. Для случайной функции ДП он задает вероятностные переходы при преобразовании множества пикселей исходного изображения $\hat{\mathbf{f}}$ во множество пикселей изображения $\hat{\mathbf{g}}$.

Оператор является сюръекцией, в том смысле, что каждый элемент деформированной функции, является образом хотя бы одного

элемента исходной. Обратное, вообще говоря, неверно. Как уже сказано, возможны незначительные выпадения пикселей исходного изображения и их замещение близко расположенными пикселями этого же изображения. Также при размещении возможны повторения пикселей исходного изображения в случае, если изменения деформации происходят с отрицательной производной.

Анализ показывает, что значения $\boldsymbol{\varepsilon}(i,k)$ в различных точках пространства малы (сопоставимы с погрешностью квантования исходной функцией деформации) и практически не коррелированы, что позволяет считать ее воздействие близким к воздействию белого шума с существенно меньшей относительно первого слагаемого амплитудой. С точки зрения решения задачи аугментации изображений это влияние будет незначительным и, скорее всего, иметь позитивное значение, так как дополнительное зашумление часто используется в стандартных схемах аугментации данных.

В ходе дальнейших исследований на основе рассмотренной математической модели были реализованы алгоритмы внесения ДП, имеющие воздействие в виде биполярных импульсов квазидетерминированной формы.

Для исключения краевых эффектов и ограничения ДП использовалась трапециевидная масочная функция, задаваемая в виде матрицы размера $N \times M$ и однозначно исключающая краевые эффекты:

$$h(i,k) = \tilde{h}_1(i)\tilde{h}_2(k),\ \tilde{h}_1(i) = \frac{h_1(i)}{\max[h_1(i)]},$$

$$\tilde{h}_2(k) = \frac{h_2(k)}{\max[h_2(k)]},\ i = \overline{1,N},\ k = \overline{1,M},$$

$$h_1(i) = \begin{cases} i-1,\ 1 \le i \le i_x+1, \\ i_x,\ i_x+1 < i \le N-i_x-1, \\ N-i-1,\ N-i_x \le i \le N. \end{cases} \quad h_2(k) = \begin{cases} k-1,\ 1 \le k \le k_y+1, \\ k_y,\ k_y+1 < k \le M-k_y-1, \\ M-k-1,\ M-k_y \le k \le M. \end{cases} \tag{6}$$

где $i_x = k_y = 2A_m - 1$ – целочисленное (в ед. пикселей) представление диапазона возможных значений деформаций, определяющее расстояние от края области определения деформируемой функции, на котором начинают действовать ограничения. Значение $A_m$ при этом фактически определяет максимальное количество позиций, на которое

может быть переставлен один элемент. Для всех граничных элементов значение маски равно нулю. В целом, анализ формы функции (6) показывает, что ни один из элементов исходного изображения при перестановке не выйдет за границы области определения. Такая масочная функция уже может быть использована в качестве функции деформации, если мы хотим ограничиться кусочно-линейным характером внесенных искажений типа сдвига и растяжения.

Для иллюстрации на рисунке 1 представлен граф переходов, описывающий перестановку элементов вдоль одной координаты (аналогичную по сути одномерной перестановке элементов последовательности) в соответствии с функцией ДП, заданной только на основе ограничивающей маски вида $h_1$. Количество переставляемых элементов равно 20. Вид функции, обеспечивающей такие перестановки, представлен штриховой линией. Здесь мы видим, что краевые элементы не меняются, часть элементов выпадает, а часть замещает выпадающие с повторениями.

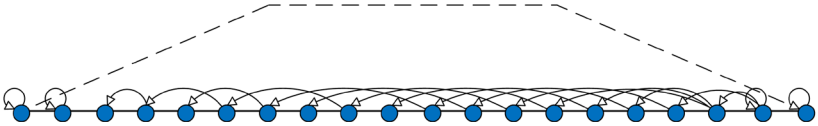


Рис. 1. Граф переходов, описывающий перестановку элементов вдоль одной координаты в соответствии с функцией ДП на основе маски вида $h_1$

На рисунке 2(а) представлен вид получаемой таким образом функции деформации, заданной $h(i,k) = \tilde{h}_1(i),\ i = \overline{1,N}$ одинаковым образом для каждой координаты с индексом $k = \overline{1,M}$ $(h_2(k) = 1,\ i = \overline{1,M})$ при $N = M = 200,\ A_m = 25.$ На рисунке 2(в) представлено эталонное изображение в виде регулярной шахматной структуры $N \times M,\ N = M = 200,$ а на рисунке 2(г) – его преобразование в соответствии с используемой таким образом функцией деформации.

Анализ рисунка 2(а, в, г) показывает, что здесь смещение элементов осуществляется только в одном направлении. В результате такого ДП, примененного к изображению, происходит «дружный» сдвиг элементов, размещенных по вертикали, в левой части изображения, и растяжение элементов, также размещенных по вертикали, в правой части изображения. Такое преобразование, по сути, соответствует схеме перестановки, представленной на рисунке 1.

Как уже отмечалось, при использовании соотношений для $h_1(i),\ i=\overline{1,N}$, $h_2(k),\ k=\overline{1,M}$ из (6) в полной мере в результате получается ограничивающая по обеим координатам маска трапециевидной формы, которая и будет далее использована во всех примерах и анализируемых моделях. Отображение ее формы для $N=M=200,\ A_m=25$ представлено на рисунке 2(б).

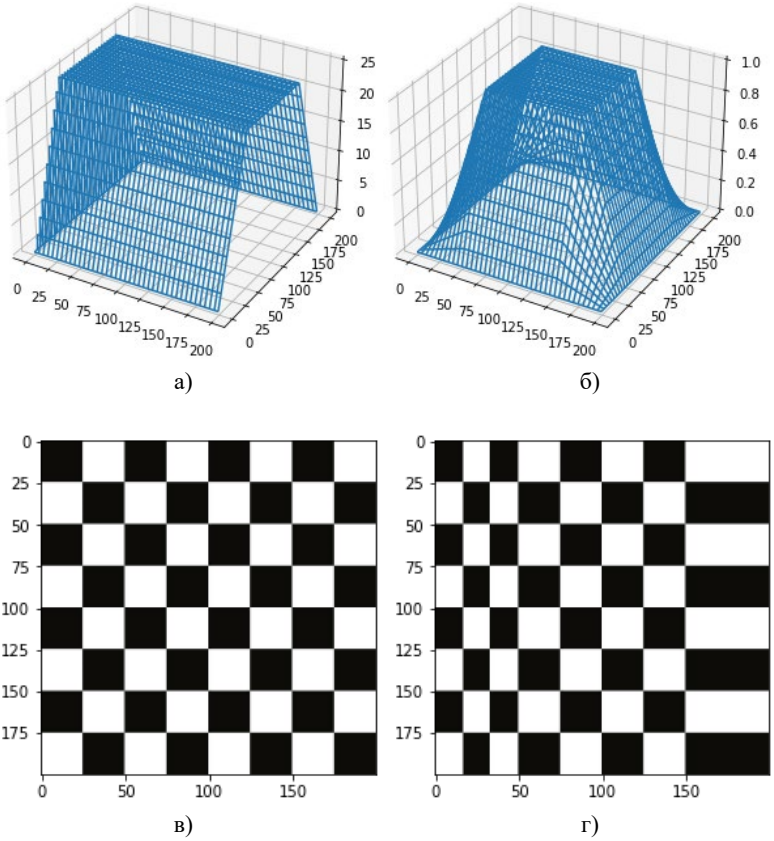



Рис. 2. Вид функции деформации, заданной: а) на основе маски $h_1$ только вдоль одной из осей координат; б) масок $h_1$ и $h_2$ в полной мере; в) эталонное изображение; г) результат его преобразования при помощи функции ДП (а)

Собственно, функция деформирующих искажений может задаваться в целочисленном виде как квазидетерминированная функция со случайными параметрами и наложенной на нее маской (6).

При этом рассматривались два варианта. Первый вариант предполагает использование функции вида:

$$\begin{gathered} v_{1,2}(\mathbf{z}) = S_{a1,2} h(\mathbf{z}) \exp[-\frac{1}{2}(\mathbf{z}-\mathbf{m}_{rnd})^{\mathrm{T}} C_{rnd}^{-1}(\mathbf{z}-\mathbf{m}_{rnd})], \\ S_{a1,2} \in \{1;-1\},\ \mathbf{m}_{rnd} \in \{\mathbf{m}_{T,L}, \mathbf{m}_{T,R}, \mathbf{m}_{B,L}, \mathbf{m}_{B,R}, \mathbf{m}_{C,C}\}, \\ C_{rnd} = \frac{NM}{3}\begin{pmatrix} 1 & \rho_{rnd} \\ \rho_{rnd} & 1 \end{pmatrix},\ -1 < \rho_{rnd} < 1, \end{gathered} \tag{7}$$

где $\mathbf{z} = (z_1, z_2)^T = (i,k)^{\mathrm{T}}$ – вектор дискретных координат точки на изображении; $S_{a1,2}$ – случайная величина, определяющая знак импульса ДИ; $\mathbf{m}_{rnd}$ – случайным образом выбираемый центр размещения максимума функции ДИ по четырем углам сетки Ψ с некоторым отступом от края, а также в ее центре, определяемом вектором $\mathbf{m}_{C,C} = (m_{C,x}, m_{C,y})^{\mathrm{T}}$, $m_{C,x} = \mathrm{round}(N/2)$, $m_{C,y} = \mathrm{round}(M/2)$; $C_{rnd}$ – положительно определенная матрица со случайным равновероятным в указанном диапазоне значений коэффициентом $\rho_{rnd}$ и параметром влияния $d_s$, определяющем фактически дисперсию импульса ДИ. Фактически данная функция определяет форму импульса ДИ, знак которой задает различные направления выполняемой деформации.

Второй вариант предполагает использование функции вида:

$$\begin{gathered} v_{1,2}(\mathbf{z}) = S_{a1,2} h(\mathbf{z}) \exp\left[-\frac{1}{2}(\mathbf{z}-\mathbf{m}_{rnd})^{\mathrm{T}} C_{rnd}^{-1}(\mathbf{z}-\mathbf{m}_{rnd})\right] a(\mathbf{z}), \\ a(\mathbf{z}) = (z_1 - m_{rnd,x}+1)\,(z_2 - m_{rnd,y}+1). \end{gathered} \tag{8}$$

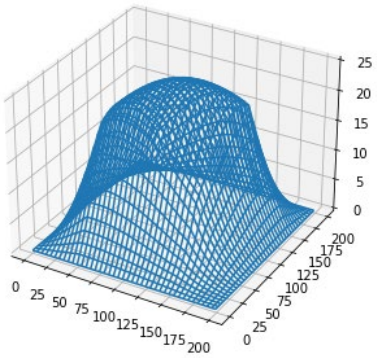
В отличие от предыдущего варианта здесь импульс ДИ обеспечивает максимальную по амплитуде деформацию в центрах квадрантов, размещающихся относительно центральной точки изображения.

После наложения маски к обеим функциям применялась нормализация с приведением максимального значения к величине амплитуды $A_m$:

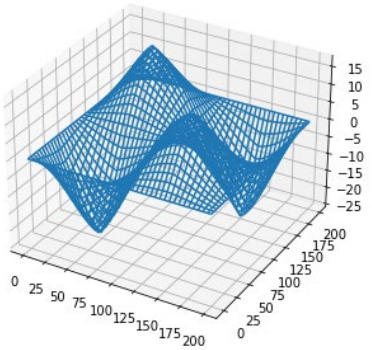
$$\hat{v}_{1,2}(\mathbf{z}) = A_m v_{1,2}(\mathbf{z}) \Big/ \max_{\Psi} v_{1,2}(\mathbf{z}) \cdot$$

На рисунке 3(а) представлен вид функции ДИ (7) с наложенной на нее стандартной маской (6) (рисунок 2(б)) и с размещением в центре $\mathbf{m}_{rnd} = \mathbf{m}_{C,C}$, а на рисунке 3в – соответствующее искаженное изображение изначально регулярной шахматной структуры (рисунок 2(в)). На рисунке 3(б) представлен вид функции ДИ (8) с наложенной маской (6) и с размещением в центре $\mathbf{m}_{rnd} = \mathbf{m}_{C,C}$, а на рисунке 3(г) – соответствующее искаженное изображение изначально регулярной шахматной структуры. Анализ представленных изображений показывает, что предложенные алгоритмы реализуют достаточно плавное их искажение. При этом все структурные элементы – сегменты исходного изображения сохраняются даже при весьма значительной амплитуде вносимых деформаций. Также и сохраняются границы между ними.

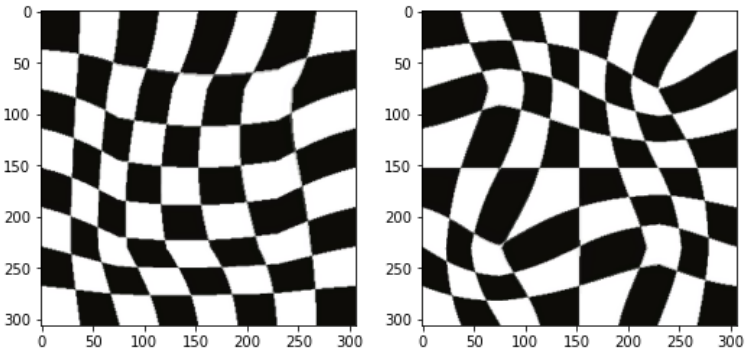
Следует отметить, что для повышения быстродействия процесса генерации ДИ при необходимости можно ограничиться использованием общей функции деформации по обеим координатам $v_1(\mathbf{z}) = v_2(\mathbf{z}) = v(\mathbf{z})$.




а)




б)




в) г)

Рис. 3. Примеры задания функций деформации и соответствующим образом деформированных эталонных изображений

Как уже упоминалось выше, одной из целей данной работы было создание относительно быстрого алгоритма ДП, ориентированного на снижение затрат времени при ИРД в задачах классификации изображений. Основная идея предлагаемого алгоритма состоит в использовании плавных деформаций импульсного характера с реализацией оператора перестановки элементов исходного изображения в элементы трансформируемого изображения в рамках наложенной ограничивающей маски. Для оценки получаемого выигрыша по времени выполнения ДП было проведено сравнительное исследование этого алгоритма с алгоритмом ElasticTransform. При этом рассмотрено две известные программные реализации последнего, размещенная в виде исходного кода в открытом доступе на ресурсе Kaggle [32] и размещенная в составе библиотеки Albumentations [17]. Предлагаемый алгоритм также был реализован в двух вариантах. Первый вариант основан на генерации различных функций деформации $v_1(\mathbf{z}), v_2(\mathbf{z})$ для каждой пространственной координаты в соответствии с (8). Второй вариант реализует использование общей функции деформации $v_1(\mathbf{z}) = v_2(\mathbf{z}) = v(\mathbf{z})$ для каждой координаты. Следуют отметить, что визуально получаемые от применения этих вариантов результаты отличаются не существенно.

Сравнение проводилось для разных размеров изображений и разных значений гиперпараметров алгоритмов, отвечающих за степень плавности и амплитуду деформации. Для предлагаемого алгоритма таким параметром является амплитуда ДП, определяемая относительно размера изображения $N = N_x = N_y$ как $A_m = k_s \times N$, где

$k_s$ – варьируемый в экспериментах коэффициент $0 < k_s < 1$. Для алгоритма ElasticTransform варьируемым параметром являлась величина $\sigma = k_s \times N$, которая определяет размеры окна сглаживающих фильтров, используемых в этом алгоритме. При этом чем больше величина $k_s$, тем получаемые деформации являются более плавными и, напротив, при малых значениях этого коэффициента вносимые деформации становятся более изрезанными вплоть до потери структурного сходства элементов. В ходе вычислительного эксперимента замеры времени производились для эталонного изображения, вид которого представлен на рисунке 2, которое 1000 раз подвергалось ДП на стационарном ПК. Полученные результаты представлены в таблице 1 и носят, естественно, относительный характер, позволяя, тем не менее, проводить сопоставительный анализ алгоритмов.

Таблица 1. Сравнение временных характеристик алгоритмов ДП при выполнении 1000 запусков алгоритма

| Используемый алгоритм | Время (с), $N = 200$, $A_m = 0{,}125N$, $\sigma = 0{,}05N$ | Время (с), $N = 400$, $A_m = 0{,}125N$, $\sigma = 0{,}05N$ | Время (с), $N = 600$, $A_m = 0{,}125N$, $\sigma = 0{,}05N$ | Время (с), $N = 400$, $A_m = 0{,}0625N$, $\sigma = 0{,}025N$ | Время (с), $N = 400$, $A_m = 0{,}25N$, $\sigma = 0{,}1N$ |
|---|---|---|---|---|---|
| ElasticTransform (Kaggle) | 37,47 | 245,52 | 753,30 | 164,54 | 415,56 |
| ElasticTransform (Albumentations) | 17,27 | 113,29 | 364,87 | 71,84 | 206,61 |
| Предлагаемый алгоритм (вариант 1) | 6,00 | 34,46 | 83,26 | 34,33 | 34,61 |
| Предлагаемый алгоритм (вариант 2) | 3,58 | 19,99 | 48,41 | 19,93 | 19,76 |

Анализ показывает, что предлагаемый алгоритм позволяет получить повышение быстродействия от 2,9 до 4,4 раз для первого варианта реализации и от 4,8 до 7,6 для второго варианта в зависимости от размера изображения при фиксированных значениях гиперпараметров (первые три колонки). Также следует отметить практическую инвариантность времени его выполнения относительно амплитуды деформации, тогда как время выполнения алгоритма ElasticTransform существенно зависит от гиперпараметра $\sigma$,

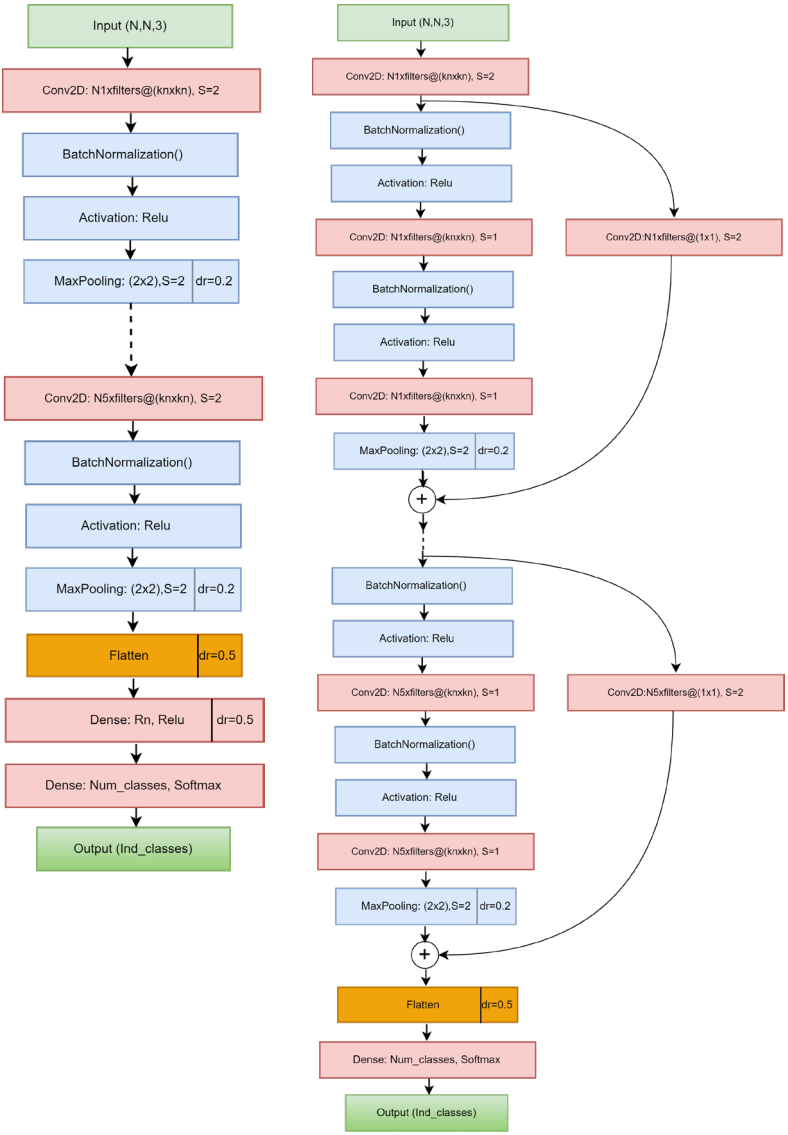
отвечающего за уровень вносимых деформаций (колонки 2, 4, 5 таблицы 1).

**3. Методика и результаты обучения нейронных сетей с использованием предлагаемого метода аугментации.** В ходе проведенных исследований решалась задача оценки влияния предлагаемого способа аугментации изображений в задаче классификации на основе обучения с нуля относительно легковесных глубоких нейронных сетей стандартной архитектуры. Таких сетей было рассмотрено две: сверточная сеть стандартной архитектуры (convolutional neural network, CNN) и сеть с остаточными связями (deep residual network, DRN). В зависимости от объема обучающих данных количественные характеристики слоев и используемые при обучении гиперпараметры подбирались различным образом.

Архитектура первой сети в виде набора повторяющихся фрагментов показана на рисунке 4(а). Здесь использованы следующие обозначения: N – размер входного изображения; kn – размер ядра свертки; filters – количество фильтров (каналов) первого сверточного слоя, определяющее количество реализуемых карт признаков; N1,...,N5 – умножающий коэффициент для количества фильтров на последующих слоях; S – величина сдвига при проведении свертки или пуллинга; dr – величина коэффициента слоя дропаута, вставленного после выполнения предшествующего слоя, отделенного чертой; Rn – количество нейронов в первом плотном слое; Num_classes – количество классов изображений; Relu, Softmax – стандартные обозначения активационных функций, используемых в слоях нейронной сети; Ind_classes – формируемый на выходе индекс класса. Во всех сверточных слоях использовалась стандартная batch-нормализация.

Архитектура второй сети показана на рисунке 4(б). Здесь использованы следующие обозначения: N – размер входного изображения; kn – размер ядра свертки; filters – количество фильтров (каналов) первого сверточного слоя, определяющее количество реализуемых карт признаков; N1,...,N5 – умножающий коэффициент для количества фильтров на последующих слоях; S – величина сдвига при проведении свертки или пуллинга; dr – величина коэффициента слоя дропаута, вставленного после выполнения предшествующего слоя, отделенного чертой; Rn – количество нейронов в первом плотном слое; Num_classes – количество классов изображений; Relu, Softmax – стандартные

обозначения активационных функций, используемых в слоях нейронной сети; Ind_classes – формируемый на выходе индекс класса.




а) б)

Рис. 4. Архитектура используемых нейронных сетей

Методика подготовки данных для обучения нейронных сетей состояла в следующем. Использовались три исходных обучающих выборки изображений, приведенных к общему размеру по обеим осям $w_{size}$:

– малая выборка $B_1$ изображений 3 классов животных, содержащая по 1000 экземпляров изображений в каждом классе (всего 3000);

– средняя по объему выборка $B_2$ изображений 3 классов животных, содержащая по 2000 экземпляров изображений в каждом классе (всего 6000);

– средняя по объему выборка $B_3$ изображений 10 классов животных, содержащая по 2000 экземпляров изображений в каждом классе (всего 20000).

Выборка $B_1$ (как наиболее важная с точки зрения исследования) была сформирована в двух вариантах с различающимися изображениями одних и тех же классов, полученными из разных источников.

Каждая выборка $B_1$, $B_2$, $B_3$ подвергалась преобразованиям, направленным на искусственное размножение обучающих данных. Сначала на основе исходной формировалась выборка увеличенного вдвое объема путем выполнения зеркального отражения каждого изображения. Таким образом, были получены выборки: $B_1^{flip}$ – 6000 изображений; $B_2^{flip}$ – 12000 изображений; $B_3^{flip}$ – 40000 изображений. Затем уже эти выборки подвергались аугментации на основе изложенного выше алгоритма внесения деформирующих искажений, что позволило, в свою очередь, получить выборки $B_1^{flip+aug}$ – 12000 изображений; $B_2^{flip+aug}$ – 24000 изображений; $B_3^{flip+aug}$ – 80000 изображений. Выборки $B_1$, $B_2$ и преобразованные от них выборки разбивались на обучающую и валидационную, исходя из соотношения 80% и 20%. Выборка $B_3$ и преобразованные от нее выборки разбивались на обучающую и валидационную, исходя из соотношения 90% и 10%. Для тестирования всегда использовалась подвыборка, содержащая, соответственно, 20% из $B_1$, $B_2$ и 10% из $B_3$, т.е. исключительно состоящая из оригинальных, не преобразованных каким-либо образом изображений. Поэтому при обучении по исходным, не преобразованным выборкам, $B_2$, $B_3$ тестирующие и валидационные выборки совпали. Для выборок $B_1$, $B_2$ и производных

от них выборок использовались два значения размера изображения $w_{size} = 128$ и $w_{size} = 200$. Для выборки $B_3$ и ее аугментированных вариантов использовалось значение $w_{size} = 128$ ввиду значительного роста времени обучения.

Аугментация ДИ осуществлялась в варианте (7) с параметром $A_m = 16$ для $w_{size} = 128$ и $A_m = 25$ для $w_{size} = 200$.

Для первой сети при обучении по малым выборкам $B_1$, $B_2$ и их производных задавались следующие гиперпараметры. Значение исходного количества filters устанавливалось равным 32. Использовалось пять вычислительных сверточных блоков (рисунок 4а) с коэффициентами умножения числа фильтров N1=1, N2=N3=2, N4=N5=4. Размер ядра фильтров свертки устанавливался равным kn=3 при $w_{size} = 128$ и kn=5 при $w_{size} = 200$. В слоях сетей использовалась L2-регуляризация с коэффициентом $l2_{val}$=0,01, а также регуляризация на основе метода стохастического исключения весовых связей (метод dropout) со значением вероятности исключения dr=0,2. При обучении использовался оптимизатор RMSprop с начальной скоростью $lr_{val}$=0,0004. Процесс обучения занимал 200 эпох и сопровождался постепенным понижением скорости обучения до нижнего предела 0,00001. При обучении по выборке $B_3$ и производным от нее выборкам задавались аналогичные параметры, с тем только отличием, что коэффициенты умножения числа фильтров задавались как N1=1, N2=2, N3=4, N4=8, N5=16. В слоях сетей использовалась L2-регуляризация с коэффициентом $l2_{val}$=0,001, при этом регуляризация на основе метода стохастического исключения весовых связей не проводилась (dr=0,0). Процесс обучения занимал 100 эпох и сопровождался постепенным понижением скорости обучения до нижнего предела 0,00001.

Для второй сети аналогично при обучении по малым выборкам $B_1$, $B_2$ и их производных задавались следующие гиперпараметры. Значение исходного количества filters устанавливалось равным 32. Использовалось пять вычислительных сверточных блоков (рисунок 4(б)) с коэффициентами умножения числа фильтров N1=1, N2=N3=2, N4=N5=4, размер ядра фильтров свертки устанавливался равным kn=3 при $w_{size} = 128$ и при $w_{size} = 200$. В слоях сетей аналогично использовалась L2-регуляризация с коэффициентом $l2_{val}$=0,01, а также регуляризация на основе метода

стохастического исключения весовых связей (метод dropout) со значением вероятности исключения dr=0,2. При обучении использовался оптимизатор RMSprop с начальной скоростью $lr_{val}$=0,0004 . Процесс обучения занимал 200 эпох и сопровождался постепенным понижением скорости обучения до нижнего предела 0,00001.

При обучении по выборке $B_3$ и производным от нее выборкам задавались аналогичные параметры, с тем только отличием, что коэффициенты умножения числа фильтров задавались как N1=1, N2=2, N3=4, N4=8, N5=16. В слоях сетей аналогично использовалась L2-регуляризация с коэффициентом $l2_{val}$=0,001, при этом регуляризация на основе метода стохастического исключения весовых связей не проводилась (dr=0,0). Процесс обучения занимал 100 эпох и сопровождался постепенным понижением скорости обучения до нижнего предела 0,00001.

Как уже отмечалось, в процессе обучения был реализован план ступенчатого изменения скорости обучения в указанных пределах с сохранением весовых коэффициентов сети, которая показала лучшие по отношению к прошлым результаты по валидационной подвыборке. Последняя из сохраненных сетей использовалась для тестирования. Несмотря на очевидные признаки переобучения сетей на завершающей части обучения, определенное улучшение качества классификации происходило практически до последней эпохи.

В таблицах 2 – 4 представлены полученные в ходе двухэтапной аугментации результаты точности классификации (метрика accuracy) для сетей классов CNN и DRN представленной архитектуры.

Таблица 2. Результаты тестирования при обучении по выборке $B_1$ и производным от нее выборкам (3 класса, l2=0,01, dr=0,2, в ячейках со значениями: вверху – первый вариант $B_1$, внизу – второй вариант $B_1$)

| Размер изображения и параметры ядер свертки | Обучение по выборке $B_1$ | | Обучение по выборке $B_1^{flip}$ | | Обучение по выборке $B_1^{flip+aug}$ | |
|---|---|---|---|---|---|---|
| | CNN | DRN | CNN | DRN | CNN | DRN |
| $w_{size}=128,$ kn=3 | 85,57<br>86,50 | 85,83<br>83,66 | 88,67<br>88,83 | 90,50<br>87,00 | **90,68**<br>**89,60** | **91,00**<br>**89,67** |
| $w_{size}=200,$ kn=5 | 88,00<br>87,83 | 88,17<br>00,00 | 90,16<br>89,00 | 90,17<br>87,17 | **92,60**<br>**89,83** | **92,67**<br>**87,83** |

Таблица 3. Результаты тестирования при обучении по выборке $B_2$ и производным от нее выборкам (3 класса, l2=0,01, dr=0,2)

| Размер изображения и параметры ядер свертки | Обучение по выборке $B_2$ | | Обучение по выборке $B_2^{flip}$ | | Обучение по выборке $B_2^{flip+aug}$ | |
|---|---|---|---|---|---|---|
| | CNN | DRN | CNN | DRN | CNN | DRN |
| $w_{size}=128$, kn=3 | 93,75 | 94,17 | 94,42 | 95,50 | **95,92** | **95,58** |
| $w_{size}=200$, kn=5 | 93,58 | 93,33 | 95,08 | 94,83 | **95,83** | **95,50** |

Таблица 4. Результаты тестирования при обучении по выборке $B_3$ и производным от нее выборкам (10 классов, l2=0,001, dr=0,0)

| Размер изображения и параметры ядер свертки | Обучение по выборке $B_3$ | | Обучение по выборке $B_3^{flip}$ | | Обучение по выборке $B_3^{flip+aug}$ | |
|---|---|---|---|---|---|---|
| | CNN | DRN | CNN | DRN | CNN | DRN |
| $w_{size}=128$, kn=3 | 89,25 | 84,85 | 91,28 | 89,90 | **92,05** | **92,60** |

Анализ представленных в таблицах 2 – 4 результатов показывает, что во всех проведенных экспериментах аугментация, выполненная только на основе стандартной процедуры зеркального отражения, дает прирост в точности (частоты правильного распознавания). Этот прирост особенно заметен при размножении малой выборки $B_1$ и, в отдельных случаях, достигает уровня 3,0%. Следующий этап ИРД, выполненный по отношению как к исходным, так и к зеркально отраженным выборкам, также показал наличие прироста точности классификации во всех экспериментах, который аналогично особенно заметен для исходной выборки $B_1$. Здесь его значение достигает значений до 2,5%. Для выборки $B_2$ прирост, естественно, менее значителен, особенно для сети DRN. Для выборки $B_3$, напротив, наиболее значимое увеличение точности 2,7% получено для сети DRN. В целом можно констатировать, что суммарный прирост точности классификации по результатам выполнения обоих этапов аугментации имеет значения, в основном доходящие до 5%. Это, на наш взгляд является вполне удовлетворительным результатом для фиксируемого диапазона значений точности классификации.

Анализ уровня статистической значимости полученных результатов проводился по стандартным соотношениям для оцененных вероятностей и представлен в таблице 5 в виде максимального значения модуля отклонения оцененной точности классификации от возможной истинной. Объем независимой тестирующей подвыборки исходных (не преобразованных) изображений составлял, соответственно, 20% из $B_1$, $B_2$ и 10% из $B_3$.

Таблица 5. Результаты анализа статистической значимости

| Объем тестирующей подвыборки | Обучение и тестирование по выборке $B_1$ (600 примеров) | | Обучение и тестирование по выборке $B_2$ (1200 примеров) | | Обучение и тестирование по выборке $B_3$ (2000 примеров) | |
|---|---|---|---|---|---|---|
| Уровень значимости | 10 % | 5 % | 10 % | 5 % | 10 % | 5 % |
| Максимальное значение модуля отклонения от истинного значения | 1,46 % | 1,74 % | 1,03 % | 1,23 % | 0,80 % | 0,96 % |

Анализ представленных результатов показывает, что полученный для каждого эксперимента суммарный по результатам выполнения двух этапов прирост точности классификации является статистически значимым. Также, с учетом смещения всех полученных оценок в большую сторону, статистически значимым является и прирост, полученный только на основе аугментации с внесением деформирующих искажений.

Следует также отметить, что различия в эффективности применения сетей различной архитектуры в проведенных экспериментах не являются принципиальными с учетом того, что поставленная задача состоит не в нахождении лучшей сети, а в демонстрации возможностей предлагаемого метода аугментации данных.

**4. Заключение.** Таким образом, в работе предложен подход к обучению глубоких нейронных сетей для классификации изображений в условиях малой и средней по объему выборки, основанный на применении техники аугментации обучающих данных с использованием зеркальных отражений и деформирующих преобразований обрабатываемых изображений. Предложена математическая модель и обоснован реализующий ее алгоритм

внесения деформирующих искажений, основанные на выполнении циклической перестановки элементов исходного изображения в соответствие с используемой функцией пространственной деформации. Показано, что его применение при любых уровнях деформации сохраняет структурную основу исходного изображения, что делает его пригодным для проведения аугментации. Показано также, что быстродействие алгоритма существенно выше по сравнению с известными реализациями аналогичного характера. В ходе всех экспериментов с обучением нейронных сетей различной архитектуры установлено наличие прироста точности классификации при применении процедуры стандартной аугментации на основе зеркального отражения исходных изображений обучающей выборки с последующим применением аугментации на основе предложенного алгоритма деформирующего преобразования. Указанный прирост для рассмотренных архитектур нейронных сетей наиболее заметен при обучении по относительно малой выборке (порядка $10^3$ примеров на каждый класс).

Дальнейшие исследования, на наш взгляд, целесообразно сосредоточить на исследовании возможностей предлагаемого подхода для повышения качества обучения в задаче семантической сегментации, а также его сравнении с известными алгоритмами, которые ранее использовались при решении данной задачи.

## Литература


1. Chawla N.V., Lazarevic A., Hall L.O., Bowyer K.W. SMOTEBoost: Improving Prediction of the Minority Class in Boosting // 7th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD). 2003. pp. 107–119. DOI: 10.1007/978-3-540-39804-2_12.
2. Minaee S., Luo P., Lin Zh., Bowyer K. Going deeper into face detection: A survey // arXiv preprint. 2021. DOI: 10.48550/arXiv.2103.14983.
3. Ciresan D.C., Meier U., Gambardella L.M., Schmidhuber J. Deep, Big, Simple Neural Nets For Handwritten Digit Recognition // Neural computation. 2010. vol. 22. no. 12. pp. 3207–3220. DOI: 10.1162/NECO_a_00052.
4. Tao X., Zhang D., Ma W., Liu X., Xu D. Automatic Metallic Surface Defect Detection and Recognition with Convolutional Neural Networks // Applied Sciences. 2018. vol. 8. no. 9. pp. 1575–1590. DOI: 10.3390/app8091575.
5. Shorten C., Khoshgoftaar T.M. Survey on Image Data Augmentation for Deep Learning // Journal of Big Data. 2019. vol. 6. no. 1. pp. 1–48. DOI: 10.1186/s40537-019-0197-0.
6. Wang W., Xie E., Li X., Fan, D. P., Song, K., Liang, D., Lu T., Luo P., Shao, L. Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction Without Convolutions // Proceedings of the IEEE/CVF international conference on computer vision. 2021. pp. 568–578. DOI: 10.1109/ICCV48922.2021.00061.
7. Качалин С.В. Повышение устойчивости обучения больших нейронных сетей дополнением малых обучающих выборок примеров-родителей, синтезированными биометрическими примерами-потомками // Труды научно-

технической конференции кластера пензенских предприятий, обеспечивающих безопасность информационных технологий. 2014. Т. 9. С. 32–35.

8. Акимов А.В., Сирота А.А. Модели и алгоритмы искусственного размножения данных для обучения алгоритмов распознавания лиц методом Виолы–Джонса // Компьютерная оптика. 2016. Т. 40. № 6. С. 911–918. DOI: 10.18287/2412-6179-2016-40-6-911-918.
9. Небаба С.Г., Захарова А.А. Алгоритм построения деформируемых 3D моделей лица и обоснование его применимости в системах распознавания личности. Труды СПИИРАН. 2017. Т. 52. С. 157–179. DOI: 10.15622/sp.52.8.
10. Сирота А.А., Донских А.О., Акимов А.В., Минаков Д.А. Смешанные ядерные оценки многомерных распределений и их применение в задачах машинного обучения для классификации биологических объектов на основе спектральных измерений // Компьютерная оптика. 2019. Т. 43. № 4. С. 677–691. DOI: 10.18287/2412-6179-2019-43-4-677-691.
11. Дагаева М.В., Сулейманов М.А., Катасева Д.В., Катасёв, А.С., Кирпичников А.П. Технология построения отказоустойчивых нейросетевых моделей распознавания рукописных символов в системах биометрической аутентификации // Вестник Технологического университета. 2018. Т. 21. № 2. С. 133–138.
12. Емельянов С.О., Иванова А.А., Швец Е.А., Николаев Д.П. Методы аугментации обучающих выборок в задачах классификации изображений // Сенсорные системы. 2018. Т. 32. № 3. С. 236–245. DOI: 10.1134/S0235009218030058.
13. Рюмина Е.В., Рюмин Д.А., Маркитантов М.В., Карпов А.А. Метод генерации обучающих данных для компьютерной системы обнаружения защитных масок на лицах людей // Компьютерная оптика. 2022. Т. 46. № 4. С. 603–611. DOI: 10.18287/2412-6179-CO-1039.
14. Камалова Ю.Б., Андриянов Н.А. Распознавание микроскопических изображений пыльцевых зерен с помощью сверточной нейронной сети VGG-16 // Вестник Южно-Уральского государственного университета. Серия: Компьютерные технологии, управление, радиоэлектроника. 2022. Т. 22. № 3. С. 39–46. DOI: 10.14529/ctcr220304.
15. Ковун В.А., Каширина И.Л. Использование нейронной сети W-Net в металлографическом анализе образцов стали // Вестник ВГУ (Системный анализ и информационные технологии). 2022. № 1. С. 101–110. DOI: 10.17308/sait.2022.1/9205.
16. Simard P.Y., Steinkraus D., Platt J.C. Best practices for convolutional neural networks applied to visual document analysis // In Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR '03). 2003. vol. 2. pp. 1–6.
17. Buslaev A., Iglovikov V.I., Khvedchenya E., Parinov A., Druzhinin M., Kalinin A.A. Albumentations: Fast and flexible image augmentations. Information. 2020. vol. 11. no. 2. pp. 1–20. DOI: 10.3390/info11020125.
18. Hasan S.M.K., Linte C.A. U-NetPlus: A modified encoder-decoder U-Net architecture for semantic and instance segmentation of surgical instruments from laparoscopic images // 41st Annual International Conference of the Ieee Engineering in Medicine and Biology Society (EMBC). 2019. pp. 7205–7211.
19. Keysers D., Deselaers T., Gollan C., Ney H. Deformation models for image recognition // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2007. vol. 29(8). pp. 1422–1435. DOI: 10.1109/TPAMI.2007.1153.
20. Felzenswalb P., McAllester D., Ramanan D. A discriminatively trained, multiscale, deformable part model // IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2008. pp. 1–8. DOI: 10.1109/CVPR.2008.4587597.

21. Wiskott L., Fellous J.-M., Kruger N., von der Malsburg C. Face Recognition by Elastic Bunch Graph Matching // Proceedings of International Conference on Image Processing. 1997. vol. 1. pp. 129–132. DOI: 10.1109/ICIP.1997.647401.
22. Li X., Li W., Ren D., Zhang H., Wang M., Zuo W. Enhanced Blind Face Restoration with Multi-Exemplar Images and Adaptive Spatial Feature Fusion // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020. pp. 2706–2715. DOI: 10.1109/CVPR42600.2020.00278.
23. Deng Y., Yang J., Tong X. Deformed Implicit Field: Modeling 3D Shapes With Learned Dense Correspondence // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021. pp. 10286–10296. DOI: 10.48550/arXiv.2011.13650.
24. Venkatesh S., Ramachandra R., Raja K., Busch Ch. Face Morphing Attack Generation and Detection: A Comprehensive Survey // IEEE Transactions on Technology and Society. 2021. vol. 2. no. 3. pp. 128–145. DOI: 10.1109/TTS.2021.3066254.
25. Scherhag U., Rathgeb C., Merkle J. Busch C. Deep Face Representations for Differential Morphing Attack Detection // IEEE Transactions on Information Forensics and Security. 2020. vol. 15. pp. 3625–3639. DOI: 10.1109/TIFS.2020.2994750.
26. Ling H., Kreis K., Li D., Kim S.W., Torralba A., Fidler S. EditGAN: High-Precision Semantic Image Editing // Advances in Neural Information Processing Systems. 2021. vol. 34. pp. 16331–16345. DOI: 10.48550/arXiv.2111.03186.
27. Wang S.Y., Bau D., Zhu J.Y. Rewriting Geometric Rules of a GAN // ACM Transactions on Graphics (TOG). 2022. vol. 41. no. 4. pp. 1–16. DOI: 10.48550/arXiv.2207.14288.
28. Акимов А.В., Дрюченко М.А., Сирота А.А. Модели и алгоритмы внесения деформирующих искажений на изображениях с использованием радиально-базисных функций // Вестник ВГУ (Системный анализ и информационные технологии). 2014. № 1. С. 130–137.
29. Захарова А.А., Небаба С.Г., Завьялов Д.А. Алгоритмическое и программное обеспечение для повышения эффективности обработки многомерных гетерогенных данных // Программирование. 2019. № 4. С. 64–70. DOI: 10.1134/S0132347419040101.
30. Buckley M.J. Fast computation of a discretized thin-plate smoothing spline for image data // Biometrika. 1994. vol. 81. no. 2. pp. 247–258. DOI: 10.2307/2336955.
31. Sastry S.P., Zala V., Kirby R.M. Thin-plate-spline curvilinear meshing on a calculus-of-variations framework // Procedia Engineering. 2015. vol. 124. pp. 135–147. DOI: 10.1016/j.proeng.2015.10.128.
32. Elastic Transform for Data Augmentation. URL: https://www.kaggle.com/code/bguberfain/elastic-transform-for-data-augmentation (accessed: 30.10.2023).

**Сирота Александр Анатольевич** — д-р техн. наук, профессор, заведующий кафедрой, кафедра технологий обработки и защиты информации факультета компьютерных наук, ФГБОУ ВО «Воронежский государственный университет». Область научных интересов: синтез и анализ систем сбора и обработки информации, методы и технологии компьютерного моделирования информационных процессов и систем, машинное обучение, компьютерная обработка изображений, нейронные сети и нейросетевые технологии в системах принятия решений. Число научных публикаций — 303. sir@cs.vsu.ru; Университетская площадь, 1, 394018, Воронеж, Россия; р.т.: +7(903)030-6943.

**Акимов Алексей Викторович** — канд. физ.-мат. наук, старший преподаватель, кафедра технологий обработки и защиты информации факультета компьютерных наук, ФГБОУ ВО «Воронежский государственный университет». Область научных интересов: распознавание изображений, машинное обучение. Число научных публикаций — 28. akimov@vsu.ru; Университетская площадь, 1, 394018, Воронеж, Россия; р.т.: +7(903)030-6943.

**Отырба Ростислав Русланович** — аспирант, кафедра технологий обработки и защиты информации факультета компьютерных наук, ФГБОУ ВО «Воронежский государственный университет». Область научных интересов: машинное обучение, глубокое обучение, компьютерное зрение, обработка естественного языка. Число научных публикаций — 6. otyrba@cs.vsu.ru; Университетская площадь, 1, 394018, Воронеж, Россия; р.т.: +7(903)854-9545.

**DOI** 10.15622/ia.23.2.4

A. SIROTA, A. AKIMOV, R. OTYRBA

# IMAGE WARPING AND ITS APPLICATION FOR DATA AUGMENTATION WHEN TRAINING DEEP NEURAL NETWORKS

*Sirota A., Akimov A., Otyrba R.* **Image Warping and Its Application for Data Augmentation when Training Deep Neural Networks.**

**Abstract.** The paper focuses on the improvement of the quality of learning for deep neural networks for a small data set in a classification task. One of the possible approaches to improve the quality of learning is researched which is based on the use of data augmentation (artificial reproduction of the data set) by image warping. The presented mathematical model and fast algorithm for warping make it possible to transform the original image while preserving its structural basis. The proposed algorithm is used to augment image data sets containing a small number of training samples. The augmentation consists of two stages including horizontal mirroring and warping of each of the samples. The effectiveness of such augmentation is tested through the training of neural networks of various types: convolutional neural networks (CNN) of a standard architecture and deep residual networks (DRN). A specific feature of the implemented approach for the solution of the problem under consideration consists in the refusal to use pre-trained neural networks with a large number of layers as well as further transfer learning, since their application incurs costs in terms of the computational resources. The paper shows that the efficiency of image classification when implementing the proposed method of augmenting training data on small and medium-sized data sets increases to statistically significant values of the metric used.

**Keywords:** deep neural networks, training data augmentation, image warping, efficiency of deep neural networks.

## References

1. Chawla N.V., Lazarevic A., Hall L.O., Bowyer K.W. SMOTEBoost: Improving Prediction of the Minority Class in Boosting. 7th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD). 2003. pp. 107–119. DOI: 10.1007/978-3-540-39804-2_12.
2. Minaee S., Luo P., Lin Zh., Bowyer K. Going deeper into face detection: A survey. arXiv preprint. 2021. DOI: 10.48550/arXiv.2103.14983.
3. Ciresan D.C., Meier U., Gambardella L.M., Schmidhuber J. Deep, Big, Simple Neural Nets For Handwritten Digit Recognition. Neural computation. 2010. vol. 22. no. 12. pp. 3207–3220. DOI: 10.1162/NECO_a_00052.
4. Tao X., Zhang D., Ma W., Liu X., Xu D. Automatic Metallic Surface Defect Detection and Recognition with Convolutional Neural Networks. Applied Sciences. 2018. vol. 8. no. 9. pp. 1575–1590. DOI: 10.3390/app8091575.
5. Shorten C., Khoshgoftaar T.M. Survey on Image Data Augmentation for Deep Learning. Journal of Big Data. 2019. vol. 6. no. 1. pp. 1–48. DOI: 10.1186/s40537-019-0197-0.
6. Wang W., Xie E., Li X., Fan D.P., Song K., Liang D., Lu T., Luo P., Shao L. Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction Without Convolutions. Proceedings of the IEEE/CVF international conference on computer vision. 2021. pp. 568–578. DOI: 10.1109/ICCV48922.2021.00061.

7. Kachalin S.V. [Improving the stability of large neural net-works by extending small training sets of parent samples with synthesized biometric descendant samples]. Trudy nauchno-tehnicheskoj konferencii klastera penzenskih predprijatij, obespechivajushhih bezopasnost' informacionnyh tehnologij – Proceedings of the Scientific and Technical Conference of The cluster of Penza Enterprises Providing Security of Information Technologies. 2014. vol. 9. pp. 32–35. (In Russ.).
8. Akimov, A.V., Sirota A.A. [Synthetic data generation models and algorithms for training image recognition algorithms using the Viola-Jones framework]. Komp'juternaja optika – Computer Optics. 2016. vol. 40. no. 6. pp. 911–918. DOI: 10.18287/2412-6179-2016-40-6-911-918. (In Russ.).
9. Nebaba S.G., Zakharova A.A. [An Algorithm for Building Deformable 3d Human Face Models and Justification of its Applicability for Recognition Systems]. Trudy SPIIRAN – SPIIRAS Proceedings. 2017. vol. 52. pp. 157–179. DOI: 10.15622/sp.52.8. (In Russ.).
10. Sirota A.A., Donskikh A.O., Akimov A.V., Minakov D.A. [Multivariate mixed kernel density estimators and their application in machine learning for classification of biological objects based on spectral measurements]. Komp'juternaja optika – Computer Optics. 2019. vol. 43. no. 4. pp. 677–691. DOI: 10.18287/2412-6179-2019-43-4-677-691. (In Russ.).
11. Dagaeva M.V., Sulejmanov M.A., Kataseva D.V., Katasyov, A.S., Kirpichnikov A.P. [Technology for building fault-tolerant neural network models for recognizing handwritten characters in biometric authentication systems]. Vestnik Tehnologicheskogo universiteta – Bulletin of the Technological University. 2018. vol. 21. no. 2. pp. 133–138. (In Russ.).
12. Emel'janov S.O., Ivanova A.A., Shvec E.A., Nikolaev D.P. [Methods of augmentation of training samples in image classification problems]. Sensornye sistemy – Sensory Systems. 2018. vol. 32. no. 3. pp. 236–245. DOI: 10.1134/S0235009218030058. (In Russ.).
13. Rjumina E.V., Rjumin D.A., Markitantov M.V., Karpov A.A. [A method for generating training data for a computer system for detecting protective masks on people's faces]. Komp'juternaja optika – Computer Optics 2022. vol. 46. no. 4. pp. 603–611. DOI: 10.18287/2412-6179-CO-1039. (In Russ.).
14. Kamalova Ju.B., Andrijanov N.A. [Recognition of microscopic images of pollen grains using the convolutional neural network VGG-16]. Vestnik Juzhno-Ural'skogo gosudarstvennogo universiteta. Serija: Komp'juternye tehnologii, upravlenie, radiojelektronika – Bulletin of South Ural State University, Series: Computer Technologies, Automatic Control and Radioelectronics. 2022. vol. 22. no. 3. pp. 39–46. DOI: 10.14529/ctcr220304. (In Russ.).
15. Kovun V.A., Kashirina I.L. [Using the W-Net neural network in metallographic analysis of steel samples]. Vestnik VGU (Sistemnyj analiz i informacionnye tehnologii) – Vestnik VSU (System Analysis and Information Technology). 2022. no. 1. pp. 101–110. DOI: 10.17308/sait.2022.1/9205. (In Russ.).
16. Simard P.Y., Steinkraus D., Platt J.C. Best practices for convolutional neural networks applied to visual document analysis. In Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR '03). 2003. vol. 2. pp. 1–6.
17. Buslaev A., Iglovikov V.I., Khvedchenya E., Parinov A., Druzhinin M., Kalinin A.A. Albumentations: Fast and flexible image augmentations. Information. 2020. vol. 11. no. 2. pp. 1–20. DOI: 10.3390/info11020125.
18. Hasan S.M.K., Linte C.A. U-NetPlus: A modified encoder-decoder U-Net architecture for semantic and instance segmentation of surgical instruments from

laparoscopic images. 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). 2019. pp. 7205–7211.
19. Keysers D., Deselaers T., Gollan C., Ney H. Deformation models for image recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2007. vol. 29(8). pp. 1422–1435. DOI: 10.1109/TPAMI.2007.1153.
20. Felzenswalb P., McAllester D., Ramanan D. A discriminatively trained, multiscale, deformable part model. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2008. pp. 1–8. DOI: 10.1109/CVPR.2008.4587597.
21. Wiskott L., Fellous J.-M., Kruger N., von der Malsburg C. Face Recognition by Elastic Bunch Graph Matching. Proceedings of International Conference on Image Processing. 1997. vol. 1. pp. 129–132. DOI: 10.1109/ICIP.1997.647401.
22. Li X., Li W., Ren D., Zhang H., Wang M., Zuo W. Enhanced Blind Face Restoration with Multi-Exemplar Images and Adaptive Spatial Feature Fusion. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020. pp. 2706–2715. DOI: 10.1109/CVPR42600.2020.00278.
23. Deng Y., Yang J., Tong X. Deformed Implicit Field: Modeling 3D Shapes With Learned Dense Correspondence. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021. pp. 10286–10296. DOI: 10.48550/arXiv.2011.13650.
24. Venkatesh S., Ramachandra R., Raja K., Busch Ch. Face Morphing Attack Generation and Detection: A Comprehensive Survey. IEEE Transactions on Technology and Society. 2021. vol. 2. no. 3. pp. 128–145. DOI: 10.1109/TTS.2021.3066254.
25. Scherhag U., Rathgeb C., Merkle J. Busch C. Deep Face Representations for Differential Morphing Attack Detection. IEEE Transactions on Information Forensics and Security. 2020. vol. 15. pp. 3625–3639. DOI: 10.1109/TIFS.2020.2994750.
26. Ling H., Kreis K., Li D., Kim S.W., Torralba A., Fidler S. EditGAN: High-Precision Semantic Image Editing. Advances in Neural Information Processing Systems. 2021. vol. 34. pp. 16331–16345. DOI: 10.48550/arXiv.2111.03186.
27. Wang S.Y., Bau D., Zhu J.Y. Rewriting Geometric Rules of a GAN. ACM Transactions on Graphics (TOG). 2022. vol. 41. no. 4. pp. 1–16. DOI: 10.48550/arXiv.2207.14288.
28. Akimov AV, Dryuchenko MA, Sirota AA. [Models and algorithms for making distorting distortion in images using radial basis functions]. Vestnik VGU (Sistemnyj analiz i informacionnye tehnologii) – Vestnik VSU (System Analysis and Information Technology). 2014. vol. 1. pp. 130–137. (In Russ.).
29. Zaharova A.A., Nebaba S.G., Zav'jalov D.A. [Algorithmic and software to improve the efficiency of processing multidimensional heterogeneous data]. Programmirovanie – Programming. 2019. no. 4. pp. 64–70. DOI: 10.1134/S0132347419040101. (In Russ.).
30. Buckley M.J. Fast computation of a discretized thin-plate smoothing spline for image data. Biometrika. 1994. vol. 81. no. 2. pp. 247–258. DOI: 10.2307/2336955.
31. Sastry S.P., Zala V., Kirby R.M. Thin-plate-spline curvilinear meshing on a calculus-of-variations framework. Procedia Engineering. 2015. vol. 124. pp. 135–147. DOI: 10.1016/j.proeng.2015.10.128.
32. Elastic Transform for Data Augmentation. Available at: https://www.kaggle.com/code/bguberfain/elastic-transform-for-data-augmentation (accessed: 30.10.2023).

**Sirota Alexander** — Ph.D., Dr.Sci., Professor, Head of the department, Department of information security and processing technologies, faculty of computer sciences, Voronezh State University. Research interests: synthesis and analysis of systems for information

collection and processing, methods and technologies for computer modeling of information processes and systems, machine learning, computer image processing, neural networks and neural network technologies in decision-making systems. The number of publications — 303. sir@cs.vsu.ru; 1, Universitetskaya Sq., 394018, Voronezh, Russia; office phone: +7(903)030-6943.

**Akimov Aleksei** — Ph.D., Senior lecturer, Department of information security and processing technologies, faculty of computer sciences, Voronezh State University. Research interests: image recognition, machine learning. The number of publications — 28. akimov@vsu.ru; 1, Universitetskaya Sq., 394018, Voronezh, Russia; office phone: +7(903)030-6943.

**Otyrba Rostislav** — Postgraduate student, Department of information security and processing technologies, faculty of computer sciences, Voronezh State University. Research interests: machine learning, deep learning, computer vision, natural language processing. The number of publications — 6. otyrba@cs.vsu.ru; 1, Universitetskaya Sq., 394018, Voronezh, Russia; office phone: +7(903)854-9545.

**DOI** 10.15622/ia.23.2.5

N. SUJATA GUPTA, K. RAMYA, R. KARNATI

# A REVIEW WORK: HUMAN ACTION RECOGNITION IN VIDEO SURVEILLANCE USING DEEP LEARNING TECHNIQUES

*Sujata Gupta N., Ramya K., Karnati R.* **A Review Work: Human Action Recognition in Video Surveillance Using Deep Learning Techniques.**

**Abstract.** Despite being extensively used in numerous uses, precise and effective human activity identification continues to be an interesting research issue in the area of vision for computers. Currently, a lot of investigation is being done on themes like pedestrian activity recognition and ways to recognize people's movements employing depth data, 3D skeletal data, still picture data, or strategies that utilize spatiotemporal interest points. This study aims to investigate and evaluate DL approaches for detecting human activity in video. The focus has been on multiple structures for detecting human activities that use DL as their primary strategy. Based on the application, including identifying faces, emotion identification, action identification, and anomaly identification, the human occurrence forecasts are divided into four different subcategories. The literature has been carried several research based on these recognitions for predicting human behavior and activity for video surveillance applications. The state of the art of four different applications' DL techniques is contrasted. This paper also presents the application areas, scientific issues, and potential goals in the field of DL-based human behavior and activity recognition/detection.

**Keywords:** face recognition, emotion recognition, action recognition, anomaly recognition, DL, human behavior and activity recognition/detection.

**1. Introduction.** Numerous actual environments have applications for human behavior identification, such as intelligent video surveillance and purchasing behavior evaluation [1]. There are many uses for surveillance footage, particularly in indoor, outdoor, and public spaces. Safety includes surveillance as a crucial component. For the sake of security and protection, surveillance cameras are now a must [2]. Among the key goals of the Indian government's growth initiative, Digital India is e-surveillance. It still includes surveillance footage in some form. Efficient surveillance, a need for less labor, cost-effective auditing capabilities, adopting of recent safety trends, etc. are all benefits of surveillance footage [3]. Until now, monitoring was done manually. We have to manage enormous amounts of video footage that can easily wear individuals out. Furthermore, omissions brought on by manual intervention will significantly reduce the structure's efficacy [4]. Video surveillance automation has provided a solution for this. Nowadays, it is difficult to manually watch every incident captured on a CCTV (Closed Circuit Television) camera. Even if the incident occurred previously, manually looking for it in the video footage is a laborious procedure [5].

Among the oldest and most active areas of computer vision and pattern identification study is monitoring footage. In earlier times, operator humans who watched dozens of displays at once were the mainstay of video-based

monitoring systems [6]. Individuals are shown to be extremely inconsistent in identifying the so-called "unusual events" while evaluating either online video clips or archival data because of the amount of data and relatively lengthy monitoring recordings [7]. The key difficulty is creating an intelligent, autonomous, video-based monitoring system that doesn't need human involvement. An emerging area in the field of automated video surveillance structures is the analysis of anomalous occurrences from video [8].

In recent times, DL-based video surveillance systems (VSSs) have produced a range of impressive outcomes when used for diverse purposes, including crowd counting (CC) [9], abnormal event detection (AED) [10], object detection (OD) [11], human action recognition (HAR) [12], etc. Deep networks mimic human vision by modeling high-level abstractions via several levels of non-linear transformations. DL algorithms must be trained on a large quantity of data to do this [13]. These techniques, meanwhile, have drawbacks to numerous other uses and only perform effectively for specific applications when getting the data is simple [14].

The most important details are that there are insufficient funds and data scales to train DL algorithms from scratch, it is costly and takes time to gather massive databases, the majority of DL algorithms rely on supervised learning, and enlisting the help of human experts to label training datasets is a significant expense and effort [15].

Many other strategies are being put forth; however, the initial research heavily depends on trajectory-based methods. These methods use visual tracking to make an effort to predict the target's trajectories, while a model is acquired to explain typical activities [16]. The activity associated with trajectory deviations from the learned model is therefore considered an anomaly. However, because of their great temporal complexity and the occlusion problem brought on by objects moving, these approaches are unsuitable for difficult and dense situations [17]. As a result, non-object-centered unsupervised techniques have become increasingly popular recently. By learning typical patterns of activity from the behavior-related traits of individuals and things in geographical and temporal settings, these techniques address the issue of recognizing anomalies [18]. Target size, gradient, speed, and direction are all typically considered to be behavioral characteristics along can be represented with low-level illustrations like dense spatial-temporal interest points (dense STIPs), histograms of optical flow (HOF), and histograms of oriented gradients (HOG). These techniques are superior to trajectory-based techniques because they operate at the pixel level, which renders them more reliable in challenging settings [19]. Although there are numerous distinct kinds of anomalous behavior, all of these approaches rely on hand-crafted characteristics that are challenging to explain a priori. They are also incapable of adapting to defects that were never seen before [20].

In the last decade, DL approaches have been primarily employed to address a variety of computer vision problems, beating the state-of-the-art in a variety of challenging scenarios, based on the depth of hidden layers we can differentiate the neural network into four various categories as illustrated in Figure 1.

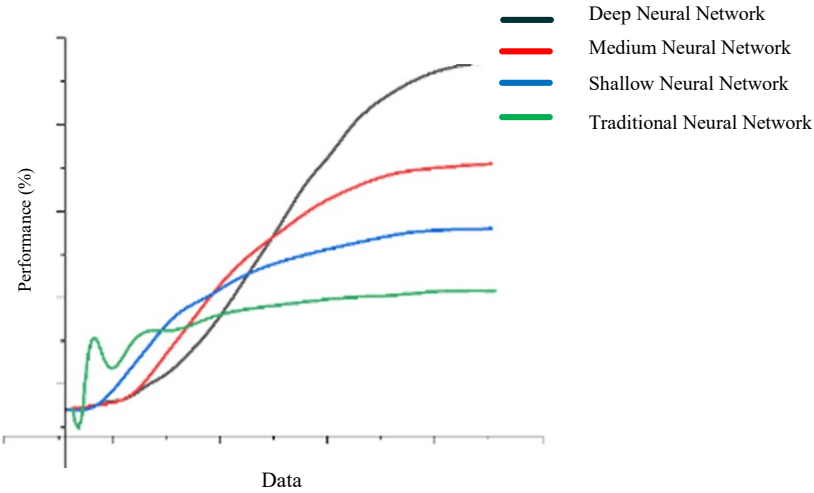



Fig. 1. Graphical representation of various neural networks' performance in human activity recognition

Traditional neural networks, also known as single-layer perceptrons, are simple linear classification models that cannot handle complicated problems. Shallow neural networks feature a limited number of hidden layers, which makes them successful for basic tasks but ineffective for more complicated ones. With a reasonable number of hidden layers, medium neural networks may learn more complicated features and patterns. With over six hidden layers, deep neural networks excel at learning highly abstract characteristics and can perform complicated tasks such as picture and speech recognition. Deep network training, on the other hand, maybe computationally costly and require a huge quantity of data to avoid overfitting. Integrating recognizing objects, object classification, and action identification. Since DL, a subtype of ML, trains to interpret the input as a hierarchy of nested concepts within various stages of the neural network, this study focuses on DL advancement to achieve outstanding results. As data volumes increase for recognizing behavior and activities, DL outperforms classical machine learning.

**2. Literature Survey.** The review has covered papers under the years 2019-2023 are provided in Figure 2. Totally 40 research articles from different sources are collected and the works are elaborated clearly in the following section with their respective pros and cons.

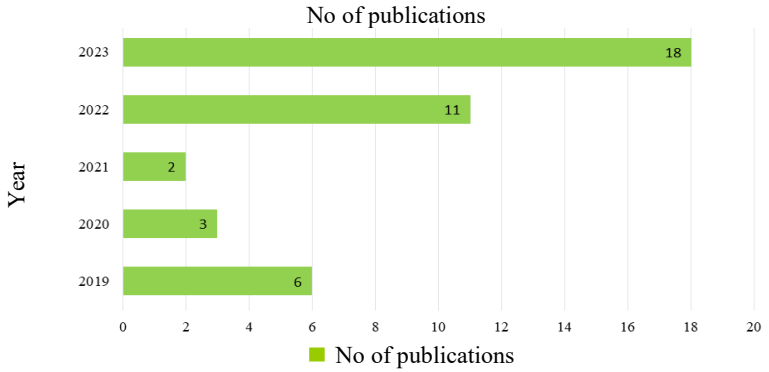



Fig. 2. Year-wise weightage of papers considered in the review

From Figure 2 it is clear that recent 2023 papers are considered in more numbers compared to the other papers. The variation in the study from 2019 to 2023 can be evaluated from this review. The collection of journals that have shown a lot of curiosity in recognizing human conduct and activities for video surveillance is depicted in Figure 3.

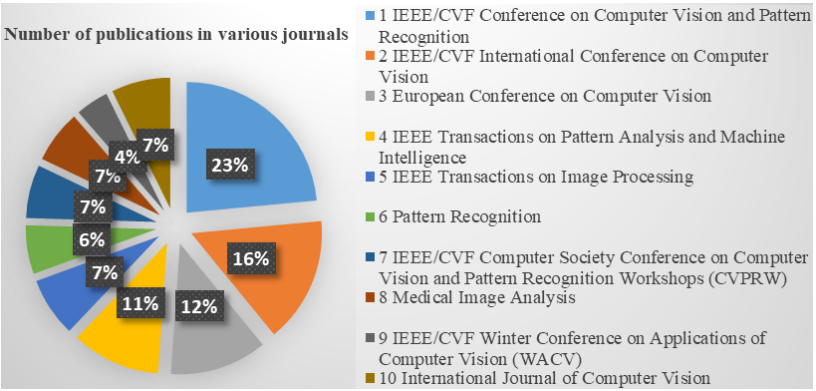



Fig. 3. Top 10 journals which have shown interest in video surveillance

**Face recognition.** In study [21] the authors suggested a technique for updating model parameters that will allow the dispersed EC

environment to synchronize the global DL model. A dynamic data movement strategy is further suggested to resolve the disparity between the workload and processing capabilities of edge nodes. The suggested DIVS system can effectively handle video surveillance and analysis duties, according to experimental results, which also demonstrated the EC architecture's ability to deliver elastic and scalable processing capacity.

In paper [22] the authors suggested an attribute-driven strategy for feature disentangling and frame re-weighting. In the study disentangling the characteristics of a single frame into sub-feature groups that individually relate to distinct semantic properties is described. The final representation is created by aggregating the sub-features at the temporal dimension after the sub-features have been reweighted by the attribute recognition confidence. This method enhances the most informative areas of each frame and helps the representation of the sequence be more discriminative. Numerous ablation experiments support the value of both feature disentangling and temporal re-weighting. The proposed strategy outperforms current state-of-the-art methods, as shown by the experimental findings on the iLIDS-VID, PRID-2011, and MARS datasets.

Paper [23] offered a straightforward method for achieving this goal using several key machine learning technologies, including TensorFlow, Keras, OpenCV, and Scikit-Learn. The proposed method detects the face in the picture or video and then assesses whether or not it is covered by a mask. It can distinguish a face and a mask in motion and a video as a surveillance task performance. The method achieves superb accuracy. We look at the most effective parameter settings for the Convolutional Neural Network model (CNN) to precisely detect the presence of masks without producing over-fitting.

In study [24] the authors state that AdaFocus is an adaptive focus technique that, although requiring a challenging three-stage training pipeline, has achieved a favorable accuracy-to-inference-speed trade-off. Through the use of an enhanced training scheme and a differentiable interpolation-based patch selection operation, this work reformulates the AdaFocus training as a straightforward one-stage approach. Extensive tests on six benchmark datasets show that the model performs far better than the original AdaFocus and other industry benchmarks while being much easier and more effective to train.

Study [25] presented a face mask identification technique for static photos and real-time videos that separates images into "with mask" and "without mask" categories. The Kaggle data set is used to train and assess the model. The collected data set has a performance accuracy rating of 98% and contains around 4,000 images. Comparing the proposed model to DenseNet-121, MobileNet-V2, VGG-19, and Inception-V3 reveals that it is more accurate and computationally economical. Schools, hospitals, banks, airports, and many other public or commercial institutions can use this work as a digital scanning tool.

In paper [26] the hybrid driver and vehicle identification module that can identify both the driver and the vehicle is presented. It can identify the driver and the car using facial recognition, voice recognition, and license plate identification. FaceNet was utilized for face identification, multi-task cascaded convolutional networks were used to crop the faces for facial recognition, and a three-layer long short-term memory model was used for speech verification. A tesseract was employed to identify vehicle license plates. The trials' findings demonstrate that the suggested method can consistently recognize both drivers and autos with zero error, which is a key advancement for guaranteeing the security of institutions.

In study [27] the authors sought to apply a model to complicated data by focusing on tasks for face identification in the picture and real-time video footage of persons wearing and without masks. The suggested technique performed well, with a 99.64% accuracy rate for images and a respectable accuracy rate for real-time video images. According to experimental findings, the suggested technique worked effectively, achieving an accuracy of 99.64% for images and a respectable accuracy for real-time video images. With an accuracy of 100%, recall of 99.28%, f1-score of 99.64%, and an error rate of 0.36%, additional measures demonstrated that the approach beat earlier models. In current technology, face mask detection is frequently employed in fields like artificial intelligence and smartphones.

In paper [28] the authors use a mix of a neural network and a genetic algorithm to pick and categorize face traits. The efficiency of the suggested technique was recently evaluated using both individual and composite elements of the face region. In experimental experiments, composite features outperform face region features. This research also includes a thorough comparison with different face recognition methods found in the FERET database. The classification accuracy achieved by the suggested approach is 94%, which represents a considerable increase and the highest classification accuracy among the findings from earlier investigations.

Study [29] presented real-time face recognition along with a DL framework for person identification and authentification in live or recorded CCTV feeds. The recommended approach is based on the VGGFace DL neural architecture and utilizes transfer learning to retrain the algorithm using a lesser originally designed dataset of 7500 photos of 26 different people. The presented approach provides the maximum level of recognition accuracy, as evidenced by a mean average of 96 percent on real-time inputs and confidence levels that vary from 78.54 percent to 100%.

In paper [30] the simple and effective facial detection method QMagFace is built using a recognition algorithm that utilizes a magnitude-aware angular margin loss and a quality-aware comparison score. The

suggested method incorporates model-specific facial picture characteristics into the comparison process to improve identification performance in unrestricted environments. The tests performed on various face recognition benchmarks and databases show that introducing quality awareness consistently improves recognition performance. The QMagFace source code is accessible to everyone. The information is presented in Table 1.

Table 1. Systematic Survey on Face Recognition

| Ref no. | Year | Technique | Significance | Limitation/Future scope |
|---|---|---|---|---|
| [21] | 2019 | Upgrading parameter values to ensure global DL network synchronization in a distributed EC scenario | – Effectively handle video surveillance and analysis duties,<br>– Deliver elastic and scalable processing capacity | Low contrast images |
| [22] | 2019 | An attribute-driven approach for feature disentangling and then frame re-weighting | – Enhances the most informative areas of each frame<br>– Represents the sequence being more discriminative | Low-resolution movies often feature background movement which is not considered |
| [23] | 2021 | Distinguish a face and a mask in motion of a video | Distinguish a face and a mask in motion and a video as a surveillance task performance | Biometric scans can be carried out in the future while wearing a face mask |
| [24] | 2022 | Differentiable interpolation-based patch selection procedure | Simple and efficient | The time-consuming phase of feature representation |
| [25] | 2022 | A face mask identification technique for static photos and real-time videos | Accuracy = 98% | Yet to increase the reliability |
| [26] | 2023 | The hybrid driver and vehicle identification module | Capable of recognizing both drivers and autos with 0% error | Scaling up the recommended technique for estimating the risk factors |
| [27] | 2023 | Face detection tasks with an effort to apply a model to complex data | Accuracy rate of 99.64% | Space complexity costs |
| [28] | 2023 | Neural network with genetic algorithm | Accuracy = 94% | Non-local change between frames increases the complexity |
| [29] | 2023 | DL system for person identification and recognition in live or recorded CCTV feeds | Confidence = 78.54 to 100 %<br>Mean average = 96 % | To reduce the dataset's size while increasing the number of photographs it contains |
| [30] | 2023 | QMagFace | Accuracy 98.98 % | Should concentrate on quality-based fusion methods |

Despite significant progress, there are still several limitations as given in Table 2 that make video-based physical detection more difficult and demanding.

Table 2. Systematic Survey on Emotion Recognition

| Ref no. | Year | Technique | Significance | Limitation/Future scope |
|---|---|---|---|---|
| [31] | 2019 | Brand-new RAN | Enhanced FER's performance with occlusion and alternative poses | Makes content interpretation complicated |
| [32] | 2019 | The DL-based emotion identification system | The effectiveness of the system is suggested using CNNs and ELMs | The future study assesses system performance in an edge-cloud computing environment |
| [33] | 2019 | The research uses the DL technique to categorize emotions through an iterative process | DL methods effectively classify emotions using a large number of sensors | Integrating sensors and modalities |
| [34] | 2020 | Self-care Network effectively suppresses uncertainty in deep networks | SCN outperforms advanced techniques with high scores | - |
| [35] | 2021 | The article introduces the DL technique for focusing on facial characteristics | Sensitivity is very high | - |
| [36] | 2022 | The article lists facial expressions and outlines four steps for execution | System performance was evaluated for databases and compared to current approaches | - |
| [37] | 2022 | Transformer-based fusion module combines static vision with dynamic multimodal properties | The performance of the model is increased | - |
| [38] | 2023 | Paper reduces processing power using a hierarchical Swin Transformer for expression recognition | Optimal speed-precision balance through high computational effort | Researchers can improve convolutional modules for expression recognition |
| [39] | 2023 | Research develops CNN for facial expression recognition using SMM dataset | The model achieves 93.94% accuracy and 67.18% FER2013 score on CK+ | Future SMM facial expression collection should include emotions like fear and contempt |
| [40] | 2023 | Low-light image enhancement, convolutional neural network for facial emotion recognition | Experimental assessment shows suggested technique outperforms others with 69.3% accuracy | Researchers developed smart glasses prototypes for vision-impaired identification |

In actuality, choosing the traits that make a moving object is a difficult process since they have a significant influence on the description and analysis of the activity. For example, when the scene's backdrop changes often or when brand-new things appear out of nowhere, it could be challenging to depict the action. Furthermore, a variety of elements, like scene location (outdoor/indoor) and outfit (dress, suit, footwear, etc.), can influence how the moving object seems.

**Survey on Emotion Recognition.** In [31] the authors looked at several in-the-wild FER datasets with pose and occlusion characteristics to answer the real-life pose with occlusion robust FER challenge. A unique Region Attention Network (RAN) was suggested in addition to the pose variation FER where an area biassed loss was added to imaginatively capture the significance of facial areas for occlusion. High attention weights for very significant locations might be encouraged by this method. Numerous investigations show that RAN and area biassed loss produce novel findings on FERPlus, AffectNet, RAF-DB, and SFEW, significantly enhancing FER performance with occlusion and changing position.

Study [32] demonstrated a DL-based emotion recognition mechanism built on emotional Big Data. Both speech and video data are recovered and fed to a CNN after being processed in the frequency domain to create a Mel-spectrogram. The outputs of the two CNNs are combined for the last classification using two ELMs and an SVM. The utility of the suggested technology was demonstrated by experimental results.

In study [33] an ongoing procedure that includes incorporating and eliminating enormous quantities of sensor data from several modalities is offered, this research applied a DL approach for emotion categorization. CNN-LSTM is employed in the approach, which reduces the requirement for human feature discovery and engineering by applying a hybrid strategy to raw sensor data. The findings show that DL algorithms are effective at classifying human emotions when a lot of sensors are being utilized (average accuracy is 95% and F-Measure is %95). In addition, hybrid models perform better than previously created Ensemble approaches that train the model through feature engineering (average accuracy 83%, F-Measure = 82%).

In study [34] it is described how to avoid deep networks from over-fitting ambiguous face images, this research proposed a simple but efficient Self-care Network (SCN) that efficiently suppresses uncertainty. Two methods that SCN particularly suppresses the uncertainty are through a self-attention mechanism over a mini-batch to weight every sample used for training with a ranking regularization with a meticulous relabeling procedure to update the labels of those specimens in the lowest-ranked group. Both

simulated FER datasets with gathered Web Emotion datasets have been used to assess the effectiveness of the proposed technique. SCN overcomes state-of-the-art approaches according to outcomes from open benchmarks, scoring 89.35% on FERPlus, 60.23% on Affect-Net, and 88.14% on RAF-DB.

In article [35] the author proposed an attentional convolutional network-based DL technique that can target important face characteristics and surpasses previous approaches on a range of datasets, such as FER-2013, CK+, FERG, and JAFFE. In addition, based on the classifier's output, we use a visualization technique that may help us pinpoint key facial features that indicate different moods. This research study's findings indicate that various emotions are responsive to various facial characteristics.

In research [36] a technique to identify facial expressions was suggested. Its four primary components are face recognition, a CNN framework founded on DL, data augmentation techniques, and a trade-off among data augmentation with DL characteristics. For thorough results from experiments, three benchmark databases – KDEF, GENKI-4k, and CK+ – have been employed. The performance of the suggested approach is being compared to currently employed state-of-the-art techniques, demonstrating its advantages.

Paper [37] offered a transformer-based fusion module that fuses static vision with dynamic multimodal features. The fusion module's cross-attention module focuses the output integrated features on the critical sections, easing downstream detection tasks. To increase model performance even further, we employ various data balance, data augmentation, and post-processing strategies. In the EXPR and AU tracks of the official ABAW3 Competition test, the model wins first place. On the Aff-Wild2 dataset, extensive quantitative evaluations and ablation experiments show how effective the recommended method is.

In article [38] the author uses the hierarchical Swin Transformer for the expression recognition job, which significantly reduces its processing power. The Swin Transformer with CNN and utilize it in an expression recognition job. At the same time, it is fused with a CNN model to suggest a network design that integrates the Transformer and CNN. We next test the suggested strategy using certain expression datasets that are available to the public, and we can achieve competitive results.

Paper [39] proposes a CNN design to distinguish facial expressions and create a facial expression dataset for the SMM. The suggested technique was evaluated for facial expression identification on two distinct benchmark datasets, FER2013 and CK+. We tested the suggested model on CK+ and

obtained accuracy for FER2013 of 93.94% and 67.18%, respectively. To investigate as well as assess the recommended algorithm's accuracy, we used the SMM Facial Expression dataset and attained 96.60% accuracy.

Study [40] offered a method for recognizing facial emotions in masked facial photographs by utilizing low-light image enhancement and feature analysis utilizing a CNN. The suggested method makes use of the AffectNet picture collection, which contains eight different types of facial emotions and 420,299 photos. The head and upper features of the face are represented using boundary and regional representation approaches. A facial landmark identification method-based feature extraction methodology is used to extract features. In an experimental test using the AffectNet dataset, the recommended method achieved an accuracy of 69.3%.

From Table 3 we can find that tracking gets challenging when analyzing photos with fluctuating light, which is a common aspect of actual environments. Outside CCTV cameras are subjected to external illumination fluctuations while gathering footage at night, which may provide low-contrast images that are challenging to comprehend. The adaptive background subtraction method also offers a consistent means of handling recurrent and long-term situation changes, as well as fluctuations in light. Noise reduction is necessary because low-resolution videos frequently have background movement brought on by camera movement or changes in lighting. While the optical flow vector's amplitude is a very effective signal for determining how much movement there is, the flow direction also may provide extra motion data.

**Action Recognition.** Study [41] presents a deep neural network that collects and categorizes activity characteristics by fusing convolutional layers with LSTM by fusing convolutional layers with LSTM, collects and categorizes activity characteristics. The proposed architecture comprises a two-layer LSTM followed by convolutional layers, a GAP level to reduce the parameters of the model, and then a BN layer to speed up convergence. The efficacy of the model was assessed using three publicly accessible datasets. The accuracy of the model was 95.78%, 95.85%, and 92.63% overall. The results demonstrate that the suggested theory looks more robust and effective in spotting activity than many of the previous results.

In paper [42] the author suggested a deep human action detection framework that is view-invariant and incorporates two crucial action cues: motion and shape temporal dynamics (STD). The motion stream encodes the motion content of the action as RGB-DIs, whereas the STD stream learns long-term view-invariant shape dynamics of action by mining view-invariant features from structural similarity index matrix (SSIM) dependent key depth human pose images. Research that employed cross-subject and

cross-view validation methodologies to measure the performance utilized three publicly accessible benchmarks. In regards to accuracy, ROC curve, and AUC, the technique greatly beat the state-of-the-art at the time.

Study [43] presented a brand-new end-to-end method for identifying unsupervised human actions using skeletons. We provide an innovative design that employs a convolutional autoencoder and graph Laplacian regularization to describe the skeletal geometry across the temporal dynamics of activities. Due to this approach including a self-supervised gradient reverse layer that provides generalization between camera perspectives, it is resistant to viewpoint fluctuations. The proposed method outperforms all earlier unsupervised skeleton-based methods on the cross-subject, cross-view, and cross-setup protocols on the large datasets NTU-60 and NTU-120. Though unsupervised, the system even outperforms a few supervised skeleton-based action recognition techniques owing to its learnable representation.

Research [44] offered a complete method for recognizing human motion in real-time from unprocessed depth picture sequences. It is based on a 3D fully CNN called the 3DFCNN, which dynamically encodes spatiotemporal patterns from raw depth data. The suggested 3DFCNN has been adjusted to operate in real-time with a respectable accuracy performance. On three well-known public datasets, it was recently compared to different state-of-the-art systems, showing that 3DFCNN surpasses other non-DNN-based current methods with an optimal precision of 83.6% yet maintains a noticeably lower computational cost of 1.09 seconds.

In paper [45] the authors used LSTM and CNN to construct a hybrid approach for activity identification. 20 individuals used the Kinect V2 sensor to build a brand-new challenging dataset with 12 distinct groups of human physical activity. A detailed ablation investigation was conducted using several traditional ML and DL neural networks to discover the optimal HAR solution. The accuracy of 90.89% achieved with the CNN-LSTM method shows that the model suggested is suitable for HAR applications.

In paper [46] the authors suggested a unique deep ConvLSTM network for skeletal-based activity identification and then fall detection. In sequence, LSTM systems, fully linked layers, and CNNs are combined. The acquisition method uses human identification and posture assessment to pre-calculate skeletal coordinates from an image/video sequence. From the raw skeleton coordinates and their unique geometrical and kinematic properties, the ConvLSTM network generates fresh directed features. On

the KinectHAR dataset, the recommended approach surpassed CNNs and LSTMs, which recorded accuracy of 93.89% and 92.75%, respectively.

In study [47] a spatially adaptive residual graph convolutional network (SARGCN) based on skeleton feature extraction was suggested for action recognition. It employs a learnable parameter matrix to decrease the number of parameters and improve feature extraction and generalization. To achieve greater accuracy at reduced computing costs and learning challenges, a residual connection is added. The effectiveness of the offered strategy has been confirmed by extensive trials on two substantial datasets.

In [48] the authors examined how well cuboid-aware feature aggregation performed when huge amounts of activity were presented. The authors also suggested monitoring actors and conducting temporal feature aggregation along the corresponding tracks to improve actor's feature representation under big motion. The intersection-over-union (IoU) between the boxes of action tubes/tracks was used by the authors to describe the actor's motion at various fixed time scales. Large-motion activities would eventually have reduced IoU, but slower actions would keep IoU higher. Researchers discover that, as compared to the cuboid-aware baseline, track-aware feature aggregation regularly produces a significant boost in action identification performance, particularly for actions with significant motion. As a result, the authors also provide the most recent findings using the extensive multi-sports dataset.

Paper [49] suggested the Spatio-TemporAl cRoss (STAR)-transformer, that is capable of successfully representing two cross-modal information as a vector. The encoder consists of a full spatio-temporal attention (FAttn) module and a proposed zigzag spatio-temporal attention (ZAttn) module, whilst the continuous decoder comprises a FAttn component with a recommended binary spatio-temporal attention (BAttn) module. Investigations show that the recommended method enhances performance excitingly on the Penn-Action, NTU-RGB+D 60, and 120 datasets.

Study [50] focused on the body occlusions for Skeleton-based One-shot Action Recognition (SOAR) in their study. It primarily takes into account two types of occlusions: arbitrary occlusions and more realistic occlusions brought on by various commonplace items. The authors formalize the first benchmark for SOAR from partly occluded body postures by using the suggested process to blend out sections of the skeleton sequences of three widely used action identification datasets. A novel transformer-based model called Trans4SOAR uses mixed attention fusion and three data streams to lessen the negative impact of occlusions. On all datasets, it performs better than alternative designs, outperforming the best-reported method on NTU-120 by 2.85%.

Table 3. Systematic Survey on Action Recognition

| Ref no. | Year | Technique | Significance | Limitation/Future scope |
|---|---|---|---|---|
| [41] | 2020 | Deep neural network combining LSTM and convolutional layers | The model achieves 95.67% accuracy | - |
| [42] | 2020 | Shape temporal dynamics in deep human behavior perception | The method outperforms the current state-of-the-art in accuracy, ROC curve, and AUC | The author seeks to improve action identification with skeletal and depth details |
| [43] | 2022 | New end-to-end approach for skeleton-based unsupervised human action identification | Improved methodology outperforms previous unsupervised skeleton-based algorithms on large datasets | The author focuses on real-time AE-L deployment and spatiotemporal connection enforcement |
| [44] | 2022 | Real-time human activity recognition method using unprocessed depth picture sequences | 3 DFCNN outperforms non-DNN techniques with 83.6% accuracy | Enhancing recognition accuracy in behaviors remains an ongoing research concern |
| [45] | 2022 | The author's mixed approach combines LSTM and CNN for activity identification | CNN-LSTM algorithm achieves 90.89% accuracy in HAR applications | Develop a model for identifying multiple people's activities and expanding advanced physical activities |
| [46] | 2022 | The article presents the ConvLSTM network for skeletal-based activities and fall identification | ConvLSTM achieves 98.89% accuracy, surpassing CNNs and LSTMs | |
| [47] | 2023 | The study proposes SARGCN for action identification using skeleton feature extraction | The efficiency of the model is high | A writer explores feature extraction and spatiotemporal graph structure analysis |
| [48] | 2023 | Investigating cuboid-aware feature aggregation performance in high activity | Track-aware feature aggregation enhances action recognition performance, particularly for significant motion actions | |
| [49] | 2023 | STAR-transformer represents cross-modal characteristics as identifiable vectors | Study shows suggested strategy improves performance compared to older methodologies | Scientists develop algorithms without overfitting using limited data |
| [50] | 2023 | The study focuses on SOAR body occlusions | NTU-120 outperforms the best SOAR technique by 2.85% | Future investigation into one-shot video identification is excluded |

Conquering the following problem will be challenging for optical flow-based motion assessments as given in Table 4. The mathematical feature points of the head, arms, legs, elbows, and shoulders form distinctive abstractions of various postures. The phase of correlation technique should be utilized to determine global motion among every pair of succeeding frames. If global motion is found, a Point Spread Function must be created using the projected slope and length of displacement, and the following frame can be deconvolved using the iterative deconvolution method.

Table 4. Systematic Survey on Anomaly Recognition

| Ref no. | Year | Technique | Significance | Limitation/Future scope |
|---|---|---|---|---|
| [51] | 2019 | Neural network for anomaly recognition | Detecting abnormal occurrences | |
| [52] | 2022 | Real-world traffic surveillance records require ongoing monitoring to ensure proper response to fatal situations. Nevertheless, maintaining constant human supervision of them is time-consuming and prone to mistakes | Real-world video traffic surveillance datasets are used to run the model, and both qualitatively and quantitatively substantial results have been obtained | The code may be implemented on PYNQ hardware in the future to process video frames more quickly for anomaly identification. The application of active learning to identify anomalies might also be the focus of the study |
| [53] | 2022 | The researcher proposed the Deep Residual Spatiotemporal Translation Network (DR-STN), an innovative unsupervised Deep Residual conditional Generative Adversarial Network (DR-cGAN) architecture using an online hard negative mining (OHNM) algorithm | The frame-level evaluation for the three benchmarks has an average AUC score of 96.73%. Between DR-STN and cutting-edge techniques, there is a 7.6% improvement in AUC at the frame level | The author's future work will concentrate on ongoing learning of unknown events, helping to determine if they are truly aberrant or merely unusual typical happenings |
| [54] | 2022 | This paper proposed a cheating detection system to deal with plagiarism and other forms of academic dishonesty | The authority is informed of the unusual conduct by an automated alarm, reducing the possibility of error that may arise from manual monitoring | |
| [55] | 2023 | In this study, researchers proposed a weakly supervised deep temporal encoding-decoding approach based on multiple instances of learning for anomaly detection in surveillance videos | The results show that the recommended method works as well as or is superior to the state-of-the-art techniques for detecting anomalies in video surveillance uses, achieving a state-of-the-art false alarm rate on the UCF-crime dataset | |

*Continuation of Table 4*

| Ref no. | Year | Technique | Significance | Limitation/Future scope |
|---|---|---|---|---|
| [56] | 2023 | The suggested method efficiently uses both geographic with temporal information by adopting a geographical branch with a temporal branch in a single network | The outcomes show that the network surpasses state-of-the-art techniques, obtaining, in terms of Area Under Curve, 97.4% for UCSD Ped2, 86.7% for CUHK Avenue, and 73.6% for the Shanghai Tech dataset | Practical surveillance needs to improve in the future |
| [57] | 2023 | This study demonstrated the creation of an automated safety mechanism that can quickly assist victims and identify suspicious activity in real time | On the test set database, the suggested approach's AUC was equal to 94.21%, and the detection accuracy was equivalent to 88.46% | Future studies will examine new feature extraction theories, feature selection strategies, and decreasing dimensionality approaches to increase the precision of the indicator |
| [58] | 2023 | This paper describes Ancilia, an end-to-end scalable, intelligent video surveillance platform for the IoT | To create safer and more secure communities, Ancilia intends to change the surveillance environment by introducing more efficient, intelligent, and fair security to the sector without asking individuals to give up their right to privacy | Future studies will examine new feature extraction theories, feature selection strategies, and decreasing dimensionality approaches to increase the precision of the indicator |
| [59] | 2023 | The author of this work using isolation tree-based unsupervised clustering divides the deep feature space of the video segments | According to experimental findings, the suggested framework outperforms state-of-the-art video anomaly detection techniques in terms of accuracy | Data training and quality must be improved in the future |
| [60] | 2023 | In video surveillance, finding frames that differed noticeably from the norm was the aim of anomaly detection. To solve this problem, the author created a unique bi-directional frame interpolation-based video anomaly recognition framework | The recommended method's value was confirmed by the excellent frame-level video anomaly detection results on open benchmarks | The suggested method's key is to interpolate normal frames with little to no mistakes, but aberrant frames with significant errors |

**Anomaly Recognition.** In study [51] the authors recommended the Anomaly Net neural network as a unique neural network for anomaly recognition because it combines feature learning, sparse representation, and dictionary learning in three joint neural processing blocks. To address the shortcomings of existing sparse coding optimizers, the researchers developed a special RNN to learn sparse representation with a sparse representation dictionary. Numerous trials demonstrate the method's cutting-edge performance in the task of detecting abnormal occurrences.

In study [52] the authors suggested that to monitor and respond appropriately in the event of tragic incidents, real-world traffic surveillance recordings need constant oversight. However, it is time-consuming and error-prone to oversee them continually with humans. As a result, a DL method for automatically detecting and localizing traffic accidents has been suggested by redefining the issue as anomaly finding. The technique uses sequence-to-sequence LSTM autoencoder and spatiotemporal autoencoder to model spatial and temporal representations in the video. Additionally, it employs a one-class categorization scheme. Real-world video traffic surveillance datasets are being used to apply the methodology, and both subjectively and numerically useful outcomes were achieved.

Paper [53] offered that the Deep Residual Spatiotemporal Translation Network (DR-STN) is a unique unsupervised Deep Residual Conditional Generative Adversarial Network (DR-cGAN) system that employs an online hard negative mining (OHNM) technique. It expands the network available for finding a mapping from spatial to temporal memories thus raising the perceived calibre of artificially created images. It has thoroughly tested against publicly accessible benchmarks and has outperformed other cutting-edge techniques. The difference in AUC between DR-STN and cutting-edge techniques at the frame level is 7.6%.

In study [54] the authors suggested a cheating detection method to deal with plagiarism and other types of academic dishonesty. During examinations, the system employs video monitoring to keep an eye on student behavior, particularly unusual behavior. The system employs three distinct methods: calculating the direction of the students' heads as they turn from their starting orientation, seeing pupil movement, and recognizing the point at which a student's hands come into touch with their faces. An automated alarm that informs the appropriate authority when any of these are found helps to reduce the possibility of error that may result from manual monitoring.

In paper [55] the authors proposed a weakly supervised deep temporal encoding-decoding approach employing multiple instances of learning for anomaly detection in surveillance videos. The proposed approach makes use of a deep temporal encoding-decoding network to record the spatiotemporal evolution of video instances across time while training using both abnormal and typical video clips. Low false alarm rates are produced by the suggested loss function, which optimizes the mean separation between predictions for normal and abnormal instance types. On the UCF-crime dataset, the suggested technique obtains a state-of-the-art false alert rate when compared to cutting-edge procedures.

Study [56] adopts a geographical branch and a temporal branch in a single network, effectively using both geographic and temporal information. It has a residual auto-encoder structure that is comprised of a deep CNN-powered encoder and a multi-stage channel attention-based decoder. System performance is estimated by utilizing three standard benchmark datasets: UCSD Ped2 (97.4%), CUHK Avenue (86.7%), and ShanghaiTech (73.6%).

Paper [57] demonstrated the creation of an autonomous security mechanism that can quickly assist victims in recognizing suspicious activity in real time. It utilizes an adaptive method based on DL (DL), PCA, and machine learning (ML). The suggested method has an experimented AUC and detection accuracy of 88.46% on the UCF-crime dataset. When compared to previously constructed systems, the suggested solution has proven to be accurate and resilient.

Study [58] introduces Ancilia, an end-to-end scalable, intelligent video surveillance solution for the IoT. Ancilia utilizes cutting-edge artificial intelligence for practical surveillance applications while upholding moral considerations and executing complex cognitive operations in real time. By bringing more efficient, intelligent, and fair security to the field, Ancilia hopes to change the surveillance environment and create safer and more secure communities without asking individuals to give up their right to privacy.

In paper [59] the deep feature space of video clips was divided using isolation tree-based unsupervised grouping. A pseudo anomaly score is produced by the RGB- -stream, while a pseudo dynamicity score, is produced by the flow stream. The majority voting method is employed to combine these scores and generate initial bags of beneficial and negative parts. Both scores are refined using a segment re-mapping and cross-branch feed-forward network refinement approach. According to experimental findings, the suggested framework

outperforms cutting-edge video anomaly identification techniques in terms of accuracy.

Paper [60] offered that in surveillance footage, finding frames that differed noticeably from the norm was the aim of identifying anomalies. The investigators used bi-directional frame interpolation to develop a brand-new model for video anomaly identification to address this issue. The proposed system includes a unique dynamic memory technique to balance memory scarcity with normality presentation variance, along with an optical flow estimating network with an interpolation system that has both been collaboratively optimized end-to-end. Numerous tests on widely used benchmarks show how much better the proposed framework is than existing solutions.

In common, the computationally and time-intensive phase of feature representation through video violence recognition acts as a substantial impediment to the deployment of violence detection in practical applications. The vast number of technologies as given in Table 5 is now in use to detect violent material in video and thus have substantial time and space complexity costs. These methods are therefore inappropriate for usage in real-world applications. As a result, denser and deeper DNN models are needed for better feature extraction and description. There is also a need for quicker, easier, and more accurate ways to identify violence. The high dimensional structure and a non-local shift among frames, nevertheless, make it more challenging for the methods employed to identify aberrant video.

**3. Summary.** After studying different paradigms of human behavior and activity recognition in video surveillance systems the following conclusions arrive which must be concentrated on in future research:

– The review of the literature reveals that certain methodologies aim to ignore the background and concentrate exclusively on foreground characteristics for anomaly identification. We believe that background knowledge might be helpful to simulate potential event-causing situations.

– The potential of human activity and behavior identification employing CNN, Deep Learning, LSTM, and GAN is bright since they provide more accuracy in most surveillance settings.

– In addition, the performance of the anomaly detection technique depends on the crowd density; as the crowd grows, its effectiveness declines, and it performs best in sparse crowds.

– In addition, while only benchmark data sets could be used for comparison, they might not be adequate to account for all real-world events.

– Edge computing is a potential strategy for delay-sensitive applications like intelligent surveillance and anomaly recognition. Since the data is handled on the device itself, it provides greater privacy and security. The burden is distributed through job offloading and ongoing improvement in edge devices, increasing total efficiency. Edge computing and human behavior and activity identification together will open up new opportunities for computer vision.

– From the research, we can find that more concentration must be provided in the arena of human action identification and anomaly detection which is the most crucial need in the current video surveillance.

– Although every group may be employed in a supervised or unsupervised way, the assessment reveals that the majority of investigators employed unsupervised learning to address the challenge of recognizing human conduct and activities since there was a dearth of huge datasets.

Table 5 compares various deep learning techniques used in detecting human activities across different applications. Convolutional Neural Networks (CNNs) excel in identifying faces, leveraging robust feature extraction and spatial hierarchies. Recurrent Neural Networks (RNNs) are pivotal in emotion identification, adept at modeling sequential data and capturing temporal dependencies. Long Short-Term Memory Networks (LSTMs) shine in action identification, effectively handling long-range dependencies in sequences. Autoencoders prove valuable in anomaly identification by self-supervised learning, although they may be sensitive to hyperparameters. Generative Adversarial Networks (GANs) show promise in generating synthetic data for anomaly detection, but their training stability can be challenging. Capsule Networks (CapsNets) offer improved handling of spatial hierarchies and resistance to certain adversarial attacks, though they're still underexplored. Lastly, Transfer Learning is a versatile approach applicable across all subcategories, leveraging pre-trained models to reduce the need for extensive data and accelerate training, but it may require task-specific fine-tuning for optimal performance.

Table 5. Comparison chart based on various deep learning methods

| **Technique** | **Used for** | **Signifi-cance** | **Advantages** | **Limitations** | **Future Recommen-dations** |
|---|---|---|---|---|---|
| Convolutional Neural Networks (CNNs) | Identifying faces | Highly effective in image recognition tasks | Robust feature extraction, spatial hierarchies | Limited to fixed-size inputs, may struggle with occlusions | Investigate multi-scale architectures for better adaptability |
| Recurrent Neural Networks (RNNs) | Emotion identification | Sequential data modeling | Captures temporal dependencies, variable-length sequences | Prone to vanishing / exploding gradients, computationally intensive | Explore variants like LSTMs, and GRUs for improved efficiency |
| Long Short-Term Memory Networks (LSTMs) | Action identification | Handling sequential data | Effective for modeling long-range dependencies, avoids vanishing gradient | Computationally expensive, harder to interpret | Investigate attention mechanisms for better context modeling |
| Autoencoders | Anomaly identification | Anomaly detection in unlabeled data | Self-supervised learning, robust to noisy data | Sensitive to choice of hyperparameters, may require large datasets | Explore unsupervised pre-training for improved anomaly detection |
| Generative Adversarial Networks (GANs) | Anomaly identification | Generate synthetic data for anomaly detection | Effective in generating realistic data distributions | Training instability, mode collapse | Investigate techniques for stable GAN training, utilize in semi-supervised setups |
| Capsule Networks (CapsNets) | Identifying faces, Emotion identification | Improved handling of spatial hierarchies | Resistant to certain types of adversarial attacks | Limited adoption, computationally intensive | Investigate hybrid architectures with CNNs for improved performance |
| Transfer Learning | Across all subcategories | Utilizing pre-trained models for specific tasks | Reduces the need for large datasets, faster training | May not always transfer well, task-specific fine-tuning needed | Investigate techniques for better model adaptation in transfer learning scenarios |

**4. Future recommendation.** Advancements in CNNs, Deep Learning, LSTM, and GANs hold great promise for improving accuracy in surveillance. Further exploration of novel architectures and techniques, such as attention mechanisms and multi-modal integration, could significantly enhance human activity identification. Some more recommendations are given below based on the state of the art:

– Future research could explore hybrid approaches that incorporate both foreground and background information, leveraging the potential benefits of simulating event-causing situations with background knowledge.

– Investigate novel architectures and techniques within CNNs, Deep Learning, LSTM, and GANs to further enhance accuracy in surveillance settings, possibly by incorporating multi-modal information or attention mechanisms.

– Develop adaptive anomaly detection techniques that dynamically adjust their sensitivity based on crowd density, potentially utilizing reinforcement learning or adaptive thresholding mechanisms.

– Encourage the creation of more diverse and realistic benchmark datasets that capture a broader range of real-world events, potentially through crowdsourcing or incorporating simulated data augmentation techniques.

– Investigate techniques for optimizing and accelerating anomaly detection algorithms specifically for edge computing environments, possibly through model compression, quantization, or specialized hardware acceleration.

– Allocate research efforts towards developing specialized models and algorithms dedicated to human action identification and anomaly detection, possibly exploring novel architectures or incorporating domain-specific knowledge.

– Encourage the creation of larger annotated datasets to support the application of supervised learning approaches. Additionally, explore techniques for semi-supervised learning that leverage limited labeled data with a larger pool of unlabeled data.

– These future directions aim to address specific areas of improvement and expansion within the field of human activity identification and anomaly detection, ultimately advancing the capabilities and effectiveness of surveillance systems.

**5. Conclusion.** This article examines DL-based methods for video surveillance that span a range of techniques and approaches for recognizing human behavior and activities. Readers should ideally be able to appreciate not just the justification for employing a particular technique, but also to

compare several approaches, generate a comparative analysis, and suggest a strategy after reading a complete overview of human behavior and activity identification. First, we divided the methods into four groups, based on how well they could identify faces, emotions, actions, and anomalies. Additionally, we listed every category's advantages and disadvantages in accordance. Future work on the DL model should focus on studying human action and emotion identification, which can improve situational knowledge for targeted video surveillance.

## References


1. Zhang J., Zi L., Hou Y., Wang M., Jiang W., Deng D. A DL-based approach to enable action recognition for construction equipment. Advances in Civil Engineering. 2020. pp. 1–14.
2. Wang X., Che Z., Jiang B., Xiao N., Yang K., Tang J., Ye J., Wang J., Qi Q. Robust unsupervised video anomaly detection by multipath frame prediction. IEEE transactions on neural networks and learning systems. 2021. vol. 33. no. 6. pp. 2301–2312.
3. Zhang H.B., Zhang Y.X., Zhong B., Lei Q., Yang L., Du J.X., Chen D.S. A comprehensive survey of vision-based human action recognition methods. Sensors. 2019. vol. 19(5). no. 1005.
4. Pervaiz M., Jalal A., Kim K. A hybrid algorithm for multi-people counting and tracking for smart surveillance. International Bhurban conference on applied sciences and technologies (IBCAST). 2021. pp. 530–535.
5. Kong Y., Fu Y. Human action recognition and prediction: A survey. International Journal of Computer Vision. 2022. vol. 130(5). pp. 1366–1401.
6. Franco A., Magnani A., Maio D. A multimodal approach for human activity recognition based on skeleton and RGB data. Pattern Recognition Letters. 2020. vol. 131. pp. 293–299.
7. Wang L., Huynh D.Q., Koniusz P. A comparative review of recent kinect-based action recognition algorithms. IEEE Transactions on Image Processing. 2019. vol. 29. pp. 15–28.
8. Zhou X., Liang W., Kevin I., Wang K., Wang H., Yang L.T., Jin Q. Deep-learning-enhanced human activity recognition for the Internet of Healthcare things. IEEE Internet of Things Journal. 2020. vol. 7(7). pp. 6429–6438.
9. Qiu Z., Yao T., Ngo C.W., Tian X., Mei T. Learning spatio-temporal representation with local and global diffusion. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019. pp. 12056–12065.
10. Sreenu G., Durai S. Intelligent video surveillance: a review through DL techniques for crowd analysis. Journal of Big Data. 2019. vol. 6(1). pp. 1–27.
11. Elharrouss O., Almaadeed N., Al-Maadeed S., Bouridane A., Beghdadi A. A combined multiple action recognition and summarization for surveillance video sequences. Applied Intelligence. 2021. vol. 51. pp. 690–712.
12. Jaouedi N., Boujnah N., Bouhlel M.S. A new hybrid DL model for human action recognition. Journal of King Saud University – Computer and Information Sciences. 2020. vol. 32. no. 4. pp. 447–453.
13. Dang L.M., Min K., Wang H., Piran M.J., Lee C.H., Moon H. Sensor-based and vision-based human activity recognition: A comprehensive survey. Pattern Recognition. 2020. vol. 108. no. 107561.

14. Saeed A., Ozcelebi T., Lukkien J. Multi-task self-supervised learning for human activity detection. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies. 2019. vol. 3(2). pp. 1–30.
15. Fu B., Damer N., Kirchbuchner F., Kuijper A. Sensing technology for human activity recognition: A comprehensive survey. IEEE Access. 2020. vol. 8. pp. 83791–83820.
16. du Toit J., du Toit T, Kruger H. Heuristic Data Augmentation for Improved Human Activity Recognition. Proceedings of the Southern Africa Telecommunication Networks and Applications Conference (SATNAC). 2019. pp. 264–269.
17. Rezaee K., Rezakhani S.M., Khosravi M.R., Moghimi M.K. A survey on DL-based real-time crowd anomaly detection for secure distributed video surveillance. Personal and Ubiquitous Computing. 2021. pp. 1–17.
18. Concone F., Re G.L., Morana M. A fog-based application for human activity recognition using personal smart devices. ACM Transactions on Internet Technology (TOIT). 2019. vol. 19(2). pp. 1–20.
19. He J.Y., Wu X., Cheng Z.Q., Yuan Z., Jiang Y.G. DB-LSTM: Densely-connected Bi-directional LSTM for human action recognition. Neurocomputing. 2021. vol. 444. pp. 319–331.
20. Beddiar D.R., Nini B., Sabokrou M., Hadid A. Vision-based human activity recognition: a survey. Multimedia Tools and Applications. 2020. vol. 79. no. 41-42. pp. 30509–30555.
21. Chen J., Li K., Deng Q., Li K., Philip S.Y. Distributed DL model for intelligent video surveillance systems with edge computing. IEEE Transactions on Industrial Informatics. 2019. DOI: 10.1109/TII.2019.2909473.
22. Zhao Y., Shen X., Jin Z., Lu H., Hua X.S. Attribute-driven feature disentangling and temporal aggregation for video person re-identification. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019. pp. 4913–4922.
23. Kaur G., Sinha R., Tiwari P.K., Yadav S.K., Pandey P., Raj R., Vashisth A., Rakhra M. Face mask recognition system using CNN model. Neuroscience Informatics. 2021. vol. 2(3). no. 100035. DOI:10.1016/j.neuri.2021.100035.
24. Wang Y., Yue Y., Lin Y., Jiang H., Lai Z., Kulikov V., Huang G. Adafocus v2: End-to-end training of spatial dynamic networks for video recognition. IEEE/CVF Computer Vision and Pattern Recognition (CVPR). 2022. pp. 20030–20040.
25. Goyal H., Sidana K., Singh C., Jain A., Jindal S. A real-time face mask detection system using a convolutional neural network. Multimedia Tools and Applications. 2022. vol. 81(11). pp. 14999–15015.
26. Sayeed A., Srizon A.Y., Hasan M.M., Shin J., Hasan M.A.M., Mahmud M.R. A Hybrid Campus Security System Combined Face, Number-Plate, and Voice Recognition. International Conference on Recent Trends in Image Processing and Pattern Recognition. 2022. pp. 356–368.
27. Kumar B.A., Bansal M. Face Mask Detection on Photo and Real-Time Video Images Using Caffe-MobileNetV2 Transfer Learning. Applied Sciences. 2023. vol. 13(2). no. 935.
28. Kamyab T., Daealhaq H., Ghahfarokhi A.M., Beheshtinejad F., Salajegheh E. Combination of Genetic Algorithm and Neural Network to Select Facial Features in Face Recognition Technique. International Journal of Robotics and Control Systems. 2023. vol. 3(1). pp. 50–58.
29. Singh A., Bhatt S., Nayak V., Shah M. Automation of surveillance systems using DL and facial recognition. International Journal of System Assurance Engineering and Management. 2023. vol. 14. pp. 236–245.

30. Terhorst P., Ihlefeld M., Huber M., Damer N., Kirchbuchner F., Raja K., Kuijper A. Qmagface: Simple and accurate quality-aware face recognition. In Proceedings of the IEEE/CVF Applications of Computer Vision. 2023. 3484–3494.
31. Wang K., Peng X., Yang J., Meng D., Qiao Y. Region attention networks for pose and occlusion robust facial expression recognition. IEEE Transactions on Image Processing. 2020. vol. 29. pp. 4057–4069.
32. Hossain M.S., Muhammad G. Emotion recognition using DL approach from audio–visual emotional big data. Information Fusion. 2019. vol. 49. pp. 69–78.
33. Kanjo E., Younis E.M., Ang C.S. DL analysis of mobile physiological, environmental, and location sensor data for emotion detection. Information Fusion. 2019. vol. 49. pp. 46–56.
34. Wang K., Peng X., Yang J., Lu S., Qiao Y. Suppressing uncertainties for large-scale facial expression recognition. Proceedings of the IEEE/CVF computer vision and pattern recognition. 2020. pp. 6897–6906.
35. Minaee S., Minaei, M., Abdolrashidi A. Deep-emotion: Facial expression recognition using the attentional convolutional network. Sensors. 2021. vol. 21(9). no. 3046.
36. Umer S., Rout R.K., Pero C., Nappi M. Facial expression recognition with trade-offs between data augmentation and DL features. Journal of Ambient Intelligence and Humanized Computing. 2022. pp. 1–15.
37. Zhang W., Qiu F., Wang S., Zeng H., Zhang Z., An R., Ma B., Ding Y. Transformer-based multimodal information fusion for facial expression analysis. Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition. 2022. pp. 2428–2437.
38. Zhu X., Li Z., Sun J. Expression recognition method combining convolutional features and Transformer. Mathematical Foundations of Computing. 2023. vol. 6. no. 2. pp. 203–217.
39. Bapat M.M., Patil C.H., Mali S.M. Database Development and Recognition of Facial Expression using DL. 2023. 20 p. DOI: 10.21203/rs.3.rs-2477808/v1.
40. Mukhiddinov M., Djuraev O., Akhmedov F., Mukhamadiyev A., Cho J. Masked Face Emotion Recognition Based on Facial Landmarks and DL Approaches for Visually Impaired People. Sensors. 2023. vol. 23(3). no. 1080.
41. Xia K., Huang J., Wang H. LSTM-CNN architecture for human activity recognition. IEEE Access. 2020. vol. 8. pp. 56855–56866.
42. Dhiman C., Vishwakarma D.K. View-invariant deep architecture for human action recognition using two-stream motion and shape temporal dynamics. IEEE Transactions on Image Processing. 2020. vol. 29. pp. 3835–3844.
43. Paoletti G., Cavazza J., Beyan C., Del Bue A. Unsupervised human action recognition with skeletal graph Laplacian and self-supervised viewpoints invariance. 2022. arXiv preprint arXiv:2204.10312.
44. Sanchez-Caballero A., de Lopez-Diz S., Fuentes-Jimenez D., Losada-Gutiérrez C., Marrón-Romera M., Casillas-Perez D., Sarker M.I. 3dfcnn: Real-time action recognition using 3d deep neural networks with raw depth information. Multimedia Tools and Applications. 2022. vol. 81. no. 17. pp. 24119–24143.
45. Khan I.U., Afzal S., Lee J.W. Human activity recognition via hybrid DL-based model. Sensors. 2022. vol. 22(1). no. 323.
46. Yadav S.K., Tiwari K., Pandey H.M., Akbar S.A. Skeleton-based human activity recognition using Conv LSTM and guided feature learning. Soft Computing. 2022. pp. 1–14.
47. Zhu Q., Deng H. Spatial adaptive graph convolutional network for skeleton-based action recognition. Applied Intelligence. 2023. pp. 1–13.
48. Singh G., Choutas V., Saha S., Yu F., Van Gool L. Spatio-Temporal Action Detection under Large Motion. Proceedings of the IEEE/CVF Applications of Computer Vision. 2023. pp. 6009–6018.

49. Ahn D., Kim S., Hong H., Ko B.C. STAR-Transformer: A Spatio-temporal Cross Attention Transformer for Human Action Recognition. In Proceedings of the IEEE/CVF Applications of Computer Vision. 2023. pp. 3330–3339.
50. Peng K., Roitberg A., Yang K., Zhang J., Stiefelhagen R. Delving Deep into One-Shot Skeleton-based Action Recognition with Diverse Occlusions. IEEE Transactions on Multimedia. 2023. arXiv preprint arXiv:2202.11423v3.
51. Zhou J.T., Du J., Zhu H., Peng X., Liu Y., Goh R.S.M. Anomalynet: An anomaly detection network for video surveillance. IEEE Transactions on Information Forensics and Security. 2019. vol. 14(10). pp. 2537–2550.
52. Pawar K., Attar V. DL-based detection and localization of road accidents from traffic surveillance videos. ICT Express. 2022. vol. 8. no. 3. pp. 379–387.
53. Ganokratanaa T., Aramvith S., Sebe N. Video anomaly detection using deep residual-spatiotemporal translation network. Pattern Recognition Letters. 2022. vol. 155. pp. 143–150.
54. Roa'a M., Aljazaery I.A., ALRikabi H.T.S., Alaidi A.H.M. Automated Cheating Detection Based on Video Surveillance in the Examination Classes. iJIM. 2022. vol. 16(08). no. 125.
55. Kamoona A.M., Gostar A.K., Bab-Hadiashar A., Hoseinnezhad R. Multiple instance-based video anomaly detection using deep temporal encoding–decoding. Expert Systems with Applications. 2023. vol. 214. no. 119079. DOI: 10.1016/j.eswa.2022.119079.
56. Le V.T., Kim Y.G. Attention-based residual autoencoder for video anomaly detection. Applied Intelligence. 2023. vol. 53(3). pp. 3240–3254.
57. Abbas Z.K., Al-Ani A.A. An adaptive algorithm based on principal component analysis-DL for anomalous events detection. Indonesian Journal of Electrical Engineering and Computer Science. 2023. vol. 29(1). pp. 421–430.
58. Pazho A.D., Neff C., Noghre G.A., Ardabili B.R., Yao S., Baharani M., Tabkhi H. Ancilia: Scalable Intelligent Video Surveillance for the Artificial Intelligence of Things. 2023. arXiv preprint arXiv:2301.03561.
59. Thakare K.V., Raghuwanshi Y., Dogra D.P., Choi H., Kim I.J. DyAnNet: A Scene Dynamicity Guided Self-Trained Video Anomaly Detection Network. Proceedings of the IEEE/CVF Applications of Computer Vision. 2023. pp. 5541–5550.
60. Deng H., Zhang Z., Zou S., Li X. Bi-Directional Frame Interpolation for Unsupervised Video Anomaly Detection. In Proceedings of the IEEE/CVF Applications of Computer Vision. 2023. pp. 2634–2643.



**Nukala Sujata Gupta** — Research scholar, Department of computer science and engineering, Koneru Lakshmaiah Education Foundation. Research interests: science and engineering. gsuji29@gmail.com; Green Fields, Vaddeswaram, 522302, Guntur, Andhra Pradesh, India; office phone: +91(8645)350-0200.

**Ramya K. Ruth** — Associate professor, Department of computer science and engineering, Koneru Lakshmaiah Education Foundation. Research interests: science and engineering. The number of publications — 12. ramya_cse@kluniversity.in; Green Fields, Vaddeswaram, 522302, Guntur, Andhra Pradesh, India; office phone: +91(8645)350-0200.

**Karnati Ramesh** — Associate professor, Department of computer science and engineering, Vardhaman College of Engineering. Research interests: data mining, machine learning, artificial intelligence, IoT. The number of publications — 17. ramesh.krnt@vardhaman.org; Kacharam, Shamshabad, 501218, Hyderabad, Telangana, India; office phone: +91(8688)901-557.

**УДК** 004.7 **DOI** 10.15622/ia.23.2.5

Н. Суджата Гупта, К.Р. Рамья, Р. Карнати

# РАСПОЗНАВАНИЕ ДЕЙСТВИЙ ЧЕЛОВЕКА В СИСТЕМАХ ВИДЕОНАБЛЮДЕНИЯ С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ ГЛУБОКОГО ОБУЧЕНИЯ – ОБЗОР

*Суджата Гупта Н., Рамья К.Р., Карнати Р.* **Распознавание действий человека в системах видеонаблюдения с использованием методов глубокого обучения – обзор.**

**Аннотация.** Несмотря на широкое применение во многих областях, точная и эффективная идентификация деятельности человека продолжает оставаться интересной исследовательской проблемой в области компьютерного зрения. В настоящее время проводится много исследований по таким темам, как распознавание активности пешеходов и способы распознавания движений людей с использованием данных глубины, трехмерных скелетных данных, данных неподвижных изображений или стратегий, использующих пространственно-временные точки интереса. Это исследование направлено на изучение и оценку подходов DL для обнаружения человеческой активности на видео. Основное внимание было уделено нескольким структурам для обнаружения действий человека, которые используют DL в качестве своей основной стратегии. В зависимости от приложения, включая идентификацию лиц, идентификацию эмоций, идентификацию действий и идентификацию аномалий, прогнозы появления людей разделены на четыре различные подкатегории. В литературе было проведено несколько исследований, основанных на этих распознаваниях для прогнозирования поведения и активности человека в приложениях видеонаблюдения. Сравнивается современное состояние методов DL для четырех различных приложений. В этой статье также представлены области применения, научные проблемы и потенциальные цели в области распознавания человеческого поведения и активности на основе DL.

**Ключевые слова:** распознавание лиц, распознавание эмоций, распознавание действий, распознавание аномалий, DL, распознавание человеческого поведения и активности /обнаружение.

## Литература

1. Zhang J., Zi L., Hou Y., Wang M., Jiang W., Deng D. A DL-based approach to enable action recognition for construction equipment. Advances in Civil Engineering. 2020. pp. 1–14.
2. Wang X., Che Z., Jiang B., Xiao N., Yang K., Tang J., Ye J., Wang J., Qi Q. Robust unsupervised video anomaly detection by multipath frame prediction. IEEE transactions on neural networks and learning systems. 2021. vol. 33. no. 6. pp. 2301–2312.
3. Zhang H.B., Zhang Y.X., Zhong B., Lei Q., Yang L., Du J.X., Chen D.S. A comprehensive survey of vision-based human action recognition methods. Sensors. 2019. vol. 19(5). no. 1005.
4. Pervaiz M., Jalal A., Kim K. A hybrid algorithm for multi-people counting and tracking for smart surveillance. International Bhurban conference on applied sciences and technologies (IBCAST). 2021. pp. 530–535.
5. Kong Y., Fu Y. Human action recognition and prediction: A survey. International Journal of Computer Vision. 2022. vol. 130(5). pp. 1366–1401.

6. Franco A., Magnani A., Maio D. A multimodal approach for human activity recognition based on skeleton and RGB data. Pattern Recognition Letters. 2020. vol. 131. pp. 293–299.
7. Wang L., Huynh D.Q., Koniusz P. A comparative review of recent kinect-based action recognition algorithms. IEEE Transactions on Image Processing. 2019. vol. 29. pp. 15–28.
8. Zhou X., Liang W., Kevin I., Wang K., Wang H., Yang L.T., Jin Q. Deep-learning-enhanced human activity recognition for the Internet of Healthcare things. IEEE Internet of Things Journal. 2020. vol. 7(7). pp. 6429–6438.
9. Qiu Z., Yao T., Ngo C.W., Tian X., Mei T. Learning spatio-temporal representation with local and global diffusion. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019. pp. 12056–12065.
10. Sreenu G., Durai S. Intelligent video surveillance: a review through DL techniques for crowd analysis. Journal of Big Data. 2019. vol. 6(1). pp. 1–27.
11. Elharrouss O., Almaadeed N., Al-Maadeed S., Bouridane A., Beghdadi A. A combined multiple action recognition and summarization for surveillance video sequences. Applied Intelligence. 2021. vol. 51. pp. 690–712.
12. Jaouedi N., Boujnah N., Bouhlel M.S. A new hybrid DL model for human action recognition. Journal of King Saud University – Computer and Information Sciences. 2020. vol. 32. no. 4. pp. 447–453.
13. Dang L.M., Min K., Wang H., Piran M.J., Lee C.H., Moon H. Sensor-based and vision-based human activity recognition: A comprehensive survey. Pattern Recognition. 2020. vol. 108. no. 107561.
14. Saeed A., Ozcelebi T., Lukkien J. Multi-task self-supervised learning for human activity detection. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies. 2019. vol. 3(2). pp. 1–30.
15. Fu B., Damer N., Kirchbuchner F., Kuijper A. Sensing technology for human activity recognition: A comprehensive survey. IEEE Access. 2020. vol. 8. pp. 83791–83820.
16. du Toit J., du Toit T, Kruger H. Heuristic Data Augmentation for Improved Human Activity Recognition. Proceedings of the Southern Africa Telecommunication Networks and Applications Conference (SATNAC). 2019. pp. 264–269.
17. Rezaee K., Rezakhani S.M., Khosravi M.R., Moghimi M.K. A survey on DL-based real-time crowd anomaly detection for secure distributed video surveillance. Personal and Ubiquitous Computing. 2021. pp. 1–17.
18. Concone F., Re G.L., Morana M. A fog-based application for human activity recognition using personal smart devices. ACM Transactions on Internet Technology (TOIT). 2019. vol. 19(2). pp. 1–20.
19. He J.Y., Wu X., Cheng Z.Q., Yuan Z., Jiang Y.G. DB-LSTM: Densely-connected Bi-directional LSTM for human action recognition. Neurocomputing. 2021. vol. 444. pp. 319–331.
20. Beddiar D.R., Nini B., Sabokrou M., Hadid A. Vision-based human activity recognition: a survey. Multimedia Tools and Applications. 2020. vol. 79. no. 41-42. pp. 30509–30555.
21. Chen J., Li K., Deng Q., Li K., Philip S.Y. Distributed DL model for intelligent video surveillance systems with edge computing. IEEE Transactions on Industrial Informatics. 2019. DOI: 10.1109/TII.2019.2909473.
22. Zhao Y., Shen X., Jin Z., Lu H., Hua X.S. Attribute-driven feature disentangling and temporal aggregation for video person re-identification. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019. pp. 4913–4922.

23. Kaur G., Sinha R., Tiwari P.K., Yadav S.K., Pandey P., Raj R., Vashisth A., Rakhra M. Face mask recognition system using CNN model. Neuroscience Informatics. 2021. vol. 2(3). no. 100035. DOI:10.1016/j.neuri.2021.100035.
24. Wang Y., Yue Y., Lin Y., Jiang H., Lai Z., Kulikov V., Huang G. Adafocus v2: End-to-end training of spatial dynamic networks for video recognition. IEEE/CVF Computer Vision and Pattern Recognition (CVPR). 2022. pp. 20030–20040.
25. Goyal H., Sidana K., Singh C., Jain A., Jindal S. A real-time face mask detection system using a convolutional neural network. Multimedia Tools and Applications. 2022. vol. 81(11). pp. 14999–15015.
26. Sayeed A., Srizon A.Y., Hasan M.M., Shin J., Hasan M.A.M., Mahmud M.R. A Hybrid Campus Security System Combined Face, Number-Plate, and Voice Recognition. International Conference on Recent Trends in Image Processing and Pattern Recognition. 2022. pp. 356–368.
27. Kumar B.A., Bansal M. Face Mask Detection on Photo and Real-Time Video Images Using Caffe-MobileNetV2 Transfer Learning. Applied Sciences. 2023. vol. 13(2). no. 935.
28. Kamyab T., Daealhaq H., Ghahfarokhi A.M., Beheshtinejad F., Salajegheh E. Combination of Genetic Algorithm and Neural Network to Select Facial Features in Face Recognition Technique. International Journal of Robotics and Control Systems. 2023. vol. 3(1). pp. 50–58.
29. Singh A., Bhatt S., Nayak V., Shah M. Automation of surveillance systems using DL and facial recognition. International Journal of System Assurance Engineering and Management. 2023. vol. 14. pp. 236–245.
30. Terhorst P., Ihlefeld M., Huber M., Damer N., Kirchbuchner F., Raja K., Kuijper A. Qmagface: Simple and accurate quality-aware face recognition. In Proceedings of the IEEE/CVF Applications of Computer Vision. 2023. 3484–3494.
31. Wang K., Peng X., Yang J., Meng D., Qiao Y. Region attention networks for pose and occlusion robust facial expression recognition. IEEE Transactions on Image Processing. 2020. vol. 29. pp. 4057–4069.
32. Hossain M.S., Muhammad G. Emotion recognition using DL approach from audio–visual emotional big data. Information Fusion. 2019. vol. 49. pp. 69–78.
33. Kanjo E., Younis E.M., Ang C.S. DL analysis of mobile physiological, environmental, and location sensor data for emotion detection. Information Fusion. 2019. vol. 49. pp. 46–56.
34. Wang K., Peng X., Yang J., Lu S., Qiao Y. Suppressing uncertainties for large-scale facial expression recognition. Proceedings of the IEEE/CVF computer vision and pattern recognition. 2020. pp. 6897–6906.
35. Minaee S., Minaei, M., Abdolrashidi A. Deep-emotion: Facial expression recognition using the attentional convolutional network. Sensors. 2021. vol. 21(9). no. 3046.
36. Umer S., Rout R.K., Pero C., Nappi M. Facial expression recognition with trade-offs between data augmentation and DL features. Journal of Ambient Intelligence and Humanized Computing. 2022. pp. 1–15.
37. Zhang W., Qiu F., Wang S., Zeng H., Zhang Z., An R., Ma B., Ding Y. Transformer-based multimodal information fusion for facial expression analysis. Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition. 2022. pp. 2428–2437.
38. Zhu X., Li Z., Sun J. Expression recognition method combining convolutional features and Transformer. Mathematical Foundations of Computing. 2023. vol. 6. no. 2. pp. 203–217.
39. Bapat M.M., Patil C.H., Mali S.M. Database Development and Recognition of Facial Expression using DL. 2023. 20 p. DOI: 10.21203/rs.3.rs-2477808/v1.

40. Mukhiddinov M., Djuraev O., Akhmedov F., Mukhamadiyev A., Cho J. Masked Face Emotion Recognition Based on Facial Landmarks and DL Approaches for Visually Impaired People. Sensors. 2023. vol. 23(3). no. 1080.
41. Xia K., Huang J., Wang H. LSTM-CNN architecture for human activity recognition. IEEE Access. 2020. vol. 8. pp. 56855–56866.
42. Dhiman C., Vishwakarma D.K. View-invariant deep architecture for human action recognition using two-stream motion and shape temporal dynamics. IEEE Transactions on Image Processing. 2020. vol. 29. pp. 3835–3844.
43. Paoletti G., Cavazza J., Beyan C., Del Bue A. Unsupervised human action recognition with skeletal graph Laplacian and self-supervised viewpoints invariance. 2022. arXiv preprint arXiv:2204.10312.
44. Sanchez-Caballero A., de Lopez-Diz S., Fuentes-Jimenez D., Losada-Gutiérrez C., Marrón-Romera M., Casillas-Perez D., Sarker M.I. 3dfcnn: Real-time action recognition using 3d deep neural networks with raw depth information. Multimedia Tools and Applications. 2022. vol. 81. no. 17. pp. 24119–24143.
45. Khan I.U., Afzal S., Lee J.W. Human activity recognition via hybrid DL-based model. Sensors. 2022. vol. 22(1). no. 323.
46. Yadav S.K., Tiwari K., Pandey H.M., Akbar S.A. Skeleton-based human activity recognition using Conv LSTM and guided feature learning. Soft Computing. 2022. pp. 1–14.
47. Zhu Q., Deng H. Spatial adaptive graph convolutional network for skeleton-based action recognition. Applied Intelligence. 2023. pp. 1–13.
48. Singh G., Choutas V., Saha S., Yu F., Van Gool L. Spatio-Temporal Action Detection under Large Motion. Proceedings of the IEEE/CVF Applications of Computer Vision. 2023. pp. 6009–6018.
49. Ahn D., Kim S., Hong H., Ko B.C. STAR-Transformer: A Spatio-temporal Cross Attention Transformer for Human Action Recognition. In Proceedings of the IEEE/CVF Applications of Computer Vision. 2023. pp. 3330–3339.
50. Peng K., Roitberg A., Yang K., Zhang J., Stiefelhagen R. Delving Deep into One-Shot Skeleton-based Action Recognition with Diverse Occlusions. IEEE Transactions on Multimedia. 2023. arXiv preprint arXiv:2202.11423v3.
51. Zhou J.T., Du J., Zhu H., Peng X., Liu Y., Goh R.S.M. Anomalynet: An anomaly detection network for video surveillance. IEEE Transactions on Information Forensics and Security. 2019. vol. 14(10). pp. 2537–2550.
52. Pawar K., Attar V. DL-based detection and localization of road accidents from traffic surveillance videos. ICT Express. 2022. vol. 8. no. 3. pp. 379–387.
53. Ganokratanaa T., Aramvith S., Sebe N. Video anomaly detection using deep residual-spatiotemporal translation network. Pattern Recognition Letters. 2022. vol. 155. pp. 143–150.
54. Roa'a M., Aljazaery I.A., ALRikabi H.T.S., Alaidi A.H.M. Automated Cheating Detection Based on Video Surveillance in the Examination Classes. iJIM. 2022. vol. 16(08). no. 125.
55. Kamoona A.M., Gostar A.K., Bab-Hadiashar A., Hoseinnezhad R. Multiple instance-based video anomaly detection using deep temporal encoding–decoding. Expert Systems with Applications. 2023. vol. 214. no. 119079. DOI: 10.1016/j.eswa.2022.119079.
56. Le V.T., Kim Y.G. Attention-based residual autoencoder for video anomaly detection. Applied Intelligence. 2023. vol. 53(3). pp. 3240–3254.
57. Abbas Z.K., Al-Ani A.A. An adaptive algorithm based on principal component analysis-DL for anomalous events detection. Indonesian Journal of Electrical Engineering and Computer Science. 2023. vol. 29(1). pp. 421–430.

58. Pazho A.D., Neff C., Noghre G.A., Ardabili B.R., Yao S., Baharani M., Tabkhi H. Ancilia: Scalable Intelligent Video Surveillance for the Artificial Intelligence of Things. 2023. arXiv preprint arXiv:2301.03561.
59. Thakare K.V., Raghuwanshi Y., Dogra D.P., Choi H., Kim I.J. DyAnNet: A Scene Dynamicity Guided Self-Trained Video Anomaly Detection Network. Proceedings of the IEEE/CVF Applications of Computer Vision. 2023. pp. 5541–5550.
60. Deng H., Zhang Z., Zou S., Li X. Bi-Directional Frame Interpolation for Unsupervised Video Anomaly Detection. In Proceedings of the IEEE/CVF Applications of Computer Vision. 2023. pp. 2634–2643.

**Суджата Гупта Нукала** — научный сотрудник, факультет компьютерных наук и инженерии, Образовательный фонд Конеру Лакшмайи. Область научных интересов: наука и техника. gsuji29@gmail.com; Зеленые поля, Ваддесварам, 522302, Гунтур, Андхра-Прадеш, Индия; р.т.: +91(8645)350-0200.

**Рамья К. Рут** — доцент, факультет компьютерных наук и инженерии, Образовательный фонд Конеру Лакшмайи. Область научных интересов: наука и техника. Число научных публикаций — 12. ramya_cse@kluniversity.in; Зеленые поля, Ваддесварам, 522302, Гунтур, Андхра-Прадеш, Индия; р.т.: +91(8645)350-0200.

**Карнати Рамеш** — доцент, факультет компьютерных наук и инженерии, Инженерный колледж Вардхамана. Область научных интересов: интеллектуальный анализ данных, машинное обучение, искусственный интеллект, интернет вещей. Число научных публикаций — 17. ramesh.krnt@vardhaman.org; Качарам, Шамшабад, 501218, Хайдарабад, Телангана, Индия; р.т.: +91(8688)901-557.

**УДК** 004.912 **DOI** 10.15622/ia.23.2.6

Д.Ю. Кравченко, Ю.А. Кравченко, А. Мансур, Ж. Мохаммад, Н.С. Павлов

# АЛГОРИТМ ОПТИМИЗАЦИИ ИЗВЛЕЧЕНИЯ КЛЮЧЕВЫХ СЛОВ НА ОСНОВЕ ПРИМЕНЕНИЯ ЛИНГВИСТИЧЕСКОГО ПАРСЕРА

*Кравченко Д.Ю., Кравченко Ю.А., Мансур А., Мохаммад Ж., Павлов Н.С.* **Алгоритм оптимизации извлечения ключевых слов на основе применения лингвистического парсера.**

**Аннотация.** В данной статье представлено аналитическое исследование особенностей двух типов парсинга, а именно синтаксический анализ составляющих (constituency parsing) и синтаксический анализ зависимостей (dependency parsing). Также в рамках проведенного исследования разработан алгоритм оптимизации извлечения ключевых слов, отличающийся применением функции извлечения именных фраз, предоставляемой парсером, для фильтрации неподходящих фраз. Алгоритм реализован с помощью трех разных парсеров: SpaCy, AllenNLP и Stazna. Эффективность предложенного алгоритма сравнивалась с двумя популярными методами (Yake, Rake) на наборе данных с английскими текстами. Результаты экспериментов показали, что предложенный алгоритм с парсером SpaCy превосходит другие алгоритмы извлечения ключевых слов с точки зрения точности и скорости. Для парсера AllenNLP и Stanza алгоритм так же отличается точностью, но требует гораздо большего времени выполнения. Полученные результаты позволяют более детально оценить преимущества и недостатки изучаемых в работе парсеров, а также определить направления дальнейших исследований. Время работы парсера SpaCy значительно меньше, чем у двух других парсеров, потому что парсеры, которые используют переходы, применяют детерминированный или машинно-обучаемый набор действий для пошагового построения дерева зависимостей. Они обычно работают быстрее и требуют меньше памяти по сравнению с парсерами, основанными на графах, что делает их более эффективными для анализа больших объемов текста. С другой стороны, AllenNLP и Stanza используют модели парсинга на основе графов, которые опираются на миллионы признаков, что ограничивает их способность к обобщению и замедляет скорость анализа по сравнению с парсерами на основе переходов. Задача достижения баланса между точностью и скоростью лингвистического парсера является открытой темой, требующей дальнейших исследований в связи с важностью данной проблемы для повышения эффективности текстового анализа, особенно в приложениях, требующих точности при работе в реальном масштабе времени. С этой целью авторы планируют проведение дальнейших исследований возможных решений для достижения такого баланса.

**Ключевые слова:** синтаксический анализ составляющих, синтаксический анализ зависимостей, извлечение ключевых слов, обработка естественного языка, NLP, SpaCy, Stanza, AllenNLP.

**1. Введение.** В динамичном мире обработки естественного языка (Natural Language Processing, NLP) синтаксический анализ (парсинг) играет ключевую роль в раскрытии сложностей естественного языка. Как основа к пониманию структуры и смысла предложений, парсеры служат незаменимыми инструментами

в различных задачах NLP, позволяя машинам воспринимать и обрабатывать естественный язык с более высокой точностью и эффективностью. От анализа настроений до машинного перевода, а также для систем вопросов и ответов, парсеры играют ключевую роль в преобразовании предложений в синтаксические структуры, что в свою очередь облегчает более точную и контекстно значимую обработку языка. Разбивая предложения на понятные единицы, парсеры создают фундамент для машинного понимания семантики и взаимосвязей между словами, делая возможным достижение более сложных и тонких результатов в различных приложениях [1 – 3], поэтому *создание эффективных текстовых парсеров является весьма актуальной научной проблемой в настоящее время*.

Основные подходы к синтаксическому парсингу включают синтаксический анализ составляющих (constituency parsing) и синтаксический анализ зависимостей (dependency parsing).

Анализ составляющих и зависимостей – это взаимодополняющие подходы, которые направлены на анализ синтаксической структуры предложений. Эти методы анализа предоставляют ценные сведения о грамматической структуре и семантических отношениях в предложении [3].

Анализ составляющих сосредоточен на определении конституентов, которые являются группами слов, выполняющими единую функцию в предложении. Эти конституенты могут быть фразами, такими как именные фразы (NP) или глагольные фразы (VP), или даже более крупными единицами, такими как предложения. Анализ составляющих представляет иерархическую схему предложения с использованием древовидной структуры, называемой деревом разбора или синтаксическим деревом. С другой стороны, анализ зависимостей сосредоточен на отношениях между отдельными словами в предложении. Он представляет эти отношения в виде направленных связей или зависимостей, где каждое слово связано со своим синтаксическим корневым или управляющим словом. Анализ зависимостей обеспечивает более линейное представление структуры предложения, акцентируя внимание на зависимостях между словами, а не на иерархической организации конституентов [4].

Одним из известных алгоритмов в конституентном анализе являются алгоритм CYK (Cocke-Younger-Kasami). Этот классический алгоритм разбора, основанный на динамическом программировании, эффективно строит дерево разбора, разбивая предложения на более мелкие конституенты с использованием контекстно-свободной грамматики. Так же известен алгоритм Эрли, который способен

обрабатывать неоднозначные грамматики и разбирать предложения с использованием предсказывающего сверху-вниз и снизу-вверх подхода, что приводит к более надежному процессу разбора [3].

С другой стороны, популярным алгоритмом анализа зависимостей является алгоритм Arc-Eager. Это алгоритм разбора на основе переходов (Transition-based), который предсказывает последовательность действий для построения дерева зависимостей, эффективно отображая отношения между корневыми и зависимыми словами. Другим подходом на основе переходов является алгоритм Arc-Standard, который строит деревья зависимостей, сводя предложение к однокоренному дереву с помощью серии действий [5].

Последние исследования в области обработки естественного языка открыли потенциал глубокого обучения для повышения эффективности синтаксического анализа зависимостей [5, 6]. Используя архитектуры нейронных сетей и обширные объемы размеченных данных, парсеры на основе глубокого обучения достигли значительного улучшения точности и эффективности. Существуют техники, включая парсинг на основе переходов с использованием нейронных сетей [5], которые улучшают традиционные парсеры путем интеграции нейронных сетей для более точного улавливания контекстуальных особенностей и зависимостей.

В [6] используют графовые нейронные сети (graph-based) для выполнения синтаксического анализа зависимостей, что позволяет более эффективно обрабатывать неявные зависимости и синтаксические структуры.

Было установлено, что глубоко контекстуализированные представления слов оказывают еще больший положительный эффект на transition-based парсинг, чем graph-based парсинг [7]. Информация о синтаксической структуре, содержащаяся в глубоко контекстуализированных представлениях слов, помогает смягчить главный недостаток transition-based алгоритмов в виде ошибок при обработке длинных предложений. Модели глубоко контекстуализированных представлений слов ELMo [2] и BERT [1] позволили значительно улучшить результаты для обоих алгоритмов парсинга, причем для transition-based парсинга улучшение оказалось более значимым.

Transition-based и graph-based подходы обладают взаимодополняющими преимуществами и недостатками. Несмотря на то, что transition-based и graph-based парсеры показывают примерно равную точность, они совершают ошибки разного рода. Transition-based парсеры чаще ошибаются в длинных предложениях,

зависимостях около корня дерева, зависимостях с глаголом и союзами, а также в определении корневого слова. Это связано с жадным алгоритмом, где ошибка в одной зависимости может привести к каскадным ошибкам в других зависимостях. С другой стороны, graph-based парсеры чаще допускают ошибки в коротких предложениях, зависимостях с существительными и местоимениями, а также в зависимостях вблизи листьев дерева. Это связано с ограниченным набором признаков. Таким образом, оба подхода имеют свои преимущества и недостатки, которые дополняют друг друга.

Чтобы подтвердить наблюдения, упомянутые выше, и с целью выяснения различий между этими двумя типами лингвистических анализаторов и изучения их влияния на производительность и скорость задач анализа текста, в данном исследовании представлено аналитическое сравнение нескольких известных лингвистических анализаторов, каждый из которых принадлежит к различным подходам: SpaCy [8], основанный на переходах (анализ зависимостей), и Stanza [9] (анализ зависимостей) и AllenNLP [10] (анализ составляющих), основанные на графах. Оценена скорость каждого парсера при анализе длинных текстов и коротких предложений. Также изучается влияние использования каждого из этих парсеров на задачу извлечения ключевых фраз, где разрабатывается алгоритм для извлечения ключевых фраз, использующий возможности этих технологий для определения именных фраз в тексте.

**2. Синтаксический парсинг составляющих.** Синтаксический анализ составляющих заключается в разбиении предложения на составные части (отдельные слова). Наиболее распространенной моделью, описывающей составную структуру предложения, является контекстно-свободная грамматика (context-free grammar).

Контекстно-свободная грамматика представляет собой набор правил, который определяет способы группировки и упорядочивания слов. Она получила свое название по причине того, что все порождающие правила в грамматике могут применяться независимо от контекста – они не зависят от каких-либо других языковых единиц, которые могут или не могут быть вокруг данной языковой единицы, к которой применяется правило.

Например, правила контекстно-свободной грамматики могут быть представлены следующим образом:

$$\text{Nominal} \rightarrow \text{Noun} \mid \text{Nominal Noun}, \quad (1)$$

$$\text{NP} \rightarrow Det\ \text{Nominal}, \quad (2)$$

$$\text{NP} \rightarrow \text{Proper Noun}. \quad (3)$$

Именная группа (Noun Phrase или NP) может быть составлена либо из определяющего слова (determiner или det) и следующего за ним одного или нескольких существительных (Nominal), что отражено в выражении 2, либо из имени собственного (Proper Noun) – в выражении 3.

Контекстно-свободная грамматика обладает иерархической структурой, то есть правила могут быть вложены друг в друга. Например, в правило 2 может быть вложено следующее правило:

$$Det \rightarrow a, \quad (4)$$

$$Det \rightarrow an, \quad (5)$$

$$Det \rightarrow the, \quad (6)$$

$$Noun \rightarrow cat. \quad (7)$$

В контекстно-свободной грамматике используются символы двух классов. Терминальные символы (terminal symbols) соответствуют словам (“a”, “good”, “dog” и т.д.) и не могут быть разделены на меньшие элементы. Вместе они составляют лексикон или словарь языка. Нетерминальные символы (non-terminal symbols) представляют собой группы или категории терминальных символов, таких как предложные группы (prepositional phrases), именные группы (noun phrases), глагольные группы (verb phrases) и т.д. Такие символы могут быть разложены на меньшие элементы, включая терминальные символы и другие нетерминальные символы. Таким образом, контекстно-свободная грамматика представляет собой генератор, преобразующий нетерминальные символы в строку символов.

Синтаксический анализ составляющих может быть представлен в виде дерева (рисунок 1).

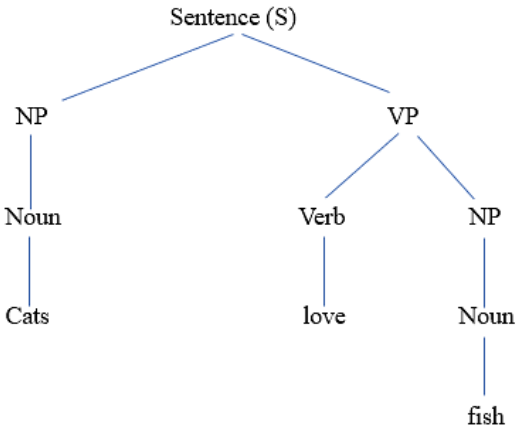



Рис. 1. Синтаксический анализ составляющих в виде дерева

Современные подходы конвертируют дерево в линейную форму, чтобы сделать возможным применение sequence-to-sequence моделей [4]. В линейном виде дерево выглядит следующим образом:

$$\big(\text{S (NP N)}\big(\text{VP V(NP N)}\big)\big). \quad (8)$$

Для применения алгоритмов парсинга составляющая контекстно-свободной грамматики должна быть приведена к нормальной форме Хомского. В нормальной форме Хомского правила грамматики могут принимать две следующие формы:

1. $A \rightarrow B\ C$, где $A$, $B$ и $C$ – нетерминальные символы;
2. $A \rightarrow a$, где $A$ – нетерминальный символ, $a$ – терминальный символ.

Ввиду данных правил, деревья, построенные из правил нормальной формы Хомского, будут бинарными. Потери информации при переходе к нормальной форме Хомского не происходит.

Любая контекстно-свободная грамматика может быть приведена к нормальной форме Хомского. Если правило контекстно-свободной грамматики содержит с правой стороны один нетерминальный символ и один терминальный, например, правило INF-VP → to VP (инфинитивная группа → to глагольная группа) может быть приведено к нормальной форме Хомского путем добавления нового нетерминального символа. Тогда правило INF-VP → to VP будет заменено на 2 правила: INF-VP → TO VP и TO → to.

Если правило контекстно-свободной грамматики содержит в правой части правила один нетерминальный символ, то оно преобразуется в правило с одним терминальным символом в правой части правила. Например, набор правил A → B, B → C, C → to преобразуется в A → to, устраняя тем самым избыточность информации.

Если в правой части правила более 2 символов, то оно преобразуется путем введения нового нетерминального символа. Например, правило A → B C a, где A, B, C – терминальные символы, a – нетерминальный, заменяется на 2 правила: A → X1 и X1 → B C.

**3. Синтаксический парсинг зависимостей.** Если парсинг составляющих основывается на контекстно-свободной грамматике, то парсинг зависимостей – на грамматике зависимостей (dependency grammar). Грамматика зависимостей представляет синтаксическую структуру предложения как совокупность направленных грамматических зависимостей между словами. В предложении выделяются главные и зависимые слова. Одно и то же слово в предложении может быть главным по отношению к одному слову и зависимым по отношению к другому. Корневым словом в предложении, которое не является зависимым по отношению ни к одному другому слову, является предикат, в роли которого, как правило, выступает глагол. На рисунке 2 отражена синтаксическая структура, предложения в соответствии с грамматикой зависимостей, где стрелки направлены от главных слов к зависимым.

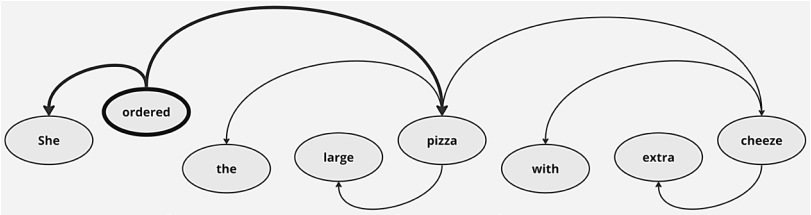



Рис. 2. Синтаксическая структура предложения в соответствии с грамматикой зависимостей

В форме синтаксического дерева данное предложение может быть представлено следующим образом (рисунок 3).

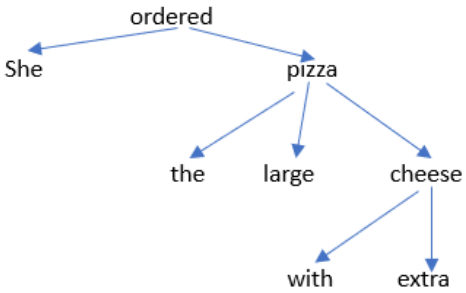



Рис. 3. Предложение в форме синтаксического дерева

Дерево зависимостей представляет собой ориентированный граф, удовлетворяющий следующим ограничениям:

1. Существует единственный определенный корневой узел, у которого нет входящих дуг.
2. За исключением корневого узла, каждая вершина имеет ровно одну входящую дугу.
3. Существует уникальный путь от корневого узла до каждой вершины в дереве зависимостей.

Такие ограничения обеспечивают связность структуры зависимостей, где у каждого слова не более одного главного слова и существует уникальный путь из корневого узла к каждому слову в предложении.

***3.1. Алгоритмы синтаксического парсинга зависимостей***. Основными алгоритмами парсинга зависимостей являются Transition-Based Dependency Parsing и Graph-Based Dependency Parsing.

Transition-Based Dependency Parsing основан на механизме shift-reduce. Алгоритм был предложен Ямада и Матцумото [11] и Нивре [12] на основе history-based parsing [13] и data-driven shift-reduce parsing [14]. Идея алгоритма заключается в сведении задачи парсинга к пошаговому прогнозированию наличия или отсутствия зависимости между двумя словами в предложении и направления выявленной зависимости. Парсер состоит из буфера входных токенов (слов), стека, предиктора и набора определенных зависимостей. В первоначальной конфигурации буфер входных токенов состоит из слов предложения в порядке, в котором они расположены в предложении, набор определенных зависимостей пуст, стек состоит из одного служебного элемента ROOT. Парсер обрабатывает предложение слева направо, последовательно сдвигая элементы из буфера в стек. На каждом шаге предиктор отправляет один токен из буфера в стек, анализирует два верхних элемента в стеке и принимает одно из следующих решений:

– назначить первое слово в стеке главным по отношению ко второму (левая дуга) и удалить второе слово из стека;

– если для верхнего слова в стеке уже назначены все зависимые слова, тогда назначить второе слово в стеке главным по отношению к первому слову (правая дуга) и удалить первое слово из стека;

– отложить обработку текущего слова, сдвинув его вниз по стеку.

Дополнительное условие для второго оператора (правая дуга) необходимо для того, чтобы слово не было извлечено из стека до того, как ему будут присвоены все его зависимые элементы.

Окончательное синтаксическое дерево будет составлено, когда буфер окажется пуст, а в стеке останется только служебный символ ROOT. Преимуществом алгоритма в сравнении с динамическими алгоритмами парсинга составляющих является его линейная сложность по длине предложения. Данный алгоритм представляет собой жадный алгоритм, так как предиктор делает один выбор на каждом шаге, данный выбор считается квазиоптимальным, повторно элементы не обрабатываются, и другие варианты построения зависимостей не рассматриваются. Некорректный выбор на одном шаге ведет к построению ошибочного дерева, без возможности вернуться назад и исправить ошибку. Кроме того, алгоритм возвращает только один вариант синтаксического дерева, в то время как, ввиду проблемы двусмысленности, возможно наличие более одного варианта корректных синтаксических деревьев.

Предиктор парсера может быть основан на классификаторе на основе признаков (classic feature-based algorithm) или на нейронном классификаторе. Алгоритм на основе признаков полагается на такие признаки, как форма слова, лемма, часть речи главного и зависимого слова; форма слова, лемма, часть речи для слов перед или между главным и зависимым словом; также учитывается состояние буфера входных токенов, стека и набора определенных зависимостей.

Признаки определяются вручную или с помощью обучения классификатора. Выбор признаков вручную несет в себе несколько проблем. Во-первых, слишком большое количество признаков может привести к переобучению и замедлению модели. Во-вторых, для корректного выбора признаков необходимы глубокие знания в области лингвистики [15].

Что касается классификатора, в последние годы произошел переход к нейронным классификаторам, который привел к значительному повышению точности предиктора [7]. Стандартный

алгоритм состоит в следующем: предложение проходит через энкодер, затем векторные представления двух первых слов из стека и первого слова из буфера конкатенируются и подаются в нейронную сеть прямого распространения. Кроме того, был разработан нетерпеливый (arc eager) transition-based алгоритм, который использует парсер SpaCy [5]. Главным отличием arc eager алгоритма от стандартного заключается в применении операторов к первому слову в стеке и первому слову в буфере. Это позволяет избавиться от условия предварительного назначения всех зависимых слов для зависимого слова в операторе «правая дуга», так как в данном алгоритме это зависимое слово вместо того, чтобы быть удаленным отправиться в стек и будет доступно для дальнейшей обработки. Такое изменение позволяет быстрее назначать зависимости слева направо (правая дуга) и ускоряет работу парсера. Для корректной работы парсера добавлен оператор «сокращение/reduce», необходимый для завершения процедуры парсинга в случае, если входной буфер оказался пуст.

Для повышения точности transition-based парсинга он может быть дополнен алгоритмом лучевого поиска (beam-search algorithm) [2], что смягчает проблему жадности у такого типа инструментов. Идея состоит в том, что вместо того, чтобы выбирать единственный оператор на каждом шаге, выбираются все операторы на каждом шаге, а затем все полученные частичные деревья оцениваются классификатором. Оценка каждого следующего дерева в одной последовательности рассчитывается как сумма оценки предшествующего дерева и оценки примененного к нему оператора:

$$TScore(i) = TScore(i-1) + OpScore(i-1), \quad (9)$$

где *TScore* – оценка дерева, *OpScore* – оценка оператора.

Количество деревьев ограничивается предустановленным лимитом – шириной луча. Чем шире луч, тем выше точность и ниже скорость. Когда ширина луча достигает лимита, новые деревья добавляются вместо худших, если оценка нового дерева выше оценки худшего дерева в луче. Процесс парсинга завершается, когда луч содержит только полные деревья входного предложения. Синтаксический анализатор выбирает из луча дерево с наивысшей оценкой и возвращает его в качестве окончательного вывода. Таким образом, парсеру не приходится принимать окончательные решения слишком рано, есть возможность вернуться на начальные этапы построения дерева и исправить ошибку, что значительно повышает точность парсера.

***3.2. Алгоритм парсинга зависимостей на основе графа (graph-based parsing).*** Данный алгоритм разработан Макдональдом [16] на основе работы Эйснера [17]. В отличие от transition-based парсинга, полагающегося на жадные локальные решения, graph-based парсинг основан на оценке полного синтаксического дерева. Идея graph-based алгоритма заключается в представлении пространства возможных синтаксических деревьев в виде ориентированного графа (вершинами которого являются слова, а направленными ребрами – зависимости) и поиске в этом графе дерева с наилучшей оценкой. Общая оценка каждого дерева вычисляется как сумма весов отдельных зависимостей, из которых оно состоит. В результате, graph-based алгоритм рассматривает и оценивает все возможные зависимости в предложении, что является предпосылкой для более высокой точности по сравнению с transition-based парсингом.

Таким образом, нахождение наилучшего дерева зависимостей сводится к нахождению максимального остовного дерева, которое представляет собой подграф с максимальной суммой весов ребер, содержащий все вершины исходного графа и корневую вершину ROOT.

Так же, как и при transition-based парсинге, оценка зависимостей и полного дерева, как суммы оценок зависимостей, может производиться классификатором на основе признаков (classic feature-based algorithm) или на нейронном классификаторе. В feature-based алгоритме оценка зависимости (ребра графа) вычисляется как взвешенная сумма признаков:

$$\text{Score(S,e)} = \sum_{i=1}^{N} w_i f_i(S, e), \tag{10}$$

где $S$ – предложение, $e$ – зависимость (ребро), $w$ – веса, $f$ – признак, $N$ – количество признаков.

Главной задачей является выявление релевантных признаков и их комбинаций. Могут использоваться следующие признаки: форма слова, лемма, часть речи главного и зависимого слова; расстояние между главным и зависимым словом, направление связи (слева направо или справа налево), векторные представления слов и т.д. По сравнению с transition-based парсингом, для graph-based парсинга возможен ограниченный набор признаков, так как алгоритм рассматривает признаки только самой пары слов (рассматриваемых

как потенциальная зависимость) и игнорирует признаки других слов в предложении, упуская тем самым глобальный контекст.

Нейронные классификаторы показывают более высокую точность. Предложение подается в энкодер, где для каждого токена строится глубокое контекстуализированное векторное представление. Ряд исследователей установили, что такие представления содержат информацию о синтаксической структуре предложения [18, 19]. Затем полученные представления передается в нейронную сеть, которая присваивает оценки каждой зависимости.

Дозат и Мэннинг предложили архитектуру нейронной сети, в которой использовали биафинное внимание (biaffine attention) вместо стандартного билинейного внимания [20]. В такой сети на вход подаются последовательность токенов, конкатенированных с тегами их частей речи:

$$x_i = v_i(word) \oplus v_i(tag), \tag{11}$$

где $x_i$ – конкатенированный вектор, $v_i^{(word)}$ – представление токена, $v_i^{(tag)}$ – представление тега части речи токена.

Затем они обрабатываются энкодером в виде многослойной двунаправленной сети долгой краткосрочной памяти (LSTM):

$$r_i = \text{BiLSTM}(r_0, (x_1, \ldots, x_n))_i, \tag{12}$$

где $r_i$ – конечное состояние, $r_0$ – первоначальное состояние, $(x_1,...,x_n)$ – конкатенированные вектора. Такой энкодер отражает глобальный контекст в локальных представлениях слов, что расширяет набор признаков за пределы непосредственно главного и зависимого слова, смягчая тем самым главный недостаток graph-based парсинга.

**4. Постановка задачи извлечения ключевых слов с помощью парсера.** Извлечение ключевых слов является фундаментальной задачей в обработке естественного языка (Natural Language Processing, NLP) и включает в себя выявление и извлечение наиболее релевантных и значимых слов или фраз из заданного текста. Парсеры играют ключевую роль в этом процессе, анализируя синтаксическую структуру текста и помогая выявлять ключевые компоненты, представляющие важные понятия или темы.

1. *Синтаксический анализ структуры для извлечения ключевых слов*. Для выполнения извлечения ключевых слов с использованием синтаксического анализа структуры, можно

сосредоточиться на конкретных синтаксических единицах, таких как именные фразы (NPs) [21] и глагольные фразы (VPs). Эти фразы часто являются хорошими кандидатами на роль ключевых слов, так как обычно содержат важную информацию о субъекте, объекте или действии в предложении. Например, рассмотрим предложение: "The swift fox jumps over the lazy dog." С помощью синтаксического анализа структуры будут выделены следующие фразы:

– Именные фразы (NPs): "The swift fox", "the lazy dog";
– Глагольная фраза (VP): "jumps over".

Из извлеченных фраз можно выделить следующие ключевые слова: "swift fox" (быстрая лиса); "lazy dog" (ленивая собака); "jumps over" (перепрыгивает через), как существенные компоненты данного предложения.

2. *Синтаксический анализ зависимостей для извлечения ключевых слов*. Синтаксический анализ зависимостей помогает определить отношения между субъектом, глаголом и объектом, а также другие существенные синтаксические зависимости, которые вносят вклад в общий смысл предложения. Ключевые слова могут быть извлечены из этих зависимостей, причем предпочтение отдается словам, несущим значительный семантический вес и играющим важные роли в структуре предложения [1].

В контексте извлечения ключевых слов текст преобразуется в граф, где вершины представляют собой возможные ключевые слова, а ребра – их отношения. Взаимосвязь между ключевыми фразами-кандидатами может быть определена по тому, как часто они встречаются вместе или насколько семантически близки.

Предположим, что строится ориентированный граф $G = (V, E)$, где – $V$ множество вершин, а $E$ – множество ребер. Оценка или важность вершины определяется как:

$$S(V_i) = (1 - d) + d * \sum_{j \in In(V_i)} \frac{1}{\left|Out(V_j)\right|} S(V_j), \quad (13)$$

где $In(V_i)$ – это набор вершин, которые указывают на $V_i$, а $Out(V_i)$ – это набор вершин, на которые указывает $V_i$. При этом $d$ – это коэффициент затухания, который устанавливается в диапазоне от 0 до 1.

**5. Разработка алгоритма оптимизации извлечения ключевых слов на основе парсера.** Исходя из предположения, что ключевое слово (ключевые фразы) обычно является существительным

или именной фразой, предлагается алгоритм оптимизации для извлечения ключевых слов, характеризующийся использованием парсера для фильтрации неподходящих фраз, которые не являются именными фразами. Цель состоит в том, чтобы проверить влияние парсера на производительность алгоритма извлечения ключевых слов.

В сформулированном листинге 1 шаги алгоритма описаны в виде псевдокода. Входными данными алгоритма является текст, а выходными данными является окончательный список ключевых слов.

Сначала функция *preprocessing* обрабатывает текст для удаления ненужных символов, знаков препинания, стоп-слов и переводов текста в нижний регистр. После этого слова-кандидаты, *n*-граммы (длиной от 2 до 4), извлекаются с помощью функции *extract_ngrams*, которая применяет весовую функцию *TF*, определенную по следующей формуле:

$$TF = \frac{n_t^i}{\sum_k n_k^i}, \tag{14}$$

где $n_k$ – общее количество *n*-грамм (с $i$ элементов) в документе, а $n_t$ – количество вхождений токена *t* в документ.

После этого текст необходимо разделить на предложения, а затем каждое предложение анализируется на компоненты для извлечения именных фраз *NP*. Для этого весь текст анализируется парсером, который, в свою очередь, возвращает предложения, составляющие текст, а также компоненты и атрибуты каждого предложения включая именных фраз.

Именные фразы в их исходной форме не подходят в качестве ключевых слов, так как они могут содержать артикли или быть слишком длинными, поэтому они обрабатываются, опуская артикли, длинные словосочетания и фразы, которые не начинаются с существительного или прилагательного (листинг 1, строки 10-20).

В итоге получается список именных фраз на уровне текста, которые используются в качестве фильтра (строки 21–32 в листинге). Если ключевая фраза-кандидат соответствует одной из именных фраз, она сохраняется. Если именная фраза является частью фразы-кандидата, в окончательном списке сохраняется именная фраза, поскольку именная фраза имеет более полную структуру.

```
1   Ввод текст
2   Вывод список ключевых фраз KWS
3:  T* = preprocessing(T) //удаление ненужных символов, знаков
      препинания, стоп-слов и перевод текста в нижний регистр
4:  Candidates = extract_ngrams(T*, (2,4)) // формула 14
5:  S = splitter(T*) // Разделить текст на список предложения S
6:  NP = [] // Пустой список для хранения именных фраз NP
7:  Foreach sentence in S do:
8:      NP ← get_noun_phrase(sentence) //Разбор предложения с
          помощью парсера и сохранение именной фразы в списке NP
9:  End
10: // преобразования именных фраз в ключевые слова
11: FNP = [] //Список для хранения отфильтрованных именных фраз.
12: Foreach np in NP do
13:     if len(np) == 1:
14:         if np[0].pos_ in ['NOUN', 'PROPN', 'ADJ']: // первое слово np
            — существительное, имя собственное или прилагательное
15:             FNP.append(np.text.strip())
16:     elif 1 <len(np)< 6: // если длина np не превышает 6 слов
17:         if np[0].pos_ in ['DET']: //первый термин - это артикль
18:             FNP.append(np[1:].text) // удалить артикль
19:         Else
20:             FNP.append(np.text.strip())
21:  // Проверка совпадения именной фразы с кандидатами на n-граммы
22: Foreach candidate in candidates do:
23:     If candidate in FNP do
24:         KWS []← candidate // добавить кандидата в список
25:     Else // проверка наличия общих термов между кандидатом и
          именной фразой
26:         Foreach np in FNP do
27:                 If np ∩ candidate ≠ ∅ do:
28:                     KWS []← np // добавить np в список
29:                 End
30:         End
31:     End
32: End
```

Листинг 1. Псевдокод алгоритма оптимизации извлечения и фильтрации ключевых слов с помощью парсеров

Этот процесс способствует повышению точности алгоритма извлечения ключевых слов, поскольку он гарантирует, что слова-кандидаты представляют собой не просто последовательность слов, а составные слова в однородном контексте. Далее, в экспериментальных исследованиях проводятся тесты для подтверждения этого утверждения.

**6. Экспериментальные исследования**. Разработанный алгоритм извлечения ключевых слов реализован с использованием трех популярных парсеров, а именно:

SpaCy – это бесплатная библиотека с открытым исходным кодом для расширенной обработки естественного языка (NLP) в Python.

Парсер SpaCy – это компонент библиотеки SpaCy, который отвечает за анализ грамматической структуры предложений. Реализация парсера SpaCy выполняет синтаксический анализ зависимостей и включает в себя методы машинного обучения для прогнозирования в процессе разбора. Он использует статистическую модель, обученную на размеченных данных, для предсказания наиболее вероятного перехода на каждом шаге. Парсер использует вариант немонотонной дуговой системы с переходами [22], с добавлением перехода "break" для выполнения сегментации предложений. Для возможности предсказания неверных разборов парсер использует псевдопроективное преобразование зависимостей, предложенное в [23].

AllenNLP [10] – это платформа для исследований в области глубокого обучения методов обработки естественного языка. AllenNLP не имеет встроенного парсера, подобного SpaCy, однако предоставляет предварительно обученные модели для синтаксического парсинга составляющих и парсинга зависимостей. Парсер зависимостей следует модели глубокого биаффинного внимания для нейронного синтаксического анализа зависимостей, который использует нейронное внимание в простом графовом парсере зависимостей [20]. Парсер составляющих построен на основе минимальной нейронной модели, основанной на независимой оценке меток. Модель использует встроенную процедуру ELMo и кодирует последовательность текста с помощью стекового Seq2SeqEncoder.

Stanza [9] – это пакет для анализа естественного языка, написанный на Python, созданный Стэнфордской группой NLP. Он предоставляет несколько инструментов для анализа естественного текста, одним из которых является предоставление синтаксической структуры в виде дерева зависимостей. Как и в AllenNLP парсер

зависимостей в Stanza реализует Bi-LSTM-сеть, основанную на глубокой биаффинной нейронной модели, которая относится к категории парсеров зависимостей на основе графов [20].

Предложенный авторами алгоритм сравнивается с двумя известными методами, Yake и Rake.

Метод Yake (Yet Another Keyword Extractor) [24] использует комбинацию статистических и лингвистических признаков для определения важности слов или фраз в тексте. Он учитывает как частоту и распределение терминов внутри документа, так и их семантическую связь. Метод использует подход скользящего окна для выявления кандидатов на ключевые фразы, а затем применяет функцию оценки для их ранжирования по степени релевантности.

Метод Rake (Rapid Automatic Keyword Extraction) [25] использует простой алгоритм, который опирается на распределение слов в тексте для определения фраз кандидатов. Сначала текст разбивается на отдельные слова, а затем генерируются фразы-кандидаты, идентифицируя последовательности слов, разделенные стоп-словами или знаками пунктуации. Алгоритм присваивает оценки этим фразам на основе их частоты и степени совместного встречаемости слов.

Для Yake из каждого текста извлекаются первые 20 *n*-грамм (от 1 до 3). То же самое и с Rake, но Rake не позволяет указать количество извлекаемых слов. Не выполнялась никакая предварительная обработка текстов, чтобы не повлиять на структуру, но после получения именных фраз выполняется некоторая обработка текста, например, такая как удаление стоп-слов.

*Набор данных* Inspec [26] состоит из 2000 рефератов статей из научных журналов по компьютерным наукам. Каждому документу присвоены два набора ключевых слов: контролируемые ключевые слова, которые являются назначенными вручную ключевыми словами и появляются в тезаурусе Inspec, но могут не появляться в документе. Неконтролируемые ключевые слова свободно назначаются редакторами.

Для оценки результатов отслеживалось точное совпадение, при котором автоматически извлеченная ключевая фраза из документа должна точно соответствовать ключевой фразе эталонного паттерна для документа. Традиционно извлечение ключевых слов по своей природе является проблемой ранжирования. Исходя из этого, для определения эффективности предлагаемого алгоритма используется одна из наиболее часто используемых мер качества для ранжирования Средняя точность на *K* элементах «*Mean average precision at K,*

*MAP@K*». Таким образом, совпадения проверяются среди первых *K* ключевых слов, возвращаемых алгоритмом. Для оценки *MAP@K* сначала подсчитывается точность на *K* элементах p@k. Базовая метрика качества ранжирования для одного объекта определяется следующим выражением:

$$pr_k = \frac{\sum_{k=1}^{K} r_{true}(e_i)}{k}, \tag{15}$$

где $e_i$ это элемент (ключевое слово) $e \in E$, который в результате перестановки оказался на *i*-ой позиции, $r_{true}(e_i)$ – функция равная 1, если *e* релевантен, 0 – в противном случае. Недостаток этой метрики заключается в том, что не учитываются позиции правильных элементов (порядок элементов). Далее на основе $pr_K$ рассчитывается average precision at K ($apr_K$):

$$apr_k = \frac{1}{k}\sum_{k=1}^{K} r_{true}\pi^{-1}(k) \cdot pr_k. \tag{16}$$

Такая мера учитывает позиции элементов, но качество ранжирования оценивается для отдельно взятого объекта. Для того чтобы посчитать *Mean average precision at K* для *N* различных объектов вычисляется среднее по apr@K для каждого:

$$map_k = \frac{1}{N}\sum_{i=1}^{N} apr_k^i, \tag{17}$$

где $apr_K^i$ – это $apr_K$ для *i*-го объекта. Под MAP подразумевается Mean average precision по всем объектами и всем элементам.

**7. Результаты экспериментальных исследований.** Результаты сравнительного анализа алгоритмов извлечения ключевых фраз представлены в таблицах 1 и 2. Сравнивается производительность предложенного алгоритма при использовании трех разных парсеров с производительностью двух исследованных алгоритмов Yake и Rake.

Разработанный алгоритм (с использованием любого парсера) превосходит методы Yake и Rake, подтверждая утверждение о том, что большинство ключевых слов являются именными фразами. Наилучшие результаты достигаются с парсером SpaCy, как с точки зрения точности, так и с точки зрения скорости.

Таблица 1. Точность алгоритмов, измеренная с помощью mAP@K

| Метод | map@k | | | |
|---|---|---|---|---|
| K | @1 | @5 | @10 | @20 |
| TF-SpaCy | **0.36** | **0.15** | **0.095** | 0.077 |
| TF-Stanza | 0.303 | 0.131 | 0.083 | 0.06 |
| TF-AllenNLP | 0,299 | 0,128 | 0,081 | 0,064 |
| Yake | 0.267 | 0.127 | 0.084 | **0.0775** |
| Rake | 0.169 | 0.103 | 0.081 | 0.074 |

Результаты Yake и Rake приближается к показателям TF-SpaCy при определении правильных ключевых слов в топ-20 ($K = 20$).

Результаты проведенного исследования проиллюстрированы в таблице 2. В строке “Gold keywords” выделены полужирным шрифтом ключевые слова, определенные из текста экспертами. Что касается слов вне текста, то в данной работе они не учитываются.

Из таблицы 2 видно, что разработанный алгоритм оптимизации определения ключевых слов находит 4 правильных слова из 5 с помощью SpaCy, в то время как AllenNLP определяет только 2. Stanza правильно определяет “Bar code labels”, в то время как SpaCy определяет часть этого как “code labels”. AllenNLP не распознает “food processing” и “Bar code labels”, потому что он рассматривает "Fresh tracks [food processing] Bar code labels" как одну фразу на основе существительного.

Таблица 2. Текст из набора данных в качестве примера, показывающего эталонные ключевые слова и ключевые слова, определенные каждым алгоритмом

| | |
|---|---|
| Текст | Fresh tracks [food processing] Bar code labels and wireless terminals linked to a centralized database accurately track meat products from receiving to customers for Farmland Foods |
| Gold keywords | [**'food processing'**, **'bar code labels', 'wireless terminals'**, **'Farmland Foods'**, 'automatic data capture', 'Intermec Technologies', **'bar codes'**, 'data acquisition', 'food processing industry', 'mobile computing', 'production control', ''] |
| TF-SpaCy | ['Fresh tracks', **'food processing'**, **'code labels'**, **'wireless terminals'**, 'centralized database', 'meat products', **'Farmland Foods'**] |
| TF-AllenNLP | [**'wireless terminals'**, **'Farmland Foods',** 'centralized database', 'Fresh tracks', 'meat products'] |
| TF-Stanza | [**'wireless terminals'**, **'Farmland Foods', 'food processing'**, 'customers for Farmland Foods', 'Fresh tracks', 'meat products', 'centralized database', **'Bar code labels'**] |
| Yake | [**'Bar code labels'**, 'wireless terminals linked', 'centralized database accurately', 'database accurately track', 'accurately track meat', 'track meat products', **'Bar code'**, **'Farmland Foods'**, 'customers for Farmland', **'food processing'**] |
| Rake | ['wireless terminals linked', **'bar code labels'**, 'fresh tracks', **'food processing'**, **'farmland foods'**, 'receiving', 'customers'] |

Алгоритмы Yake и Rake неправильно определили глагол «linked» как часть ключевого слова «wireless terminals», тогда как трем парсерам удалось идентифицировать его правильно и исключить из фразы.

Хотя SpaCy превосходит другие парсеры в задаче определения именных фраз, нельзя однозначно утверждать, что парсеры, основанные на переходах, всегда превосходят парсеры, основанные на графах. Это зависит от нескольких факторов, включая структуру и сложность анализируемого текста, эффективность реализации парсера и его внутренних алгоритмов.

Парсеры на основе переходов постепенно строят дерево и принимают решения о разборе на основе локального контекста, что может помочь распознавать и извлекать именные фразы, имеющие явные зависимости в пределах короткого расстояния. Таким образом, поскольку текст, использованный в этих экспериментах, относительно короткий, он содержит простые предложения, в которых преобладают локальные зависимости, подходящие для данного типа синтаксического анализатора. С другой стороны, парсеры на основе графов учитывают всю структуру предложения при построении дерева зависимостей. Этот глобальный контекст может быть полезен для определения неявных зависимостей между словами, которые охватывают несколько слов или предложений, как в примере AllenNLP в таблице 2.

*Анализ времени выполнения.* Рисунки 4 и 5 и таблица 3 демонстрируют различие в скорости работы предложенного алгоритма по сравнению с методами Yake и Rake при обработке разного количества документов. Поскольку время выполнения алгоритмов Rake, Yake и SpaCy очень мало по сравнению с остальными алгоритмами, для наглядности их временные характеристики продемонстрированы в другом масштабе (рисунок 5).

Время выполнения предложенного авторами алгоритма представляет собой сумму времен выполнения нескольких операций (листинг 1): время извлечения *n*-грамм в качестве ключевых слов-кандидатов; время анализа текста на компоненты предложений и именных фраз; время фильтрации этих именных фраз; время фильтрации ключевых слов-кандидатов по именным фразам.

Таблица 3. Анализ времени выполнения алгоритмов

| | Время выполнения в секундах | | | | | | |
|---|---|---|---|---|---|---|---|
| Кол-во документов | 10 | 20 | 30 | 40 | 50 | 100 | 2000 |
| TF-SpaCy | 1 | 1,33 | 2,2 | 3,58 | 3,56 | 8 | 102 |
| TF-AllenNLP | 71 | 80 | 147 | 222 | 295 | 481 | 10614 |
| TF-Stanza | 223 | 414 | 585 | 820 | 1071 | 1954 | 14625 |
| Yake | 2,39 | 3,08 | 3,61 | 5,6 | 9 | 14 | 207 |
| Rake | 0,012 | 0,035 | 0,047 | 0,056 | 0,071 | 0,094 | 2 sec |

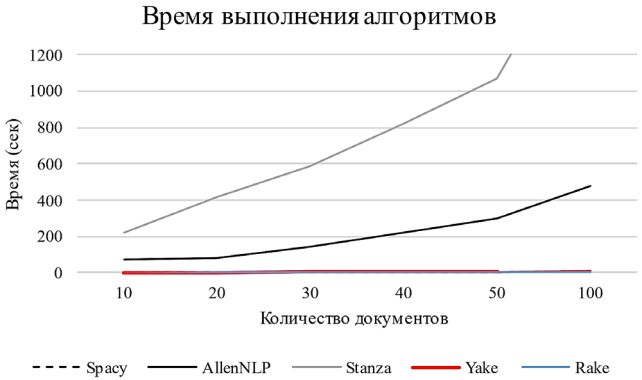



Рис. 4. Сравнение времени выполнения алгоритмов извлечения ключевых слов

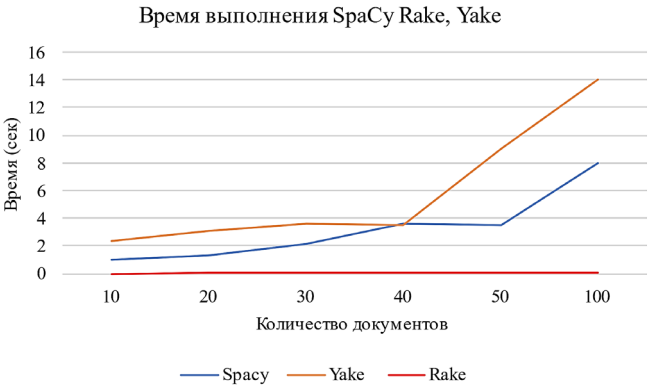



Рис. 5. Сравнение времени выполнения алгоритмов SpaCy, Yake и Rake

Хотя статистические алгоритмы работают быстрее, чем алгоритмы на основе парсера, поскольку их работа не требует анализа текста, предложенный в данной статье алгоритм с парсером Spacy (TF-SpaCy) превзошёл все рассмотренные канонические алгоритмы кроме Rake.

Время выполнения предложенного алгоритма с парсером SpaCy значительно меньше, чем у двух других парсеров, потому что парсеры, которые используют переходы, применяют детерминированный или машинно-обучаемый набор действий для пошагового построения дерева зависимостей. Они обычно работают быстрее и требуют меньше памяти по сравнению с парсерами, основанными на графах, что делает их более эффективными для анализа больших объемов

текста. С другой стороны, AllenNLP и Stanza используют модели парсинга на основе графов, которые полагаются на миллионы созданных вручную признаков, что ограничивает их способность к обобщению и замедляет скорость анализа по сравнению с парсерами на основе переходов.

Задача достижения баланса между точностью и скоростью лингвистического парсера является открытой темой, требующей дальнейших исследований в связи с важностью данной проблемы для повышения эффективности текстового анализа, особенно в приложениях, требующих точности при работе в реальном масштабе времени. С этой целью авторы планируют проведение дальнейших исследований возможных решений для достижения такого баланса.

Планируется рассмотреть дистилляцию нейронных сетей для быстрого разбора зависимостей, а также использование атомарных признаков, таких как униграммы слов и униграммы POS-тегов, вместо использования большого количества признаков, созданных вручную. Кроме того, планируется рассмотреть интеграцию подходов на основе графов и переходов с использованием преимуществ глубокого контекстного представления.

**Заключение.** В данной статье исследованы два основных типа лингвистического анализа: синтаксический анализ составляющих и анализ зависимостей. Представлено аналитическое сравнение нескольких известных парсеров (SpaCy, Stanza, AllenNLP), принадлежащих к различным подходам обработки текстовой информации. Также был разработан алгоритм извлечения ключевых слов на основе частоты, отличающийся применением функции извлечения именных фраз, предоставляемой парсером, для извлечения уточненного списка именной фразы. Этот список используется в качестве фильтра для отсеивания неподходящих ключевых слов-кандидатов, извлеченных на основе частоты, что позволяет повысить точность извлечения ключевых слов.

Множество экспериментов проведено с целью изучения влияния использования парсера на время и эффективность извлечения ключевых фраз по сравнению с двумя известными методами YAKE и RAKE. Результаты экспериментов подтвердили тот факт, что использование парсера существенно повышает точность алгоритма извлечения ключевых слов. Также отмечается, что эффективность зависит от типа парсера, контекста и длины обрабатываемого текста.

В данной работе основное внимание уделялось производительности парсера в качестве инструмента для извлечения ключевых фраз без углубления в анализ внутренних алгоритмов. Это

не дает полной картины различий в производительности анализаторов, а лишь показывает различия с точки зрения их применения. В будущем требуется провести дополнительный анализ данных алгоритмов с целью поиска решений проблемы снижения их временной сложности.

## Литература


1. Brown T., Mann B., Ryder N., Subbiah M., Kaplan J.D., Dhariwal P., Neelakantan A., Shyam P., Sastry G., Askell A., et al. Language models are few-shot learners // Advances in neural information processing systems. 2020. vol. 33. pp. 1877–1901.
2. Zhang Y., Clark S. A tale of two parsers: Investigating and combining graph-based and transition-based dependency parsing // Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing. 2008. pp. 562–571.
3. Gao L., Madaan A., Zhou S., Alon U., Liu P., Yang Y., Callan J., Neubig G. Pal: Program aided language models. 2023. pp. 10764–10799.
4. Kravchenko Yu.A., Bova V.V., Kuliev E.V., Rodzin S.I. Simulation of the semantic network of knowledge representation in intelligent assistant systems based on ontological approach // Futuristic Trends in Network and Communication Technologies: Third International Conference, FTNCT. 2021. pp. 241–252.
5. Chen D., Manning C.D. A fast and accurate dependency parser using neural networks // Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014. pp. 740–750.
6. Kiperwasser E., Goldberg Y. Simple and accurate dependency parsing using bidirectional LSTM feature representations // Transactions of the Association for Computational Linguistics. 2016. vol. 4. pp. 313–327.
7. Kulmizev A., de Lhoneux M., Gontrum J., Fano E., Nivre J. Deep Contextualized Word Embeddings in Transition-Based and Graph-Based Dependency Parsing – A Tale of Two Parsers Revisited // arXiv preprint arXiv: 07397. 2019.
8. Vasiliev Y. Natural language processing with Python and SpaCy: A practical introduction. No Starch Press, 2020. 216 p.
9. Qi P., Zhang Y., Zhang Y., Bolton J., Manning C.D. Stanza: A Python natural language processing toolkit for many human languages // arXiv preprint arXiv: 07082. 2020.
10. Gardner M., Grus J., Neumann M., Tafjord O., Dasigi P., Liu N., Peters M., Schmitz M., Zettlemoyer L. Allennlp: A deep semantic natural language processing platform // arXiv preprint arXiv: 07640. 2018.
11. Yamada H., Matsumoto Y. Statistical dependency analysis with support vector machines // Proceedings of the eighth international conference on parsing technologies. 2003. pp. 195–206.
12. Nivre J. An efficient algorithm for projective dependency parsing // Proceedings of the eighth international conference on parsing technologies. 2003. pp. 149–160.
13. Kim G., Baldi P., McAleer S. Language models can solve computer tasks. arXiv preprint arXiv:2303.17491. 2023.
14. Liu B., Jiang Y., Zhang X., Liu Q., Zhang S., Biswas J., Stone P. Llm+p: Empowering large language models with optimal planning proficiency. arXiv preprint arXiv:2304.11477. 2023.
15. Pei W., Ge T., Chang B. An effective neural network model for graph-based dependency parsing // Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. 2015. vol. 1. pp. 313–322.

16. McDonald R., Crammer K., Pereira F. Online large-margin training of dependency parsers // Proceedings of the 43rd annual meeting of the association for computational linguistics (ACL'05). 2005. pp. 91–98.
17. Eisner J. Three new probabilistic models for dependency parsing: An exploration // arXiv preprint cmp-lg/ 9706003. 1997.
18. Tenney I., Das D., Pavlick E. BERT rediscovers the classical NLP pipeline // arXiv preprint arXiv: 05950. 2019.
19. Hewitt J., Manning C.D. A structural probe for finding syntax in word representations // Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019. vol. 1. pp. 4129–4138.
20. Dozat T., Manning C.D. Deep biaffine attention for neural dependency parsing // arXiv preprint arXiv: 01734. 2016.
21. Mao X., Huang S., Li R., Shen L. Automatic keywords extraction based on co-occurrence and semantic relationships between words // IEEE Access. 2020. vol. 8. pp. 117528–117538.
22. Yang S., Nachum O., Du Y., Wei J., Abbeel P., Schuurmans D. Foundation models for decision making: Problems, methods, and opportunities. arXiv preprint arXiv:2303.04129. 2023.
23. Honnibal M., Johnson M. An Improved Non-monotonic Transition System for Dependency Parsing. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing – Lisbon, Portugal: Association for Computational Linguistics. 2015. pp. 1373–1378. DOI: 10.18653/v1/D15-1162.
24. Campos R., Mangaravite V., Pasquali A., Jorge A., Nunes C., Jatowt A. YAKE! Keyword extraction from single documents using multiple local features // Information Sciences. 2020. vol. 509. pp. 257–289.
25. Rose S., Engel D., Cramer N., Cowley W. Automatic keyword extraction from individual documents // Text mining: applications theory. 2010. pp. 1–20.
26. Hulth A. Improved automatic keyword extraction given more linguistic knowledge // Proceedings of the 2003 conference on Empirical methods in natural language processing. 2003. pp. 216–223.

**Кравченко Даниил Юрьевич** — аспирант, кафедра систем автоматизированного проектирования, Федеральное государственное автономное образовательное учреждение высшего образования «Южный федеральный университет». Область научных интересов: технологии искусственного интеллекта, инженерия знаний. Число научных публикаций — 40. dkravchenko@sfedu.ru; переулок Некрасовский, 44, 347922, Таганрог, Россия; р.т.: 8(8634)371-651.

**Кравченко Юрий Алексеевич** — профессор, кафедра систем автоматизированного проектирования, Федеральное государственное автономное образовательное учреждение высшего образования «Южный федеральный университет». Область научных интересов: технологии искусственного интеллекта, инженерия знаний. Число научных публикаций — 240. krav-jura@yandex.ru; переулок Некрасовский, 44, 347922, Таганрог, Россия; р.т.: 8(8634)371-651.

**Мансур Али** — аспирант, кафедра систем автоматизированного проектирования, Федеральное государственное автономное образовательное учреждение высшего образования «Южный федеральный университет». Область научных интересов: технологии искусственного интеллекта, инженерия знаний. Число научных публикаций — 15. mansur@sfedu.ru; переулок Некрасовский, 44, 347922, Таганрог, Россия; р.т.: 8(9880)158-697.

**Мохаммад Жуман** — аспирант, кафедра систем автоматизированного проектирования, Федеральное государственное автономное образовательное учреждение высшего образования «Южный федеральный университет». Область научных интересов: технологии искусственного интеллекта, инженерия знаний. Число научных публикаций — 14. zmohammad@sfedu.ru; переулок Некрасовский, 44, 347922, Таганрог, Россия; р.т.: +7(918)543-3526.

**Павлов Николай Сергеевич** — аспирант, кафедра систем автоматизированного проектирования, Федеральное государственное автономное образовательное учреждение высшего образования «Южный федеральный университет». Область научных интересов: технологии искусственного интеллекта, инженерия знаний. npavlov@sfedu.ru; переулок Некрасовский, 44, 347922, Таганрог, Россия; р.т.: +7(8634)371-651.

**Поддержка исследований.** Исследование выполнено за счет гранта Российского научного фонда № 23-21-00089, https://rscf.ru/project/23-21-00089/ в Южном федеральном университете.

**DOI** 10.15622/ia.23.2.6

D. KRAVCHENKO, YU. KRAVCHENKO, A. MANSOUR, J. MOHAMMAD, N. PAVLOV

# ALGORITHM FOR OPTIMIZATION OF KEYWORD EXTRACTION BASED ON THE APPLICATION OF A LINGUISTIC PARSER

*Kravchenko D., Kravchenko Yu., Mansour A., Mohammad J., Pavlov N.* **Algorithm for Optimization of Keyword Extraction Based on the Application of a Linguistic Parser.**

**Abstract.** This article presents an analytical comparison between constituency parsing and dependency parsing – two types of parsing used in the field of natural language processing (NLP). The study introduces an algorithm to enhance keyword extraction, employing the noun phrase extraction feature of the parser to filter out unsuitable phrases. This algorithm is implemented using three different parsers: Spacy, AllenNLP and Stazna. The effectiveness of this algorithm was compared with two popular methods (Yake, Rake) on a dataset of English texts. Experimental results show that the proposed algorithm with the SpaCy parser is superior to other keyword extraction algorithms in terms of accuracy and speed. For the AllenNLP and Stanza parsers, our algorithm is also more accurate, but requires much longer execution time. The results obtained allow us to evaluate in more detail the advantages and disadvantages of the parsers studied in the work, as well as to determine directions for further research. The running time of the SpaCy parser is significantly less than the other two parsers because parsers that use transitions for deterministic or machine-learned set of actions to build the dependency tree step by step. They are typically faster and require less memory than graph-based parsers, making them more efficient for parsing large amounts of text. On the other hand, AllenNLP and Stanza use graph-based parsing models that rely on millions of features, which limits their ability to generalize and slows down the speed of analysis compared to transition-based parsers. The task of achieving a balance between the accuracy and speed of a linguistic parser is an open topic that requires further research due to the importance of this problem for improving the efficiency of text analysis, especially in applications that require real-time accuracy. To this end, the authors plan to conduct further research into possible solutions to achieve this balance.

**Keywords:** constituency parsing, dependency parsing, keyword extraction, natural language processing, SpaCy, Stanza, AllenNLP.

## References

1. Brown T., Mann B., Ryder N., Subbiah M., Kaplan J.D., Dhariwal P., Neelakantan A., Shyam P., Sastry G., Askell A., et al. Language models are few-shot learners. Advances in neural information processing systems. 2020. vol. 33. pp. 1877–1901.
2. Zhang Y., Clark S. A tale of two parsers: Investigating and combining graph-based and transition-based dependency parsing. Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing. 2008. pp. 562–571.
3. Gao L., Madaan A., Zhou S., Alon U., Liu P., Yang Y., Callan J., Neubig G. Pal: Program aided language models. 2023. pp. 10764–10799.
4. Kravchenko Yu.A., Bova V.V., Kuliev E.V., Rodzin S.I. Simulation of the semantic network of knowledge representation in intelligent assistant systems based on ontological approach. Futuristic Trends in Network and Communication Technologies: Third International Conference, FTNCT. 2021. pp. 241–252.

5. Chen D., Manning C.D. A fast and accurate dependency parser using neural networks. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014. pp. 740–750.
6. Kiperwasser E., Goldberg Y. Simple and accurate dependency parsing using bidirectional LSTM feature representations. Transactions of the Association for Computational Linguistics. 2016. vol. 4. pp. 313–327.
7. Kulmizev A., de Lhoneux M., Gontrum J., Fano E., Nivre J. Deep Contextualized Word Embeddings in Transition-Based and Graph-Based Dependency Parsing – A Tale of Two Parsers Revisited. arXiv preprint arXiv: 07397. 2019.
8. Vasiliev Y. Natural language processing with Python and SpaCy: A practical introduction. No Starch Press, 2020. 216 p.
9. Qi P., Zhang Y., Zhang Y., Bolton J., Manning C.D. Stanza: A Python natural language processing toolkit for many human languages. arXiv preprint arXiv: 07082. 2020.
10. Gardner M., Grus J., Neumann M., Tafjord O., Dasigi P., Liu N., Peters M., Schmitz M., Zettlemoyer L. Allennlp: A deep semantic natural language processing platform. arXiv preprint arXiv: 07640. 2018.
11. Yamada H., Matsumoto Y. Statistical dependency analysis with support vector machines. Proceedings of the eighth international conference on parsing technologies. 2003. pp. 195–206.
12. Nivre J. An efficient algorithm for projective dependency parsing. Proceedings of the eighth international conference on parsing technologies. 2003. pp. 149–160.
13. Kim G., Baldi P., McAleer S. Language models can solve computer tasks. arXiv preprint arXiv:2303.17491. 2023.
14. Liu B., Jiang Y., Zhang X., Liu Q., Zhang S., Biswas J., Stone P. Llm+p: Empowering large language models with optimal planning proficiency. arXiv preprint arXiv:2304.11477. 2023.
15. Pei W., Ge T., Chang B. An effective neural network model for graph-based dependency parsing. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. 2015. vol. 1. pp. 313–322.
16. McDonald R., Crammer K., Pereira F. Online large-margin training of dependency parsers. Proceedings of the 43rd annual meeting of the association for computational linguistics (ACL'05). 2005. pp. 91–98.
17. Eisner J. Three new probabilistic models for dependency parsing: An exploration. arXiv preprint cmp-lg/ 9706003. 1997.
18. Tenney I., Das D., Pavlick E. BERT rediscovers the classical NLP pipeline. arXiv preprint arXiv: 05950. 2019.
19. Hewitt J., Manning C.D. A structural probe for finding syntax in word representations. Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019. vol. 1. pp. 4129–4138.
20. Dozat T., Manning C.D. Deep biaffine attention for neural dependency parsing. arXiv preprint arXiv: 01734. 2016.
21. Mao X., Huang S., Li R., Shen L. Automatic keywords extraction based on co-occurrence and semantic relationships between words. IEEE Access. 2020. vol. 8. pp. 117528–117538.
22. Yang S., Nachum O., Du Y., Wei J., Abbeel P., Schuurmans D. Foundation models for decision making: Problems, methods, and opportunities. arXiv preprint arXiv:2303.04129. 2023.
23. Honnibal M., Johnson M. An Improved Non-monotonic Transition System for Dependency Parsing. Proceedings of the 2015 Conference on Empirical Methods in

Natural Language Processing – Lisbon, Portugal: Association for Computational Linguistics. 2015. pp. 1373–1378. DOI: 10.18653/v1/D15-1162.
24. Campos R., Mangaravite V., Pasquali A., Jorge A., Nunes C., Jatowt A. YAKE! Keyword extraction from single documents using multiple local features. Information Sciences. 2020. vol. 509. pp. 257–289.
25. Rose S., Engel D., Cramer N., Cowley W. Automatic keyword extraction from individual documents. Text mining: applications theory. 2010. pp. 1–20.
26. Hulth A. Improved automatic keyword extraction given more linguistic knowledge. Proceedings of the 2003 conference on Empirical methods in natural language processing. 2003. pp. 216–223.

**Kravchenko Daniil** — Postgraduate, Department of computer-aided design, Southern Federal University. Research interests: artificial intelligence technologies, knowledge engineering. The number of publications — 40. dkravchenko@sfedu.ru; 44, Nekrasovsky Lane, 347922, Taganrog, Russia; office phone: 8(8634)371-651.

**Kravchenko Yury** — Professor, Department of computer-aided design, Southern Federal University. Research interests: artificial intelligence technologies, knowledge engineering. The number of publications — 240. krav-jura@yandex.ru; 44, Nekrasovsky Lane, 347922, Taganrog, Russia; office phone: 8(8634)371-651.

**Mansour Ali Mahmoud** — Postgraduate, Department of computer-aided design, Southern Federal University. Research interests: artificial intelligence technologies, knowledge engineering. The number of publications — 15. mansur@sfedu.ru; 44, Nekrasovsky Lane, 347922, Taganrog, Russia; office phone: 8(9880)158-697.

**Mohammad Juman** — Postgraduate, Department of computer-aided design, Southern Federal University. Research interests: artificial intelligence technologies, knowledge engineering. The number of publications — 14. zmohammad@sfedu.ru; 44, Nekrasovsky Lane, 347922, Taganrog, Russia; office phone: +7(918)543-3526.

**Pavlov Nikolai** — Postgraduate, Department of computer-aided design, Southern Federal University. Research interests: artificial intelligence technologies, knowledge engineering. npavlov@sfedu.ru; 44, Nekrasovsky Lane, 347922, Taganrog, Russia; office phone: +7(8634)371-651.

**Acknowledgements.** The study was performed with the grant from the Russian Science Foundation № 23-21-00089, https://rscf.ru/project/23-21-00089/ in the Southern Federal University.

**DOI** 10.15622/ia.23.2.7

D. BALONI, D. RAI, P. SIVAGAMINATHAN, H. ANANDARAM, M. THAPLIYAL, K. JOSHI

# H-DETECT: AN ALGORITHM FOR EARLY DETECTION OF HYDROCEPHALUS

*Baloni D., Rai D., Sivagaminathan P., Anandaram H., Thapliyal M., Joshi K.* **H-Detect: an Algorithm for Early Detection of Hydrocephalus.**

**Abstract.** Hydrocephalus is a central nervous system disorder which most commonly affects infants and toddlers. It starts as an abnormal build-up of cerebrospinal fluid in the ventricular system of the brain. Hence, early diagnosis becomes vital, which may be performed by Computed Tomography (CT), one of the most effective diagnostic methods for diagnosing Hydrocephalus (CT), where the enlarged ventricular system becomes apparent. However, most disease progression assessments rely on the radiologist's evaluation and physical measures, which are subjective, time-consuming, and inaccurate. This paper develops an automatic prediction utilizing the H-detect framework for enhanced accurate hydrocephalus prediction. This paper uses a pre-processing step to normalize the input image and remove unwanted noises, which can help extract valuable features easily. The feature extraction is done by segmenting the image based on edge detection using triangular fuzzy rules. Thereby, the exact information on the nature of CSF inside the brain is highlighted. These segmented images are saved and again given to the CatBoost algorithm. The Categorical feature processing allows for quicker training. When necessary, the overfitting detector will stop model training and thus efficiently predicts Hydrocephalus. The outcomes demonstrate that the new H-detect strategy outperforms the traditional approaches.

**Keywords:** Hydrocephalus, Computed Tomography (CT), H-detect technique, Cerebrospinal fluid (CSF), Triangular fuzzy rules, Edge detect.

**1. Introduction.** Hydrocephalus is a typical central nervous system disorder engendered by abnormalities in Cerebrospinal Fluid (CSF) circulation. It is caused by an aberrant development of dynamic CSF balance inside the brain's ventricular system [1]. As a result, the ventricles bulge and compress the surrounding brain tissue, resulting in potentially dangerous intracranial hypertension. The degree of ventricular enlargement is frequently considerable, necessitating neurosurgery. It has some of the most severe conditions affecting the central nervous system in children and calls for early neurosurgical treatment [2]. Although this disorder can affect patients of any age, it most commonly affects newborns and babies in their early period of living. Hydrocephalus is predicted to affect one out of every 500 newborns [3]. Hydrocephalus has been examined and scanned; nevertheless, there is no standard solution or efficient strategy for the precise detection and quantitative assessment. Existing measuring methods are predominantly qualitative and produce unsatisfactory results [4]. Radiological methods, such as Computer Tomography (CT) and Magnetic Resonance Imaging (MRI), play a major role in the evaluation of Hydrocephalus (MRI). These tests yield three-dimensional (3D), volumetric

pictures of the brain. However, Hydrocephalus is still primarily assessed manually. It is often based on a qualitative study of the lesion size and other distinguishing characteristics [5].

These methods and procedures use various biophysical factors to depict anatomical features and pathological changes in the human brain. The advancement of medical imaging technology can significantly improve the identification and treatment of numerous lesions and pathological alterations [6]. Radiological diagnostics enable the precise identification of the lesion. A comprehensive analysis of the condition is required before making a treatment approach. The development of tools for automated pathogenic change recognition and classification is thus one of the trickiest problems in contemporary clinical image processing and analysis [7]. There are several methods for segmenting the CSF and the brain ventricular system from CT and MRI imaging. However, only a few papers have been undertaken regarding image processing and analysis in the quantitative assessment of Hydrocephalus [8]. These works are typically significant, and there has been no comprehensive research in this field. It is difficult because of the intricacy and wide range of brain regions. As a result, most known algorithms are based on manual or semi-manual CSF extraction or automated segmentation utilizing basic image processing methods [9].

As a result, this paper aims to propose an H-detect framework which aims:

– To remove the noise existing in the raw data and normalize that can be understandable by the model developed.

– To develop an edge detection using a triangular fuzzy rules model for a significant feature extraction process that can contribute highly to a better understanding of the predicting algorithm.

– To utilize the efficient CatBoost algorithm that can learn the extracted features and predict the disease, thereby improving the accuracy of hydrocephalus diagnosis.

This essay is divided into five distinct sections: Section 1, which discusses the introduction of the research; Section 2, which highlights earlier work completed with the same goal; Section 3, which elaborates the proposed method with three subsections; Section 4, which discusses the implementation of the model and the findings achieved; and Section 5, which wraps up the essay.

**2. Literature Survey.** In paper [10] analysis of the DL-based MRI image recognition system for bone fracture diagnosis. The study demonstrated that the MRI pictures could be categorised and arranged using deep convolutional neural networks. The CNN was able to gather

information at a high rate and was not hindered by the surrounding tissues of the hydrocephalus when collecting the 3D images of the hydrocephalus. The research still has significant drawbacks, such as the fact that deep learning depends on data.

According to the authors in [11], convolutional neural networks were used to extract patient-specific information from pictures, create dimensions for the lesion location in the picture, and apply a predetermined recognition system. Besides a classifier, the questionable components were categorised. The properties of the conventional physically constructed identification entity were altered by a small modification to the recognition system. The deep convolutional network can precisely identify the lesion's site and has a wealth of characteristics. But still, there is a need to enhance the accuracy of the paper by utilizing the data from the features in an efficient manner.

Convolutional network segment brain MRI semantic pictures were employed by the authors in [12]. In accordance with the findings, the brain MRI separation study indicated good precision, as well as the area of technology, had great reliability in the anatomical outcomes of the classification of brain MRI. The MRI characteristics of CI patients were retrieved using a convolutional network, and the results were outstanding. Since the network has not saved the segmented image and processed it, hence the features are not clearly estimated.

Brain tumours were classified using DL characteristics and machine learning algorithms by the authors in [13]. The support vector machine (SVM) with the basis function kernel outperformed other machine learning classifiers, and the incorporation of DL greatly enhanced effectiveness. Additionally, the properties of the sensors were examined, and the WHGO descriptor demonstrated outstanding recognition accuracy for the model classifiers. But it is also time-consuming and operator-dependent.

In [14] the objective of this study is to create a screening method to identify hydrocephalus cases from head MRIs. A 3D convolutional neural network was utilised to autonomously partition the other 480 exams and retrieve volumetric anatomical information after being trained on 16 manual segmentation exams (ten of which had hydrocephalus). On 240 exams, a logistic learner of these variables was developed to spot instances of hydrocephalus that needed surgical treatment for therapy. This approach can speed up and improve neuroradiology reads as well as help with the diagnosis of probable hydrocephalus. Still, this method needs to be further enhanced to be automated.

The fuzzy brain-storm optimal solution, which combines fuzzy and brain-storm objective functions, was suggested by the authors in [15] for the segmentation and classification of medical images. Brainstorm optimizing

prioritises the cluster centroids and focuses on them; like other swarm techniques, it may fall into local optimization. The brain-storm optimizer is interesting and surpasses the other strategies with superior outcomes in this investigation. The fuzzy runs numerous cycles to propose an ideal network model. But it can only be used to detect high-grade hydrocephalus.

Study [10] still needs to rely on data. In [11] there is a need to enhance the accuracy of the paper by utilizing the data from the features in an efficient manner. In [12], the network has not saved the segmented image and processed it, hence the features are not clearly estimated. Paper [13] is time-consuming and operator-dependent, [14] must be further enhanced to be automated and [15] can only be used to detect high-grade hydrocephalus. So, there is a need to develop a model which can overcome all the above-mentioned issues in an accurate manner.

**3. H-Detect Framework for Prediction of Hydrocephalus.** This research work proposes a code-based H-detect model that pre-processes MRI brain images, segments them using Fuzzy and extracts the necessary features, based on the features classifies them, and predicts them for early recognition of hydrocephalus. We used an original dataset of 100 patients from several testing facilities to test the algorithm. In order to extract better features and weed out incorrect predictions, the input photos are first pre-processed to reduce noise and normalise the image into a similar comprehensible format. Then the normalized images are segmented using the triangular membership function in fuzzy rules for Edge detection. The edge-featured images are then used as a training and testing set. For training and testing, we use datasets from the UCI machine learning laboratory and mridata.org for our larger requirements. The training and testing set includes characteristics of healthy individuals and cancer patients at various stages. CatBoost is an open-source platform that is tailored in this paper to predict hydrocephalus. CatBoost is an excellent choice because of its resilience, capacity to handle various datasets from various sources, work on non-numeric data and lack of knowledge of rigorous data preparation. The algorithm can also accept categorical variables without displaying the conversion type mistake, allowing the programmer to fine-tune the model rather than correcting trivial errors.

The H-detect framework is a combination of segmentation techniques, triangular fuzzy rules and the CatBoost algorithm. Feature extraction is done by segmenting the image based on edge detection using triangular fuzzy rules. Therefore, the exact information on the structure of CSF inside the brain is efficiently highlighted. The segmented images are given to the CatBoost algorithm. Thus, the H-detect strategy efficiently predicts hydrocephalus.

***Steps followed in the proposed method:***

Step 1. Pre-process image to remove noise.

Step 2. Design fuzzy input and outputs.

Step 3. Define membership functions.

Step 4. Define Fuzzy rules.

Step 5. Fuzzified images to get the segmented images.

Step 6. Label the segmented images and add them to the digitized dataset.

Step 7. Construct a structure and append labels and images to that.

Step 8. Split the data into training and testing sets.

Step 9. Apply the CatBoost algorithm to generate results.

Step 10. Take an MRI of a new patient and segment using fuzzy rules in step 2. Create labelled data.

Step 11. Apply predictive analysis.

The proposed method, H-detect, can be categorized into three sections: pre-processing, segmentation, and classification as shown in Figure 1. The subsequent sub-sections explain the overall methodology under the three steps mentioned earlier.

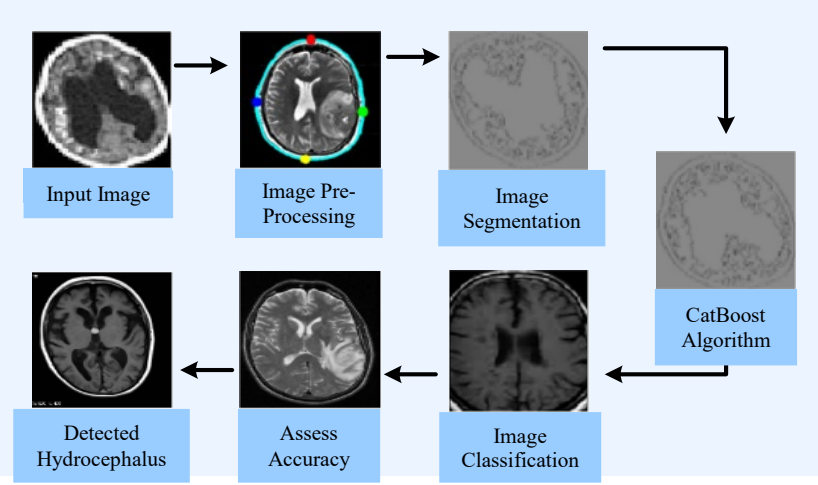



Fig. 1. Process flow diagram of the proposed H-detect method

**3.1. Pre-processing.** In computer-aided medical diagnostics, image pre-processing is a crucial component, specifically in hydrocephalus-related classification, thereby the segmentation and feature extraction algorithms can perform efficiently. Accurate hydrocephalus detection and segmentation lead to precise feature extraction and classification of

hydrocephalus. If the image is pre-processed according to image size and quality, precise hydrocephalus segmentation is feasible which is essential since most real-world data is noisy, inconsistent, and incomplete.

The quality of the acquisition equipment to capture the scene being imaged, such as the structures of the human body, is fundamental for visual interpretation and analysis of digital pictures. It is standard procedure to pre-process images to adjust or improve them before feeding them into more sophisticated processing steps. When deep machine learning is used to edit images, common pre-processing tasks include adjusting their initial dimensions and the augmentation of their intensity. Before feeding the learning machines, the dimensions of the input images are normally reduced to an appropriate size. The basic assumption of size reduction is to reduce the learning machines' compute times at the rate of image quality.

There is a requirement to establish a baseline dimension for all photos input into our AI algorithms in order to extract the features quickly without generating any incorrect predictions because the size of some images captured by the camera and provided to our AI technology fluctuates. The proposed framework employs the following approach to eliminate noise and provide a suitable scale to the input image shown in Figure 2.

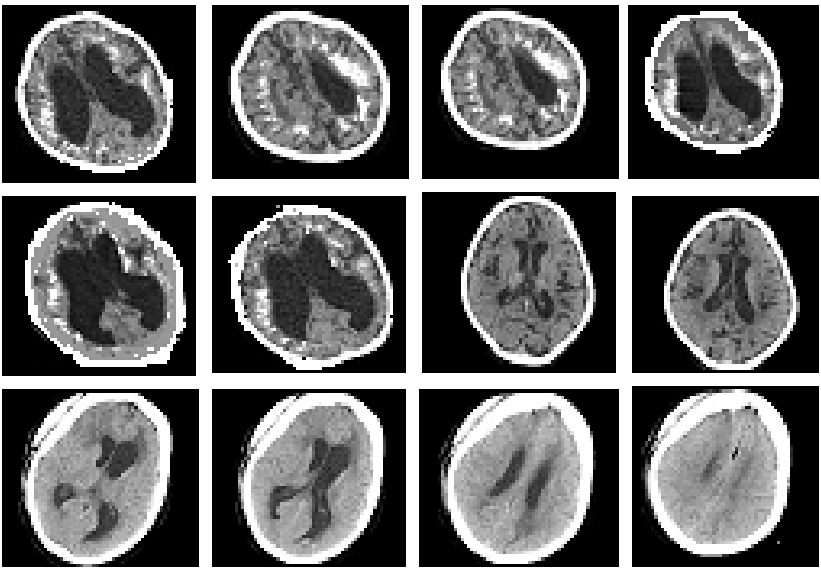


Fig. 2. MRI Image of the brain with Hydrocephalus

Algorithm 1. Pre-processing

Setup: Initialize required variables
Start
Step 1: Read the MRI image
Step 2: Get the dimensions of the image I
Step 3: Get a grey threshold of Ig
Step 4: Get the class type of red channel of I
Step 5: Determine the scaling factor
Step 6: Get a red channel of I
Step 7: $I_{scaled}$ = Ig/scaling Factor

In this paper, the input image is first imparted to the H-Detect algorithm, which retrieves the image information first. Sound interruption, offset field effects, and short-channel impacts could happen while processing brain images. The converted grey threshold function is then used to compute a threshold value. As just one channel is evaluated in each iteration, the algorithm determines the original image's grey threshold value. The image's scaling factor is also determined using the threshold value and Class type.

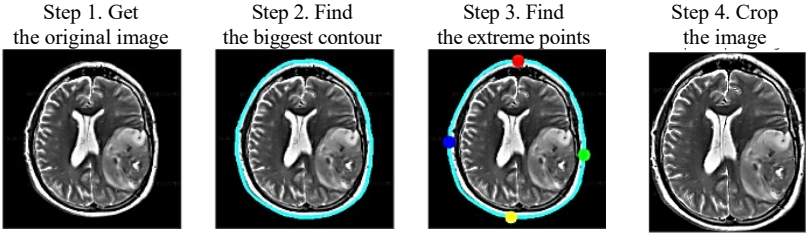



Fig. 3. Pre-processing steps involved in the proposed method

Figure 3 clearly shows the normalization of the image by finding the biggest contour, finding the extreme points and scaling the image. Finally, one image channel is retrieved and split by the scaling factor to generate a normalized image, allowing the algorithm to gather features through correct segmentation and eliminating erroneous classification. The next phase in identifying Hydrocephalus is feature extraction with segmentation, detailed in the next section.

**3.2. Segmentation based feature extraction.** The process of simplifying and modifying the representation of an image into a much more relevant and simpler form to analyse is known as segmentation. Image segmentation is commonly used to find objects and borders in images. Medical image segmentation is critical in demarcating areas of interest

under investigation. It is required in nearly all medical imaging applications and is necessary for automated disease state identification in diagnostic imaging. Nevertheless, because of separate variations and the difficulty of human organs like the brain, segmentation outputs from medical images, particularly those of brain hydrocephalus, are insufficient.

Medical brain pictures are ambiguous by nature and are therefore rife with ambiguity in diagnosis and prediction. The pixel grayscale border between the limit of the brain image and the background becomes hazy and overlapped due to the interaction between light and spatial resolution with brain pictures. It is also challenging to accurately depict the connections between the borders, points, and areas of the locations in the scene due to the influence of equipment elements, which increases the uncertainty, since voxels on a boundary often include two substances, such as border and item.

In order to improve the effectiveness of brain image classification as well as diagnostics, this work makes use of the boundary information of unlabelled and labelled data in brain medical imaging. The human brain MRI image is divided into various situations. Finally, the enhanced algorithm creates a brain disorder medical image segmentation system via fuzzy rules with a triangular membership function.

Algorithm 2. Segmentation

Setup: Initialize required variables from the pre-processed image

Start

Step 1: Create a new fuzzy structure based on the triangular membership function.

$$f(x) = \mathrm{m} \quad \left(\mathrm{m} \quad \left(\frac{x-a}{b-a}, \frac{c-x}{c-b}\right), 0\right), \tag{1}$$

where $f(x)$ is the Triangular membership function,

$a, b, c$ are the parameters,

$x$ is the input value.

Step 2: Add image gradients as input and add output variables.

Step 3: Declare the membership function as $P_{1_{2in}}, P_i$.

$$\mu(F_s) = \begin{cases} 0 \text{ if Fs} \le w \\ \dfrac{(F_s - w)}{(b - w)} \text{ if w} \le F_s \le \text{b} \\ \dfrac{(e - Fs)}{(e - b)} \text{ if b} \le F_s \le \text{e} \\ 0 \text{ if Fs} \ge \text{e} \end{cases}, \quad (2)$$

where $\mu(F_s)$ is the membership function,

Fs is universe of discourse,

$w$ is white,

$b$ is black,

$e$ is edge.

Step 4: If all the membership function equals 0, then black.

Step 5: If anyone or more numbers of the membership function are not equal to 0, then edge.

Step 6: If all the membership functions are not equal to 0, then white.

Step 7: Create an empty matrix for the output.

Step 8: Collect the rules, and create the array.

Step 9: Evaluate fuzzy rules and add the rules to FIS based on the following function:

$$\sum_{\forall R} \sum_{\forall G} \sum_{\forall B} F(I_{mri})_{\forall RGB}, \quad (3)$$

where $\forall R$: Red Channel,

$\forall G$: Green Channel,

$\forall B$: Blue Channel,

$I_{mis}$: Pre-processed MRI image, (dcom converted to jpg).

Step 10: Segment the image based on fuzzy rule-based feature extracted image:

$$\sum_{\nabla_{\text{mi}}} \text{I}_{\text{mri}} \cap \text{F}(\text{I}_{\text{mri}}), \quad (4)$$

where $\forall_{mri}$: All the elements of the MRI image,

$S_{seg}$: Edges detected using (3),

$I_{mri}$: Pre-processed MRI image (converted into jpg),

$F(I_{mri})$: Fuzzified image.

The segmentation is computed with fuzzy membership values using the triangular membership function. The augmented image shown in Figure 4 obtained after pre-processing for separate channels is taken as input. Then the proposed H-detect has designed a fuzzy model to segment

the RGB MRI images with a triangular membership function. The mathematical representation of the triangular membership function is explained in steps 1 and 3. The images are segmented with different membership functions to choose the right membership function as shown in Figure 5. The other parameters remained the same. The functions in steps 9 and 10 are employed to design the Fuzzy model, and then the designed rules are applied to detect the edges. The whole algorithm is repeated for green and blue channels also. Thus finally, the segmented RGB image is obtained by overlaying the segmented R, B and G channel images.

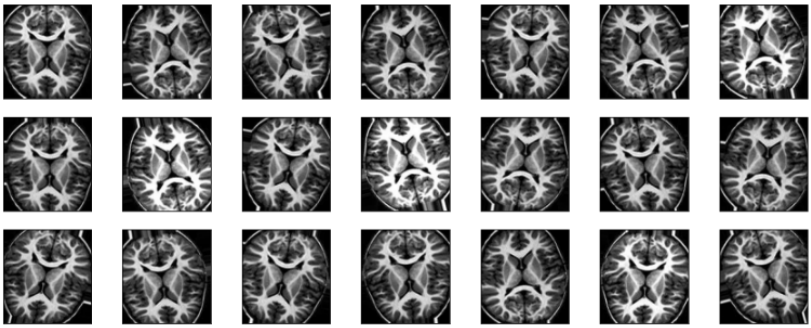


Fig. 4. Augmented image with a single channel for edge detection

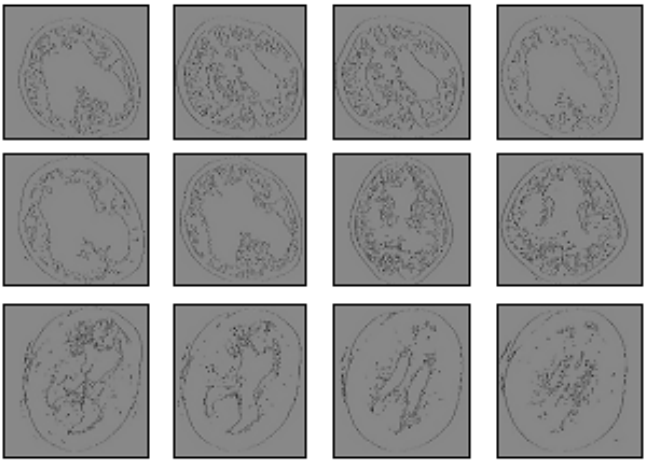


Fig. 5. Segmented Hydrocephalus Images

**3.3. Classification.** All the segmented images are stored in a folder and are scaled to bring uniformity. The scaled images are read one by one

from the folder, and a threshold value is predicted by calculating the mean of the minima of an Image. The threshold of all the images is summed up and then divided by the number of images in the folder. It gives us the scaling factor. The folder is again iterated, and every image is divided by the scaling factor and forms the image dataset; thus, the image will be free from all falsifying factors, enhancing accuracy.

Then a label is assigned to each dataset for classification. Before applying CatBoost, the data type is changed to float. It is then flattened and then divided by 255 (pixel value). We use steps 4 and 5 to design the model and then fit the data on the model. Once the model fits, a new image is taken, and predictive analysis is done, as explained in Step 7.

Setup: Initialize required variables

Start

Step 1. Read folder having Edge featured images.

Step 2. Create a learning dataset using fn (3):

$$\sum_{i=1}^{\nabla S_{eg}} \left\{F\left(S_{seg}\right)_i\right\}, \tag{5}$$

where $\forall S_{seg}$: All the elements of the segmented image from (4)

$\left\{F\left(S_{seg}\right)_i\right\}$: The segmented image taken as features

Step 3. Label the data.

Step 4. Create the CatBoost classifier model:

$$\sum_{i=1}^{\nabla S_{\text{see}}} \left\{F\left(S_{seg}\right)_i\right\}, \tag{6}$$

where $\forall S_{seg}$: All the elements of the segmented image from (4),

$\left\{F\left(S_{seg}\right)_i\right\}$: The segmented image taken as features.

Step 5. Train the CatBoost classifier model with training data:

$$\sum_{\nabla F} D_c(\{F_{\text{train}}, \mathrm{F}_{\text{test}}\}), \tag{7}$$

where $\forall F$: All the features,

$D_c\left(\left\{F_{\text{train}}, F_{\text{test}}\right\}\right)$: Apply CatBoost on the Training and testing set Created after (6).

Step 6. Fit the model on Dataset, Data Label.

Step 7. Test the model using (8):

$$\sum_{\nabla R} \mathrm{P}(\forall \mathrm{R}, \mathrm{N}_{\text{feature}}), \tag{8}$$

where $\forall R$: Results of CatBoost,

$\mathrm{P}(\forall \mathrm{R}, \mathrm{N}_{\text{feanive}})$: Predictive analysis of New Features.

The CatBoost classifier is one machine learning technique that's also effective in forecasting classified variables. Gradient boosting is carried out via CatBoost, which uses binary decision trees as baseline forecasts [16]. Suppose we observe data with samples $D = \{(X_j, y_j)\}_{j=1,\dots,m}$, where $x_j = x_j^1, x_j^2, \dots\dots\dots, x_j^n$ vector of n features and response feature $y_j \in R$, which can be binary (i.e., yes or no) or encoded as a numerical feature (0 or 1). Samples $(X_j, y_j)$ are independently and identically distributed according to some unknown distribution $(\cdot,\cdot)$. The goal of the learning task is to train a function $H : R^n \rightarrow R$, which minimizes the expected loss given as:

$$\mathcal{L}(H) = EL\big(y, H(X)\big), \tag{9}$$

where $L(\cdot,\cdot)$ is a smooth loss function, and (X, y) is testing data sampled from the training data D.

The procedure for gradient boosting constructs iteratively a sequence of approximations $H^t: R^m \rightarrow R, t = 0, 1, \dots$ in a greedy fashion. From the previous approximation $H^{t-1}$, $H^t$ is obtained in an additive process, such that $H^t = H^{t-1} + \alpha g^t$. With a step size α and function $g^t : R^n \rightarrow R$, which is a base predictor, is selected from a set of functions G to reduce or minimize the expected loss defined below:

$$g^t = \arg min_{g \in G} \mathcal{L}(H^{t-1} + g), \tag{10}$$

$$= \arg min_{g \in G} EL\big(y, H^{t-1}(X) + g(X)\big). \tag{11}$$

Often, the minimization problem is approached by the Newton method using a second-order approximation of $L(H^{t-1} + g^t)\ at H^{t-1}$ or by taking a (negative) gradient step.

Thus, the suggested H-detect algorithms accurately identify Hydrocephalus in brain MRI images with minimal processing time. One of the predictions is depicted in Figure 6. The suggested technique effectively eliminates incorrect predictions and overfitting. The next part evaluates and discusses the results gained by applying the proposed strategy and its performance.

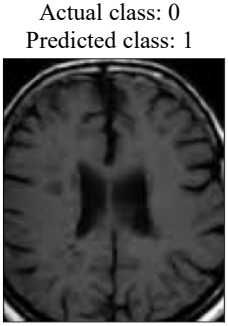



Fig. 6. Prediction of labelled image classes

**4. Result and Discussion.** The proposed H-detect technique is implemented in MATLAB. MRI images are converted from DCOM to jpg format using a third-party tool. If we discuss about pattern recognition we may use the YOLO technique but in this section, we initialized the basic approach to segmentation for hydrocephalus detection. 3D image processing technique is used for visualization, processing, and analysis of 3D image data through geometric transformations. The 3D approach can be also applicable for such tasks since CT is a sequence of images of a brain and the 3D approach can also be useful for the diagnosis of hydrocephalus with a couple of datasets. 3D is an older technique. Here, we use the Catboost algorithm which is used for prediction and classification. It is much better than 3D for various factors such as segmentation, classification, decision-making, precision and accuracy. By applying Catboost with fuzzy logic, the proposed system works better than the existing work. The output of the implementation and the obtained results are discussed in this section.

**4.1. System Configuration.**

Processor : Intel Core i5, V generation
RAM : 16 GB
Graphics : Nvidia
HDD : 1 TB
OS : Windows 10

**4.2. Dataset description.** Hydrocephalus datasets from three testing labs have been taken. Due to the non-disclosure agreement, further details cannot be shared. The dataset also includes records of Tumor, Malignant, Benign, and Hydrocephalus which were taken from Brigham and Women's Hospital, Surgical Planning Laboratory, Department of Radiology, Harvard Medical School (Boston, MA, USA), BRATS, BITE, metadata.org, and cancerimagingarchive.net.

**4.3. Implementation results.** The image of Figure 7, 8 shows the input dataset containing both brain MRI images with and without hydrocephalus. The distribution of the images with respective ratio values is given below in Figure 9. The H-detect model predicts the hydrocephalus efficiently.

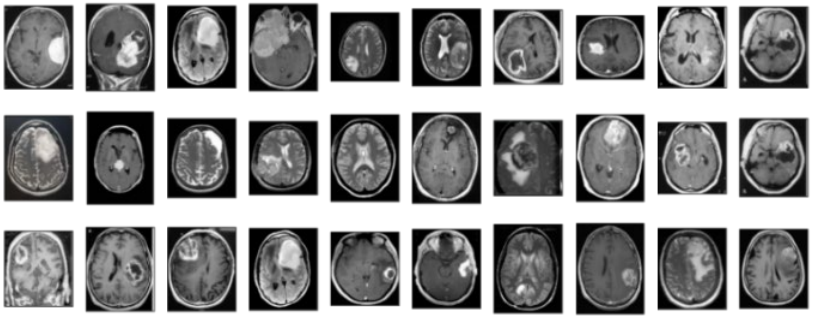


Fig. 7. Dataset images with hydrocephalus

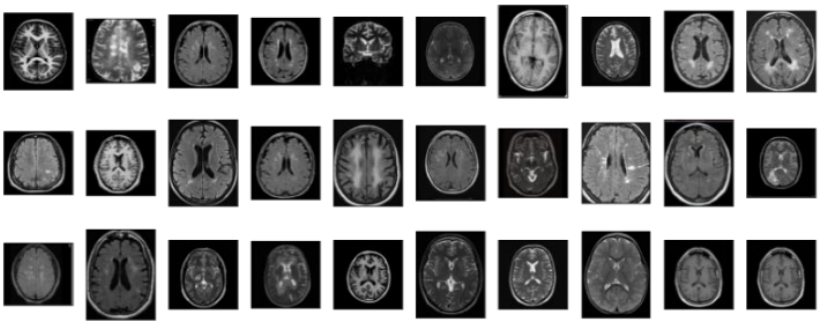


Fig. 8. Dataset images without hydrocephalus

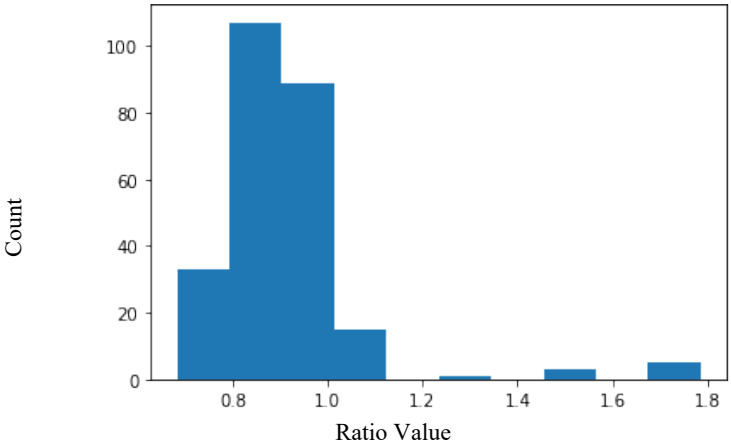



Fig. 9. Distribution of Image Ratios

The cat boost model has 34 Iterations, 0.05 Learning rate, depth of 12 and multi-class loss function.

**4.4. Performance metrics.** The following formulae from (12) to (15) are calculated for checking the robustness of H-detect Precision, Recall, Accuracy, and F1 scores. True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) are the metrics used to calculate the scores:

$$Precision = \frac{TP}{(TP + FN)}, \tag{12}$$

$$Recall = \frac{TN}{(TN + FP)}, \tag{13}$$

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}, \tag{14}$$

$$F1 = \frac{2 \times (Precision \times Recall)}{(Precision + Recall)}. \tag{15}$$

The confusion matrix that was produced after using the proposed approach is shown in Figure 10. A True Positive (TP) value of 19, a True Negative (TN) value of 0, a False Positive (FP) value of 1, and a False

Negative (FN) value of 30 are displayed in the matrix. Thus, utilizing the proper extracted features obtained from the fuzzy triangular membership function increases the accuracy which is clear as the model has just a 1% loss.

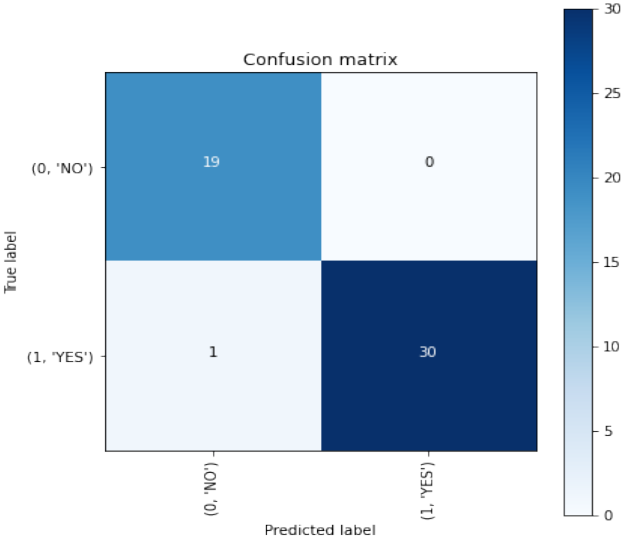



Fig. 10. Confusion matrix obtained for the proposed method

The proposed model's accuracy and loss function are depicted above in Figure 11. The validation set performs significantly better than the training set as the CatBoost algorithm performs efficient prediction. With 40 epochs, the maximum accuracy of 0.99 was achieved. The validation set's loss is similarly lower than the training sets. With 40 epochs, only 0.1 of loss was observed.

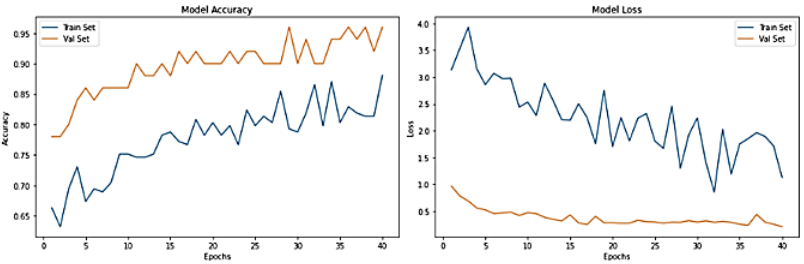



Fig. 11. Performance Graph obtained for the proposed method

Table 1 displays the performance of the proposed H-detect method based on its precision, accuracy, F1 score and recall. As this research removed the unwanted noises, normalized and extracted features from fuzzy logic and utilized the CatBoost algorithm for prediction the findings indicate that the proposed technique has obtained 99% precision, accuracy, F1 score, and 100% recall.

Table 1. Performance metrics of the proposed method

| Performance measures | Value |
|---|---|
| Precision | 0.99009901 |
| Recall | 1 |
| Accuracy | 0.995 |
| F1-score | 0.995025 |

**4.5. Comparison Metrics.** Table 2 shows the variation in time taken for the segmentation proposed method compared with Consecutive deep encoder-decoder networks, Morphological adaptive fuzzy thresholding and Fuzzy c-means. The graph shows that a Consecutive deep encoder-decoder network takes 3021 ms, Morphological adaptive fuzzy thresholding takes 1894 ms, and Fuzzy c-means takes 189 ms. Our proposed method, H-detect, takes just 62 ms, as fuzzy logic uses the triangular membership function which is much faster than the conventional methods.

Table 2. Comparison of time taken in segmentation with different techniques

| Model | Time (in ms) |
|---|---|
| [17] Consecutive Deep Encoder-Decoder Network | 3021 |
| [18] fuzzy c-means | 189 |
| [19] Morphological Adaptive Fuzzy Thresholding | 1894 |
| H-detect | 62 |

Table 3 above exhibits how the proposed methodology compares to the Consecutive deep encoder-decoder network, Morphological adaptive fuzzy thresholding, and Fuzzy c-means in terms of classification duration. The graph shows that the Deep Convolutional Neural Network requires 261 milliseconds, Spark-based parallel fuzzy c-means requires 278 milliseconds, F score-based method requires 176 milliseconds. Our proposed method, H-detect requires only 98 milliseconds, which is significantly superior to the conventional methods by utilization of the CatBoost algorithm.

Table 3. Comparison of time taken in classification with different techniques

| Model | Time (in ms) |
|---|---|
| [20] Deep Convolutional Neural Network | 261 |
| [18] Spark-based parallel fuzzy c-means | 278 |
| [21] F score-based method | 176 |
| H-detect | 98 |

Table 4 tends to parallel the accuracy of the proposed method founded on the membership function used. The proposed method uses triangular membership, compared with singleton, gaussian, generalized bell, and sigmoidal functions. The results show that the proposed membership function has 100% exactness, accurateness, F1 score and recall, which shows that the proposed triangular membership function is the most superior function to the traditional one in all aspects.

Table 4. Performance metrics based on membership function

| Membership Functions | Precision | Accuracy | F1-score | Recall |
|---|---|---|---|---|
| Singleton | 0.900901 | 0.945 | 0.947867 | 1 |
| Gaussian | 0.925926 | 0.96 | 0.961538 | 1 |
| Generalized Bell | 0.884956 | 0.935 | 0.938967 | 1 |
| Sigmoidal | 0.943396 | 0.97 | 0.970874 | 1 |
| Trapezoidal | 0.961538 | 0.98 | 0.980392 | 1 |
| Triangular | 1 | 1 | 1 | 1 |

Table 5 compares the suggested method's performance dependent on various types of cancer data like tumour, benign, malignant and Hydrocephalus. The suggested technique compares the exactness, correctness, F1 score and recall. The findings indicate that 100 percent exactness, accurateness, F1 score, and recall have been obtained for Hydrocephalus, demonstrating that the proposed H-detect method is outstanding to the conventional one.

Table 5. Performance based on the type of cancer

| Type | Precision | Accuracy | F1-score | Recall |
|---|---|---|---|---|
| Tumour | 0.99009901 | 0.995 | 0.9950249 | 1 |
| Benign | 0.981 | 0.99 | 0.989 | 1 |
| Malignant | 0.980392157 | 0.99 | 0.990099 | 1 |
| Hydrocephalus | 1 | 1 | 1 | 1 |

Table 6 shows the accuracy of 7 various methods compared with our proposed method with and without noise. The graphs depict that the proposed method has shown outstanding results than the previously proposed method, with 99.9% accuracy with and without noise as the pre-processing step efficiently contributes to the accuracy.

Table 6. Accuracy comparison with different techniques

| Method | With Noise | Without Noise |
|---|---|---|
| [22] KM | 0.9720 | 0.6239 |
| [23] RKM | 0.9743 | 0.7832 |
| [24] FCM | 0.9728 | 0.7698 |
| [25] RFCM | 0.9782 | 0.7806 |
| [26] GRFCM | 0.9679 | 0.7622 |
| [27] SFRCM | 0.9264 | 0.7786 |
| [28] RIFCM | 0.8992 | 0.9016 |
| Proposed | 0.99 | 0.99 |

Table 7 compares three other conventional models based on precision, accuracy, F1 score and recall. The results obtained showed that the proposed model has the best precision of 99%, the accuracy of 99.5%, the F1 score of 99.5 %, and the recall of 100% proving that the proposed method is the best one with précised pre-processing step, fuzzy logic with triangular membership function-based segmentation, edge-based features and the CatBoost classification methodology.

Table 7. Comparison of performance with different techniques

| Algorithm | Precision | Accuracy | F1-score | Recall |
|---|---|---|---|---|
| [29] Fuzzy Reasoning Model | 0.917431193 | 0.955 | 0.956938 | 1 |
| [30] Modified Timed Automata Model | 0.934579439 | 0.965 | 0.966184 | 1 |
| [31] Gaussian Mixture Model | 0.952380952 | 0.975 | 0.97561 | 1 |
| H-Detect | 0.99009901 | 0.995 | 0.995025 | 1 |

**5. Conclusion.** This research adopted MRI data to detect various kinds of Hydrocephalus early. The brain and hydrocephalus volumes are heavily influenced by the spatial resolution of successive brain cross-sections and the slice thickness employed during CT imaging. As a result, the focus of this study is on two major enhancements. Firstly, the image segmentation method will be enhanced, allowing for better separation of the targeted brain areas by utilizing a neural fuzzy method with a triangular

membership function. Then for predictive analysis, a classifier based on the cat boost method is presented to classify Hydrocephalus. For novelty, we aggregate the fuzzy rules and CatBoost for better results. According to the observations, the suggested model has a precision, accuracy, and an F1 score of 99% and a recall of 100%. In each aspect, the comparative findings have shown to be superior to the conventional ones. As a result, the proposed H-detect approach can reliably diagnose Hydrocephalus early without any false predictions or overfitting issues, allowing numerous people's lives to be saved. In future studies, we may impose a particular feedback system to monitor the diagnosis system for hydrocephalus in a medical era so that the processing time may be reduced.

## References


1. Zhang X.J., Guo J., Yang J. Cerebrospinal fluid biomarkers in idiopathic normal pressure hydrocephalus. Neuroimmunology and Neuroinflammation. 2020. vol. 7. no. 2. pp. 109–119.
2. Karimy J.K., Reeves B.C., Damisah E., Duy P.Q., Antwi P., David W., Kahle K.T. Inflammation in acquired Hydrocephalus: pathogenic mechanisms and therapeutic targets. Nature Reviews Neurology. 2020. vol. 16. no. 5. pp. 285–296.
3. Paulsen A.H. Adult outcome in pediatric Hydrocephalus. 2018. 58 p.
4. Saygili G., Yigin B.O., Guney G., Algin O. Exploiting lamina terminalis appearance and motion in the prediction of Hydrocephalus using convolutional LSTM network. Journal of Neuroradiology. 2022. vol. 49. no. 5. pp. 364–369.
5. Nakajima M., Kawamura K., Akiba C., Sakamoto K., Xu H., Kamohara C., Miyajima M. Differentiating comorbidities and predicting prognosis in idiopathic normal pressure hydrocephalus using cerebrospinal fluid biomarkers. Croatian Medical Journal. 2021. vol. 62. no. 4. pp. 387–398.
6. Yigin B.O., Algin O., Saygili G. Comparison of morphometric parameters in prediction of Hydrocephalus using random forests. Computers in Biology and Medicine. 2020. vol. 116. no. 103547.
7. Chiarelli P.A., Hauptman J.S., Browd S.R. Machine learning and the prediction of Hydrocephalus: Can quantitative image analysis assist the clinician? JAMA paediatric. 2018. vol. 172. no. 2. pp. 116–118.
8. Chen J., He W., Zhang X., Lv M., Zhou X., Yang X., Xia J. Value of MRI-based semi-quantitative structural neuroimaging in predicting the prognosis of patients with idiopathic normal pressure hydrocephalus after shunt surgery. European Radiology. 2022. vol. 32. no. 11. pp. 7800–7810.
9. Sotoudeh H., Sadaatpour Z., Rezaei A., Shafaat O., Sotoudeh E., Tabatabaie M., Tanwar M. The Role of Machine Learning and Radiomics for Treatment Response Prediction in Idiopathic Normal Pressure Hydrocephalus. Cureus. 2021. vol. 13. no. 10.
10. Mao Y., Shen Z., Wang J., Zhu H., Yu Z., Chen X., Cheng H. Deep Learning-Based MR Imaging for Analysis of Relation between Cerebrospinal Fluid Variation and Communicating Hydrocephalus after Decompressive Craniectomy for Craniocerebral Injury. Scientific Programming. 2022. vol. 2022.
11. Brito C., Machado A., Sousa A.L. Electrocardiogram beat classification based on a Res-Net network. Studies in Health Technology and Informatics. 2019. vol. 264. pp. 55–59.

12. Hu Y., Zhao H., Li W., Li J. Semantic image segmentation of brain MRI with deep learning. Zhong nan da XueXueBao. Yi Xue ban Journal of Central South University. Medical sciences. 2021. vol. 46. no. 8. pp. 858–864.
13. Kang J., Ullah Z., Gwak J. MRI-based brain tumour classification using ensemble of deep features and machine learning classifiers. Sensors. 2021. vol. 21(6). no. 2222.
14. Huang Y., Moreno R., Malani R., Meng A., Swinburne N., Holodny A.I., Young R.J. Deep Learning Achieves Neuroradiologist-Level Performance in Detecting Hydrocephalus Requiring Treatment. Journal of Digital Imaging. 2022. vol. 35. no. 6. pp. 1662–1672.
15. Narmatha C., Eljack S.M., Tuka A.A.R.M., Manimurugan S., Mustafa M. A hybrid fuzzy brain-storm optimization algorithm for the classification of brain tumour MRI images. Journal of ambient intelligence and humanized computing. 2020. pp. 1–9.
16. Prokhorenkova L., Gusev G., Vorobev A., Dorogush A.V, Gulin A. CatBoost: Unbiased Boosting with Categorical Features. Advances in Neural Information Processing Systems. 2018. vol. 31. pp. 6638–6648.
17. Nguyen N.Q., Lee S.W. Robust Boundary Segmentation in Medical Images Using a Consecutive Deep Encoder-Decoder Network. IEEE Access. 2019. vol. 7. pp. 33795–33808. DOI: 10.1109/ACCESS.2019.2904094.
18. Liu B., He S., He D., Zhang Y., Guizani M. A Spark-based Parallel Fuzzy $c$-Means Segmentation Algorithm for Agricultural Image Big Data. IEEE Access. 2019. vol. 7. pp. 42169–42180. DOI: 10.1109/ACCESS.2019.2907573.
19. Almotiri J., Elleithy K., Elleithy A. A Multi-Anatomical Retinal Structure Segmentation System for Automatic Eye Screening Using Morphological Adaptive Fuzzy Thresholding. IEEE Journal of Translational Engineering in Health and Medicine. 2018. vol. 6. pp. 1–23. DOI: 10.1109/JTEHM.2018.2835315.
20. Liu M., Jiang J., Wang Z. Colonic Polyp Detection in Endoscopic Videos with Single Shot Detection Based Deep Convolutional Neural Network. IEEE Access. 2019. vol. 7. pp. 75058–75066. DOI: 10.1109/ACCESS.2019.2921027.
21. Raweh A.A., Nassef M., Badr A. A Hybridized Feature Selection and Extraction Approach for Enhancing Cancer Prediction Based on DNA Methylation. IEEE Access. 2018. vol. 6. pp. 15212–15223. DOI: 10.1109/ACCESS.2018.2812734.
22. Gonzalez R., Tou J. Pattern recognition principles. Applied Mathematics and Computation. Reading, MA: Addison-Wesley. 1974. 377 p.
23. Lingras P., West C. Interval set clustering of web users with rough k-means. Journal of Intelligent Information Systems. 2004. vol. 23. no. 1. pp. 5–16. DOI: 10.1023/B:JIIS.0000029668.88665.1a.
24. Chuang K.S., Tzeng H.L., Chen S., Wu J., Chen T.J. Fuzzy c-means clustering with spatial information for image segmentation. Comput. Med. Imag. Graph., Jan. 2006. vol. 30. no. 1. pp. 9–15. DOI: 10.1016/j.compmedimag.2005.10.001.
25. Lingras P., Peters G. Applying rough set concepts to clustering. In Rough Sets: Selected Methods and Applications in Management and Engineering. London: Springer. 2012. pp. 23–37.
26. Ji Z., Sun Q., Xia Y., Chen Q., Xia D., Feng D. Generalized rough fuzzy c-means algorithm for brain MR image segmentation. Computer methods and programs in biomedicine. 2012. vol. 108. no. 2. pp. 644–655.
27. Namburu A., Srinivas Kumar S., Srinivasa Reddy E. Review of Set-Theoretic Approaches to Magnetic Resonance Brain Image Segmentation. IETE Journal of Research. 2022. vol. 68. no. 1. pp. 350–367. DOI: 10.1080/03772063.2019.1604176.
28. Dubey Y.K., Mushrif M.M., Mitra K. Segmentation of brain MR images using rough set based intuitionistic fuzzy clustering. Biocybern. Biomedical engineering. 2016. vol. 36. no. 2. pp. 413–426. DOI: 10.1016/j.bbe.2016.01.001.

29. Liu J., Peng Y., Zhang Y.A Fuzzy Reasoning Model for Cervical Intraepithelial Neoplasia Classification Using Temporal Grayscale Change and Textures of Cervical Images during Acetic Acid Tests. IEEE Access. 2019. vol. 7. pp. 13536–13545. DOI: 10.1109/ACCESS.2019.2893357.
30. Brunese L., Mercaldo F., Reginelli A., Santone A. Prostate Gleason Score Detection and Cancer Treatment through Real-Time Formal Verification. IEEE Access. 2019. vol. 7. pp. 186236–186246. DOI: 10.1109/ACCESS.2019.2961754.
31. Yin S., Zhang Y., Karim S. Large Scale Remote Sensing Image Segmentation Based on Fuzzy Region Competition and Gaussian Mixture Model. IEEE Access. 2018. vol. 6. pp. 26069–26080. DOI: 10.1109/ACCESS.2018.2834960.

**Baloni Dev** — Ph.D., Dr.Sci., Associate professor, Quantum school of technology, Quantum University. Research interests: image processing, machine learning, deep learning, artificial intelligence. The number of publications — 20. devbaloni1982@gmail.com; Dehradun Highway, Mandawar, 247167, Roorkee, Uttarakhand, India; office phone: +91(8755)507-830.

**Rai Dhajvir Singh** — Assistant professor, School of engineering and computing, Dev Bhoomi Uttarakhand University. Research interests: cloud computing, machine learning, deep learning, artificial intelligence. dhajvirrai123@gamil.com; Chakrata Road, Manduwala, 248007, Naugaon, Uttarakhand, India; office phone: +91(9557)891-499.

**Sivagaminathan PG** — Professor, School of engineering, Ajeenkya D.Y Patil University. Research interests: information retrieval, pattern recognition, AI, machine learning, deep learning. The number of publications — 31. sai.sivagaminathan@gmail.com; City Road via Lohegaon, 412105, Charholi Budruk, Pune, Maharashtra, India; office phone: +91(9500)903-614.

**Anandaram Harishchander** — Ph.D., Dr.Sci., Assistant professor, Amrita school of artificial intelligence, Amrita Vishwa Vidyapeetham (Amrita University). Research interests: computational systems biology, functional genomics, developmental biology. The number of publications — 43. a_harishchander@cb.amrita.edu; Amritanagar, Ettimadai Village, 641112, Coimbatore, Tamil Nadu, India; office phone: +91(9940)066-227.

**Thapliyal Madhur** — Assistant professor, Graphic Era Hill University. Research interests: cyber security, image processing, deep learning, machine learning. The number of publications — 3. madhurthapliyal@gehu.ac.in; Bell Road, Clement Town, 248002, Dehradun, Uttarakhand, India; office phone: +91(7017)747-897.

**Joshi Kapil** — Ph.D., Dr.Sci., Assistant professor, Uttaranchal institute of technology, Uttaranchal University. Research interests: image fusion, image processing, deep learning, CNN. The number of publications — 132. Kapilengg0509@gmail.com; Prem Nagar, 248007, Dehradun, Uttarakhand, Russia; office phone: +91(8979)799-289.

**DOI** 10.15622/ia.23.2.7

Д. БАЛОНИ, Д. РАЙ, П. СИВАГАМИНАТАН, Х. АНАНДАРАМ, М. ТАПЛИЯЛ, К. ДЖОШИ

# H-DETECT: АЛГОРИТМ РАННЕГО ВЫЯВЛЕНИЯ ГИДРОЦЕФАЛИИ

*Балони Д., Рай Д., Сивагаминатан П., Анандарам Х., Таплиял М., Джоши К.* **H-Detect: алгоритм раннего выявления гидроцефалии.**

**Аннотация.** Гидроцефалия – это заболевание центральной нервной системы, которое чаще всего поражает младенцев и детей ясельного возраста. Оно начинается с аномального накопления спинномозговой жидкости в желудочковой системе головного мозга. Следовательно, жизненно важной становится ранняя диагностика, которая может быть выполнена с помощью компьютерной томографии (КТ), одного из наиболее эффективных методов диагностики гидроцефалии (КТ), при котором становится очевидным увеличение желудочковой системы. Однако большинство оценок прогрессирования заболевания основаны на оценке рентгенолога и физических показателях, которые являются субъективными, отнимающими много времени и неточными. В этой статье разрабатывается автоматическое прогнозирование с использованием фреймворка H-detect для повышения точности прогнозирования гидроцефалии. В этой статье используется этап предварительной обработки для нормализации входного изображения и удаления нежелательных шумов, что может помочь легко извлечь ценные признаки. Выделение признаков осуществляется путем сегментации изображения на основе определения границ с использованием треугольных нечетких правил. Таким образом, выделяется точная информация о природе ликвора внутри мозга. Эти сегментированные изображения сохраняются и снова передаются алгоритму CatBoost. Обработка категориальных признаков позволяет ускорить обучение. При необходимости детектор переобучения останавливает обучение модели и, таким образом, эффективно прогнозирует гидроцефалию. Результаты демонстрируют, что новая стратегия H-detect превосходит традиционные подходы.

**Ключевые слова:** гидроцефалия, компьютерная томография (КТ), метод H-детекции, спинномозговая жидкость (ликвор), треугольные нечеткие правила, обнаружение краев.

## Литература

1. Zhang X.J., Guo J., Yang J. Cerebrospinal fluid biomarkers in idiopathic normal pressure hydrocephalus. Neuroimmunology and Neuroinflammation. 2020. vol. 7. no. 2. pp. 109–119.
2. Karimy J.K., Reeves B.C., Damisah E., Duy P.Q., Antwi P., David W., Kahle K.T. Inflammation in acquired Hydrocephalus: pathogenic mechanisms and therapeutic targets. Nature Reviews Neurology. 2020. vol. 16. no. 5. pp. 285–296.
3. Paulsen A.H. Adult outcome in pediatric Hydrocephalus. 2018. 58 p.
4. Saygili G., Yigin B.O., Guney G., Algin O. Exploiting lamina terminalis appearance and motion in the prediction of Hydrocephalus using convolutional LSTM network. Journal of Neuroradiology. 2022. vol. 49. no. 5. pp. 364–369.
5. Nakajima M., Kawamura K., Akiba C., Sakamoto K., Xu H., Kamohara C., Miyajima M. Differentiating comorbidities and predicting prognosis in idiopathic normal pressure hydrocephalus using cerebrospinal fluid biomarkers. Croatian Medical Journal. 2021. vol. 62. no. 4. pp. 387–398.

6. Yigin B.O., Algin O., Saygili G. Comparison of morphometric parameters in prediction of Hydrocephalus using random forests. Computers in Biology and Medicine. 2020. vol. 116. no. 103547.
7. Chiarelli P.A., Hauptman J.S., Browd S.R. Machine learning and the prediction of Hydrocephalus: Can quantitative image analysis assist the clinician? JAMA paediatric. 2018. vol. 172. no. 2. pp. 116–118.
8. Chen J., He W., Zhang X., Lv M., Zhou X., Yang X., Xia J. Value of MRI-based semi-quantitative structural neuroimaging in predicting the prognosis of patients with idiopathic normal pressure hydrocephalus after shunt surgery. European Radiology. 2022. vol. 32. no. 11. pp. 7800–7810.
9. Sotoudeh H., Sadaatpour Z., Rezaei A., Shafaat O., Sotoudeh E., Tabatabaie M., Tanwar M. The Role of Machine Learning and Radiomics for Treatment Response Prediction in Idiopathic Normal Pressure Hydrocephalus. Cureus. 2021. vol. 13. no. 10.
10. Mao Y., Shen Z., Wang J., Zhu H., Yu Z., Chen X., Cheng H. Deep Learning-Based MR Imaging for Analysis of Relation between Cerebrospinal Fluid Variation and Communicating Hydrocephalus after Decompressive Craniectomy for Craniocerebral Injury. Scientific Programming. 2022. vol. 2022.
11. Brito C., Machado A., Sousa A.L. Electrocardiogram beat classification based on a Res-Net network. Studies in Health Technology and Informatics. 2019. vol. 264. pp. 55–59.
12. Hu Y., Zhao H., Li W., Li J. Semantic image segmentation of brain MRI with deep learning. Zhong nan da XueXueBao. Yi Xue ban Journal of Central South University. Medical sciences. 2021. vol. 46. no. 8. pp. 858–864.
13. Kang J., Ullah Z., Gwak J. MRI-based brain tumour classification using ensemble of deep features and machine learning classifiers. Sensors. 2021. vol. 21(6). no. 2222.
14. Huang Y., Moreno R., Malani R., Meng A., Swinburne N., Holodny A.I., Young R.J. Deep Learning Achieves Neuroradiologist-Level Performance in Detecting Hydrocephalus Requiring Treatment. Journal of Digital Imaging. 2022. vol. 35. no. 6. pp. 1662–1672.
15. Narmatha C., Eljack S.M., Tuka A.A.R.M., Manimurugan S., Mustafa M. A hybrid fuzzy brain-storm optimization algorithm for the classification of brain tumour MRI images. Journal of ambient intelligence and humanized computing. 2020. pp. 1–9.
16. Prokhorenkova L., Gusev G., Vorobev A., Dorogush A.V, Gulin A. CatBoost: Unbiased Boosting with Categorical Features. Advances in Neural Information Processing Systems. 2018. vol. 31. pp. 6638–6648.
17. Nguyen N.Q., Lee S.W. Robust Boundary Segmentation in Medical Images Using a Consecutive Deep Encoder-Decoder Network. IEEE Access. 2019. vol. 7. pp. 33795–33808. DOI: 10.1109/ACCESS.2019.2904094.
18. Liu B., He S., He D., Zhang Y., Guizani M. A Spark-based Parallel Fuzzy $c$-Means Segmentation Algorithm for Agricultural Image Big Data. IEEE Access. 2019. vol. 7. pp. 42169–42180. DOI: 10.1109/ACCESS.2019.2907573.
19. Almotiri J., Elleithy K., Elleithy A. A Multi-Anatomical Retinal Structure Segmentation System for Automatic Eye Screening Using Morphological Adaptive Fuzzy Thresholding. IEEE Journal of Translational Engineering in Health and Medicine. 2018. vol. 6. pp. 1–23. DOI: 10.1109/JTEHM.2018.2835315.
20. Liu M., Jiang J., Wang Z. Colonic Polyp Detection in Endoscopic Videos with Single Shot Detection Based Deep Convolutional Neural Network. IEEE Access. 2019. vol. 7. pp. 75058–75066. DOI: 10.1109/ACCESS.2019.2921027.
21. Raweh A.A., Nassef M., Badr A. A Hybridized Feature Selection and Extraction Approach for Enhancing Cancer Prediction Based on DNA Methylation. IEEE Access. 2018. vol. 6. pp. 15212–15223. DOI: 10.1109/ACCESS.2018.2812734.

22. Gonzalez R., Tou J. Pattern recognition principles. Applied Mathematics and Computation. Reading, MA: Addison-Wesley. 1974. 377 p.
23. Lingras P., West C. Interval set clustering of web users with rough k-means. Journal of Intelligent Information Systems. 2004. vol. 23. no. 1. pp. 5–16. DOI: 10.1023/B:JIIS.0000029668.88665.1a.
24. Chuang K.S., Tzeng H.L., Chen S., Wu J., Chen T.J. Fuzzy c-means clustering with spatial information for image segmentation. Comput. Med. Imag. Graph., Jan. 2006. vol. 30. no. 1. pp. 9–15. DOI: 10.1016/j.compmedimag.2005.10.001.
25. Lingras P., Peters G. Applying rough set concepts to clustering. In Rough Sets: Selected Methods and Applications in Management and Engineering. London: Springer. 2012. pp. 23–37.
26. Ji Z., Sun Q., Xia Y., Chen Q., Xia D., Feng D. Generalized rough fuzzy c-means algorithm for brain MR image segmentation. Computer methods and programs in biomedicine. 2012. vol. 108. no. 2. pp. 644–655.
27. Namburu A., Srinivas Kumar S., Srinivasa Reddy E. Review of Set-Theoretic Approaches to Magnetic Resonance Brain Image Segmentation. IETE Journal of Research. 2022. vol. 68. no. 1. pp. 350–367. DOI: 10.1080/03772063.2019.1604176.
28. Dubey Y.K., Mushrif M.M., Mitra K. Segmentation of brain MR images using rough set based intuitionistic fuzzy clustering. Biocybern. Biomedical engineering. 2016. vol. 36. no. 2. pp. 413–426. DOI: 10.1016/j.bbe.2016.01.001.
29. Liu J., Peng Y., Zhang Y.A Fuzzy Reasoning Model for Cervical Intraepithelial Neoplasia Classification Using Temporal Grayscale Change and Textures of Cervical Images during Acetic Acid Tests. IEEE Access. 2019. vol. 7. pp. 13536–13545. DOI: 10.1109/ACCESS.2019.2893357.
30. Brunese L., Mercaldo F., Reginelli A., Santone A. Prostate Gleason Score Detection and Cancer Treatment through Real-Time Formal Verification. IEEE Access. 2019. vol. 7. pp. 186236–186246. DOI: 10.1109/ACCESS.2019.2961754.
31. Yin S., Zhang Y., Karim S. Large Scale Remote Sensing Image Segmentation Based on Fuzzy Region Competition and Gaussian Mixture Model. IEEE Access. 2018. vol. 6. pp. 26069–26080. DOI: 10.1109/ACCESS.2018.2834960.

**Балони Дев** — Ph.D., Dr.Sci., доцент, школа квантовых технологий, Университет Квантум. Область научных интересов: обработка изображений, машинное обучение, глубокое обучение, искусственный интеллект. Число научных публикаций — 20. devbaloni1982@gmail.com; шоссе Дехрадун, Мандавар, 247167, Рурки, Уттаракханд, Индия; р.т.: +91(8755)507-830.

**Рай Дханвир Сингх** — доцент, школа инженерии и вычислительной техники, Университет Дев Бхуми Уттаракханд. Область научных интересов: облачные вычисления, машинное обучение, глубокое обучение, искусственный интеллект. Число научных публикаций — 0. dhajvirrai123@gamil.com; Чакрата-роуд, Мандувала, 248007, Наагон, Уттаракханд, Индия; р.т.: +91(9557)891-499.

**Сивагаминатан П.Г.** — профессор, инженерная школа, Университет Аджинкья Д.Я. Патила. Область научных интересов: поиск информации, распознавание образов, искусственный интеллект, машинное обучение, глубокое обучение. Число научных публикаций — 31. sai.sivagaminathan@gmail.com; Городская дорога через Лохегаон, 412105, Чархоли Бадрек, Пуна, Махараштра, Индия; р.т.: +91(9500)903-614.

**Анандарам Харишчандер** — Ph.D., Dr.Sci., доцент, школа искусственного интеллекта "амрита", Амрита Вишва Видьяпитхам (Университет Амриты). Область научных интересов: биология вычислительных систем, функциональная геномика, биология

развития. Число научных публикаций — 43. a_harishchander@cb.amrita.edu; Амританагар, деревня Эттимадай, 641112, Коимбатур, Тамилнад, Индия; р.т.: +91(9940)066-227.

**Таплиял Мадхур** — доцент, Университет Graphic Era Hill. Область научных интересов: кибербезопасность, обработка изображений, глубокое обучение, машинное обучение. Число научных публикаций — 3. madhurthapliyal@gehu.ac.in; Белл-роуд, Клемент-Таун, 248002, Дехрадун, Уттаракханд, Индия; р.т.: +91(7017)747-897.

**Джоши Капил** — Ph.D., Dr.Sci., доцент, технологический институт уттаранчала, Университет Уттаранчала. Область научных интересов: слияние изображений, обработка изображений, глубокое обучение, CNN. Число научных публикаций — 132. Kapilengg0509@gmail.com; Прем Нагар, 248007, Дехрадун, Уттаракханд, Россия; р.т.: +91(8979)799-289.

**DOI** 10.15622/ia.23.2.8

V. ROMANIUK, A. KASHEVNIK

# INTELLIGENT EYE GAZE LOCALIZATION METHOD BASED ON EEG ANALYSIS USING WEARABLE HEADBAND

*Romaniuk V., Kashevnik A.* **Intelligent Eye Gaze Localization Method Based on EEG Analysis Using Wearable Headband.**

**Abstract.** In the rapidly evolving digital age, human-machine interface technologies are continuously being improved. Traditional methods of computer interaction, such as a mouse and a keyboard, are being supplemented and even replaced by more intuitive methods, including eye-tracking technologies. Conventional eye-tracking methods utilize cameras to monitor the direction of gaze but have their limitations. An alternative and promising approach for eye-tracking involves the use of electroencephalography, a technique for measuring brain activity. Historically, EEG was primarily limited to laboratory conditions. However, mobile and accessible EEG devices are entering the market, offering a more versatile and effective means of recording bioelectric potentials. This paper introduces a gaze localization method using EEG obtained from a mobile EEG recorder in the form of a wearable headband (provided by BrainBit). The study aims to decode neural patterns associated with different gaze directions using advanced machine learning methods, particularly neural networks. Pattern recognition is performed using both ground truth data collected from wearable camera-based eye-tracking glasses and unlabeled data. The results obtained in this research demonstrate a relationship between eye movement and EEG, which can be described and recognized through a predictive model. This integration of mobile EEG technology with eye-tracking methods offers a portable and convenient solution that can be applied in various fields, including medical research and the development of more intuitive computer interfaces.

**Keywords:** eye-tracking, EEG, neural networks, wearable EEG, supervised learning, unsupervised learning.

**1. Introduction.** In an ever-evolving world humans predominantly rely on vision as the primary conduit for gathering information and making decisions. This centrality of vision is mirrored in modern computing interfaces, which are predominantly graphical and designed for interaction through screens. As technology advances, new methods of control – ranging from body movements to eye movements, speech, and even brain activity – are being developed to foster more natural and intuitive human-computer interactions.

Eye-tracking technologies have witnessed significant advancements in terms of accessibility and ease of use. These technologies record eye movements to pinpoint an individual's focal point and are increasingly being employed in both academic research and commercial applications. Traditional eye-tracking methods often utilize video cameras to capture the shape of the pupil or other markers. While effective, these methods come with limitations, such as sensitivity to light levels and the necessity for open eyes [1].

An alternative and promising approach to eye tracking is the use of electroencephalography, a technique for measuring brain activity. Eye

movements affect EEG recordings by adding muscle and eye dipole potentials to signals recorded by EEG electrodes [2]. This effect can be used to extract eye movements from recording. Like traditional eye-tracking methods, EEG does come with its own set of challenges. It is highly constrained by environmental factors, such as electromagnetic interference, but it doesn't require specific lighting conditions or opened eyes, making it versatile in different scenarios. Historically, EEG was predominantly limited to laboratory settings and required specialized equipment and trained personnel. However, mobile and affordable EEG devices are revolutionizing this domain, offering a more versatile yet effective means of capturing biopotentials [3].

This study aims to investigate the correlation between EEG collected by BrainBit – wearable headband [4], and eye movements recorded by eye tracker. We hypothesize that there exists a correlation between eye gaze direction based on a change of coordinates for 0.1 second and electrical activity from O1, O2, T3, and T4 leads recorded by wearable EEG. We aim to decode the neural signatures associated with different gaze directions using advanced machine learning techniques, particularly neural networks. Our approach combines the strengths of both EEG and eye-tracking technologies, offering a comprehensive perspective on gaze localization. The use of a wearable EEG headband facilitates data collection in more naturalistic settings, enhancing the ecological validity of our findings.

Objectives of the study:

1. Collect and find recordings of eye activity recorded by camera-based devices and EEG recorded by wearable devices;

2. Preprocess and normalize the data;

3. Develop predictive models using supervised and unsupervised machine learning methods.

The scientific novelty of the paper includes employing a wearable EEG headband for collecting a unique dataset with an uncommonly low number of EEG channels and a correlational research of this data with eye movements to localize gaze.

The rest of the paper is structured as follows: Section II provides a comprehensive review of existing literature on eye movements, EEG data, and the challenges posed by artifacts in EEG data. It also discusses the integration of EEG and eye-tracking, the challenges and solutions associated with mobile EEG systems, and the combination of EEG and eye-tracking in mobile scenarios. Section III explains the methodology of the study. It begins with a general description of the study's objectives and approach. The dataset subsection provides details about the participants, experimental setup, and methodology for the two used in the study datasets. The first dataset is collected

during this study using the BrainBit EEG headband and PupilLabs eye tracker. The second one is the open NeuMa dataset [5]. The preprocessing subsection discusses the challenges and solutions for handling eye movements in EEG data. The neural network architecture subsection describes the supervised and unsupervised machine learning models used in the study. Supervised learning is presented by feedforward and recurrent neural networks. They use previously discussed datasets to build predictive models based on known ground truth. Unsupervised learning is presented by clustering datasets with time-specific distance calculation. Section IV presents the results of the study. It provides a detailed analysis of the performance of different machine-learning models. The results from both supervised and unsupervised learning approaches are discussed. Section V summarizes the main findings of the study, discusses its implications, and suggests directions for future research.

**2. Related work.** The integration of EEG and eye-tracking technologies has garnered significant attention in the realm of cognitive and neuroscientific research over the past years. Numerous studies have sought to harness the complementary strengths of these two methodologies.

In study [6] the authors provide a comprehensive discussion on the challenges posed by eye movements in EEG data. They emphasize the importance of identifying and effectively correcting them to ensure the accurate interpretation of underlying neural signals.

The use of Independent Component Analysis (ICA) combined with a high temporal resolution eye tracking has been highlighted as a promising approach for identifying and correcting ocular artifacts in laboratory EEG data [7]. In study [8] the authors introduced the VME-DWT algorithm, which efficiently detects and eliminates eye blinks from short segments of single EEG channels using Variational Mode Extraction (VME) and the automatic Discrete Wavelet Transform (DWT) algorithm. In [9] the authors developed the "Optimized Fingerprint Method" that utilizes spatial, temporal, spectral, and statistical features to automatically classify artifactual independent components in EEG, achieving over 90% accuracy in identifying artifacts of physiological origin. In study [10] the authors proposed a framework combining unsupervised machine learning with singular spectrum analysis (SSA) to remove eye blink artifacts without altering the uncontaminated EEG regions.

The integration of EEG and eye tracking provides a comprehensive understanding of cognitive processes during visual tasks. In paper [7] the authors developed a system that captures EEG signals during eye movement and employs a random forests classification algorithm to categorize them into 6 classes – eyes open, close, left, right, up, down.

In [11] the authors introduced the BeMoBIL Pipeline, a MATLAB-based solution that supports the synchronized handling of multimodal data, including EEG and eye tracking. It presents a new robust method for region-of-interest-based group-level clustering of independent EEG components.

In paper [12] the authors explored a multimodal approach for identifying Autism Spectrum Disorders of children by fusing EEG and eye-tracking data, demonstrating the potential of such integrative methods in clinical applications. The approach consists of extracting EEG and eye-tracking features from data and using two separate deep learning models for feature processing accordingly at the first step and one deep learning model for processing the outputs of the first step.

In study [13] the authors used an eye tracker to improve the detection of evoked responses to complex visual stimuli during EEG by excluding moments when the gaze was disoriented. This approach increased the accuracy of detection by 15%.

In paper [14] the authors employed a wearable EEG headset with stationary eye tracker for prediction of decisions made during the product design selection. They concluded that the fusion of eye movements and EEG characteristics can significantly improve the efficiency of decision-making in projects compared to using a single data processing method. In their experiment, the accuracy improvement was more than 10%.

Mobile EEG systems, while offering the advantage of capturing brain activity in naturalistic settings, come with inherent challenges.

One of the challenges associated with mobile EEG is the reduced number of channels compared to traditional stationary EEG systems. This limitation can potentially impact the quality and interpretability of the recorded data.

In [15] the authors conducted a study where participants performed an auditory oddball task while concurrently completing various motor tasks outdoors. The study utilized a 30-channel mobile EEG montage and observed that increased movement complexity imposed a higher workload on the cognitive system, effectively reducing the availability of cognitive resources for the cognitive task.

In paper [16] the authors explored a human EEG-based emergency stop interface designed to activate when the human operator detects or foresees a potential emergency. The study employed a mobile EEG recorder with 14 channels and utilized a decision tree for classifying the operator's state. While the use of mobile EEG introduced complexities to the classification task, the findings indicated consistent EEG signal patterns across various potential emergencies.

In study [17] the authors specifically addressed the challenges of Independent Component Analysis (ICA) decomposition in both mobile and stationary EEG experiments. They found that while commonly used settings (like stationary experiments with 64 channels and a 0.5 Hz filter) yield acceptable ICA results, mobile experiments with fewer channels require higher high-pass filter cutoff frequencies for optimal decomposition.

In study [18] the authors researched the existence of cardiogenic artifacts in EEG recorded by single-channel mobile EEG and proposed an algorithm for automated artifact detection and removal.

These studies underscore the importance of considering the limitations and specific requirements of mobile EEG systems, especially when working with a reduced number of channels. While mobile EEG offers unique opportunities for research in real-world settings, careful preprocessing and data analysis are crucial to account for the challenges posed by the limited channel count.

The combination of EEG and eye tracking in mobile scenarios offers valuable insights into cognitive processes during visual tasks.

In [19] the authors explored this domain by investigating the automatic detection of visual attention using pre-trained computer vision models in conjunction with human gaze in mobile eye-tracking scenarios.

In [20] the authors investigated the impact of swiping direction on the interaction performance using mobile EEG and eye-tracking technology.

Table 1 provides a summary of these works. Many researchers are affected by eye activity in their data and remove it as well as use of additional devices to collect such activity. There is no one-size-fits-all solution. The choice of method often depends on the specific application, the nature of the artifacts, and the constraints of the mobile device. Continuous research and development in the area of mobile EEG and eye tracking are essential to further enhance the reliability and utility in real-world scenarios for improving and extending existing ways of working with human activity.

**3. Method.** The primary objective of this study is to investigate the intricate relationship between eye movements and neural activity as captured by EEG recordings. The method of this study includes dataset collection and selection, preprocessing of this data and applying two different deep learning techniques – supervised and unsupervised learning of predictive models.

Table 1. Summary of the related work

| Paper | Eye-tracking | Wearable EEG | Identifying eye activity in EEG | Removing eye activity from EEG | Description of found activity | Combining eye tracking and EEG features | Find relation between eye activity and EEG | Research differences of mobile EEG |
|---|---|---|---|---|---|---|---|---|
| Plöchl et al. [6] | - | - | ICA, regression | + | + | - | + | - |
| Antoniou et al. [7] | + | - | + | + | + | - | + | - |
| Shahbakhti et al. [8] | - | - | VME | Blinks only | + | - | + | - |
| Stone et al. [9] | - | - | + | + | - | - | + | - |
| Maddirala and Veluvolu [10] | - | + | SSA | Blinks only | + | - | + | - |
| Klug et al. [11] | - | - | ICA | Blinks only | - | - | + | - |
| Han et al. [12] | - | - | - | - | - | + | - | - |
| Ahtola et al. [13] | + | - | - | Bandpass filter | - | + | - | - |
| Wang et al. [14] | + | + | - | Bandpass filter | - | + | - | - |
| Reiser et al. [15] | - | + | ICA | + | - | - | + | - |
| Buerkle et al. [16] | - | + | - | Bandpass filter | - | - | - | - |
| Klug and Gramann [17] | - | + | ICA | + | - | - | - | + |
| Chiu at el. [18] | - | + | - | - | - | - | - | + |
| Barz and Sonntag [19] | + | - | - | - | - | - | - | - |
| Zhou et al. [20] | + | + | - | - | - | - | + | - |

**3.1. General Description.** The study involves the collection and analysis of EEG and eye-tracking data from participants engaged in predefined tasks. The data is then subjected to a series of preprocessing steps to extract meaningful patterns and eliminate potential noise or artifacts. Subsequent to preprocessing, the data is fed into neural network models, both supervised and unsupervised, to discern patterns and relationships.

Figure 1 provides a visual overview of the entire process, showing the flow from data collection to analysis. The subsequent sections explain the specifics of the dataset, the experimental setup, and the models employed. The first section describes data collection in terms of devices, tasks and information that we collected. The second section gives an overview of datasets with the recordings of information. The third section lists the models and classes that were used to do predictions.

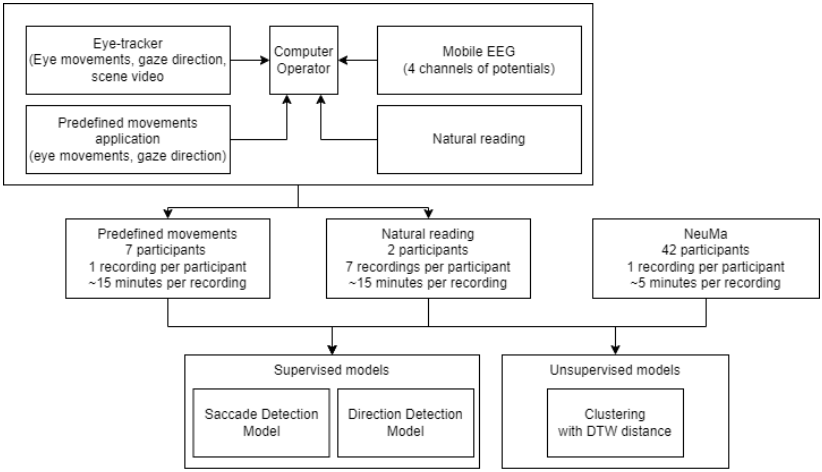



Fig. 1. A general description of the proposed method (from data collection with the signal collecting sources to deep learning models that use it)

**3.2. Dataset.** The dataset used in a study plays a pivotal role in shaping the outcomes and conclusions drawn from the research. In this section, we provide a comprehensive overview of two datasets employed in this study and preprocessing of these data. The first dataset is collected for this study using wearable devices. The second an one is open NeuMa dataset that can be used in the same way as the first o ne. The preprocessing part includes specifics of work and extracting valuable information from EEG and eye-tracking data.

**3.2.1. Our Dataset.** A new dataset was recorded for the purpose of this study. Data were collected during two tasks: predefined movements

and natural reading. The recording involved the use of a wearable EEG recorder in the form of a headband and a wearable eye tracker in the form of glasses.

Seven (6 males; 1 female; age $21 \pm 3$ years) healthy volunteers with no neurological or vision deficits participated in this study. All seven participants were recorded in an experiment with predefined movements for $\approx 15$ minutes, in total 90 minutes of recording. Two participants were recorded in 6 sessions of natural reading experiments 15 minutes each, in total of 180 minutes of recording.

EEG was recorded using a 4-channel mobile band (BrainBit, 100Hz sampling rate). Dry electrodes were placed at O1, O2, T3, and T4 points according to the 10/20 system. Eye gaze data was collected using glasses with video cameras that point to the pupils (Pupil Invisible [21], 60Hz sampling rate). The computer showing tasks with a Full HD resolution 23-inch display was positioned at a distance of 1 meter in front of the person with a refresh rate of 60Hz.

The structure of the experiments remained consistent regardless of the specific task assigned to the participants. The typical procedure for each experiment is as follows:

1. The participant sits in a comfortable position in front of a computer screen, at a distance of 1 meter. Relaxation and minimization of body movements are emphasized.

2. Connection and calibration of the glasses are performed.

3. Connection and verification of electrode contact with the mobile recorder are conducted.

4. Instructions specific to the current experiment are provided.

5. The participant performs the assigned task.

6. Devices are disconnected and the experiment concludes.

For the predefined movement experiment, a graphical application was developed. This application displayed fixation points alternately to the participant. These fixation points were located centrally and at 16 surrounding points as shown in Figure 2. The participants would fixate on each point in a randomized sequence, and then return to the central point. The points alternate each other in the interval from 2.5 to 3.5 seconds.

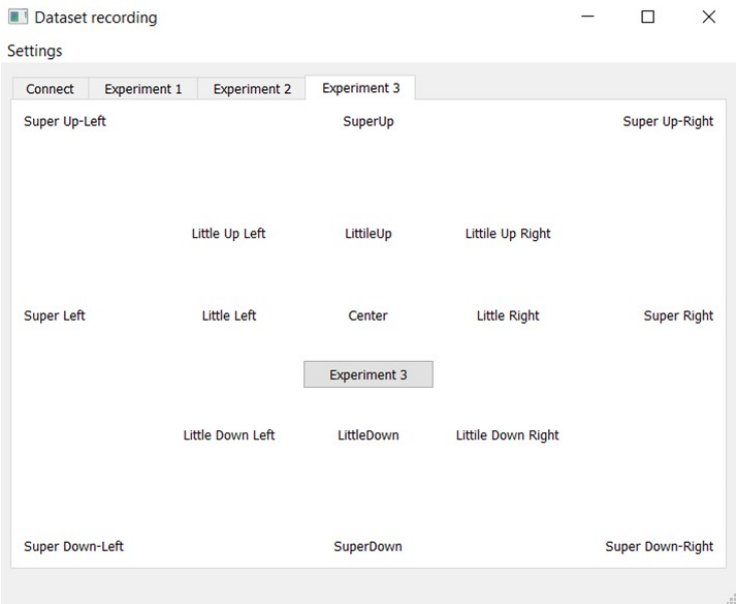



Fig. 2. The interface of an application for predefined movement experiments showing the position of points before the start

**3.2.2. NeuMa Dataset.** In addition to our dataset, an open dataset with the recorded EEG and eye tracker data named NeuMa was also utilized [5].

The NeuMa dataset stores raw experimental data for 42 subjects (23 males and 19 females, aged $31.5 \pm 8.84$).

Structure of dataset:

1. EEG Data: Continuous mode brain activity recording. This includes a time series of the 128 channels of EEG activity and corresponding timestamps recorded at 600 Hz.

2. Eye Tracker Data: Gaze data metrics for both left and right eyes recorded in 200Hz.

3. Mouse Clicks: Sequence of mouse clicks.

4. Mouse Positions: 2D screen coordinates corresponding to each mouse click.

5. Markers: Information regarding alterations among brochure pages, initiation, and completion of the experiment.

During the NeuMa dataset experimental procedure the participants were seated comfortably in an armchair positioned 50 cm away from a 28-inch LCD monitor. Although the participants had the freedom to move their heads

during the procedure, they were advised to restrict their movements, including those of the head, to reduce potential artifacts in the EEG signals. However, they were also encouraged to ensure their comfort to prevent any negative impact on their overall experience. Before the presentation of the products, a resting state EEG was recorded for a duration of two minutes. Following the resting state recording, the participants were presented with brochures. They were allowed to navigate freely through these brochures using the left and right arrow keys on the keyboard to move forward and backward.

**3.2.3. Dataset comparison.** A wearable EEG recorder and a camera-based eye tracker were used to record both of these datasets according to our objectives. Despite this, it had different additional channels of information and different amounts of the recorded EEG channels. A comparison is presented in Table 2.

For the purpose of this article, only 4 channels (O1, O2, T3, T4) from the NeuMa dataset were selected and used. This decision was based on the fact that our dataset only contained these channels, ensuring consistency and comparability.

Table 2. The comparison of our and NeuMa datasets

| Name | Eye tracking | Wearable EEG | EEG channels | EEG rate | Mouse data | Electrodes placement system |
|---|---|---|---|---|---|---|
| Our | + | + | 4 | 100 Hz | - | 10/20 |
| NeuMa | + | + | 128 | 600 Hz | + | 10/20 |

**3.2.4. Preprocessing.** In the recordings obtained with eye-tracking glasses, two key moments are distinguished: fixations and saccades.

Fixations refer to the concentration of a person's attention on a specific point in the visual field, indicated by reduced eye movement amplitudes. During fixations, the brain processes the visual information from the point of focus, making it a crucial moment for understanding cognitive processes and attention. Saccades, on the other hand, are rapid eye movements that shift the gaze from one fixation point to another. These movements are essential for redirecting the line of sight to new areas of interest. The amplitudes of movements during saccades and fixations differ by an order of magnitude. In data processing, the period between saccades is considered as fixations because the brain is actively processing visual information during these periods, while saccades themselves are represented by significant changes in gaze coordinates, indicating shifts in attention. The detected saccades along with EEG data are presented in Figure 3 as vertical lines denoting the start of the saccade on the eye gaze coordinates graph.

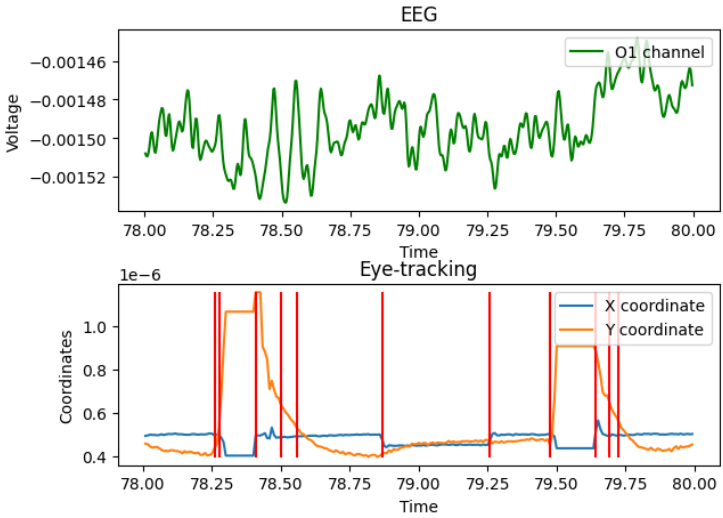



Fig. 3. Saccades on the graph of eye gaze coordinates along with EEG data

Eye movements, particularly saccades, induce large electrical potentials due to the movement of the eye's retinal dipole. The retina has a natural electrical polarity, with the front of the eye being positively charged and the back being negatively charged. When the eyes move, this retinal dipole also shifts its orientation. This movement generates electrical fields that can propagate through the tissues of the head and influence the electrical recordings on the scalp, including those of EEG. Because the eyes are anatomically close to the frontal EEG electrodes, these electrical fields generated by the retinal dipole can have a significant impact on EEG recordings. The influence of the retinal dipole's movement is so substantial that it can sometimes be mistaken for brain activity if not properly accounted for.

The EEG data is divided into series with a duration of 300 ms – the average duration of a saccade. EEG is highly dependent on the physiological state of the participant. The values of the potentials are not constant even within a single individual throughout the day due to factors like fatigue, caffeine intake, or even time of day. Therefore, each series is normalized to the average change in the potential to account for these variations and ensure that the data are comparable across different time points and participants. Normalization helps in emphasizing the relative changes in EEG signals, which are more informative than the absolute values. It can be shown as $X_{\text{normalized}} = \frac{X - median(X)}{mean(X - median(X))}$, where $X$ is a series of one EEG channel signal in the form of voltage value.

Additionally, EEG is influenced by physical parameters of the surrounding environment, such as electromagnetic oscillations, high-frequency signals, and other phenomena. Such noise can be eliminated using frequency filters. The relevant brain activity signals are typically found within the range of 1 to 40 Hz; it is common to remove other frequencies. However, it is necessary to retain low frequencies from 0 to 2 Hz. These are EEG oscillations of sufficient duration that are associated with eye movement and provide valuable insights into the correlation between eye movements and brain activity.

**3.3. Model Architecture.** Collected and preprocessed datasets are used to build predictive models of eye movements. This can be done in two ways: with the use of ground truth (known eye movements) or without it. These approaches are called supervised and unsupervised learning, respectively. Supervised learning includes the use of two types of neural networks to classify data into different classes based on direction and amplitude. Unsupervised learning consists of clustering data using the k-means algorithm with time-specific distance calculations.

**3.3.1. Supervised learning.** Supervised learning models require a structured data for training and testing. The dataset was split into the train, validation and test parts as 64%, 16% and 20% accordingly. Classes in every part have a balanced amount of entries. So, the saccade classification task has a dataset with 50% of saccades and 50% of fixations, direction classification has 12.5% of each direction.

Feedforward neural networks and recurrent neural networks are commonly used to determine eye activity due to their architecture and effectiveness in pattern recognition. The most effective configuration of a feedforward neural network consisted of a network with an input of 120 our dataset or 800 NeuMa dataset EEG points – they represent 300 ms of 4-channel EEG recording from the dataset, 1 hidden layer with 8 neurons, and output with 8 neurons for direction classification and 2 for saccade classification. ReLu is used as an activiation function. Optimization is done by the Adam algorithm.

Unlike the previous approach, a recurrent neural network can feed its output back as input in addition to the new signal. This allows for processing sequences of variable lengths, rather than strictly fixed networks. There are several types of neural networks that implement this principle, such as RNN, LSTM, and GRU. The key difference among them is their ability to remember and forget data from previous iterations. The best results were obtained using a GRU network with 2 hidden layers of 16 neurons each, where 4 potential values from different electrodes were sequentially provided as an input.

Support Vector Machine offers a distinct approach to the classification of EEG data, complementing the feedforward and recurrent neural networks

discussed earlier. Its core principle involves finding the optimal hyperplane that distinctly classifies the data points into different categories. The model would treat each EEG series as a feature vector in a high-dimensional space. The SVM would then find the hyperplane that best separates saccades from fixations or classifies the direction of eye movement.

The obtained results are presented in Table 3.

Table 3. Comparison of the results of supervised learning models

| Model | Dataset | Classes | Accuracy | Recall | Precision | F1 |
|---|---|---|---|---|---|---|
| FF | Our | 2 | 73% | 72% | **66%** | 69% |
| | | 8 | **66%** | **67%** | **56%** | **62%** |
| | NeuMa | 2 | 73% | 74% | 73% | 71% |
| | | 8 | 62% | 62% | 62% | 62% |
| RNN | Our | 2 | **76%** | **75%** | 65% | **70%** |
| | | 8 | 61% | 69% | 55% | 62% |
| | NeuMa | 2 | 73% | 64% | **81%** | 66% |
| | | 8 | **68%** | **68%** | **68%** | **67%** |
| SVM | Our | 2 | 54% | 66% | 51% | 58% |
| | | 8 | 62% | 59% | 56% | 57% |
| | NeuMa | 2 | **81%** | **81%** | **81%** | **81%** |
| | | 8 | 55% | 40% | 39% | 35% |

### 3.4. Unsupervised Learning

**3.4.1. Unsupervised learning.** Data clustering is a pivotal technique in the realm of data analytics and machine learning. It involves grouping data points into distinct clusters or sets based on intrinsic patterns or similarities. Unlike supervised learning paradigms where data is labeled, and models are trained to recognize these labels, clustering operates in the unsupervised domain. In unsupervised data clustering, the algorithm sifts through datasets without the guidance of ground truth labels. Instead, it relies solely on the intrinsic differences and similarities between the data points or series.

Among the many clustering techniques available, the k-means clustering algorithm has garnered widespread acceptance and utilization. The crux of the k-means algorithm lies in partitioning the dataset into $k$ distinct clusters. These clusters are formed by minimizing the distance between data points within the cluster and maximizing the distance to data points in other clusters.

While the k-means algorithm predominantly utilizes the Euclidean distance to ascertain the difference between series, our dataset mandated a slightly nuanced approach. Given the temporal nature of our data, traditional distance measures might fail to capture the underlying intricacies. For instance, physiological factors such as reaction time, event duration, and amplitude can variably affect the series, leading to potential discrepancies in clustering outcomes.

To mitigate these potential inconsistencies, we employed the Dynamic Time Warping (DTW) algorithm. At its core, DTW is a time-scale transformation technique. Unlike traditional distance computation methods that compare data points in a point-to-point fashion, DTW aligns the two series in a way that the alignment minimizes the overall distance. Essentially, DTW can stretch or compress the series along the temporal axis to achieve an optimal alignment. This transformation accounts for time-dependent variances like elongated reactions or variations in amplitude, ensuring a more robust clustering outcome.

## 4. Results

**4.1. Supervised Learning.** The performance of different machine learning models on both the Our and NeuMa datasets is summarized in Table 3. The metrics used for evaluation include Accuracy, Recall, Precision, and F1 Score. These metrics provide a comprehensive view of the model performance, taking into account both the true positive rate and the false positive rate.

Among the three models, RNN showed the most balanced performance across different metrics, making it a strong candidate for further optimization and real-world testing. However, the SVM's strong performance on the NeuMa dataset suggests that with sufficient data, simpler models can also achieve high accuracy.

In the segmentation of the dataset, we discerned nine distinct clusters.

The unsupervised learning approach, specifically clustering using the k-means algorithm combined with Dynamic Time Warping (DTW) for distance computation, resulted in distinct clusters representing different gaze trajectories.

The number of clusters was determined using a widely accepted method for finding the optimal number, known as the 'elbow method.' In the context of k-means, the elbow method involves plotting the total within-cluster variation against the number of clusters. As the number of clusters increases, this variation decreases. However, there is a point, resembling the bend in an elbow, where the rate of decline sharply changes, indicating the optimal number of clusters for the dataset. The critical points for the dataset are illustrated in Figure 5. There are multiple points on the graph and interpretations of each clusterization result remain ambiguous. Upon scrutinizing video segments associated with each cluster, following the segmentation to every $k$ amount of clusters shown in Figure 5, we observed congruent behaviors among the participants. Notably a consistent shift in gaze in a uniform direction when $k$ equals 9. This observation is substantiated by an evaluation of the mean displacement along both the vertical and horizontal axes.

As depicted in Figure 4, the x and y axes represent these respective displacements. Each cluster has a distinction from all others at a minimum

of 0.1 distance along one or two axes. This distance represents 10% human visual field. The visualization distinctly demarcates eight saccade trajectories and a central fixation, which collectively characterize the identified clusters.

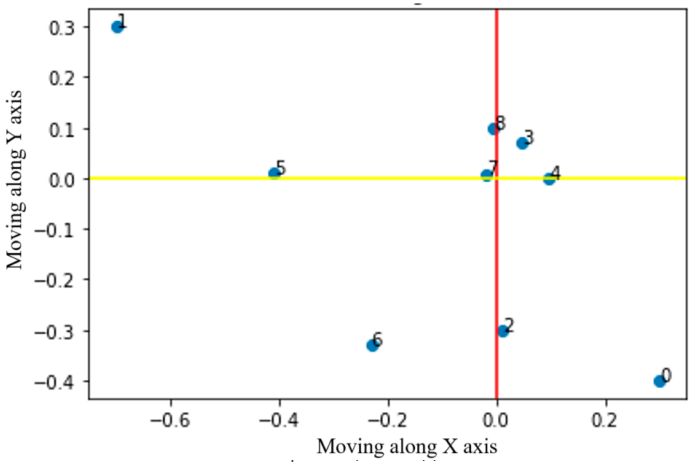



Fig. 4. Mean movement distance for every cluster along vertical and horizontal axis

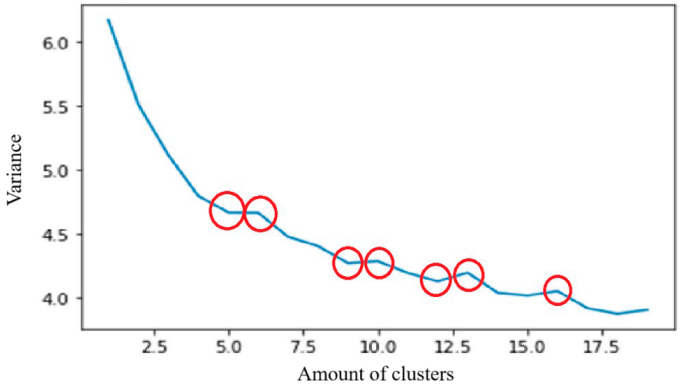



Fig. 5. Graph delineating the relationship between the number of clusters and the corresponding total within-cluster variation

The unsupervised learning approach, specifically the clustering using the k-means algorithm combined with Dynamic Time Warping (DTW) for distance computation, yielded distinct clusters that represent different gaze trajectories.

**5. Conclusion.** This study represents an endeavor in the realm of eye-tracking technologies, specifically focusing on the integration of wearable EEG headbands and machine-learning techniques. The results affirm that there is a discernible correlation between eye movements and EEG signals. This opens new directions for non-intrusive eye-tracking methods that can operate in various environmental conditions. The use of BrainBit as a mobile EEG recorder has proven to be effective for tracking eye activity. This is a significant step towards making eye-tracking technology more accessible and versatile. The study also introduces a methodology for the automatic labeling of EEG datasets, which can significantly expedite the data analysis process. Both supervised and unsupervised machine-learning techniques were employed to analyze the EEG data, demonstrating promising results in terms of accuracy, precision, and F1 scores.

This research has multiple implications. In the medical field, such technology could be used for diagnosing and monitoring neurological conditions. In human-computer interaction, it could lead to the development of more intuitive and responsive interfaces. Moreover, the technology has the potential to be used in safety-critical applications, such as fatigue detection in drivers.

While this study lays the groundwork for mobile EEG-based eye tracking, there are several avenues for future research:

– Optimizing Machine Learning Models: Further tuning of the neural network architectures could lead to even more accurate results.

– Real-world Applications: Testing the technology in real-world scenarios, such as driving or operating machinery, would provide valuable data on its effectiveness and limitations.

– Multi-modal Approaches: Combining EEG with other biometric data could offer a more comprehensive understanding of human behavior and cognitive states.

In conclusion, this study marks a significant step forward in the integration of EEG technology and eye-tracking, offering a potentially transformative approach to understanding human cognition and behavior.

## References

1. Holmqvist K., Nystrom M., Mulvey F. Eye tracker data quality: What it is and how to measure it. Proceedings of the Eye Tracking Research and Applications Symposium (ETRA). 2012. pp. 45–52.

2. Jagla F., Jergelova M., Riecansky I. Saccadic eye movement related potentials. Physiological Research. 2007. vol. 56. no. 6. pp. 707–713. DOI: 10.33549/physiolres.931368.
3. Krigolson O.E., Williams C.C., Norton A., Hassall C.D., Colino F.L. Choosing MUSE: Validation of a Low-Cost, Portable EEG System for ERP Research. Frontiers in Neuroscience. 2017. vol. 11. DOI: 10.3389/fnins.2017.00109.
4. Brainbit. Brainbit Manual. Available at: http://brainbit.com/ (accessed 09/01/2023).
5. Georgiadis K., Kalaganis F.P., Riskos K., Matta E., Oikonomou V.P., Yfantidou I., Chantziaras D., Pantouvakis K., Nikolopoulos S., Laskaris N.A., Kompatsiaris I.. NeuMa – the absolute Neuromarketing dataset en route to an holistic understanding of consumer behavior. Scientific Data. 2023. vol. 10(1). no. 508. DOI: 10.1038/s41597-023-02392-9.
6. Plochl M., Ossandon J., Konig P. Combining EEG and eye tracking: identification, characterization, and correction of eye movement artifacts in electroencephalographic data. Frontiers in Human Neuroscience. 2012. vol. 6. no. 278. URL: https://www.frontiersin.org/articles/10.3389/fnhum.2012.00278.
7. Antoniou E., Bozios P., Christou V., Tzimourta K.D., Kalafatakis K.G. Tsipouras M., Giannakeas N., Tzallas A.T. EEG-Based Eye Movement Recognition Using Brain-Computer Interface and Random Forests. Sensors. 2021. vol. 21. no. 7. no. 2339. DOI: 10.3390/s21072339.
8. Shahbakhti M., Beiramvand M., Nazari M., Broniec-Wojcik A., Augustyniak P., Rodrigues A.S., Wierzchon M., Marozas V. VME-DWT: An Efficient Algorithm for Detection and Elimination of Eye Blink From Short Segments of Single EEG Channel. IEEE Transactions on Neural Systems and Rehabilitation Engineering. 2021. vol. 29. pp. 408–417.
9. Stone D.B., Tamburro G., Fiedler P., Haueisen J., Comani S. Automatic Removal of Physiological Artifacts in EEG: The Optimized Fingerprint Method for Sports Science Applications. Frontiers in Human Neuroscience. 2018. vol. 12. no. 96.
10. Maddirala A.K., Veluvolu K. Eye-blink artifact removal from single channel EEG with k-means and SSA. Scientific Reports. 2021. vol. 11(1). no. 11043.
11. Klug M., Jeung S., Wunderlich A., Gehrke L., Protzak J., Djebbara Z., Argubi-Wollesen A., Wollesen B., Gramann K. The BeMoBIL Pipeline for automated analyses of multimodal mobile brain and body imaging data. bioRxiv. 2022. DOI: 10.1101/2022.09.29.510051.
12. Han J., Jiang G., Ouyang G., Li X. A Multimodal Approach for Identifying Autism Spectrum Disorders in Children. IEEE Transactions on Neural Systems and Rehabilitation Engineering. 2022. vol. 30. pp. 2003–2011.
13. Ahtola E., Stjerna S., Stevenson N., Vanhatalo S. Use of eye tracking improves the detection of evoked responses to complex visual stimuli during EEG in infants. Clinical Neurophysiology Practice. 2017. vol. 2. pp. 81–90. URL: https://www.sciencedirect.com/science/article/pii/S2467981X17300070.
14. Wang Y., Yu S., Ma N., Wang J., Hu Z., Liu Z., He J. Prediction of product design decision Making: An investigation of eye movements and EEG features. Advanced Engineering Informatics. 2020. vol. 45. no. 101095. DOI: 10.1016/j.aei.2020.101095.
15. Reiser J., Wascher E., Arnau S. Recording mobile EEG in an outdoor environment reveals cognitive-motor interference dependent on movement complexity. Scientific Reports. 2019. vol. 9(1). no. 13704.
16. Buerkle A., Bamber T., Lohse N., Ferreira P. Feasibility of Detecting Potential Emergencies in Symbiotic Human-Robot Collaboration with a mobile EEG. Robotics and Computer-Integrated Manufacturing. 2021. vol. 72. no. 102179.

17. Klug M., Gramann K. Identifying key factors for improving ICA-based decomposition of EEG data in mobile and stationary experiments. European Journal of Neuroscience. 2021. vol. 54. no. 12. pp. 8406–8420.
18. Chiu N.-T., Huwiler S., Ferster M.L., Karlen W., Wu H.-T., Lustenberger C. Get rid of the beat in mobile EEG applications: A framework towards automated cardiogenic artifact detection and removal in single-channel EEG. Biomedical Signal Processing and Control. 2022. vol. 72. no. 103220. DOI: 10.1016/j.bspc.2021.103220.
19. Barz M., Sonntag D. Automatic Visual Attention Detection for Mobile Eye Tracking Using Pre-Trained Computer Vision Models and Human Gaze. Sensors. 2021. vol. 21(12). no. 4143. DOI: 10.3390/s21124143.
20. Zhou C., Shi Z., Huang T., Zhao H., Kaner J. Impact of swiping direction on the interaction performance of elderly-oriented smart home interface: EEG and eye-tracking evidence. Frontiers in Psychology. 2023. vol. 14.
21. Tonsen M., Baumann C., Dierkes K. A High-Level Description and Performance Evaluation of Pupil Invisible. arXiv preprint arXiv:2009.00508. 2020.

**Romaniuk Vladimir** — Junior researcher, Laboratory of integrated automation systems, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS). Research interests: neural networks, human state, EEG. romaniukvr@yandex.ru; 39, 14-th Line V.O., 199178, St. Petersburg, Russia; office phone: +7(812)328-8071.

**Kashevnik Alexey** — Ph.D., Senior researcher, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS). Research interests: intelligent transport systems, human monitoring, knowledge management, cloud computing, human-computer interaction, user profiling, ontologies, smart spaces. The number of publications — 250. alexey@iias.spb.su; 39, 14-th Line V.O., 199178, St. Petersburg, Russia; office phone: +7(812)328-8071.

**УДК** 004.93 **DOI** 10.15622/ia.23.2.8

В.Р. Романюк, А.М. Кашевник

# МЕТОД ИНТЕЛЛЕКТУАЛЬНОЙ ЛОКАЛИЗАЦИИ ВЗГЛЯДА НА ОСНОВЕ АНАЛИЗА ЭЭГ С ИСПОЛЬЗОВАНИЕМ НОСИМОЙ ГОЛОВНОЙ ПОВЯЗКИ

*Романюк В.Р., Кашевник А.М.* **Метод интеллектуальной локализации взгляда на основе анализа ЭЭГ с использованием носимой головной повязки.**

**Аннотация.** В стремительно развивающейся цифровой эпохе интерфейсы человеко-машинного взаимодействия непрерывно совершенствуется. Традиционные методы взаимодействия с компьютером, такие как мышь и клавиатура, дополняются и даже заменяются более интуитивными способами, которые включают технологии отслеживания глаз. Обычные методы отслеживания глаз используют камеры, которые отслеживают направление взгляда, но имеют свои ограничения. Альтернативным и многообещающим подходом к отслеживанию глаз является использование электроэнцефалографии, техники измерения активности мозга. Исторически ЭЭГ была ограничена в основном лабораторными условиями. Однако мобильные и доступные устройства для ЭЭГ появляются на рынке, предлагая более универсальное и эффективное средство для регистрации биопотенциалов. В данной статье представлен метод локализации взгляда с использованием электроэнцефалографии, полученной с помощью мобильного регистратора ЭЭГ в виде носимой головной повязки (компании BrainBit). Это исследование направлено на декодирование нейрональных паттернов, связанных с разными направлениями взгляда, с использованием продвинутых методов машинного обучения, в частности, нейронных сетей. Поиск паттернов выполняется как с использованием данных, полученных с помощью носимых очков с камерой для отслеживания глаз, так и с использованием неразмеченных данных. Полученные в исследовании результаты демонстрируют наличие зависимости между движением глаз и ЭЭГ, которая может быть описана и распознана с помощью предсказательной модели. Данная интеграция мобильной технологии ЭЭГ с методами отслеживания глаз предлагает портативное и удобное решение, которое может быть применено в различных областях, включающих медицинские исследования и разработку более интуитивных компьютерных интерфейсов.

**Ключевые слова:** отслеживание глаз, ЭЭГ, нейронные сети, носимый ЭЭГ, обучение с учителем, обучение без учителя.

## Литература

1. Holmqvist K., Nystrom M., Mulvey F. Eye tracker data quality: What it is and how to measure it // Proceedings of the Eye Tracking Research and Applications Symposium (ETRA). 2012. pp. 45–52.
2. Jagla F., Jergelova M., Riecansky I. Saccadic eye movement related potentials. Physiological Research. 2007. vol. 56. no. 6. pp. 707–713. DOI: 10.33549/physiolres.931368.
3. Krigolson O.E., Williams C.C., Norton A., Hassall C.D., Colino F.L. Choosing MUSE: Validation of a Low-Cost, Portable EEG System for ERP Research // Frontiers in Neuroscience. 2017. vol. 11. DOI: 10.3389/fnins.2017.00109.
4. Brainbit. Brainbit Manual. URL: http://brainbit.com/ (accessed 09/01/2023).
5. Georgiadis K., Kalaganis F.P., Riskos K., Matta E., Oikonomou V.P., Yfantidou I., Chantziaras D., Pantouvakis K., Nikolopoulos S., Laskaris N.A., Kompatsiaris I.. NeuMa

– the absolute Neuromarketing dataset en route to an holistic understanding of consumer behaviour // Scientific Data. 2023. vol. 10(1). no. 508. DOI: 10.1038/s41597-023-02392-9.

6. Plochl M., Ossandon J., Konig P. Combining EEG and eye tracking: identification, characterization, and correction of eye movement artifacts in electroencephalographic data // Frontiers in Human Neuroscience. 2012. vol. 6. no. 278. URL: https://www.frontiersin.org/articles/10.3389/fnhum.2012.00278.
7. Antoniou E., Bozios P., Christou V., Tzimourta K.D., Kalafatakis K.G. Tsipouras M., Giannakeas N., Tzallas A.T. EEG-Based Eye Movement Recognition Using Brain-Computer Interface and Random Forests // Sensors. 2021. vol. 21. no. 7. no. 2339. DOI: 10.3390/s21072339.
8. Shahbakhti M., Beiramvand M., Nazari M., Broniec-Wojcik A., Augustyniak P., Rodrigues A.S., Wierzchon M., Marozas V. VME-DWT: An Efficient Algorithm for Detection and Elimination of Eye Blink From Short Segments of Single EEG Channel // IEEE Transactions on Neural Systems and Rehabilitation Engineering. 2021. vol. 29. pp. 408–417.
9. Stone D.B., Tamburro G., Fiedler P., Haueisen J., Comani S. Automatic Removal of Physiological Artifacts in EEG: The Optimized Fingerprint Method for Sports Science Applications // Frontiers in Human Neuroscience. 2018. vol. 12. no. 96.
10. Maddirala A.K., Veluvolu K. Eye-blink artifact removal from single channel EEG with k-means and SSA // Scientific Reports. 2021. vol. 11(1). no. 11043.
11. Klug M., Jeung S., Wunderlich A., Gehrke L., Protzak J., Djebbara Z., Argubi-Wollesen A., Wollesen B., Gramann K. The BeMoBIL Pipeline for automated analyses of multimodal mobile brain and body imaging data // bioRxiv. 2022. DOI: 10.1101/2022.09.29.510051.
12. Han J., Jiang G., Ouyang G., Li X. A Multimodal Approach for Identifying Autism Spectrum Disorders in Children // IEEE Transactions on Neural Systems and Rehabilitation Engineering. 2022. vol. 30. pp. 2003–2011.
13. Ahtola E., Stjerna S., Stevenson N., Vanhatalo S. Use of eye tracking improves the detection of evoked responses to complex visual stimuli during EEG in infants // Clinical Neurophysiology Practice. 2017. vol. 2. pp. 81–90. URL: https://www.sciencedirect.com/science/article/pii/S2467981X17300070.
14. Wang Y., Yu S., Ma N., Wang J., Hu Z., Liu Z., He J. Prediction of product design decision Making: An investigation of eye movements and EEG features // Advanced Engineering Informatics. 2020. vol. 45. no. 101095. DOI: 10.1016/j.aei.2020.101095.
15. Reiser J., Wascher E., Arnau S. Recording mobile EEG in an outdoor environment reveals cognitive-motor interference dependent on movement complexity // Scientific Reports. 2019. vol. 9(1). no. 13704.
16. Buerkle A., Bamber T., Lohse N., Ferreira P. Feasibility of Detecting Potential Emergencies in Symbiotic Human-Robot Collaboration with a mobile EEG // Robotics and Computer-Integrated Manufacturing. 2021. vol. 72. no. 102179.
17. Klug M., Gramann K. Identifying key factors for improving ICA-based decomposition of EEG data in mobile and stationary experiments // European Journal of Neuroscience. 2021. vol. 54. no. 12. pp. 8406–8420.
18. Chiu N.-T., Huwiler S., Ferster M.L., Karlen W., Wu H.-T., Lustenberger C. Get rid of the beat in mobile EEG applications: A framework towards automated cardiogenic artifact detection and removal in single-channel EEG // Biomedical Signal Processing and Control. 2022. vol. 72. no. 103220. DOI: 10.1016/j.bspc.2021.103220.

19. Barz M., Sonntag D. Automatic Visual Attention Detection for Mobile Eye Tracking Using Pre-Trained Computer Vision Models and Human Gaze // Sensors. 2021. vol. 21(12). no. 4143. DOI: 10.3390/s21124143.
20. Zhou C., Shi Z., Huang T., Zhao H., Kaner J. Impact of swiping direction on the interaction performance of elderly-oriented smart home interface: EEG and eye-tracking evidence // Frontiers in Psychology. 2023. vol. 14.
21. Tonsen M., Baumann C., Dierkes K. A High-Level Description and Performance Evaluation of Pupil Invisible. arXiv preprint arXiv:2009.00508. 2020.

**Романюк Владимир Русланович** — младший научный сотрудник, лаборатория интегрированных систем автоматизации, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук (СПб ФИЦ РАН). Область научных интересов: машинное обучение, определение состояния человека. romaniukvr@yandex.ru; 14-я линия В.О., 39, 199178, Санкт-Петербург, Россия; р.т.: +7(812)328-8071.

**Кашевник Алексей Михайлович** — канд. техн. наук, старший научный сотрудник, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук (СПб ФИЦ РАН). Область научных интересов: интеллектуальные транспортные системы, мониторинг человека, управление знаниями, облачные вычисления, человеко-машинное взаимодействие, профилирование пользователей, онтологии, интеллектуальные пространства. Число научных публикаций — 250. alexey@iias.spb.su; 14-я линия В.О., 39, 199178, Санкт-Петербург, Россия; р.т.: +7(812)328-8071.

**DOI** 10.15622/ia.23.2.9

A.E. ASFHA, A. VAISH

# INFORMATION SECURITY RISK ASSESSMENT IN INDUSTRY INFORMATION SYSTEM BASED ON FUZZY SET THEORY AND ARTIFICIAL NEURAL NETWORK

*Asfha A.E., Vaish A.* **Information Security Risk Assessment in Industry Information System Based on Fuzzy Set Theory and Artificial Neural Network.**

**Abstract.** Information security risk assessment is a crucial component of industrial management techniques that aids in identifying, quantifying, and evaluating risks in comparison to criteria for risk acceptance and organizationally pertinent objectives. Due to its capacity to combine several parameters to determine an overall risk, the traditional fuzzy-rule-based risk assessment technique has been used in numerous industries. The technique has a drawback because it is used in situations where there are several parameters that need to be evaluated, and each parameter is expressed by a different set of linguistic phrases. In this paper, fuzzy set theory and an artificial neural network (ANN) risk prediction model that can solve the issue at hand are provided. Also an algorithm that may change the risk-related factors and the overall risk level from a fuzzy property to a crisp-valued attribute is developed. The system was trained by using twelve samples representing 70%, 15%, and 15% of the dataset for training, testing, and validation, respectively. In addition, a stepwise regression model has also been designed, and its results are compared with the results of ANN. In terms of overall efficiency, the ANN model (R2= 0.99981, RMSE=0.00288, and MSE=0.00001,) performed better, though both models are satisfactory enough. It is concluded that a risk-predicting ANN model can produce accurate results as long as the training data accounts for all conceivable conditions.

**Keywords:** risk, risk assessment, artificial neural network, fuzzy set theory, industry information system, cement industry.

**1. Introduction.** Over the past few decades, industrial digitalization has altered conventional procedures and practices in virtually every industry, and numerous digitalization solutions have been included in manufacturing assets [1]. The facility and processing of industry is no exception, and since the early 2000s, it has undergone a rapid digitalization process. For example, the Cement industry infrastructure in particular is subject to large and growing cybersecurity threats in the form of threat actors, vulnerabilities, and potential consequences.

Cybercriminals and others could potentially conduct cyberattacks against the industrial infrastructure, and all industries are always targets of malicious attacks. Modern exploration and production industry techniques depend more and more on remotely connected operational equipment, which is frequently essential for security and susceptible to cyberattacks. Because its operational technologies may have fewer cybersecurity protective measures, older infrastructure is equally prone to attack [2]. Thus, a successful cyberattack on industry infrastructure could cause physical, environmental, and economic harm.

Therefore, over time, the complexity of information systems is increasing, and the issues of information security are becoming increasingly important for any industry information system. Information security is concerned with protecting data, particularly electronic data, from unwanted use [3]. The security of the information at their disposal must be evaluated by every industry that uses information. Consequently, information security analysis is required. The first step in the risk management process is to assess the potential for information security breaches. The assessment of a system's information security or the design phase typically involves the analysis of information security threats [4]. Assessing the capability and efficacy of control mechanisms used on information technology components and the architecture of information systems in general is the primary goal of an information security evaluation.

An information security assessment includes many **tasks**, such as evaluating the effectiveness of the information processing system, evaluating the security of the technologies used, the processing process, and the management of the automated system [5]. The overarching goal of an information security assessment is to ensure the confidentiality, integrity, and availability of an organization’s assets. There are numerous risk assessment tools, and they can be used in either of two ways. Therefore, approaches for analyzing information security threats can be either quantitative or qualitative, depending on the outcome of their assessment. The numerical value of risk is produced by the algorithm of a quantitative technique [6]. Information concerning unfavorable or unexpected events in the information security system that could endanger the protection of information (information security incidents) is often gathered using the input data for evaluation. However, the results are less accurate and relevant because there are frequently insufficient statistics.

The use of overly basic scales with three degrees of risk assessment (low, medium, and high) makes qualitative procedures more prevalent. Experts are interviewed for the assessment, but there is still limited use of intelligent methods [7]. It is clear that both of the aforementioned choices have a number of fundamental flaws. In order to overwhelm them, the latest research focused on identifying alternative techniques that would be both more accurate and more adaptive, as the constant emergence of new sources of threats often renders existing approaches inaccurate and ineffective. Among the promising approaches are models based on solving uncertainty problems, such as fuzzy logic models and artificial neural networks (ANN).

Finally, fuzzy logic and artificial neural network approaches have been recommended as the appropriate tools to improve the industry

information system and may help analyze complex conditions. Thus, the main purpose of this paper is to evaluate risk values in a more reliable, flexible, and objective manner by using this proposed method and prioritizing the level of risk value.

**1.1. Problem Descriptions.** Every processing industry performs a large number of operations and tasks on a daily basis. Each activity and procedure comes with its own set of hazards, which must be identified and ranked. The sector has numerous difficulties and costs as a result of its failure to identify accessible dangers, which can lead to a lack of competitiveness, a lack of greatness, a loss of representative trust, and, ultimately, a departure from the basic goal of adequacy. Thus, the aim of this section is to identify the existing problems and evaluate the efficiency and accuracy of information security risk analysis output in industry information system.

One of the primary research problems in information security risk analysis in the industrial processing system is the lack of appropriate and standardized methodologies for industry risk analysis in different stages of the risk management process, especially the shortcomings of qualitative and quantitative risk analysis methodologies, as well as the use of old techniques. In short, the criticism of the approaches is as follows.

By ensuring that the limitations of one form of data are balanced by the strengths of another. Thus, using or integrating both a fuzzy inference system (FIS) and an artificial neural network (ANN) will result in more accurate and efficient results in industry processing systems.

**1.2. Research Goals.** This research paper aims to increase the efficiency and accuracy of information security risk analysis result in industry information systems by developing an ANN model for determining the risk of critical information security incidents based on an ISO 27005 standard. To achieve this goal, the following research objectives are set:

**1.3. Research Objective.** The objectives are listed below:

**Obj. #1:** Analysis of the existing and most recent risk analysis methods and tools in industry information systems.

**Obj. #2:** The authors have identified the different information security risks that may exist during the early developmental phases of the industrial system. Experts' opinions have been collated for compiling this list. Then develop a solution to address the identified problem(s).

**Obj. #3:** To design and implement a fuzzy inference system and artificial neural network (ANN) technique to estimate the information security risk in industry information systems.

**Obj. #4:** Evaluating the efficiency and accuracy of the proposed ANN model. To validate the applicability and effectiveness of the proposed ANN

model in industry information systems through fuzzy multiple regression modeling (MRM).

The aim of this paper is to develop a novel method for conducting risk assessments in industry information systems. Thus, this paper presented a fuzzy inference system and artificial neural network (ANN) model for estimating, evaluating, and prioritizing a more accurate and efficient risk level that minimizes the limitations of the existing methods.

**2. Literature review.** Existing risk assessment approaches mostly differ in the applied risk assessment scales: quantitative or qualitative. The output of the algorithm of the quantitative approach is the numerical value of the risk [5]. The information on unforeseen occurrences and threats is typically used as assessment input. But the frequent absence of adequate statistics reduces the sufficiency of the outcomes. The most prevalent qualitative processes, however, employ too straightforward scales that typically have three degrees of risk assessment (low, medium, and high). The assessment is conducted through expert interviews and the use of clever techniques is still insufficient [8]. Furthermore, such outcomes are not reusable.

Due to the aforementioned flaws, experts are actively seeking a method that would produce high-quality results while being able to adapt to the threat landscape's ongoing changes, omit ineffective and irrelevant expert assessments, and allow for the reuse of earlier assessments [9]. Although it takes a lot of time and intellectual energy, the fuzzy logic and artificial neural network (ANN) approach is the most promising way in this research because it addresses the problems with current approaches, notably in terms of flexibility and adaptability [10]. Additionally, the ANN has cognitive features like self-learn, making it possible to identify the optimum solution while gathering knowledge of both internal and external processes.

Fuzzy logic is a valuable method for dealing with complexity and uncertainty, providing a way to model the systems by simulating human thinking without relying on quantitative and qualitative data in computation [11]. Due to the ambiguous and complex nature of the characteristics, evaluating industrial information systems utilizing sustainable decision-making processes is difficult. Due to a lack of knowledge and a high degree of domain-related uncertainty, it is challenging to quantify risks using standard mathematics. Simple risk assessment, ranking, and prioritization based on the expertise, experience, and opinions of experts are made possible by fuzzy logic-based methodologies [12].

The key to fuzzy logic is to find appropriate fuzzy rules. For example, fuzzy IF-THEN rules are IF-THEN statements. The amount of

rules needed varies depending on the particulars of the problem [13]. Membership functions are used to characterize specific linguistic labels in a problem. The complexity of the components or information, the cross-interaction and effect between various elements, and the subjectivity of some aspects all contribute to the fuzziness in the risk assessment of the industry information system, making it challenging to precisely quantify and characterize. Fuzzy logic offers a more adaptable technique of evaluation because it doesn't rely on exact mathematical models to define and process problems [14].

Fuzzy logic can handle fuzzy and uncertain situations by introducing membership functions to characterize the relationships between variables and mapping variables to the interval between 0 and 1 [15]. In the risk assessment process, fuzzy logic is divided into fuzzy inference, fuzzy clustering, and fuzzy decision-making. Fuzzy inference is the process of deducing one or more conclusions from fuzzy rules, and it can solve the problem of uncertainty and vagueness in decision systems [16]. For instance, when customs officers receive report information most of which are inaccurate linguistic information, fuzzy logic can be used to make the information fuzzy and analyze the risk by fuzzy IF-Then rules.

To help customs agents make wiser decisions, it is helpful to model each data point to each cluster using membership functions to represent similarity degrees between data and clusters, fuzzy clustering is used to group data based on comparable features and thoroughly conduct risk assessment [17]. The process of choosing the best course of action from a variety of alternatives is known as fuzzy decision-making, and it can take both the abruptness and smoothness of variables into account. It can be a useful sensitivity analysis method for determining how variables interact with one another and how this affects the output results [18].

A collection of neurons that are organized into layers and placed in a particular configuration makes up an artificial neural network (ANN). A multilayer network is one that has an input layer, one or more hidden layers, and an output layer. The number of parameters that are provided to the network as input in the input layer corresponds to the number of neurons in the output layer. The neurons in the buried layer increase dimensionality and are in charge of feature extraction. They support activities like classification and pattern recognition [19].

The structure of ANN depicts a schematic of a fully connected, three-layer neural network consisting of input neuron layers (or nodes, units), one or more hidden neuron layers, and a final layer which consists of the output neurons. There are several approaches to categorize neural networks, with the training method-based classification being the most

common. A neural network is trained when it has had its weights, biases, and maybe other parameters updated. Once trained, ANNs may implicitly identify novel patterns and generalize output based on previously learnt patterns [20].

The two main categories of training techniques are supervised and unsupervised. While the unsupervised training of neural networks, also known as self-organizing maps, primarily uses the classification and clustering algorithms, the supervised training method enables learning based on feedback [21]. Unsupervised networks are those that are not given any feedback and are typically requested to categorize the input vectors into groups and clusters. They are widely used in the industry for lithology identification and well log interpretation. The majority of neural network applications in the industry sector, however, are based on supervised training methods [22].

The methods for assessing risk have significantly advanced, and neural network techniques are now often used. Neural networks are able to automatically learn and extract nonlinear correlations between input data through extended training on vast volumes of data because of the numerous components and their complicated relationships in the risk assessment of import and export firms [23]. Because of the neural network's adaptive characteristics, it is possible to recognize these complicated relationships and constantly alter the model parameters until the best result is reached. Back-propagation (BP), multilayer perception (MLP), recurrent neural network (RNN), and radial basis function network (RBF) are a few of the regularly utilized artificial neural network architectures.

**3. Methodology.** This research methodology was implemented to evaluate the efficient and more reliable risk analysis in industry information systems. In order to collect data, a questionnaire was developed to identify different risks. This method offers sufficient results for all the research questions and objectives of the study to be addressed. The relevant areas of data collection were identified, and interviews were conducted with different management and expert staff of the cement industry to secure an accurate account of information about the risks. An opinion was also made by the researcher so as to obtain useful information that will yield results that can address the problem identified in the study.

The participants in this study were experts and staff from different sections of the cement industry information system ($N = 81$). The participants were executive management, regular staff, technical and asset operators, and third-party consulting companies.

Participants were asked to evaluate twelve different information assets based on a scale of five points (one, two... and five) to estimate the

likelihood and consequence of the threat and group them into a five-point Likert scale (very low, low, average, high, and very high), as shown in Table 1. The collected data was analyzed to calculate the likelihood of related threats and their consequences. Some specialists in the field of cement industry information systems confirmed the reliability of the questionnaires.

Table 1. Likert-scale questionnaires

| Likelihood | | | | |
|---|---|---|---|---|
| Very Low | Low | Average | High | Very High |
| 1 | 2 | 3 | 4 | 5 |
| Consequence | | | | |
| Very Low | Low | Average | High | Very High |
| 1 | 2 | 3 | 4 | 5 |

**3.1. Risk factors identification**. Identifying the industrial information system risk factors, and this is the process of identifying, assigning, and characterizing the types of risks. All aspects of the risk assessment process are included.

***Asset identification***. In the process of identifying assets and their value, we consider the value placed on assets (including information). What work was required to develop them, how much it costs to maintain, what damage would result if it were lost or destroyed and what benefit another party would gain if it were to obtain it.

***Vulnerabilities identification***. It is a weakness or absence in information systems, system security procedures, internal control, or implementation that could be exploited by a threat of sources. So this means in short control is absent, not efficient, and no longer relevant…etc.

***Threat Identification***. After identifying the assets that require protection, the threat to those assets must be identified and examined to determine the loss. Finally, the estimation of the likelihood and consequence of risk factors based on vulnerability and threat identification based on data collection.

**3.2. Fuzzy Inference System (FIS) Model.** Because of the uncertainty of the risk factors, the fuzzy logic method and a fuzzy inference system are used in this study. First, membership functions are determined for all likelihood and consequence. Hence, it could be deduced that the membership function is a curve showing a point mapping points of inputting data into membership values, whose interval is between zero and one. Figure 1 shows the FIS process.

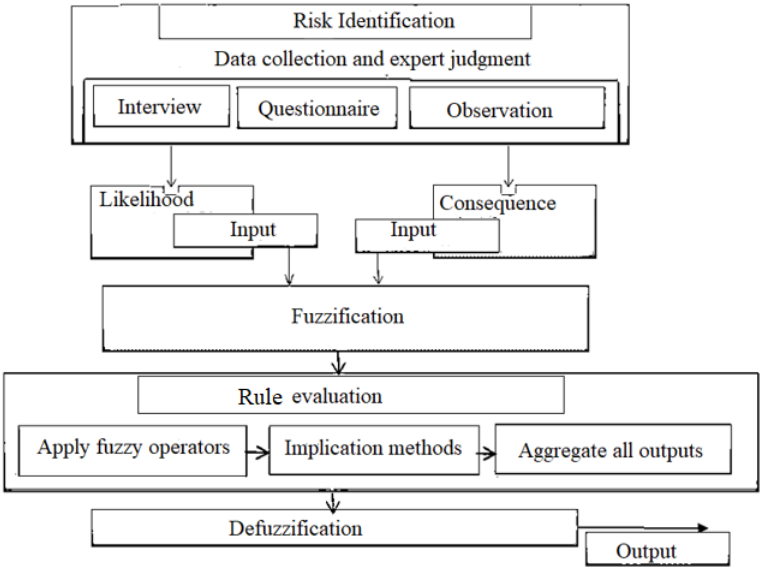



Fig. 1. Fuzzy inference system process

**Fuzzification (Fuzzify Inputs).** The first step is to take the inputs and determine the degree to which they belong to each of the appropriate fuzzy sets via membership functions (*fuzzification*) as noted in Table 2.

Table 2. Fuzzification table

| Level | Linguistic Value | Fuzzy value |
|---|---|---|
| | Linguistic Variables (Likelihood of Security Risk Occurrence: 0-1) | |
| 1 | Very Low | (0.000, 0.125, 0.250) |
| 2 | Low | (0.200, 0.325, 0.450) |
| 3 | Averages | (0.350, 0.500, 0.650) |
| 4 | High | (0.550, 0.675, 0.800) |
| 5 | Very High | (0.750, 0.875, 1.000) |
| | Linguistic Variables (Consequence of Security Risk Occurrence: 0-10) | |
| 1 | Very Low | (0.000, 1.000, 2.000) |
| 2 | Low | (2.000, 3.250, 4.500) |
| 3 | Average | (3.500, 5.000, 6.000) |
| 4 | High | (5.500, 6.750, 8.000) |
| 5 | Very High | (7.500, 0.875, 10.000) |
| | Linguistic Variables (Security Risk Value: 0-1) | |
| 1 | Low | (0.000, 0.125, 0.250) |
| 2 | Very Low | (0.200, 0.325, 0.450) |
| 3 | Average | (0.350, 0.500, 0.650) |
| 4 | High | (0.550, 0.675, 0.800) |
| 5 | Very High | (0.750, 0.875, 1.000) |

In this case, the likelihood (**L**) and consequence **(C) were** used as **crisp inputs (CI)** to the FIS (these values were taken from data collection and expert judgment).

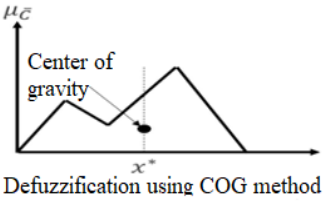
**Fuzzy Rule.** subsequently defining fuzzy membership functions, in this paper, Table 3 shows the 25 fuzzy rules constructed for the FIS.

Table 3. Risk matrix

| Likelihood / Consequence | Very Low | Low | Average | High | Very High |
|---|---|---|---|---|---|
| Very Low | VL | VL | L | L | A |
| Low | VL | L | A | A | A |
| Average | L | A | A | H | H |
| High | L | A | H | H | VH |
| Very High | A | A | H | VH | VH |
| VL= Very Low, L=Low, A= Average, H= High, and VH=Very High | | | | | |

**Aggregation.** It is the process of combining all of the fuzzy sets that symbolize each rule's outputs into a single fuzzy set. Interconversion occurs only once for each output variable, just prior to the final defuzzification phase.

**Defuzzification**. The last step in the fuzzy-molecular inference model is the defuzzification process, which is used to resolve a crisp value from the results of the inference process. Figure 2 indicates the defuzzification process using the center of gravity to finalize the FIS output.




Defuzzification using COG method

If $\mu_{\bar{C}}$ is defined with **continuous MF**: If $\mu C$ is defined with **discrete MF**:

$$x^* = \frac{\int \mu_{\bar{C}}(x) \cdot x \, dx}{\int \mu_{\bar{C}}(x) \, dx} \qquad x^* = \frac{\sum_{i=1}^{n} \mu_{\bar{C}}(x_i) \cdot x_i}{\sum_{i=1}^{n} \mu_{\bar{C}}(x_i)}$$

Fig. 2. Defuzzification processes using the Center of Gravity Method

Table 4 presents the likelihood and consequence given by Expert 1 for each security risk factor. The same procedure is then repeated for 80 experts, and the knowledge database is created. Here, the authors have assumed that the data is normally distributed. We know that if the data are

assumed to be normally distributed, Table 5 also presents the risk factors of all 81 experts in Raw Material Processing (RMP)".

Table 4. Likelihood and Consequence Given by Expert 1 for Each Security Risk Factor

| | Risk factor (Asset) | Coded Linguistic Variable | | | Numerical Value | | |
|---|---|---|---|---|---|---|---|
| | | L | C | Risk | L | C | Risk |
| RMP | Raw Material Processing | 3 | 5 | 4 | 0.500 | 8.750 | 0.675 |
| HRS | Hardware and Software | 3 | 3 | 3 | 0.500 | 5.000 | 0.500 |
| NWF | Network and Firmware | 2 | 2 | 2 | 0.325 | 3.250 | 0.325 |
| HRM | Human Resource and Data | 5 | 4 | 5 | 0.875 | 6.750 | 8.750 |
| RPT | Reputation | 4 | 2 | 3 | 0.675 | 3.250 | 0.500 |
| RMP | Raw Material Processing | 3 | 3 | 3 | 0.500 | 5.000 | 0.500 |
| ST | Storage and Transportation | 5 | 4 | 5 | 0.875 | 6.750 | 0.875 |
| RMM | Raw Material Milling | 2 | 1 | 1 | 0.325 | 1.250 | 0.125 |
| CP | Clinker Production | 4 | 3 | 4 | 0.675 | 5.000 | 0.675 |
| CM | Cement Milling | 1 | 2 | 1 | 0.125 | 3.250 | 0.125 |

Table 5. Available data for risk of «**Raw Material Processing (RMP)**» in Database

| No. | Numerical value | | | Coded linguistic variable | | |
|---|---|---|---|---|---|---|
| | Likelihood | Consequence | Risk Level | L | C | Risk Level |
| **1.** | 3 (Average) | 5 (Very High) | 4 (High) | 0.500 | 8.750 | 0.675 |
| **2.** | 4 (High) | 3 (Average) | 4 (High) | 0.675 | 5.000 | 0.675 |
| **3.** | 1 (Very Low) | 3 (Average) | 2 (Low) | 0.125 | 5.000 | 0.325 |
| **4.** | 4 (High) | 5 (Very High) | 5 (Very High) | 0.675 | 8.750 | 0.875 |
| **5.** | 3 (Average) | 4 (High) | 4 (High) | 0.500 | 6.750 | 0.675 |
| **6.** | 4 (High) | 1 (Very Low) | 2 (Low) | 0.675 | 1.250 | 0.325 |
| **7.** | 5 (Very High) | 3 (Average) | 4 (High) | 0.875 | 5.000 | 0.675 |
| **8.** | 3 (Average) | 3 (Average) | 3 (Average) | 0.500 | 5.000 | 0.500 |
| **9.** | 2 (Low) | 2 (Low) | 2 (Low) | 0.325 | 3.250 | 0.325 |
| **10.** | 4 (High) | 4 (High) | 4 (High) | 0.675 | 6.750 | 0.675 |
| **11.** | 1 (Very High) | 2 (Low) | 1 (Very Low) | 0.125 | 3.250 | 0.125 |
| **12.** | 4 (High) | 4 (High) | 4 (High) | 0.675 | 6.750 | 0.675 |
| **13.** | 3 (Average) | 4 (High) | 4 (High) | 0.500 | 6.750 | 0.675 |
| **14.** | 5 (Very High) | 5 (Very High) | 5 (Very High) | 0.875 | 8.750 | 0.875 |
| **15.** | 3 (Average) | 1 (Very High) | 2 (Low) | 0.500 | 1.250 | 0.325 |
| **16.** | 5 (Very High) | 5 (Very High) | 5 (Very High) | 0.875 | 8.750 | 0.875 |
| **17.** | 4 (High) | 3 (Average) | 4 (High) | 0.675 | 5.000 | 0.675 |
| **18.** | 1 (Very Low) | 2 (Low) | 1 (Very Low) | 0.125 | 3.250 | 0.125 |
| **19.** | 4 (High) | 4 (High) | 4 (High) | 0.675 | 6.750 | 0.675 |
| **20.** | 3 (Average) | 5 (Very High) | 4 (High) | 0.500 | 8.750 | 0.675 |
| **21.** | 2 (Low) | 1 (Very Low) | 1 (Very Low) | 0.325 | 1.250 | 0.125 |
| **22.** | 1 (Very Low) | 3 (Average) | 2 (Low) | 0.125 | 5.000 | 0.325 |
| **23.** | 3 (Average) | 3 (Average) | 3 (Average) | 0.500 | 5.000 | 0.500 |
| **24.** | 4 (High) | 1 (Very Low) | 2 (Low) | 0.675 | 1.250 | 0.325 |
| **25.** | 3 (Average) | 4 (High) | 4 (High) | 0.500 | 6.750 | 0.675 |
| **26.** | 3 (Average) | 2 (Low) | 3 (Average) | 0.500 | 3.250 | 0.500 |
| **27.** | 4 (High) | 5 (Very High) | 5 (Very High) | 0.675 | 8.750 | 0.875 |
| **28.** | 3 (Average) | 1 (Very Low) | 2 (Low) | 0.500 | 1.250 | 0.325 |
| **29.** | 5 (Very High) | 3 (Average) | 4 (High) | 0.875 | 5.000 | 0.675 |

*Continuation of Table 5*

| | | | | | | |
|---|---|---|---|---|---|---|
| **30.** | 3 (Average) | 3 (Average) | 3 (Average) | 0.500 | 5.000 | 0.500 |
| **31.** | 4 (High) | 4 (High) | 4 (High) | 0.675 | 6.750 | 0.675 |
| **32.** | 1 (Very Low) | 3 (Average) | 2 (Low) | 0.125 | 5.000 | 0.325 |
| **33.** | 4 (High) | 4 (High) | 4 (High) | 0.675 | 6.750 | 0.675 |
| **34.** | 4 (High) | 4 (High) | 4 (High) | 0.675 | 6.750 | 0.675 |
| **35.** | 5 (Very High) | 5 (Very High) | 5 (Very High) | 0.875 | 8.750 | 0.875 |
| **36.** | 1 (Very Low) | 2 (Low) | 1 (Very Low) | 0.125 | 3.250 | 0.125 |
| **37.** | 5 (Very High) | 5 (Very High) | 5 (Very High) | 0.875 | 8.750 | 0.875 |
| **38.** | 4 (High) | 3 (Average) | 4 (High) | 0.675 | 5.000 | 0.675 |
| **39.** | 3 (Average) | 2 (Low) | 3 (Average) | 0.500 | 3.250 | 0.500 |
| **40.** | 4 (High) | 4 (High) | 4 (High) | 0.675 | 6.750 | 0.675 |
| **41.** | 1 (Very Low) | 2 (Low) | 1 (Very Low) | 0.125 | 3.250 | 0.125 |
| **42.** | 3 (Average) | 5 (Very High) | 4 (High) | 0.500 | 8.750 | 0.675 |
| **43.** | 2 (Low) | 1 (Very Low) | 1 (Very Low) | 0.325 | 1.250 | 0.125 |
| **44.** | 3 (Average) | 3 (Average) | 3 (Average) | 0.500 | 5.000 | 0.500 |
| **45.** | 1 (Very Low) | 3 (Average) | 2 (Low) | 0.125 | 5.000 | 0.325 |
| **46.** | 4 (High) | 5 (Very High) | 5 (Very High) | 0.675 | 8.750 | 0.875 |
| **47.** | 4 (High) | 2 (Low) | 3 (Average) | 0.675 | 3.250 | 0.500 |
| **48.** | 3 (Average) | 4 (High) | 4 (High) | 0.500 | 6.750 | 0.675 |
| **49.** | 3 (Average) | 5 (Very High) | 4 (High) | 0.500 | 8.750 | 0.675 |
| **50.** | 4 (High) | 1 (Very Low) | 2 (Low) | 0.675 | 1.250 | 0.325 |
| **51.** | 3 (Average) | 4 (High) | 4 (High) | 0.500 | 6.750 | 0.675 |
| **52.** | 3 (Average) | 3 (Average) | 3 (Average) | 0.500 | 5.000 | 0.500 |
| **53.** | 2 (Low) | 2 (Low) | 2 (Low) | 0.325 | 3.250 | 0.325 |
| **54.** | 4 (High) | 4 (High) | 4 (High) | 0.675 | 6.750 | 0.675 |
| **55.** | 2 (Low) | 1 (Very Low) | 1 (Very Low) | 0.325 | 1.250 | 0.125 |
| **56.** | 4 (High) | 2 (Low) | 3 (Average) | 0.675 | 3.250 | 0.500 |
| **57.** | 4 (High) | 3 (Average) | 4 (High) | 0.675 | 5.000 | 0.675 |
| **58.** | 4 (High) | 4 (High) | 4 (High) | 0.675 | 6.750 | 0.675 |
| **59.** | 5 (Very High) | 2 (Low) | 3 (Average) | 0.875 | 3.250 | 0.500 |
| **60.** | 4 (High) | 4 (High) | 4 (High) | 0.675 | 6.750 | 0.675 |
| **61.** | 3 (Average) | 1 (Very Low) | 2 (Low) | 0.500 | 1.250 | 0.325 |
| **62.** | 3 (Average) | 4 (High) | 4 (High) | 0.500 | 6.750 | 0.675 |
| **63.** | 2 (Low) | 1 (Very Low) | 1 (Very Low) | 0.325 | 1.250 | 0.125 |
| **64.** | 3 (Average) | 3 (Average) | 3 (Average) | 0.500 | 5.000 | 0.500 |
| **65.** | 1 (Very Low) | 3 (Average) | 2 (Low) | 0.125 | 5.000 | 0.325 |
| **66.** | 5 (Very High) | 2 (Low) | 3 (Average) | 0.875 | 3.250 | 0.500 |
| **67.** | 3 (Average) | 3(Average) | 3 (Average) | 0.500 | 5.000 | 0.500 |
| **68.** | 2 (Low) | 2 (Low) | 2 (Low) | 0.325 | 3.250 | 0.325 |
| **69.** | 4 (High) | 4 (High) | 4 (High) | 0.675 | 6.750 | 0.675 |
| **70.** | 5 (Very High) | 3 (Average) | 4 (High) | 0.875 | 5.000 | 0.675 |
| **71.** | 3 (Average) | 3 (Average) | 3 (Average) | 0.500 | 5.000 | 0.500 |
| **72.** | 4 (High) | 1 (Very Low) | 2 (Low) | 0.675 | 1.250 | 0.325 |
| **73.** | 3 (Average) | 5 (Very High) | 4 (High) | 0.500 | 8.750 | 0.675 |
| **74.** | 2 (Low) | 2 (Low) | 2 (Low) | 0.325 | 3.250 | 0.325 |
| **75.** | 3 (Average) | 1 (Very Low) | 2 (Low) | 0.500 | 1.250 | 0.325 |
| **76.** | 2 (Low) | 5 (Very High) | 3 (Average) | 0.325 | 8.750 | 0.500 |
| **77.** | 4 (High) | 4 (High) | 4 (High) | 0.675 | 6.750 | 0.675 |
| **78.** | 1 (Average) | 1 (Low) | 1 (Average) | 0.125 | 1.250 | 0.125 |
| **79.** | 4 (High) | 2 (Low) | 3 (Average) | 0.675 | 3.250 | 0.500 |
| **80.** | 5 (Very High) | 4 (High) | 5 (Very High) | 0.875 | 6.750 | 0.875 |
| **81.** | 5 (Very High) | 4 (High) | 5 (Very High) | 0.875 | 6.750 | 0.875 |

Figure 3 notes the number of 25 if-then rules in order to provide a better understanding of the proposed fuzzy inference system framework, and with the input of the likelihood of occurrence and consequence, the risk size can be calculated. For instance, with 0.125 and 3.25 for likelihood and consequence, respectively, the risk size would be 0.125. A likelihood of 0.125 is related to rules 1–5, and a consequence of 3.25 is related to rules 2, 7, 12, 17, and 22.

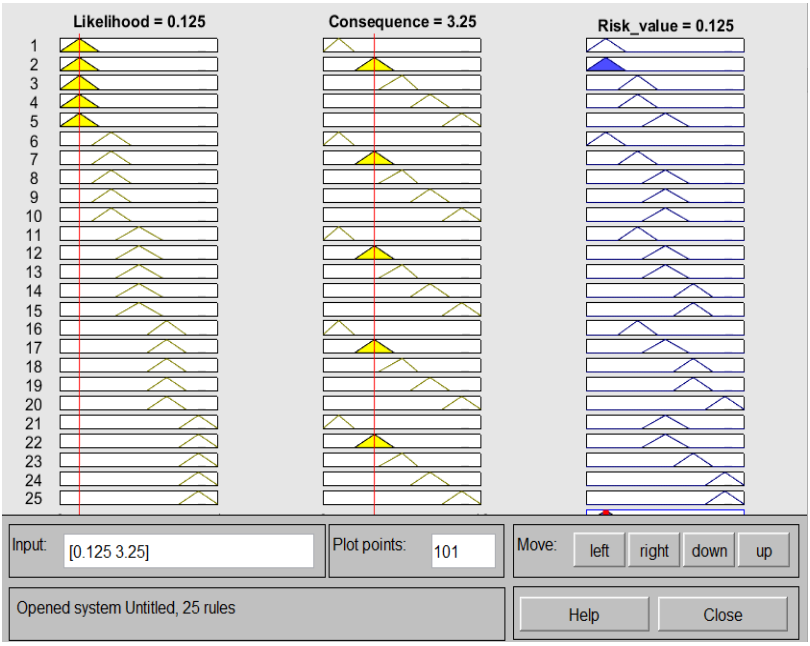



Fig. 3. Fuzzy rules according to Mamdani method

The fuzzy model designed by combining these rules estimates the risk value. The authors generated and plotted an output surface map for the industry information system fuzzy model using a surface viewer to visualize the dependence of one of the outputs on any one or two of the inputs. According to Mamdani, Figure 4 presents the processing industrial fuzzy model's output surface viewer.

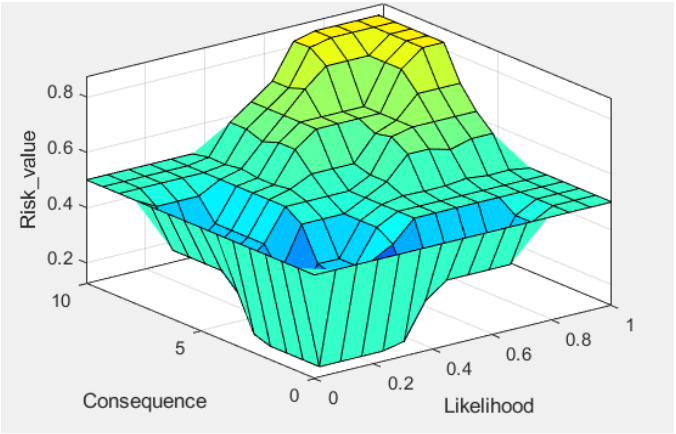



Fig. 4. 3D plots for 9 rules according to Mamdani method

Figure 5 indicates the normal probability illustration and the probability diagram of residuals for the criterion of "risk likelihood". Figure 6 indicates normal probability and residual illustrations for the criterion of "risk consequence" for the first factor "*Raw Material Processing (RMP)"*.

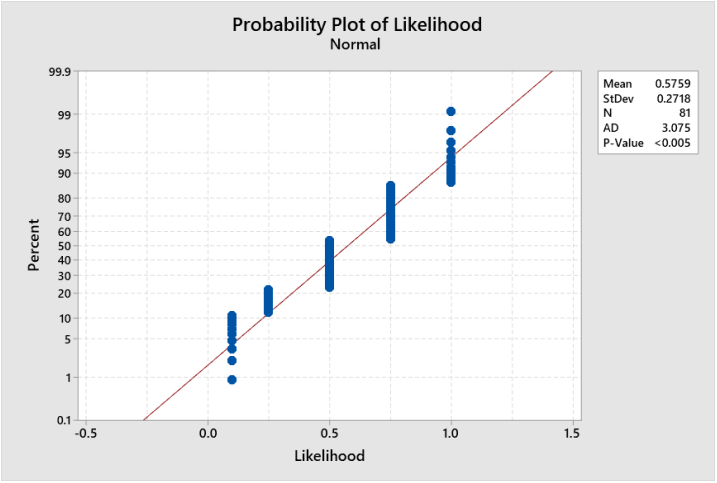



Fig. 5. Probability plot of likelihood

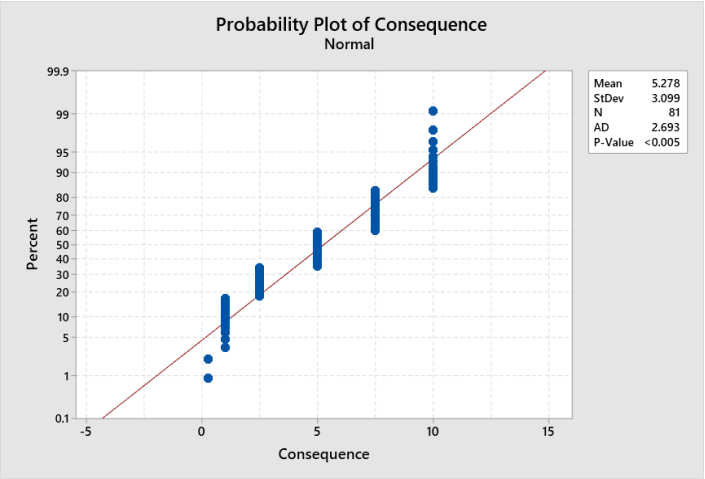



Fig. 6. Probability plot of consequence

**4. Result and Discussion.** This part uses a variety of statistical approaches to evaluate the quantitative data and provide the results of the data analysis in order to test the research hypotheses generated for the current study.

**4.1. Data collection.** Considering the chosen strategy of handing out the questionnaires to specific individuals one at a time, 95 were distributed. As a consequence, 81 of the 85 questionnaires received were complete and functional, yielding a response rate of 95.29%, which is regarded as excellent in research using a survey method and is displayed in Table 6. However, 10 employees failed to submit their surveys, and the remaining four representing 4.71% of the impractical forms were incomplete and contained inconsistent answers.

Table 6. The response rate of the participant

| Questionnaire | Number | Percentage |
|---|---|---|
| Distributed | 95 | 100 % |
| Received | 85 | 89.47% |
| practical | 81 | 95.29% |
| Impractical | 4 | 4.71% |

**4.2. Performance evaluation.** Minimum error occurrence has been considered as the basis for the selection of the best membership function. The performance of the designed fuzzy system has been evaluated on the basis of two types of errors, such as: – MSE (Mean Squared Error), and

RMSE (Root Mean Squared Error). According to the provided formulas, the correlation coefficient R between the data that were acquired and the data that ANN predicted has been determined (Equations (1) to (3)).

**MSE (Mean Squared Error)**: it is the average squared difference between the value observed in a statistical study and the values predicted from a model.

$$\text{MSE}=\frac{1}{n}\sum_{t=1}^{n}(A_t - F_t)^2. \quad (1)$$

**Root Mean Square Error (RMSE).** It is a common method for calculating a model's error in predicting quantitative data. One of the most widely used indices in performance evaluations, the RMSE index, could explain the discrepancy between the model output and the real result. It is a non-negative number that has no upper bound and can be 0 when the projected and recorded outputs coincide exactly.

$$\text{RMSE}=\sqrt{\text{MSE}} \text{ (square root of MSE)}. \quad (2)$$

**The correlation coefficient ($\mathbf{R^{2)}}$** is a positive number that indicates how much of the variability in the dependent variable can be explained by the independent variable(s) and how well the model fits the data. $\mathbf{R^2}$ can take values between 0 and 1; 1 indicates the model can acquire all the variability of the output variable, while 0, which indicates a weak correlation between predicted and actual results, expresses this.

$$\text{R}=\frac{\sum_{t=1}^{n}(A_t-\bar{A})(F_t-\bar{F})}{\sqrt{\sum_{t=1}^{n}(A_t-\bar{A})^2 * \sum_{t=1}^{n}(F_t-\bar{F})^2}}, \quad (3)$$

$$\bar{A}=(\sum_{t=1}^{n} A_t)/n \text{ and } \bar{F}=(\sum_{t=1}^{n} F_t)/n,$$

where $A_t$, $F_t$, and n represent real data (Actual) data, estimate (Predicted) data, and the number of data, respectively.

**4.3. Data prediction by ANN.** In this research, a two-layer feed-forward with a backpropagation learning algorithm was used for the risk analysis model. Based on Figure 7, the input data consisted of 81 likelihood and consequence factors, and the output data from the FIS model was used as the target data to define the ANN output. To determine with ANN, the gray color (57= 70%) data points were selected for training, the green color (12=15%) for testing, and the remaining (12=15%) for data validation. The

number of hidden neurons was defined in different ways. The model was trained using Levenberg-Margardt with a backpropagation algorithm as noted in Table 7. In this paper, the authors used MATLAB software to evaluate the efficient results based on the ANN flowchart and FIS process. The outputs of the program which include the optimum membership functions for likelihood of occurrence, risk consequence, errors of training, test and validation, procedure of inference rules, and correlation between predicted data by network and training, test and validation data, are obtained.

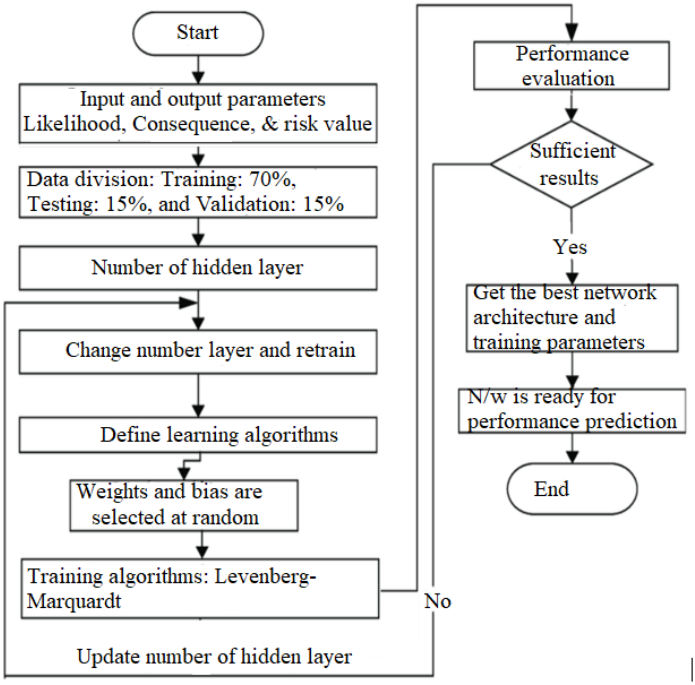



Fig. 7. ANN flowchart

Figure 8 indicates the function-fitting neural network. It is the process of training a neural network on a set of inputs in order to produce an associated set of target outputs. After you build the network with the preferred hidden layers and the training algorithm, you must train it using a set of training data. This research risk analysis was applied with different hidden layers of ANN (n = 10, 15, 25, and 50) and then the authors have selected the lowest error and best fit with the data.

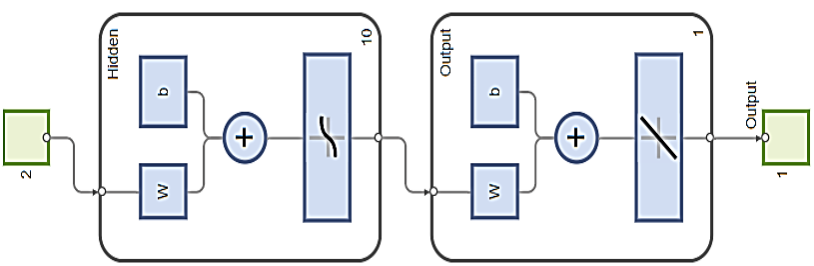



Fig. 8. Function fitting neural network (view)

Table 7. The specifications of the proposed ANFIS model

| Parameters | Description/Value |
|---|---|
| Number of layers | 3 (Input, output, and hidden layer) |
| Number of inputs( Predicators) | (2*81 double) |
| Number of outputs (Responses) | 1 (1*81 double) |
| Hidden layer | 10 |
| Number of iteration | 1000 |
| Training Algorithm | Levenberg-Marquardt |
| Data Division | Random |

The neural network regression has been shown in Figure 9, which demonstrates the interaction of the network with the training, test, and validation data. The correlation coefficient was found to be 1.00000, 0.99991, and 1.00000 for training, test, and validation data, respectively. Moreover, the straight line illustrates the linear relationship between the model-predicted and target output data. These results imply that there is a good match between the observed and model-predicted data. As a result, the model is adequate to forecast the data with high precision. The overall correlation coefficient (0.99998) confirms the outstanding prediction performance of the developed ANN model.

The plot for the best validation performance against the training data has been 6.9154e-18 at epoch 5 as shown in Figure 10. The circle in the plot clearly depicts that the validation plot lies exactly between the actual data plot and the observed data plot. Therefore, the research work is said to be validated.

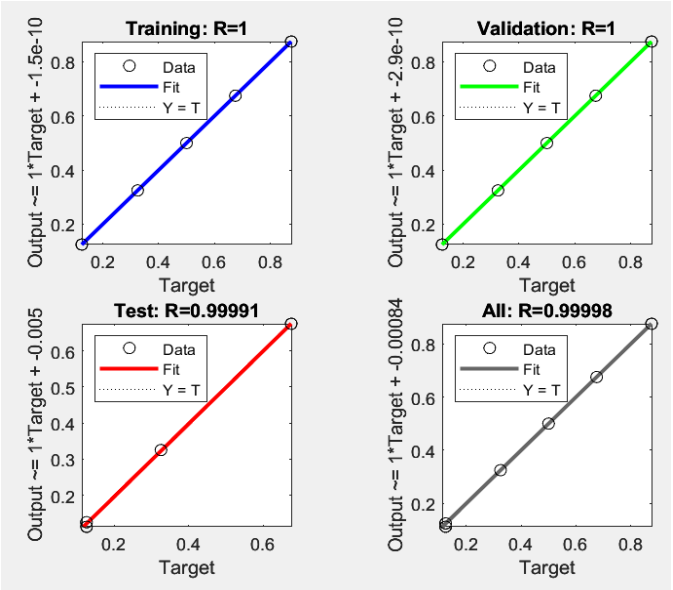



Fig. 9. ANN regression plot

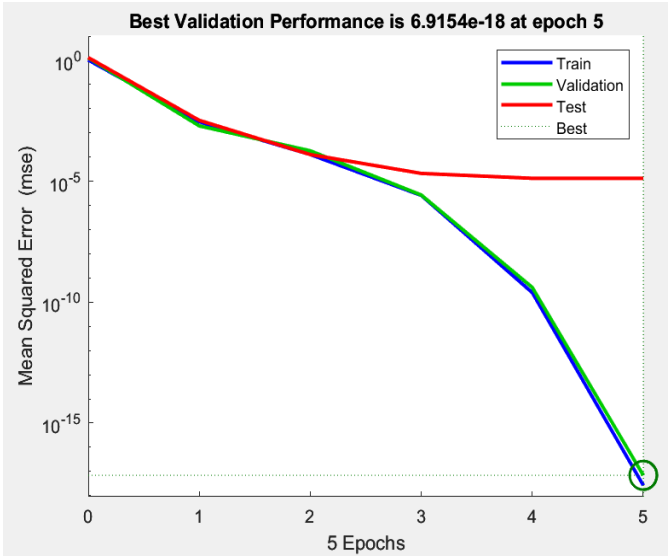



Fig. 10. ANN validation performance

Plotting the gradient values, mu, and validation fail has been shown in Figure 11. Gradient represents the slope of the tangent of a graph of a function. It points to the direction in which there is a high rate of increase for the considering function. The momentum constant or momentum parameter **(mu)** is the control parameter for the back-propagation neural network that we modeled, and the choice of mu directly affects the error convergence. **A validation check** is used to terminate the learning of the neural network. The number of validation checks will depend on the number of successive iterations of the neural network. Thus, gradient, mu, and validation check are 4.1263e-09, 1e-10, and 0 respectively at epoch 31 as shown in Figure 11.

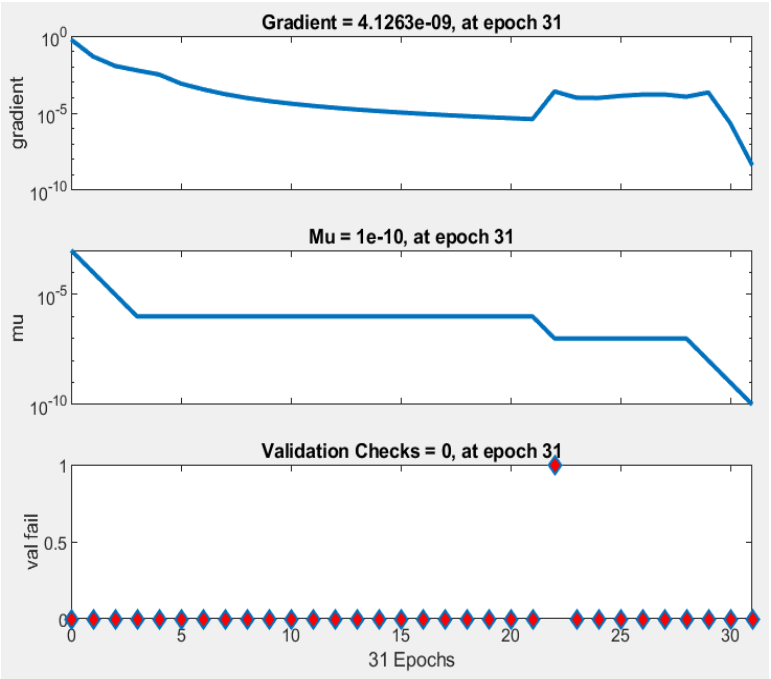



Fig. 11. ANN training state

Table 8 noted the information security risk prediction of "**Raw Material Processing (RMP)"**, and the coefficient of determination ($R^2$), RMSE, and MSE were also found. The results imply a good fit between the model-predicted data and the experimental data, indicating the models' aptness and coherence.

Table 8. ANN InfoSec risk prediction of «**Raw Material Processing (RMP)**»

| | | | | **Risk Prediction** |
|---|---|---|---|---|
| **Risk factor (Asset)** | **Likelihood** | **Consequence** | **InfoSec Risk** | **ANN** |
| **Raw Material Processing (RMP)** | 0.675 | 6.750 | 0.675 | 0.674 |
| | 0.875 | 3.250 | 0.500 | 0.500 |
| | 0.675 | 6.750 | 0.675 | 0.675 |
| | 0.500 | 1.250 | 0.325 | 0.327 |
| | 0.500 | 6.750 | 0.675 | 0.675 |
| | 0.325 | 1.250 | 0.125 | 0.126 |
| | 0.500 | 5.000 | 0.500 | 0.500 |
| | 0.125 | 5.000 | 0.325 | 0.325 |
| | 0.875 | 3.250 | 0.500 | 0.500 |
| | 0.500 | 5.000 | 0.500 | 0.500 |
| | 0.325 | 3.250 | 0.325 | 0.325 |
| | 0.675 | 6.750 | 0.675 | 0.682 |
| | 0.875 | 5.000 | 0.675 | 0.675 |
| | 0.500 | 5.000 | 0.500 | 0.500 |
| | 0.675 | 1.250 | 0.325 | 0.325 |
| | 0.500 | 8.750 | 0.675 | 0.675 |
| | 0.325 | 3.250 | 0.325 | 0.325 |
| | 0.500 | 1.250 | 0.325 | 0.328 |
| | 0.325 | 8.750 | 0.500 | 0.500 |
| | 0.675 | 6.750 | 0.675 | 0.675 |
| | 0.125 | 1.250 | 0.125 | 0.113 |
| | 0.675 | 3.250 | 0.500 | 0.500 |
| | 0.875 | 6.750 | 0.875 | 0.875 |
| | 0.875 | 6.750 | 0.875 | 0.875 |
| | | | RMSE | **0.00288** |
| | | | MSE | **0.00001** |
| | | | ***R*** | **0.99991** |
| | | | $R^2$ | **0.99981** |

**4.4. Validation of InfoSec Risk analysis via Fuzzy Multiple Regression Modeling (MRM).** A comparison of the findings acquired is necessary for confirming and validating the efficacy of the technique being used to solve any problem. The current method and the alternative procedures that were previously applied in the prior research investigations must be compared in this comparison. The authors used the ANN to evaluate the security risk in the aforementioned case study.

Multiple Regression Analysis (MRA) is a statistical technique that predicts the outcome of a response variable using a variety of explanatory variables. This technique will be heavily employed to represent the causal relationships between inputs and outputs. Equation 4 serves as a presentation of the multiple regression approach.

$$\boldsymbol{Risk = X_0 + X_1 * Likelihood + X_2 * Consequence}, \quad (4)$$

where $X_0$ is a fixed and $X_1$ and $X_2$ are regression coefficients.

The stepwise regression method has been applied for the first risk factor of *"Raw Material Processing (RMP)*" by using MINITAB 19 software to choose the best regression method for the prediction of risk size. Stepwise regression models have been presented in this paper. These models are shown in Tables 9 – 12.

Table 9. Correlation Coefficient among Input and Output Factors

| | Likelihood | Consequence/Impact | **Security Risk** |
|---|---|---|---|
| Likelihood | 1.0000 | | |
| Consequence | 0.219 | 1.0000 | |
| Security Risk | 0.709 | 0.793 | 1.0000 |

Table 10. Consists of the Multiple Regression Equation for security risk through the hierarchy

| **Multiple Regression Equation** | | $R^2$ |
|---|---|---|
| Security Risk Evaluation Model | **Regression Equation**<br>Risk Level = -0.0419+ 0.5033 Likelihood + 0.05699 Consequence | 0.93104 % |

Table 11. Multiple Regression (MRL) Equations for each identified security risk factor

| **Risk Factor** | **Multiple Regression Equation** | **RMSE** | **MSE** | $R^2$ |
|---|---|---|---|---|
| RMP | -0.0715 + 0.5191 * Likelihood + 0.05997 * Consequence | **0.05250** | **0.00276** | **0.93104** |
| HRS | -0.0386 + 0.5843 * L + 0.05286 * C | 0.05672 | 0.00322 | 0.91832 |
| NWF | -0.0281 + 0.4858 * L + 0.05749 * C | 0.03993 | 0.00159 | 0.97120 |
| HRM | -0.0109 + 0.6597 * L + 0.05594 * C | 0.04738 | 0.00224 | 0.96112 |
| RPT | -0.0535 + 0.4693 * L + 0.06453 * C | 0.04164 | 0.00173 | 0.96216 |
| RMP | -0.0178 + 0.4941 * L + 0.05398 * C | 0.06825 | 0.00466 | 0.90721 |
| ST | -0.1065 + 0.5837 * L + 0.05915 * C | 0.04015 | 0.00161 | 0.97082 |
| RMM | -0.1356 + 0.6471 * L + 0.06000 * C | 0.06106 | 0.00373 | 0.91012 |
| CP | -0.0904 + 0.5987 * L + 0.05971 * C | 0.04738 | 0.00225 | 0.95919 |
| CM | -0.0130 + 0.5530 * L + 0.05104 * C | 0.05168 | 0.00267 | 0.94401 |

Table 12. Risk prediction using the MRM model

| | | | | Risk Prediction |
|---|---|---|---|---|
| **Risk factor (Asset)** | **Likelihood** | **Consequence** | **InfoSec Risk** | **MRM** |
| **Raw Material Processing (RMP)** | 0.675 | 6.750 | 0.675 | 0.683 |
| | 0.875 | 3.250 | 0.500 | 0.584 |
| | 0.675 | 6.750 | 0.675 | 0.683 |
| | 0.500 | 1.250 | 0.325 | 0.281 |
| | 0.500 | 6.750 | 0.675 | 0.594 |
| | 0.325 | 1.250 | 0.125 | 0.193 |
| | 0.500 | 5.000 | 0.500 | 0.495 |
| | 0.125 | 5.000 | 0.325 | 0.306 |
| | 0.875 | 3.250 | 0.500 | 0.584 |
| | 0.500 | 5.000 | 0.500 | 0.495 |
| | 0.325 | 3.250 | 0.325 | 0.307 |
| | 0.675 | 6.750 | 0.675 | 0.683 |
| | 0.875 | 5.000 | 0.675 | 0.683 |
| | 0.500 | 5.000 | 0.500 | 0.495 |
| | 0.675 | 1.250 | 0.325 | 0.369 |
| | 0.500 | 8.750 | 0.675 | 0.708 |
| | 0.325 | 3.250 | 0.325 | 0.307 |
| | 0.500 | 1.250 | 0.325 | 0.281 |
| | 0.325 | 8.750 | 0.500 | 0.620 |
| | 0.675 | 6.750 | 0.675 | 0.683 |
| | 0.125 | 1.250 | 0.125 | 0.092 |
| | 0.675 | 3.250 | 0.500 | 0.483 |
| | 0.875 | 6.750 | 0.875 | 0.783 |
| | 0.875 | 6.750 | 0.875 | 0.783 |
| | | | RMSE | **0.05250** |
| | | | MSE | **0.00276** |
| | | | ***R*** | **0.96491** |
| | | | $R^2$ | **0.93104** |

**4.5. Comparison between actual and model-predicted results.** In this section, to prove the effectiveness of the proposed method, we compare our proposed algorithm with different methods. The authors compare our proposed ANN classifier with fuzzy regression modeling (MRM). The comparison and statistical analysis of the actual values and the model-predicted values of risk analysis in industry information systems are presented in Table 13. It was found that both models have sufficient capability to predict the properties of the industry.

Table 13. Comparison between actual and model-predicted results

| | | | | ***Risk Prediction*** | |
|---|---|---|---|---|---|
| **Risk factor (Asset)** | **Likelihood** | **Consequence** | **InfoSec Risk** | ANN model predicted | MRM model predicted |
| **Raw Material Processing (RMP)** | 0.675 | 6.750 | 0.675 | 0.674 | 0.683 |
| | 0.875 | 3.250 | 0.500 | 0.500 | 0.584 |
| | 0.675 | 6.750 | 0.675 | 0.675 | 0.683 |
| | 0.500 | 1.250 | 0.325 | 0.327 | 0.281 |
| | 0.500 | 6.750 | 0.675 | 0.675 | 0.594 |
| | 0.325 | 1.250 | 0.125 | 0.126 | 0.193 |
| | 0.500 | 5.000 | 0.500 | 0.500 | 0.495 |
| | 0.125 | 5.000 | 0.325 | 0.325 | 0.306 |
| | 0.875 | 3.250 | 0.500 | 0.500 | 0.584 |
| | 0.500 | 5.000 | 0.500 | 0.500 | 0.495 |
| | 0.325 | 3.250 | 0.325 | 0.325 | 0.307 |
| | 0.675 | 6.750 | 0.675 | 0.682 | 0.683 |
| | 0.875 | 5.000 | 0.675 | 0.675 | 0.683 |
| | 0.500 | 5.000 | 0.500 | 0.500 | 0.495 |
| | 0.675 | 1.250 | 0.325 | 0.325 | 0.369 |
| | 0.500 | 8.750 | 0.675 | 0.675 | 0.708 |
| | 0.325 | 3.250 | 0.325 | 0.325 | 0.307 |
| | 0.500 | 1.250 | 0.325 | 0.328 | 0.281 |
| | 0.325 | 8.750 | 0.500 | 0.500 | 0.620 |
| | 0.675 | 6.750 | 0.675 | 0.675 | 0.683 |
| | 0.125 | 1.250 | 0.125 | 0.113 | 0.092 |
| | 0.675 | 3.250 | 0.500 | 0.500 | 0.483 |
| | 0.875 | 6.750 | 0.875 | 0.875 | 0.783 |
| | 0.875 | 6.750 | 0.875 | 0.875 | 0.783 |
| | | | RMSE | **0.00288** | **0.05250** |
| | | | MSE | **0.00001** | **0.00276** |
| | | | $R$ | **0.99991** | **0.96491** |
| | | | $R^2$ | **0.99981** | **0.93104** |

As represented in Figures 12, 13, 14 in terms of overall efficiency, the ANN model ($R^2 = 0.99981$, RMSE = 0.00288, MSE = 0.00001) performed better than the MRM model ($R^2 = 0.93104$, RMSE = 0.05250, MSE = 0.00276), though both are satisfactory enough. Figure 15 shows the time series plot of actual observed values versus the values predicted by the ANN and MRM models on the test dataset.

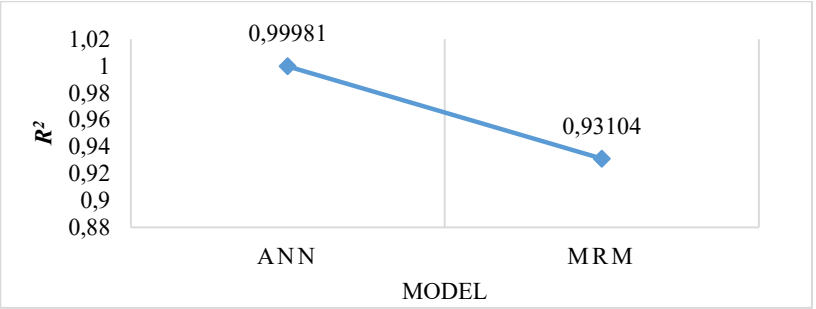



Fig. 12. Correlation Coefficient of ANN and Stepwise Regression

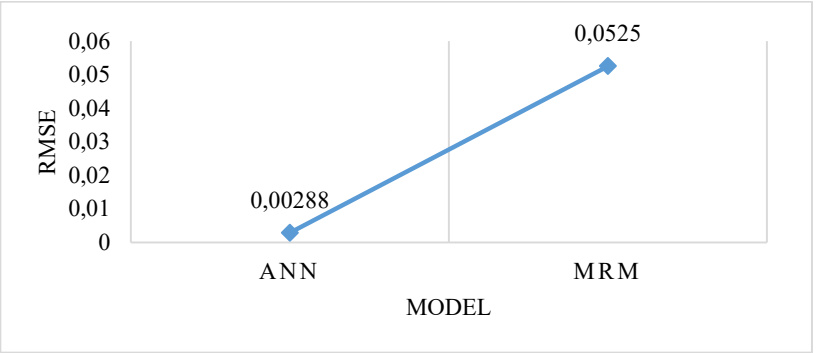



Fig. 13. RMSE of ANN and Stepwise Regression

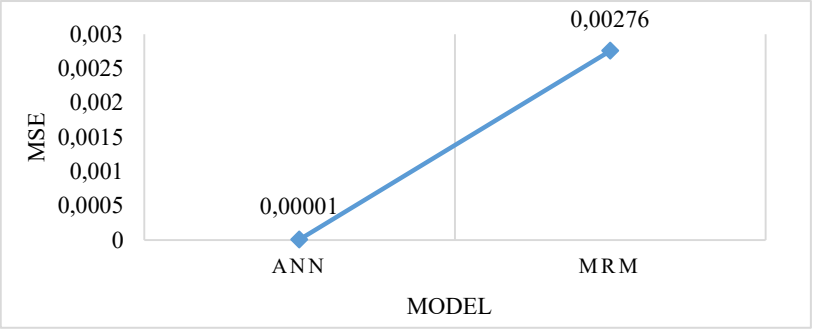



Fig. 14. MSE of ANN and Stepwise Regression

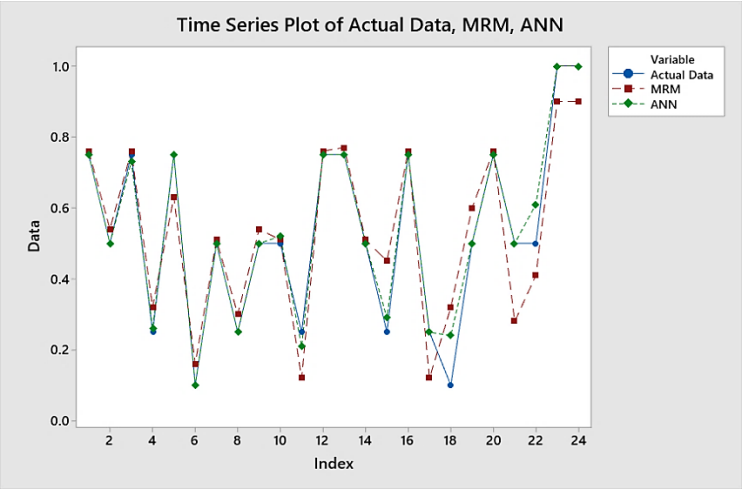



Fig. 15. Time Series Plot of ANN and MRM based on Actual data

**5. Conclusion.** Due to their shortcomings, both qualitative and quantitative methods are considered non-complete, subjective, including an element of randomness, and difficult to update or reuse. At this time many papers provide the necessary horizon scanning, focusing on AI-based methods, fuzzy logic, adaptive neural fuzzy inference system (ANFIS), and artificial neural networks (ANNs) and their usage for a more effective calculation of risk, considering the mix of qualitative input parameters such as likelihood and consequence. Thus, in this study, an information security risk assessment model based on fuzzy logic and an artificial neural network (ANN) is proposed to evaluate and calculate both qualitative and quantitative risks in a more reliable, flexible, and objective manner. The application of an artificial neural network can be used to assess information security risk since they have self-learn ability, can solve uncertain problems, and are appropriate for quantity data processing.

After fuzzy membership, functions are constructed for likelihood, consequence, and risk value. In order to obtain a more reliable and less subjective approach to the risk assessment process, an ANN has been used in this new model. Finally, in terms of overall efficiency, the ANN model ($R^2$= 0.99981, RMSE=0.00288, and MSE=0.00001,) performed better performance, though both models are satisfactory enough.

## References

1. Verhoef P.C., Broekhuizen T., Bart Y., Bhattacharya A., Dong J.Q., Fabian N., Haenlein M. Digital transformation: A multidisciplinary reflection and research

agenda. Journal of business research. 2021. vol. 122. pp. 889–901. DOI: 10.1016/j.jbusres.2019.09.022.
2. Mazhar T., Irfan H.M., Khan S., Haq I., Ullah I., Iqbal M., Hamam H. Analysis of Cyber Security Attacks and Its Solutions for the Smart grid Using Machine Learning and Blockchain Methods. Future Internet. 2023. vol. 15(2). no. 83. DOI: 10.3390/fi15020083.
3. Alhassan M.M., Adjei-Quaye A. Information Security in an Organization. International Journal of Computer. 2017. T. 24. № 1. C. 100–116. [Online]. URL: https://ijcjournal.org/index.php/InternationalJournalOfComputer/article/view/820.
4. Shaikh F.A., Siponen M. Information security risk assessments following cybersecurity breaches: The mediating role of top management attention to cybersecurity. Comput. Secur. 2023. vol. 124. no. 102974. DOI: 10.1016/j.cose.2022.102974.
5. Cruz S.T. Information security risk assessment. Information Security Management Handbook. 2007. pp. 243–250. DOI: 10.3390/encyclopedia1030050.
6. Yevseiev S., Shmatko O., Romashchenko N. Algorithm of Information Security Risk Assessment Based on Fuzzy-Multiple Approach. Adv. Inf. Syst. 2019. vol. 3. no. 2. pp. 73–79. DOI: 10.20998/2522-9052.2019.2.13.
7. By I. et al. Implementing of qualitative risk assessment procedures. 2021. pp. 1–275.
8. Aven T. Risk assessment and risk management: Review of recent advances on their foundation. European Journal of Operational Research. 2016. vol. 253. no. 1. pp. 1–13. DOI: 10.1016/j.ejor.2015.12.023.
9. Tariq U., Ahmed I., Bashir A.K., Shaukat K. A Critical Cybersecurity Analysis and Future Research Directions for the Internet of Things: A Comprehensive Review. Sensors. 2023. vol. 23(8). no. 4117. DOI: 10.3390/s23084117.
10. de Campos Souza P.V., Lughofer E. Evolving fuzzy neural classifier that integrates uncertainty from human-expert feedback. 2023. vol. 14. pp. 319–341.
11. Bozic V. Fuzzy Approach to Risk Management: Enhancing Decision-Making Under Uncertainty. 2023. DOI: 10.13140/RG.2.2.13517.82405.
12. Kaka S., Hussin H., Khan R., Akbar A., Sarwar U., Ansari J. Fuzzy Logic-Based Quantitative Risk Assessment Model for Hse in Oil and Gas Industry. Journal of Tianjin University Science and Technology. 2022. pp. 93–109. DOI: 10.17605/OSF.IO/WVG2H.
13. Nikmanesh M., Feili A., Sorooshian S. Employee Productivity Assessment Using Fuzzy Inference System. Information. 2023. vol. 14(7). no. 423. DOI: 10.3390/info14070423.
14. Crnogorac L., Tokalic R., Gutic K., Jovanovic S., Dukanovic D. Fuzzy logic model for stability assessment of underground facilities. Podzemni radovi. 2020. no. 36. pp. 29–48. DOI: 10.5937/podrad2036029c.
15. Parra-Dominguez J., Alonso-Garcia M., Corchado J.M. Fuzzy Logic to Measure the Degree of Compliance with a Target in an SDG –The Case of SDG 11. Mathematics. 2023. vol. 11(13). no. 2967. DOI: 10.3390/math11132967.
16. Madanda V.C., Sengani F., Mulenga F. Applications of Fuzzy Theory-Based Approaches in Tunnelling Geomechanics: a State-of-the-Art Review. Mining, Metallurgy and Exploration. 2023. vol. 40. no. 3. pp. 819–837. DOI: 10.1007/s42461-023-00767-5.
17. Xie J., Deng Q., Xia S., Zhao Y., Wang G., Gao X. Research on Efficient Fuzzy Clustering Method Based on Local Fuzzy Granular balls. 2023. pp. 1–10. [Online]. URL: http://arxiv.org/abs/2303.03590.
18. Aliyeva K., Aliyeva A., Aliyev R., Ozdeser M. Application of Fuzzy Simple Additive Weighting Method in Group Decision-Making for Capital Investment. Axioms. 2023. vol. 12(8). no. 797. DOI: 10.3390/axioms12080797.

19. Alaloul W., Qureshi A.H. Data Processing Using Artificial Neural Networks. IntechOpen. 2020. 26 p. DOI: 10.5772/intechopen.91935.
20. Yang G.R., Wang X.J. Artificial Neural Networks for Neuroscientists: A Primer. Neuron. 2020. vol. 107. no. 6. pp. 1048–1070. DOI: 10.1016/j.neuron.2020.09.005.
21. Sarker I.H. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Computer Science. 2021. vol. 2(3). no. 160. DOI: 10.1007/s42979-021-00592-x.
22. Zhang J., He Y., Zhang Y., Li W., Zhang J. Well-Logging-Based Lithology Classification Using Machine Learning Methods for High-Quality Reservoir Identification: A Case Study of Baikouquan Formation in Mahu Area of Junggar Basin, NW China. Energies. 2022. vol. 15. no. 10. DOI: 10.3390/en15103675.
23. Sarker I.H. Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. SN Computer Science. 2021. vol. 2(6). no. 420. DOI: 10.1007/s42979-021-00815-1.

**Asfha Amanuel** — Post-graduate student, Department of information technology security (FBI), ITMO University; Eritrea Institute of Technology. Research interests: information security methods and systems, information and cyber security, risk management. The number of publications — 4. baquesti2003@gmail.com; 49, Kronverksky Av., 197101, St. Petersburg, Russia; office phone: +7(952)378-2147.

**Vaish Abhishek** — Assistant professor, It department, Indian Institute of Information Technology, Allahabad. Research interests: information security, information security laws and regulations, cyber diplomacy, network security, IT Governance, enterprise recourses planning. The number of publications — 67. abhishek@iiita.ac.in; Uttar Pradesh, 211015, Deghat Jhalwa, India; office phone: +91(790)535-6150.

**УДК** 004.056 **DOI** 10.15622/ia.23.2.9

А.Э. Асфха, А. Вайш

# ОЦЕНКА РИСКОВ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ В ОТРАСЛЕВОЙ ИНФОРМАЦИОННОЙ СИСТЕМЕ НА ОСНОВЕ ТЕОРИИ НЕЧЕТКИХ МНОЖЕСТВ И ИСКУССТВЕННОЙ НЕЙРОННОЙ СЕТИ

*Асфха А.Э., Вайш А.* **Оценка рисков информационной безопасности в отраслевой информационной системе на основе теории нечетких множеств и искусственной нейронной сети.**

**Аннотация.** Оценка рисков информационной безопасности является важнейшим компонентом методов промышленного менеджмента, который помогает выявлять, количественно определять и оценивать риски в сравнении с критериями принятия рисков и целями, относящимися к организации. Благодаря своей способности комбинировать несколько параметров для определения общего риска традиционный метод оценки рисков, основанный на нечетких правилах, используется во многих отраслях промышленности. Этот метод имеет недостаток, поскольку он используется в ситуациях, когда необходимо оценить несколько параметров, и каждый параметр выражается различным набором лингвистических фраз. В этой статье представлены теория нечетких множеств и модель прогнозирования рисков с использованием искусственной нейронной сети (ANN), которые могут решить рассматриваемую проблему. Также разработан алгоритм, который может изменять факторы, связанные с риском, и общий уровень риска с нечеткого свойства на атрибут с четким значением. Система была обучена с использованием двенадцати выборок, представляющих 70%, 15% и 15% набора данных для обучения, тестирования и валидации соответственно. Кроме того, также была разработана пошаговая регрессионная модель, и ее результаты сравниваются с результатами ANN. С точки зрения общей эффективности, модель ANN (R2= 0,99981, RMSE=0,00288 и MSE=0,00001) показала лучшую производительность, хотя обе модели достаточно удовлетворительны. Делается вывод, что модель ANN, прогнозирующая риск, может давать точные результаты до тех пор, пока обучающие данные учитывают все мыслимые условия.

**Ключевые слова:** риск, оценка риска, искусственная нейронная сеть, теория нечетких множеств, отраслевая информационная система, цементная промышленность.

## Литература

1. Verhoef P.C., Broekhuizen T., Bart Y., Bhattacharya A., Dong J.Q., Fabian N., Haenlein M. Digital transformation: A multidisciplinary reflection and research agenda. Journal of business research. 2021. vol. 122. pp. 889–901. DOI: 10.1016/j.jbusres.2019.09.022.
2. Mazhar T., Irfan H.M., Khan S., Haq I., Ullah I., Iqbal M., Hamam H. Analysis of Cyber Security Attacks and Its Solutions for the Smart grid Using Machine Learning and Blockchain Methods. Future Internet. 2023. vol. 15(2). no. 83. DOI: 10.3390/fi15020083.
3. Alhassan M.M., Adjei-Quaye A. Information Security in an Organization. International Journal of Computer. 2017. Т. 24. № 1. С. 100–116. [Online]. URL: https://ijcjournal.org/index.php/InternationalJournalOfComputer/article/view/820.
4. Shaikh F.A., Siponen M. Information security risk assessments following cybersecurity breaches: The mediating role of top management attention to

cybersecurity. Comput. Secur. 2023. vol. 124. no. 102974. DOI: 10.1016/j.cose.2022.102974.
5. Cruz S.T. Information security risk assessment. Information Security Management Handbook. 2007. pp. 243–250. DOI: 10.3390/encyclopedia1030050.
6. Yevseiev S., Shmatko O., Romashchenko N. Algorithm of Information Security Risk Assessment Based on Fuzzy-Multiple Approach. Adv. Inf. Syst. 2019. vol. 3. no. 2. pp. 73–79. DOI: 10.20998/2522-9052.2019.2.13.
7. By I. et al. Implementing of qualitative risk assessment procedures. 2021. pp. 1–275.
8. Aven T. Risk assessment and risk management: Review of recent advances on their foundation. European Journal of Operational Research. 2016. vol. 253. no. 1. pp. 1–13. DOI: 10.1016/j.ejor.2015.12.023.
9. Tariq U., Ahmed I., Bashir A.K., Shaukat K. A Critical Cybersecurity Analysis and Future Research Directions for the Internet of Things: A Comprehensive Review. Sensors. 2023. vol. 23(8). no. 4117. DOI: 10.3390/s23084117.
10. de Campos Souza P.V., Lughofer E. Evolving fuzzy neural classifier that integrates uncertainty from human-expert feedback. 2023. vol. 14. pp. 319–341.
11. Bozic V. Fuzzy Approach to Risk Management: Enhancing Decision-Making Under Uncertainty. 2023. DOI: 10.13140/RG.2.2.13517.82405.
12. Kaka S., Hussin H., Khan R., Akbar A., Sarwar U., Ansari J. Fuzzy Logic-Based Quantitative Risk Assessment Model for Hse in Oil and Gas Industry. Journal of Tianjin University Science and Technology. 2022. pp. 93–109. DOI: 10.17605/OSF.IO/WVG2H.
13. Nikmanesh M., Feili A., Sorooshian S. Employee Productivity Assessment Using Fuzzy Inference System. Information. 2023. vol. 14(7). no. 423. DOI: 10.3390/info14070423.
14. Crnogorac L., Tokalic R., Gutic K., Jovanovic S., Dukanovic D. Fuzzy logic model for stability assessment of underground facilities. Podzemni radovi. 2020. no. 36. pp. 29–48. DOI: 10.5937/podrad2036029c.
15. Parra-Dominguez J., Alonso-Garcia M., Corchado J.M. Fuzzy Logic to Measure the Degree of Compliance with a Target in an SDG –The Case of SDG 11. Mathematics. 2023. vol. 11(13). no. 2967. DOI: 10.3390/math11132967.
16. Madanda V.C., Sengani F., Mulenga F. Applications of Fuzzy Theory-Based Approaches in Tunnelling Geomechanics: a State-of-the-Art Review. Mining, Metallurgy and Exploration. 2023. vol. 40. no. 3. pp. 819–837. DOI: 10.1007/s42461-023-00767-5.
17. Xie J., Deng Q., Xia S., Zhao Y., Wang G., Gao X. Research on Efficient Fuzzy Clustering Method Based on Local Fuzzy Granular balls. 2023. pp. 1–10. [Online]. URL: http://arxiv.org/abs/2303.03590.
18. Aliyeva K., Aliyeva A., Aliyev R., Ozdeser M. Application of Fuzzy Simple Additive Weighting Method in Group Decision-Making for Capital Investment. Axioms. 2023. vol. 12(8). no. 797. DOI: 10.3390/axioms12080797.
19. Alaloul W., Qureshi A.H. Data Processing Using Artificial Neural Networks. IntechOpen. 2020. 26 p. DOI: 10.5772/intechopen.91935.
20. Yang G.R., Wang X.J. Artificial Neural Networks for Neuroscientists: A Primer. Neuron. 2020. vol. 107. no. 6. pp. 1048–1070. DOI: 10.1016/j.neuron.2020.09.005.
21. Sarker I.H. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Computer Science. 2021. vol. 2(3). no. 160. DOI: 10.1007/s42979-021-00592-x.
22. Zhang J., He Y., Zhang Y., Li W., Zhang J. Well-Logging-Based Lithology Classification Using Machine Learning Methods for High-Quality Reservoir Identification: A Case Study of Baikouquan Formation in Mahu Area of Junggar Basin, NW China. Energies. 2022. vol. 15. no. 10. DOI: 10.3390/en15103675.

23. Sarker I.H. Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. SN Computer Science. 2021. vol. 2(6). no. 420. DOI: 10.1007/s42979-021-00815-1.

**Асфха Амануэль Эстифанос** — аспирант, факультет безопасности информационных технологий (ФБИТ), Университет ИТМО; Эритрейский технологический институт. Область научных интересов: методы и системы защиты информации, информационная и кибербезопасность, управление рисками. Число научных публикаций — 4. baquesti2003@gmail.com; Кронверкский проспект, 49, 197101, Санкт-Петербург, Россия; р.т.: +7(952)378-2147.

**Вайш Абхишек** — доцент, факультет информационных технологий, Индийский институт информационных технологий, Аллахабад. Область научных интересов: информационная безопасность, законы и нормативные акты в области информационной безопасности, кибердипломатия, сетевая безопасность, управление ИТ, планирование ресурсов предприятия. Число научных публикаций — 67. abhishek@iiita.ac.in; Уттар-Прадеш, 211015, Дегхат Джалва, Индия; р.т.: +91(790)535-6150.

**УДК** 004.95 **DOI** 10.15622/ia.23.2.10

Г.Р. Воробьева, А.В. Воробьев, Г.О. Орлов

# КОНЦЕПЦИЯ ОБРАБОТКИ, АНАЛИЗА И ВИЗУАЛИЗАЦИИ ГЕОФИЗИЧЕСКИХ ДАННЫХ НА ОСНОВЕ ЭЛЕМЕНТОВ ТЕНЗОРНОГО ИСЧИСЛЕНИЯ

*Воробьева Г.Р., Воробьев А.В., Орлов Г.О.* **Концепция обработки, анализа и визуализации геофизических данных на основе элементов тензорного исчисления.**

**Аннотация.** Одним из основных подходов к обработке, анализу и визуализации геофизических данных является применение геоинформационных систем и технологий, что обусловлено их геопространственной привязкой. Вместе с тем, сложность представления геофизических данных связана с их комплексной структурой, предполагающей множество составляющих, которые имеют одну и ту же геопространственную привязку. Яркими примерами данных такой структуры и формата являются гравитационные и геомагнитные поля, которые в общем случае задаются трех и четырехкомпонентными векторами с разнонаправленными осями координат. При этом на сегодняшний день отсутствуют решения, позволяющие визуализировать указанные данные в комплексе, не декомпозируя их на отдельные скалярные значения, которые, в свою очередь, могут быть представлены в виде одного или многих пространственных слоев. В этой связи в работе предложена концепция, использующая элементы тензорного исчисления для обработки, хранения и визуализации информации такого формата. Формализован механизм тензорного представления компонент поля с возможностью его комбинирования с другими данными такого же формата, с одной стороны, и свертки при сочетании с данными более низкого ранга. На примере гибридной реляционно-иерархической модели данных предложен механизм хранения информации по тензорным полям, предусматривающий возможность описания и применения инструкций по трансформации при переходе между различными системами координат. В работе рассматривается применение подхода при переходе от декартовой к сферической системе координат при представлении параметров геомагнитного поля. Для комплексной визуализации параметров тензорного поля предложен подход, основанный на применении тензорных глифов. В качестве последних при этом используются суперэллипсы с осями, соответствующими рангу тензора. При этом атрибутивные значения предлагается визуализировать относительно осей графического примитива таким образом, что распределение данных может быть задано посредством варьирования градиента монохромного представления параметра вдоль оси. Работоспособность концепции была исследована в ходе сравнительного анализа тензорного подхода с решениями, основанными на скалярной декомпозиции соответствующих комплексных значений с последующим их представлением в виде одного или многих пространственных слоев. Проведенный анализ показал, что применение предложенного подхода позволит в значительной степени повысить наглядность формируемого геопространственного изображения без необходимости сложного перекрывания пространственных слоев.

**Ключевые слова:** тензорные поля, тензорное исчисление, геоинформационные технологии, глифы, суперэллипсы.

**1. Введение.** В настоящее время обработка, анализ и визуализация геофизических данных являются важным фундаментом

при решении прикладных задач, а также проведении разноплановых исследований в области наук о Земле. Так, к примеру, информация такого рода используется для диагностирования геоиндуцированных токов, наводимых в сетях электропередач и способных привести к критическим сбоям в их работе [1]. Другая известная задача – моделирование и анализ пространственно-временного распределения параметров геомагнитного поля и его вариаций, в основе которых лежат результаты наблюдений наземных магнитных станций, с одной стороны, и расчетные данные, полученные на базе соответствующих геофизических моделей, с другой [2].

Объем геофизической информации непрерывно растет. При этом различные типы данных доступны с различным шагом дискретизации во времени и пространстве. Эффективная визуализация таких данных является важным инструментом как для анализа природных и техногенных процессов, так и для принятия решений в соответствующих прикладных областях.

Геопространственная привязка геофизической информации обуславливает применение геоинформационных технологий для ее обработки, анализа и визуализации. Так, к примеру, анализ геомагнитных аномалий, графически представленных в формате изолиний, позволяет даже визуально определить место для размещения лаборатории по поверке магниточувствительного оборудования. Аналогичным образом обеспечивается поддержка принятия решений и исследование различных процессов / явлений в известных геоинформационных решениях на основе геофизической информации.

Особое место в ряду задач обработки и визуализации геофизической информации занимает проблема представления геофизических полей в том формате, который позволит пользователю эффективно использовать эту информацию в своих целях. Важно при этом отметить, что общие тенденции решения обозначенной проблемы таковы, что анализируемые параметры полей декомпозируются на скалярные величины, для визуализации которых используются известные инструменты, модели и методы геоинформационных систем [3]. К ним, в частности, относятся изолинии, тепловые карты, простые геопространственные объекты и их комбинации, дополненные соответствующими цветовыми схемами и пр.

Указанное упрощенное представление геофизических полей в виде скалярных геопространственных изображений в ряде случаев недостаточно информативно. Так, в частности, основной характеристикой геомагнитного поля является вектор, представленный

тремя составляющими. При этом известные приложения, реализующие визуализацию геомагнитного поля, представляют соответствующие данные в виде скалярного поля, которое имеет место при фиксации одного из компонент или полного вектора геомагнитного поля [4].

При этом специфика геофизических полей заключается в том, что они относятся либо к векторным, либо тензорным полям. Упрощение их представления путем сведения к отображению одного компонента из совокупности составляющих вектор неизбежно приводит к снижению информативности и эффективности применения соответствующих геоинформационных решений. В прикладных областях и в процессе проведения научных исследований компоненты вектора (или тензора, в отдельных случаях) геофизического поля, имеет смысл рассматривать в совокупности, без отделения их друг от друга с последующей выделенной визуализацией.

Анализ известных решений в области геопространственной визуализации геофизических полей показал, что в их основе лежит представление поля цифровой моделью на регулярной сетке (GRID-модель) или в виде триангуляции (TIN-модель), поддерживаемые многими распространенными геоинформационными системами [5]. При этом такие подходы неприменимы именно для отображения векторных и тензорных полей. В этой связи необходимо разработать подход, который позволит использовать преимущества геоинформационных систем, с одной стороны, а также полноценно продемонстрировать специфику параметров полей, с другой.

Представляется целесообразным провести сравнительный анализ известных подходов к визуализации геофизических полей (в настоящее время доступы результаты исследований по решению обозначенной задачи для геомагнитного и гравитационного полей), выделить их ключевые положительные аспекты. На основании проведенного анализа требуется разработать подход, который с учетом обозначенных преимуществ рассмотренных методов позволит сформулировать концепцию обработки, анализа и визуализации геофизических полей именно с учетом их векторной / тензорной специфики. Указанные особенности должны в значительной степени быть ориентированы на потенциал современных программных интерфейсов, реализующих возможности двух- и трехмерной графики. Представляется целесообразным, с технической точки зрения, учитывать особенности веб-ориентированной реализации соответствующих программных решений, что позволит существенно расширить круг его потенциальных пользователей.

**2. Состояние вопроса.** В настоящее время подавляющая часть исследований посвящена разработке новых эффективных способов визуализации векторных полей. В первую очередь, известны решения на разработку методов визуализации гидродинамических полей.

В работе [6] проведен сравнительный обзор методов визуализации многомерных векторных полей. Выделенные методы разделены на четыре укрупненные группы:

– прямые методы, предполагающие использование стрелочных пиктограмм, которые, в свою очередь, размещаются в заданных пространственных координатах (географические широта / долгота). Прямой метод визуализации является достаточно наглядным решением при обработке данных касательно двумерных векторных полей. Например, стрелочные пиктограммы в пространственных точках, представленные под различным углом к земной поверхности, эффективны при двумерной визуализации направлений океанских течений, распространения ураганов и пр. Вместе с тем такой подход слабо применим к визуализации многомерных векторных полей. Разнонаправленные и привязанные к одной пространственной точке векторы могут вызвать некорректную интерпретацию результата визуализации конечным пользователем, с одной стороны, а также значительную перегруженность геопространственного изображения, с другой. В случае высокой плотности визуализируемых пространственных данных ситуация еще более усугубляется.

– методы, основанные на характерных признаках, предполагают выделение подмножества данных с определенным набором характеристик, актуальных для конкретного конечного пользователя и / или в рамках определенной решаемой прикладной / научно-технической задачи. Фактически весь обрабатываемый и анализируемый объем данных сужается до уровня отдельных подмножеств, что в целом позволяет повысить эффективность рассматриваемого метода, по сравнению, к примеру, с ранее описанным. Недостатком рассматриваемого метода является прежде всего его трудоемкость, поскольку непосредственно выделение подмножеств из всей совокупности данных требует дополнительных, зачастую больших, вычислительных затрат.

– текстурные методы визуализации векторных полей основаны на принципе искажения геопространственной структуры в соответствии с локальными свойствами векторного поля таким образом, чтобы отобразить непосредственно векторное поле. Фактически имеет место деформация цвета и формы исходной

геопространственной поверхности. Текстурные методы достаточно информативны, согласованны и детализированы и могут быть успешно применены для визуализации сложных многомерных векторных полей. Поскольку речь идет только о деформации уже готового геопространственного изображения непосредственно средствами графического процессора, текстурным методам свойственна высокая производительность.

– геометрические методы, в основе применения которых лежит выделение набора опорных точек, последующего расчета общей характерной для них траектории и подбора геометрической фигуры. Сложность метода, очевидно, заключается в том, что неправильно выбранные опорные точки могут внести существенные загромождения и искажения визуализируемого геопространственного изображения. В целом, геометрические методы относятся в настоящее время к группе наиболее развитых и эффективных с точки зрения информативности получаемого результата визуализации. Недостатком геометрических методов является большая вычислительная сложность.

Отдельно представляется целесообразным отметить два основных наиболее часто практикуемых в геоинформационном моделировании способа визуализации векторных полей [5]. Первый из них, обозначаемый покомпонентной визуализацией, предполагает выделение отдельных компонент векторных полей и их представление в виде скалярных поверхностей. Метод прост в применении, однако замена вектора одним из его компонент приводит к потере важной для понимания исследуемого процесса / явления информации, поскольку в подавляющем большинстве случаев требуется информация о полном векторе в совокупности его компонент.

Второй подход основан на применении реализованных в современных геоинформационных системах простых алгоритмов визуализации посредством стрелочных диаграмм. Данный подход во многом перекликается с описанным выше методом (прямые методы), который использует стрелочные пиктограммы с однозначной привязкой к связке пространственных координат (географические широта и долгота, к примеру). Метод характеризуется теми же недостатками, что и прямые методы. Прежде всего, здесь необходимо отметить низкую разрешающую способность при отображении векторных полей ввиду зашумления визуализации большим количеством взаимных пересечений стрелочных пиктограмм.

Резюмируя, представляется целесообразным отметить, что перечисленные в обзоре методы ориентированы главным образом на визуализацию векторных полей. При этом в случае тензорных полей

такие методы либо неприменимы, либо малоэффективны. В этой связи необходимо разработать подход, который позволит визуализировать такого рода поля без потери информации, важной для исследования соответствующего процесса / явления.

**3. Обзор и характеристика типов полей.** В общем виде принято выделять три группы полей, каждое из которых в геопространственном выражении может быть задано соответствующими поверхностями.

Скалярными будем называть поля, которые заданы функцией поверхности, где каждой точке ставится в однозначное соответствие некоторый скаляр, представленный в виде действительного или комплексного числа [7]. Типичными примерами таких полей и поверхностей являются визуализация пространственно-временного распределения мгновенного / среднесуточного значения температуры воздуха, влажности, количества осадков и пр. (иными словами, с учетом сказанного скаляры должны быть представлены в каждой пространственной точке в заданный момент времени только простым атомарным значением). Визуализация скалярных полей предусматривает использование стандартных геопространственных графических примитивов, представленных в соответствии с заданной цветовой схемой (рисунок 1). К их числу относятся, в частности, пространственные точки, полигоны, полилинии (к примеру, выраженные в виде совокупности пространственных изолиний и/или изобар) и пр.

К векторным полям будем относить заданные функцией пространственные поверхности, каждой точке которой ставится в соответствие вектор с началом в этой точке (рисунок 1) [8]. В качестве примера представляется целесообразным привести такие параметры, как направление ветра, потоки миграции населения в различных территориальных областях и пр. Как правило, при задании векторных полей обязательными являются два параметра: точка начала вектора (пространственная точка, представленная соответствующей парой пространственных координат в заданной системе координат), а также направление вектора в соответствии с углом к нормали по отношению к заданной поверхности. В общем виде визуализация векторных полей может быть представлена одним из двух способов. Первый из них ориентирован на упрощение представления пространственных данных, поскольку сводит вектор к покомпонентному разложению на соответствующие скаляры (в подавляющем большинстве случаев такой подход так или иначе сопряжен с потерей значимой для исследования / принятия решений

информации). Второй вариант визуализации пространственных векторных полей связан с применением графических примитивов, имитирующих вектор.

Тензорным полем будем называть заданную функцией поверхность, каждой точке которой ставится в соответствие тензор, привязанный к соответствующей системе координат и имеющий начало в заданной пространственной точке [9]. Примерами тензорных полей являются, в частности, различного рода геофизические поля (геомагнитное, гравитационное и др.). В зависимости от ранга соответствующего тензора, представляющего исследуемый атрибутивный параметр, каждый тензор задается исходной точкой и совокупностью характеризующих его векторов. В общем виде визуализация тензорного поля традиционно выполняется одним из способов, предусмотренных для представления данных по векторным полям. Тензор можно покомпонентно разложить на скалярные значения, представив каждое из них, к примеру, в виде отдельного пространственного слоя. Другой вариант – использовать множество геопространственных примитивов для графического представления составляющих тензора, что, в свою очередь, сопряжено с чрезвычайной перегруженностью геопространственного изображения и, как следствие, его низкой эффективностью при проведении исследований и / или принятии решений в соответствующих областях (рисунок 1).

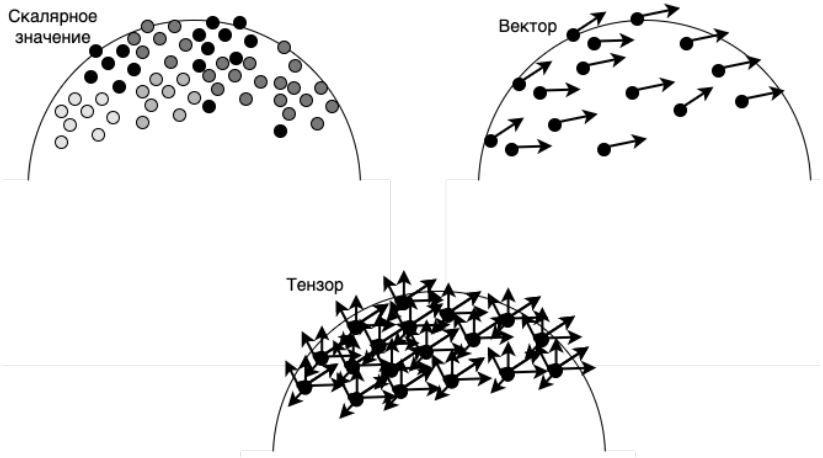



Рис. 1. Известные подходы к визуализации: а) скалярных; б) векторных; в) и тензорных полей

Таким образом, результаты проведенного обзора и анализа свидетельствуют о том, что наименее проработанной с точки зрения геопространственной визуализации является задача отображения тензорных полей. Соответственно требуется разработать подход, повышающий наглядность визуализации тензорных полей и расширяющий возможности их представления для конечного пользователя в процессе проведения исследований и / или принятия решений в прикладных областях.

**4. Геомагнитное поле как пример тензорного поля.** Магнитное поле Земли представляет собой поле, которое генерируется внутриземными источниками. В любой точке околоземного пространства оно определяется полным вектором напряженности, т.е. направлением действия и соответствующим модулем. При этом на магнитное поле Земли также оказывает воздействие переменное магнитное поле (годовые и суточные магнитные вариации, магнитные бури различной интенсивности и пр.), обусловленное внешними процессами, которые под воздействием космической погоды происходят в ионосфере.

Для разложения вектора напряженности $F$ на составляющие обычно используется декартова система координат, представленная тремя осями ($x$, $y$ и $z$ соответственно). Ось $x$ при этом ориентирована по направлению географического меридиана, ось $y$ – по направлению географической параллели, ось $z$ – направлена сверху вниз к центру Земли. Важно отметить, что для оси $x$ положительным считается направление к северу, а для оси $y$ – направление к востоку [10] (рисунок 2).

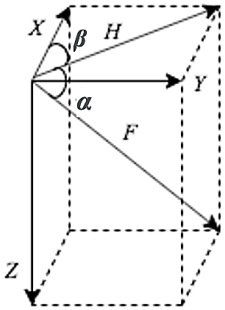



Рис. 2. Взаимосвязь компонент вектора геомагнитного поля

Пусть $X$, $Y$, $Z$ – соответствующие компоненты вектора геомагнитного поля. Проекция полного вектора $F$ на горизонтальную

плоскость называется горизонтальной составляющей и обозначается $H$. Тогда взаимосвязь между указанными компонентами геомагнитного поля может быть представлена посредством соотношений следующего вида:

$$\begin{aligned} &F = \sqrt{X^2 + Y^2 + Z^2} = \sqrt{H^2 + Z^2}, \\ &H = F \cdot \cos\alpha, \qquad Z = F \cdot \sin\alpha\,, \\ &X = H \cdot \cos\beta\,, Y = H \cdot \sin\beta\,; \end{aligned} \tag{1}$$

где $\beta$ – магнитное склонение, характеризующее угол между осью $x$ и горизонтальной составляющей $H$; $\alpha$ – магнитное наклонение, характеризующее угол между вектором $F$ и горизонтальной плоскостью.

С учетом сказанного и соотношений (1) тензор геомагнитного градиента $G$ характеризует скорость изменения параметров геомагнитного вектора по трем направлениям (соответственно $x$, $y$ и $z$) в декартовой системе координат [11]:

$$G = \begin{bmatrix} \frac{\partial F_x}{\partial x} & \frac{\partial F_x}{\partial y} & \frac{\partial F_x}{\partial z} \\ \frac{\partial F_y}{\partial x} & \frac{\partial F_y}{\partial y} & \frac{\partial F_y}{\partial z} \\ \frac{\partial F_z}{\partial x} & \frac{\partial F_z}{\partial y} & \frac{\partial F_z}{\partial z} \end{bmatrix}, \tag{2}$$

где $F_x$, $F_y$ и $F_z$ – три компонента вектора в своих проекциях на оси $x$, $y$ и $z$ соответственно (в (1) им соответствуют обозначения $X$, $Y$, $Z$).

Для упрощения записи представим (2) в виде свертки следующего вида [11]:

$$G = \begin{bmatrix} g_{xx} & g_{xy} & g_{xz} \\ g_{yx} & g_{yy} & g_{yz} \\ g_{zx} & g_{zy} & g_{zz} \end{bmatrix}. \tag{3}$$

Таким образом, тензор градиента геомагнитного поля представляет собой тензор второго ранга, который при этом состоит из $3 \times 3 = 9$ соответствующих пространственных производных.

Здесь же представляется целесообразным отметить, что ввиду того, что дивергенция и вращение геомагнитного поля равны нулю, могут быть получены следующие соотношения:

$$g_{xx} + g_{yy} + g_{zz} = 0,$$
$$g_{xy} = g_{yx}, g_{xz} = g_{zx}, g_{yz} = g_{zy}. \quad (4)$$

Следовательно, тензор магнитного градиента представляет собой симметричную матрицу размером 3 × 3, для которой могут быть выделены пять независимых компонентов, обозначенных соответственно, как $g_{xx}$, $g_{yy}$, $g_{xy}$, $g_{yz}$ и $g_{zx}$ [11]. При этом в соответствии с уравнениями Лапласа сумма размещенных по диагонали матрицы элементов равна нулю.

В общем виде тензор вообще и градиента геомагнитного поля в частности традиционно задается тремя основными характеристиками [11]: форма *C*, ранг *R* и размер *I*, которые задаются выражением следующего вида:

$$G = [C, R, I]. \quad (5)$$

Так, для рассматриваемого в случае геомагнитного поля тензора запись (5) будет иметь вид: $G = [3, 3, 9]$. Иными словами, имеет место прямоугольный тензор, в котором длина каждой из осей (форма) равна 3, количество осей (ранг) также равен 3, общее количество элементов в тензоре (размер) равно 9.

В каждой точке пространства геомагнитное поле задается тензором одного и того же (второго) ранга, также называемого диадами [11]. В этом случае для двух и более пространственных точек возможно получить новый тензор так же второго ранга, полученный алгебраическим суммированием каждого компонента тензора одного слагаемого с соответствующим компонентом тензора другого слагаемого. Таким образом, возможно рассмотреть общий тензор градиента поля как результат сложения его компонент, что в общем виде выражается следующим образом:

$$G_A = \begin{bmatrix} g_{xx_A} & g_{xy_A} & g_{xz_A} \\ g_{yx_A} & g_{yy_A} & g_{yz_A} \\ g_{zx_A} & g_{zy_A} & g_{zz_A} \end{bmatrix}; G_B = \begin{bmatrix} g_{xx_B} & g_{xy_B} & g_{xz_B} \\ g_{yx_B} & g_{yy_B} & g_{yz_B} \\ g_{zx_B} & g_{zy_B} & g_{zz_B} \end{bmatrix};$$
$$G_{AB} = \begin{bmatrix} g_{xx_A} + g_{xx_b} & g_{xy_A} + g_{xy_B} & g_{xz_A} + g_{xz_B} \\ g_{yx_A} + g_{yx_B} & g_{yy_A} + g_{yy_B} & g_{yz_A} + g_{yz_B} \\ g_{zx_A} + g_{zx_B} & g_{zy_A} + g_{zy_B} & g_{zz_A} + g_{zz_B} \end{bmatrix}. \quad (5)$$

При этом ввиду неоднородности геофизических полей формирование новых тензоров на основе сложения нескольких известных должно быть ограничено сравнительно небольшими пространственными областями, размеры которых могут быть определены на предварительных этапах исследования. Как правило, указанные операции могут быть применимы только для статистически однородных пространственных точек, изменение исследуемого параметра в которых однозначно и равномерно определяется, к примеру, воздействием одних и тех же внешних факторов (в частности, параметров космической погоды).

Представляется целесообразным обозначить подтензор тензора $G$ градиента геомагнитного поля как его подмножество, заданное соответствующими параметрами формы, ранга и размера. В соответствии с принципами тензорного исчисления, подтензор должен иметь меньшую размерность по сравнению с исходным тензором. Так, для геомагнитного поля представляется целесообразным выделить, к примеру, подтензор, представленный значениями градиента одного из компонент вектора по всем осям. Обозначим подтензор геомагнитного поля $G$ как $G'$ с характеристиками вида [1, 3, 3]. Иными словами, имеет место новый тензор, в котором длина каждой из осей (форма) равна 1, количество осей (ранг) также равен 3, общее количество элементов в тензоре (размер) равно 3.

Введем для подтензора $G'$ следующее соотношение, характеризующее градиент по оси $x$ для трех заданных в исходном тензоре осей:

$$G' = \begin{bmatrix} g_{xx} \\ g_{yx} \\ g_{zx} \end{bmatrix}; G \ni G'. \quad (6)$$

Аналогичным образом для каждой пространственной точки возможно выполнить свертку тензора таким образом, чтобы свести тензор ранга $N$ к тензору ранга $M$, при этом $N > M$. Свертка тензора, в соответствии с принципами тензорного исчисления [11], сводится к понижению валентности (ранга) тензора на 2. Иными словами, формируется новый тензор ранга $M$, такого что $M = N - 2$. Поскольку в случае градиента геомагнитного поля имеет место тензор второго ранга, то результатом свертки является скаляр, называемый первым главным инвариантом или следом тензора [11]:

$$G_{ii} = I(\mathbf{G}) = \mathrm{tr}\, G. \tag{7}$$

Свертка всегда производится по паре разновариантных индексов (один индекс должен быть верхним, а другой нижним). При этом след для двухрангового тензора является скаляром. Так, для выражения (3) выполним свертку вида по его единственной паре индексов:

$$\mathrm{tr}\, G = \mathrm{tr}\left(\begin{bmatrix} g_{xx} & g_{xy} & g_{xz} \\ g_{yx} & g_{yy} & g_{yz} \\ g_{zx} & g_{zy} & g_{zz} \end{bmatrix}\right) = g_{xx} + g_{yy} + g_{zz}. \tag{8}$$

Таким образом, для оптимизации хранения и вариативности представления параметров геомагнитного поля целесообразно использовать в том числе и свертку соответствующего тензора, который для совокупности пространственных точек также демонстрирует картину пространственно-временного распределения соответствующих параметров.

**5. Трансформационный тензор.** Особенностью описания исследуемых параметров поля в виде тензоров представляется целесообразным отметить возможность трансформации значений при изменении соответствующих систем координат. При этом ни ранг тензора, ни его размерность, ни форма в результате такой трансформации изменений не претерпевают.

Так, к примеру, параметры геомагнитного поля могут быть и обычно представлены как в декартовой, так и сферической системе координат. Начало этой системы координат помещено в центре Земли, полярная ось направлена по оси вращения Земли, координата отсчитывается вдоль радиус-вектора, проведенного из центра Земли. В общем виде сферические координаты задаются в формате ($r$, θ, λ), где $r$ – радиус-вектор объекта, θ – полярное расстояние (коширота) в диапазоне от $0^o$ до $180^o$, λ – долгота в диапазоне от $0^o$ до $360^o$.

Для примера представляется целесообразным привести соотношения, характеризующие связь сферической и декартовой систем координат:

$$\begin{aligned} x &= r \cdot \sin(\theta)\cos(\lambda), \\ y &= r \cdot \sin(\theta)\sin(\lambda), \\ z &= r \cdot \cos(\theta). \end{aligned} \tag{9}$$

Кроме того, при решении научных и прикладных геофизических задач часто используются магнитные координаты (соответственно магнитные широта и долгота, а также магнитное время (Magnetic Local Time, MLT)), что также, в свою очередь, сопряжено с рядом необходимых преобразований.

В общем виде процесс трансформации координат и атрибутивных значений из одной системы координат в другую может быть задан функцией отображения вида:

$$f: B \to B', \tag{10}$$

где $B$ – исходная система координат, $B'$ – целевая система координат.

Более детальное развертывание функции с учетом преобразования отдельных атрибутивных значений при изменении соответствующих осей представляется целесообразным описать следующим образом:

$$f: B \to B': \forall a_i \; \exists \, f(a_{iB}) \overset{T}{\to} f(a_{iB'}), \tag{11}$$

где $B$ – исходная система координат, $B'$ – целевая система координат, $a_i$ – значение по $i$-й оси, $a_{iB}$ – значение по $i$-й оси в системе координат $B$, $a_{iB'}$ – значение по $i$-й оси в системе координат $B'$, $T$ – правила трансформации.

Здесь представляется целесообразным ввести понятия базового и трансформационного тензора. Целью первого из них является представление соответствующих параметров поля в заданной системе координат. Так, к примеру, градиент $G$ из выражения (3) может рассматриваться как базовый тензор градиента геомагнитного поля, выраженный в декартовой системе координат (рисунок 2). При этом правила преобразования значений из одной координатной системы в другую предлагается задавать в формате трансформационного тензора того же ранга, формы и размера, что исходный (рисунок 2).

В общем виде трансформационный тензор 2-го ранга с характеристиками [3, 3, 9] предлагается представить следующим образом:

$$G_{B \to B'} = \begin{bmatrix} f(a_{1B}) \overset{T}{\to} f(a_{1B'}) & f(a_{2B}) \overset{T}{\to} f(a_{2B'}) & f(a_{3B}) \overset{T}{\to} f(a_{3B'}) \\ f(a_{4B}) \overset{T}{\to} f(a_{4B'}) & f(a_{5B}) \overset{T}{\to} f(a_{5B'}) & f(a_{6B}) \overset{T}{\to} f(a_{6B'}) \\ f(a_{7B}) \overset{T}{\to} f(a_{7B'}) & f(a_{8B}) \overset{T}{\to} f(a_{8B'}) & f(a_{9B}) \overset{T}{\to} f(a_{9B'}) \end{bmatrix}. \tag{12}$$

Для геомагнитного поля указанное преобразование можно представить так, как показано в выражении, характеризующем трансформационный тензор для перехода от декартовой к сферической системе координат:

$$G_{C\to S} = \begin{bmatrix} \frac{\partial(f(F_{xC})\overset{T}{\to}f(F_{xS}))}{\partial r} & \frac{\partial(f(F_{xC})\overset{T}{\to}f(F_{xS}))}{\partial \theta} & \frac{\partial(f(F_{xC})\overset{T}{\to}f(F_{xS}))}{\partial \lambda} \\ \frac{\partial(f(F_{yC})\overset{T}{\to}f(F_{yS}))}{\partial r} & \frac{\partial(f(F_{yC})\overset{T}{\to}f(F_{yS}))}{\partial \theta} & \frac{\partial(f(F_{yC})\overset{T}{\to}f(F_{yS}))}{\partial \lambda} \\ \frac{\partial(f(F_{zC})\overset{T}{\to}f(F_{zS}))}{\partial r} & \frac{\partial(f(F_{zC})\overset{T}{\to}f(F_{zS}))}{\partial \theta} & \frac{\partial(f(F_{zC})\overset{T}{\to}f(F_{zS}))}{\partial \lambda} \end{bmatrix}. \quad (13)$$

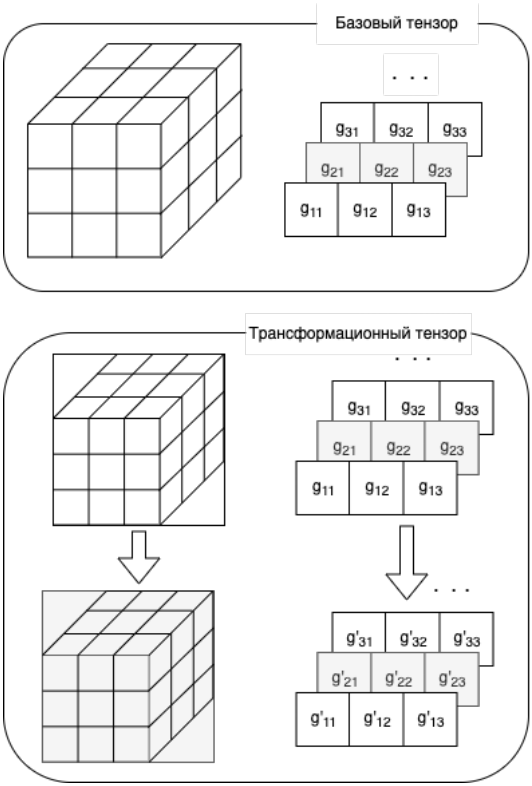



Рис. 2. Обобщенные схемы базового и трансформационного тензоров (кубы используются для условного обозначения тензора 2 ранга с 9 элементами, каждому из которых соответствует своя ячейка; фрагмент тензора (срез по одной из сторон) размещен в правой части рисунка)

При этом предполагается, что для каждой новой системы координат вводится собственный трансформационный тензор, в котором формулируются правила преобразования значений поля из одной системы координат в другую. Фактически при смене системы координат происходит «наложение» двух тензоров: базового и трансформационного. При этом сформированный новый тензор заменяет имеющийся базовый.

**6. Информационное обеспечение тензоров.** Сложность представления тензоров в соответствии с предложенным подходом заключается в необходимости хранения метаданных, базовых тензоров и выражений для их преобразования в составе трансформационного тензора для последующего оперативного доступа к ним с возможностью интеграции в единый пространственный слой. Любая известная модель данных (сетевая, иерархическая, реляционная и пр.) позволяет частично решить данную задачу, ориентируя разработчика преимущественно на представление базового тензора, с одной стороны, и метаданных, с другой. Каждому из указанных компонент данных может быть выделена собственная структура в соответствии с принятой моделью, например, отдельная таблица в случае реляционной модели. В таком случае сама суть трансформации тензора (в частности, при переходе от одной системы координат к другой) при таком подходе остается нераскрытой. В первом приближении соответствующие трансформационные операции должны быть заданы на уровне бизнес-логики соответствующего приложения и выполнятся по мере обращения потребителя данных к ним. Это сужает возможности применения решения различными потребителями, поскольку требует повторной реализации выражений трансформации отдельно для каждого программного решения. В результате увеличивается число процессов обработки данных и, как следствие, снижается реактивность соответствующих программных решений. Представляется целесообразным разработка такого подхода к представлению тензоров в рамках предложенной концепции, которая позволит задать в единой структуре тензоры и правила их трансформации с возможностью динамического изменения информационной структуры таким образом, чтобы в зависимости от метаданных тензора формировать соответствующие экземпляры объектов для их хранения. По сути имеет место динамическая структура, адаптируемая под конкретные тензоры. Извлечение искомых данных осуществляется единой операцией по обработке в составе одной структуры и может быть реализовано многими потребителями под управлением единого метода доступа.

Базовому и трансформационному тензору ставятся в соответствие метаданные, на основании которых возможно идентифицировать базис исходного тензора, а также определить, к какой системе координат приведет сопоставления базового и трансформационного тензоров. Метаданные по каждому тензору могут быть представлены в произвольном формате. Однако, накопленный авторами опыт разработки и исследований в области обработки пространственной информации показывает, что наиболее эффективным в данном случае является использование XML-формата [11].

Непосредственно данные по каждому тензору независимо от его вида в общем случае представляют собой матрицу. Ее физическое хранение может быть выполнено посредством любых доступных реализаций моделей представления данных, например, реляционных, иерархических, объектно-ориентированных и пр. Вместе с тем, представляется целесообразным использовать так называемый гибридный формат представления информации, сочетающий в себе реляционный и иерархический подходы к организации данных.

Предлагается организовать представление данных следующим образом. Каждому тензору ставится в соответствие отдельный кортеж в реляционной таблице. При этом представляется целесообразным выделить базовый тензор в одну таблицу, а трансформационный тензор – в другую. Таблицы связаны друг с другом неявным отношением вида «один-ко-многим» («1:M»), в котором на основании связки «первичный ключ (Primary Key, PK) – внешний ключ (Foreign Key, FK)» возможно определить правила трансформации из некоторой системы координат в тот, который задан непосредственно в базовом тензоре. Поскольку к базовому тензору могут привести преобразования из многих других систем координат, соответствующая реляционная таблица помечается как реляционная.

Характеристики тензора любого вида формируются частично автоматически и помещаются в поле соответствующего кортежа реляционной таблицы. Метаданные, о представлении которых в XML-формате сообщалось выше, заполняют одноименное поле таблицы в составе заданного кортежа в неформатированном виде таким образом, чтобы в дальнейшем их можно было обработать стандартными средствами работы с XML-данными (в частности, с использованием соответствующих DOM-объектов [12]).

При формировании и загрузке базового тензора часть его метаданных должна быть вычислена автоматически. К ним относятся, в частности, форма, ранг и размер соответствующего тензора.

Эта информация получается в результате парсинга данных, представленных в том же кортеже и характеризующих непосредственно значимую часть тензора.

Каждая ось базового тензора также размещается в соответствующем кортеже реляционной таблицы и занимает одну ветку XML-представления. Количество веток, соответственно, определяет количество осей тензора и является основанием для вычисления ранга и других метаданных. Фактически сам тензор в своем содержательном представлении задается отдельным XML-документом, в котором число узлов первого уровня соответствует осям, каждый дочерний XML-элемент, в свою очередь, показывает элементы тензора в родительской оси и пр. Соответственно парсинг такой XML-структуры позволяет вычислить размерность тензора и оставшиеся метапараметры для другого поля в том же кортеже реляционной таблицы.

Для трансформационного тензора в целом схема представления / физического хранения такая же, как для базового тензора. Одно поле кортежа реляционной таблицы выделено под XML-представление метаданных, другое – под XML-представление непосредственно выражения для трансформации (например, в соответствии с выражением (9) применительно к геомагнитным данным). Непосредственно трансформация затрагивает поэлементное преобразование компонент тензора из базового кортежа в соответствии с выражениями, представленными в соответствующей связанной дочерней записи трансформационного тензора. При этом во избежание возможных коллизий при обработке XML-структуры тензора, в частности, его узлов, содержащих правила трансформации атрибутивных значений из одной системы координат в другую, последние должны быть представлены в формате, условно игнорируемой XML-парсерами и соответствующими программными библиотеками. В качестве такого формата могут быть использованы, в частности, комментарии, а также блоки необрабатываемого текстового содержимого типа CDATA. Для представления трансформационных инструкций авторами был выбран второй из указанных вариантов: соответствующие конструкции помещаются между ограничителями секции CDATA и извлекаются посредством дополнительных манипуляций, что позволяет в целом избежать их автоматической обработки и запуска на выполнение, способных привести к различных ошибкам при работе с тензором.

Еще один важный момент касательно физического реляционно-иерархического представления обозначенных тензоров связан

с реализацией отношения между таблицами, соответствующими разным типам тензоров. С этой целью в дочерней реляционной таблице, соответствующей трансформационным тензорам, вводится идентификатор – первичный ключ в составе отдельного поля, например, числового типа со свойствами инкрементируемого счетчика. Альтернативным вариантом, в частности, является использование суррогатного ключа (по усмотрению разработчика). В соответствующем кортеже родительской таблицы размещается ссылка на данный идентификатор – внешний ключ.

Предложенная схема организации физического хранения и представления тензоров обоих видов представлена в общем виде на рисунке 3.

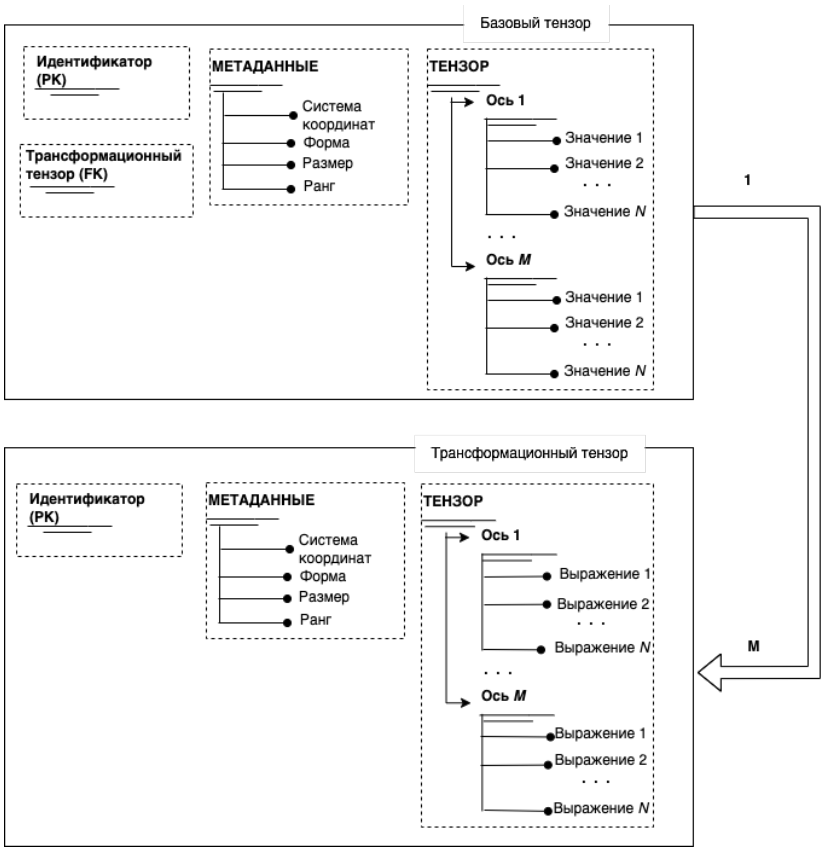



Рис. 3. Общая схема информационной модели тензоров

Платформа для реализации обозначенной модели ограничивается только реляционными СУБД ввиду того, основой соответствующей модели является именно реляционное представление. Конкретная реализация зависит от возможностей и комплекса технических средств разработчика.

**7. Подход к визуализации тензорных полей.** Анализ описанных выше подходов к визуализации тензорных полей показал их невысокую эффективность и информативность с точки зрения комплексного рассмотрения составляющих его параметров по различным осям заданной системы координат. Наиболее распространенным при этом является вариант послойного представления скалярного разложения тензорного поля на компоненты, что при большом количестве данных, характеризующих сложно организованный процесс / явление, может привести к потере значимой информации ввиду перегруженности итогового пространственного изображения.

В настоящей работе предлагается подход, учитывающий поосевое распределение параметров поля в составе соответствующего тензора и позволяющий продемонстрировать пространственное (или в ряде задач пространственно-временное) распределение данных в удобной для конечного пользователя форме. При этом ожидается, что возможность комплексного визуального анализа на уровне пространственного изображения позволит пользователям повысить оперативность принятия решений в соответствующих прикладных областях.

В основе предлагаемого подхода лежит предположение, что каждый тензор можно представить в виде геопространственного примитива, форма которого адаптирована под соответствующие значения ранга тензора и его формы. В качестве такого базового графического примитива представляется выбрать специализированную фигуру – глиф (glyth). В терминах научной визуализации глифы или графические символы отображают несколько значений данных, характеризуя их форму, размер, ориентацию и внешний вид поверхности базового геометрического примитива [14, 15]. Здесь представляется целесообразным отметить, что непосредственно пространственные данные по тензорам геомагнитного поля представлены в соответствии с информационной моделью, предложенной в предыдущем разделе. Единым запросом с заданными пространственно-временными параметрами формируется обращение к соответствующему информационному хранилищу тензоров. Формируемый при этом результат передается модулю визуализации

для последующего формирования соответствующего пространственного слоя.

При этом представляется целесообразным отметить, что при визуализации глифов используются строго определенные графические примитивы, среди которых выделяют эллипсоид, кубоид, цилиндрический глиф, а также суперквадрикс [16]. Предполагается, что выбор базовой формы глифа напрямую зависит от количества составляющих тензор осей, что, очевидно, определяется рангом соответствующего глифа. Представляется целесообразным в качестве такой основы выбрать простейший вариант представления глифа – эллипсоид. Для поддержки представления эллипсоида для визуализации сложных глифов представляется целесообразным использовать суперэллипсоиды, базирующиеся на использовании кривых Ламе [17].

Для возможности применения суперэллипсоида для решения поставленной задачи представляется дополнить его соответствующими осями с центром в центроиде графического примитива. Внутри каждой оси глифа представлена конкретная составляющая тензора. Например, в случае тензора градиента геомагнитного поля, такими составляющим являются соответствующие компоненты вектора магнитного поля Земли. Здесь и далее представляется целесообразным использовать термин «тензорный глиф» для описания подхода к визуализации рассматриваемого типа полей.

Далее размер и цвет каждой составной части вдоль соответствующей виртуальной оси, исходящей из центроида, можно использовать для отображения информации о скалярных величинах. В частности, для информативного изображения конкретных составляющих представляется использовать монохромное представление значений по каждой виртуальной оси глифа. При этом по заранее сформированной цветовой маске в зависимости от конкретного визуализируемого значения итоговый цвет должен быть представлен в виде соответствующего градиента с шириной, определяемой размерами глифа и количеством предусмотренных в нем осей. В результате интенсивность градиента по каждой из осей тензорного глифа в составе единой фигуры позволит визуально оценить их распределение с учетом аналогичных изображений в соседних пространственных точках.

Для указания направленной информации о тензорном поле необходимо дополнить глиф любым изображением указателя, представляющим вектор. Поскольку в рассматриваемом подходе

к визуализации предлагается использовать неявные оси, то и указание направления должно быть реализовано посредством дополнительных пиктограмм, составляющих части исходного глифа.

В общем виде итоговое пространственное изображение складывается из множества суперэллипсов, координатно привязанных к соответствующим пространственным точкам в заданных системах координат. Центроид суперэллипса геометрически совмещен с соответствующей пространственной точкой. В результате сформированное пространственное изображение визуально не отличается от традиционного пространственного слоя и при необходимости может быть также дополнено дополнительными пространственными изображениями и соответствующими пространственными слоями без потери информативности. Пояснительная информация по глифу должна быть документирована соответствующей цветовой схемой (легендой).

Во избежание перегруженности предлагаемого визуального представления полей представляется ограничиться только тензорами нулевого, первого и второго порядка, чтобы количество возможных осей в тензорном глифе не превышало трех.

Особенность предложенного подхода к визуализации на основе суперэллипсов заключается в том, что в случае нечетного количества осей эллипс теряет собственную симметричность. Известно, что в общем виде простой эллипс представляет собой фигуру, заданную кривыми второго порядка и выражаемую отношением вида:

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}=1, \tag{14}$$

где коэффициенты $a$ и $b$ определяют соответственно сжатость («приплюснотость») эллипса вдоль осей координат.

Вместе с тем, такой подход неприменим в том случае, если осей более двух и невозможно определить искомые коэффициенты для построения соответствующего симметричного эллипса. В этом случае его замена на суперэллипс выражается с помощью соотношения следующего вида:

$$\left(\frac{x}{a}\right)^n+\left(\frac{y}{b}\right)^n=1, \tag{15}$$

где $n$ – коэффициент, значение которого определяется количеством осей, которые должны быть визуализированы в составе суперэллипса. Так, к примеру, при $n = 1$, результатом визуализации суперэллипса в общем виде является ромб с вершинами на осях координат. В случае, если значение указанного коэффициента лежат в диапазоне $1 < n < 2$, то результатом визуализации суперэллипса является ромб с выпуклыми сторонами. При $n = 2$ результатом визуализации является эллипс (или, если $a$ и $b$ равны – окружности). По мере приближения $n$ к бесконечности результат визуализации приближается внешне к прямоугольнику.

С учетом особенностей градации тензоров по рангам представляется подбирать коэффициент $n$ таким образом, что в общем виде его можно представить как:

$$\forall\, G = [l, r, N]\ \exists S_G = \left(\frac{o_1}{a}\right)^n + \left(\frac{o_2}{b}\right)^n + \cdots + \left(\frac{o_{r-1}}{z}\right)^n = 1 : n = r + 1, \quad (16)$$

где $G$ – тензор формы $l$ ранга $r$ с общим количеством элементов, равным $N$. Здесь и далее $o_1$, $o_2$, …, $o_{r\text{-}1}$ соответствует осям суперэллипса, выделенным в соответствии с рангом визуализируемого тензора $r$.

Представляется целесообразным также отметить, что подбор коэффициентов $a$, $b$,…, $z$ возможно выполнить пропорционально значениям соответствующим осям тензора. При этом для каждой оси коэффициенты должны быть подобраны с учетом масштабирования соответствующих значений. Аналогичные расчеты представляется целесообразным выполнять в том числе и на уровне соответствующих градиентов монохрома по рассматриваемым осям тензора и непосредственно характеризующего его суперэллипса.

Здесь следует отметить, что в отдельных случаях получаемые в результате подбора соответствующих коэффициентов суперэллипсы могут выглядеть так, как будто они имеют прямые стороны по каждой из осей. Вместе с тем все точки на пересечении с осями суперэллипса соединены кривыми, которые изогнуты фактически по всему периметру. Иными словами, даже в тех фрагментах, где сегмент суперэллипса внешне выглядит прямым, в действительности он слегка изогнут. При этом кривизна соответствующих линий суперэллипса изменяется повсюду непрерывно [18, 19].

На рисунке 4 приведены обобщенные примеры использования суперэллипсов для визуализации тензоров различного ранга.

Результаты отличаются количеством осей, вдоль которых размещаются эллипсы.

Информативность предложенного варианта визуализации тензоров проявляется двояко. С одной стороны, вытянутость вдоль определенной оси суперэллипсов, изображенных на рисунке 4, показывает ранг тензора. С другой стороны, цвет тех же суперэллипсов отражает значения градиентов тензоров геомагнитного поля. В целом это позволяет оперативно оценить визуализируемые данные (атрибутивные параметры) как количественно, так и качественно в едином глифе.

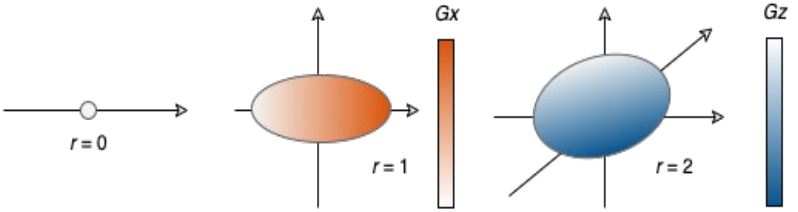


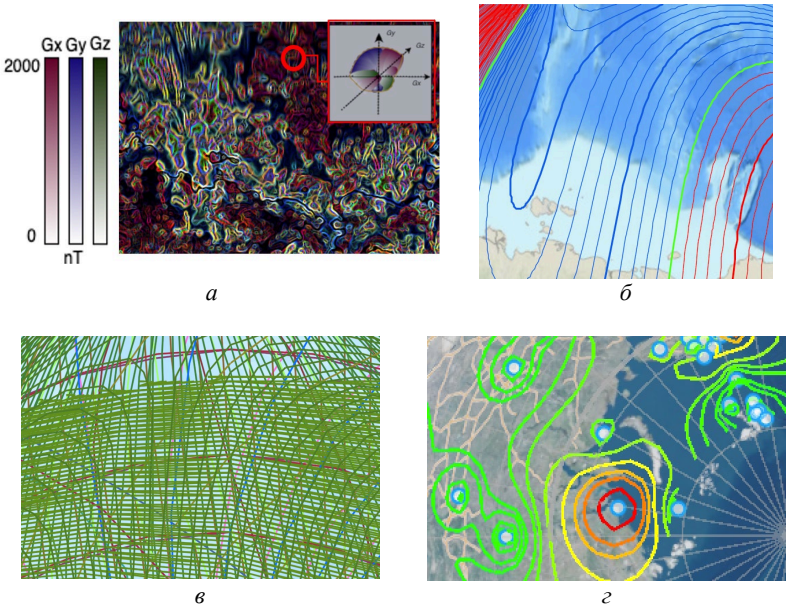

Рис. 4. Примеры визуализации симметричных суперэллипсов для тензоров различных рангов

Для простоты и наглядности в каждом из вариантов на рисунке 4 рассматривается частный случай суперэллипса, представленный симметричным эллипсом. В действительности такая ситуация возможна крайне редко и в данном случае является просто искусственно синтезированным примером. В случае, если ранг тензора не нулевой, то подбор коэффициентов $n$, $a$, $b$,… $z$ приводит к тому, что итоговый суперэллипс сильно искривляется по отношению к той или иной оси.

**8. Апробация предложенного подхода.** Для подтверждения работоспособности предложенного подхода к обработке, анализу и визуализации тензорных полей в качестве тестового был использован тестовый набор данных, характеризующих пространственное распределение градиента магнитного поля в виде набора соответствующих тензоров второго ранга.

Основным критерием оценки эффективности стала наглядность результата визуализации. Для сравнения были выполнены аналогичные преобразования на примере однослойной и многослойной визуализации посредством системы пространственных изолиний, а также смешанного подхода на основе изолиний и пространственных точек [20, 21] (рисунок 4).

Представляется целесообразным отметить, что во всех отличных от предлагаемого подхода случаях решение задачи визуализации поля сводится к его преобразованию на уровне отдельных скалярных значений. К примеру, в контексте геомагнитного поля таковым является один из компонент соответствующего вектора. Результат визуализации пространственно-временного распределения по одному заданному компоненту представлен в виде системы пространственных изолиний на рисунке 5(б).




*а* *б*

*в* *г*

Рис. 5. Возможности визуализации тензора произвольного градиента магнитного поля: а) предлагаемый подход (градиенты по трем заданным в исходном тензоре осям); б) однослойные изолинии, в) многослойные изолинии, )однослойные изолинии и пространственные точки (градиент по оси x для трех заданных в исходном тензоре осей)

В действительности необходимая информация о пространственном распределении параметров поля должна быть получена только в комплексном рассмотрении составляющих его компонент. Так, на рисунке 5(в) показано, каким образом выполнена визуализация всех компонент геомагнитного поля с учетом того, что каждый из них представлен собственным пространственным слоем, заданным соответствующим набором геопространственных изолиний

в различной цветовой схеме. Из рисунка видно, что наблюдается плотное перекрывание пространственных слоев, что, в итоге, крайне затрудняет их визуальный анализ и интерпретацию, фактически не позволяя отделить один компонент вектора от другого.

На рисунке 5(г) приведен результат визуализации поля, также представляющий собой гибридный подход. Один из параметров поля представлен в виде совокупности пространственный изолиний, а другой визуализируемый параметр задан множеством пространственных точек. При этом нет принципиальной разницы в послойном способе реализации предложенного подхода, который может задавать как отдельные слои, соответствующие отдельным визуализируемым параметрам, так и единый слой, в котором сочетаются все визуализируемые параметры, в том числе, представленные посредством различных геопространственных примитивов (в данном случае, это пространственные полилинии и точки соответственно).

Фрагмент результатов визуализации, полученных посредством предложенного подхода к графической интерпретации тензорных полей, представлен на рисунке 5(а). В каждой доступной пространственной точке проведена визуализация суперэллипса с тремя осями, вдоль каждой из которых соответствующие визуализируемые значения компонента вектора геомагнитного поля характеризуются градиентом заданного цвета (монохром). Здесь показаны градиенты по осям *x, y, z* (соответственно Gx, Gy, Gz в соответствии с выражением (3) и подтензорами выражения (6) для трех заданных в исходном тензоре вектора геомагнитного поля осей (данные извлекаются из хранилища, построенного в соответствии с предложенной гибридной моделью на базе СУБД PostgreSQL и Sedna XML). Полученный в результате пространственный слой в общем виде несет информацию о характере пространственного распределения совокупности составляющих визуализируемое тензорное поле компонент с распределением по соответствующим осям. Дополненное легендой с пояснениями касательно используемой цветовой схемы, а также дополнительной контекстной информацией о каждой составляющей слоя такой подход позволяет составить общую (единую) картину распределения всех компонент рассматриваемого поля (в данном случае данные были сформированы произвольно).

Для обобщения результатов сравнительного анализа особенностей визуализации компонент тензорного поля (на примере вектора геомагнитного поля) различными способами был выделен ряд критериев, характеризующих их результативность. Здесь

представляется целесообразным отметить, что оценить эффективность визуализации в рассматриваемых случаях возможно только качественно. Однако, небольшая группа показателей может быть условно оценена количественно. Соответствующие результаты сравнительного анализа по выделенным критериям эффективности визуализации приведены в таблице 1.

Таблица 1. Результаты сравнительного анализа подходов к визуализации тензора произвольного градиента магнитного поля

| Критерий / Подход | Количество визуализируемых в слое атрибутивных параметров / количество пространственных слоев | Время рендеринга пространственного слоя (слоев) (с) | Время отклика (с) |
|---|---|---|---|
| Подход на основе тензорного исчисления | ≥ 1 / 1 | 4 | 2 |
| Однослойные изолинии | 1 / 1 | 2 | 2 |
| Многослойные изолинии | ≥ 1 / ≥ 1 | 5 | 6 |
| Однослойные изолинии и пространственные точки | ≥ 1 / ≥ 1 | 4 | 5 |

Примечание: вычислительные эксперименты были проведены на клиентской стороне с применением ЭВМ (CPU Intel Core i5 10300H ГГц, оперативная память 4 ГБ, скорость интернет-соединения ~52.4 Мбит/с) и на серверной стороне – на базе веб-сервера с процессором 72 * Intel(R) Xeon(R) Gold 6140 CPU @ 2.30GHz.

Анализ представленных в таблице 1 результатов позволяет заключить, что при минимальном времени отклика и не превышающем максимум времени рендеринга предлагаемый подход на основе тензорного исчисления обеспечивает возможность визуализации множества атрибутивных параметров в рамках одного пространственного слоя. Ожидается, что это позволит упростить анализ данных для специалистов в соответствующей области, в частности, для интерпретации геомагнитной информации.

**9. Заключение.** В настоящее время задача обработки, анализа и визуализации полей различной природы происхождения успешно

решается только для скалярных значений. Известные решения обеспечивают широкий спектр инструментов, позволяющих визуализировать скалярные значения, распределенные по земной поверхности.

Вместе с тем данные, описывающие те или иные процессы и / или явления, имеют более сложную, отличную от атомарных значений структуру. К ним относятся, в частности, векторные данные, которые помимо атрибутивных значений характеризуются соответствующим направлением, вектором, исходящим из заданной пространственной точки. Еще более сложной структурой обладают данные, относящиеся к категории тензорных полей: в этом случае увеличивается количество направлений / осей / векторов, вдоль которых анализируются соответствующие атрибутивные значения. Так, примерами таких данных являются, в частности, гравитационное, геомагнитное поле, которые задаются многокомпонентными векторами или тензорами (так называемые тензорные поля). На сегодняшний день обработка и визуализация таких данных осуществляются в подавляющем большинстве случаев их декомпозицией на скалярные составляющие, каждый из которых рассматривается отдельно от других. В результате теряется значимая для принятия решений и / или исследования некоторого процесса / явления информация, которая должна быть рассмотрена и проанализирована именно в комплексной своей форме.

В этой связи в работе был предложен подход, ориентированный на обработку, хранение и визуализацию данных тензорных полей на основе принципов тензорного исчисления. На примере геофизических данных было формализовано тензорное представление соответствующих атрибутивных значений, определены типовые операции объединения / декомпозиции отдельных тензорных полей и их составляющих.

На основе смешанного подхода, сочетающего в себе элементы реляционных и иерархических моделей данных, был предложен способ физического хранения информации по тензорным полям. В частности, была сформулирована концептуальная схема представления тензорного поля именно по данному предложенному гибридному подходу.

Проведенные формализованные результаты послужили основой для создания подхода к визуализации тензорных полей. При этом проведенный анализ известных подходов позволил выявить их преимущества для визуализации сложно организованных пространственных данных и использовать их для формулировки

нового, усовершенствованного подхода. Предложенный авторами подход получил название тензорного глифа. При этом в качестве геопространственного примитива предложено использовать суперэллипсы, оси которого соответствуют рангу визуализируемого тензора, а атрибутивные значения выражаются варьированием цветового градиента в его монохромном представлении.

На примере произвольных значений вектора геомагнитного поля предложенный подход был сравнен с точки зрения наглядности с известными практикуемыми подходами: однослойным представлением одного скалярного значения, многослойным представлением нескольких скалярных значений, а также сочетанием различных геопространственных примитивов для представления разнородных скалярных значений. Проведенный анализ показал, что применение предложенного подхода позволят заметно «разгрузить» итоговое пространственное изображение без необходимости многослойного представления. Дополненное легендой с пояснениями касательно используемой цветовой схемы, а также дополнительной контекстной информацией о каждой составляющей слоя такой подход позволяет составить общую (единую) картину распределения всех компонент рассматриваемого поля.

## Литература


1. Vorobev A.V., Pilipenko V.A., Sakharov Ya.A., Selivanov V.N. Statistical relationships between variations of the geomagnetic field, auroral electrojet, and geomagnetically induced currents // Solar-Terrestrial Physics. 2019. vol. 5. no. 1. pp. 35–42.
2. Vorobev A.V., Pilipenko V.A., Enikeev T.A., Vorobeva G.R. Geoinformation system for analyzing the dynamics of extreme geomagnetic disturbances from observations of ground stations // Computer Optics. 2020. vol. 44. no. 5. pp. 782–790.
3. Fleming J., Marvel S., Supak S., Motsinger-Reif A., Reif D. ToxPi*GIS Toolkit: creating, viewing, and sharing integrative visualizations for geospatial data using ArcGIS // Journal of Exposure Science & Environmental Epidemiology. 2022. vol. 32. no. 6. pp. 900–907. DOI: 10.1038/s41370-022-00433-w.
4. Simonyan A., Ohanyan M. Refined Spatio-Temporal Model of Accelerations of the Main Geomagnetic Field on the Earth's Surface and Geomagnetic Jerks // Geomagnetism and Aeronomy. 2023. vol. 63. no. 3. pp. 325–348. DOI: 10.1134/S0016793223600078.
5. Boyarchuk M.A., Zhurkin I.G., Nepoklonov V.B. Concept of a visualization method for Earth's gravity field on plain maps // Scientific Visualization. 2019. vol. 11. no. 1. pp. 70–79. DOI: 10.26583/sv.11.1.06.
6. Peng Z, Laramee S. Higher Dimensional Vector Field Visualization. A Survey // Theory and Practice of Computer Graphics (TPCG '09). 2009. pp. 149–163.
7. Meuschke M., Vob S., Gaidzik F., Preim B., Lawonn K. Skyscraper Visualization of Multiple Time-Dependent Scalar Fields on Surfaces // Computers & Graphics. 2021. vol. 99. pp. 22–42. DOI: 10.1016/j.cag.2021.05.005.

8. Lobo M.-J., Telea A., Hurter C. Feature Driven Combination of Animated Vector Field Visualizations // Computer Graphics Forum. 2020. vol. 39. no. 3. pp. 429–441. DOI: 10.1111/cgf.13992.
9. Hergl C., Blecha C., Kretzschmar V., Raith F., Gunther F., Stommel M., Jankowai J., Hotz I., Nagel T., Scheuermann G. Visualization of Tensor Fields in Mechanics // Computer Graphics Forum. 2021. vol. 40. no. 6. pp. 135–161. DOI: 10.1111/cgf.14209.
10. He Z., Hu X., Teng Yu., Zhang X., Shen X. Data agreement analysis and correction of comparative geomagnetic vector observations // Earth, Planets and Space. 2022. vol. 74. DOI: 10.1186/s40623-022-01583-9.
11. Huang Y., Wu L., Li D. Theoretical Research on Full Attitude Determination Using Geomagnetic Gradient Tensor // The Journal of Navigation. 2015. no. 68(5). pp. 951–961. DOI: 10.1017/S0373463315000259.
12. Vorobev A.V., Vorobeva G.R., Yusupova N.I. Conception of geomagnetic data integrated space // SPIIRAS Proceedings. 2019. vol. 18. no. 2. pp. 390–415. DOI: 10.15622/sp.18.2.390-415.
13. Reddy B., Bommala H., Bhyrapuneni S. Strategies and Approaches for Generating Identical Extensive XML Tree Instances // International Journal on Recent and Innovation Trends in Computing and Communication. 2023. vol. 11. pp. 559–564. DOI: 10.17762/ijritcc.v11i8s.7238.
14. Yu Q., Zhang X., Huang Zh.-H. Tensor Factorization-Based Method for Tensor Completion with Spatio-temporal Characterization // Journal of Optimization Theory and Applications. 2023. vol. 119. pp. 337–362. DOI: 10.1007/s10957-023-02287-0.
15. Xia S., Qiu D., Zhang X. Tensor factorization via transformed tensor-tensor product for image alignment // Numerical Algorithms. 2023. vol. 22. pp. 1251–1289. DOI: 10.1007/s11075-023-01607-9.
16. Tomasevic D., Peer P., Solina F., Jaklic A., Struc V. Reconstructing Superquadrics from Intensity and Color Images // Sensors. 2022. vol. 22(14). no. 5332. DOI: 10.3390/s22145332.
17. Mamieva I. Ruled algebraic surfaces with a main frame from three superellipses // Structural Mechanics of Engineering Constructions and Buildings. 2022. vol. 18. no. 4. pp. 387–395. DOI: 10.22363/1815-5235-2022-18-4-387-395.
18. Borisenko V., Ustenko S., Ustenko I. Constructing a method for the geometrical modeling of the lame superellipses in the oblique coordinate systems // Eastern-European Journal of Enterprise Technologies. 2020. vol. 2. no. 4. pp. 51–59. DOI: 10.15587/1729-4061.2020.201760.
19. Olayiwola T., Choi S.-J. Superellipse model: An accurate and easy-to-fit empirical model for photovoltaic panels // Solar Energy. 2023. vol. 262. DOI: 10.1016/j.solener.2023.05.026.
20. Vorobev A.V., Pilipenko V.A., Enikeev T.A., Vorobeva G.R. Geoinformation system for analyzing the dynamics of extreme geomagnetic disturbances from observations of ground stations // Computer Optics. 2020. vol. 44. no. 5. pp. 782–790.
21. Vorobev A.V., Pilipenko V.A., Enikeev T.A., Vorobeva G.R., Khristodulo O.I. System for dynamic visualization of geomagnetic disturbances according to the data of ground magnetic stations // Scientific Visualization. 2021. vol. 13. no. 1. pp. 162–176. DOI: 10.26583/sv.13.1.11.

**Воробьева Гульнара Равилевна** — д-р техн. наук, профессор кафедры, кафедра вычислительной математики и кибернетики института математики, информатики и робототехники, Уфимский университет науки и технологий. Область научных интересов: геоинформационные и веб-технологии, системы хранения и обработки

информации. Число научных публикаций — 158. gulnara.vorobeva@gmail.com; улица Карла Маркса, 12, 450077, Уфа, Россия; р.т.: +7(917)417-4111.

**Воробьев Андрей Владимирович** — д-р техн. наук, доцент, профессор кафедры, кафедра информатики института математики, информатики и робототехники, Уфимский университет науки и технологий; научный сотрудник, Геофизический центр РАН. Область научных интересов: геоинформационные технологии, цифровая обработка сигналов. Число научных публикаций — 172. geomagnet@list.ru; улица Карла Маркса, 12, 450077, Уфа, Россия; р.т.: +7(917)345-2299.

**Орлов Глеб Олегович** — аспирант, кафедра вычислительной математики и кибернетики института математики, информатики и робототехники, Уфимский университет науки и технологий. Область научных интересов: геоинформационные и веб-технологии, системы защиты ПО от несанкционированного копирования. Число научных публикаций — 2. orlovgleb99@mail.ru; улица Карла Маркса, 12, 450077, Уфа, Россия; р.т.: +7(919)145-5147.

**Поддержка исследований.** Исследование выполнено при финансовой поддержке РНФ, проект № 21-77-30010.

**DOI** 10.15622/ia.23.2.10

G. VOROBEVA, A. VOROBEV, G. ORLOV

# THE CONCEPT OF PROCESSING, ANALYSIS AND VISUALIZATION OF GEOPHYSICAL DATA BASED ON ELEMENTS OF TENSOR CALCULUS

*Vorobeva G., Vorobev A., Orlov G.* **The Concept of Processing, Analysis and Visualization of Geophysical Data Based on Elements of Tensor Calculus.**

**Abstract.** One of the main approaches to processing, analysis and visualization of geophysical data is the use of geographic information systems and technologies, which is due to their geospatial reference. At the same time, the complexity of presenting geophysical data is associated with their complex structure, which involves many components that have the same geospatial reference. Vivid examples of data of such a structure and format are gravitational and geomagnetic fields, which in the general case are specified by three and four-component vectors with multidirectional coordinate axes. At the same time, today there are no solutions that allow visualizing these data in a complex without decomposing them into individual scalar values, which, in turn, can be presented in the form of one or many spatial layers. In this regard, the work proposes a concept that uses elements of tensor calculus for processing, storing and visualizing information of this format. In particular, a mechanism for tensor representation of field components has been formalized with the possibility of combining it with other data of the same format, on the one hand, and convolution when combined with data of a lower rank. Using the example of a hybrid relational-hierarchical data model, a mechanism for storing information on tensor fields is proposed, which provides for the possibility of describing and subsequently applying transformation instructions when transitioning between different coordinate systems. The paper discusses the use of this approach in the transition from the Cartesian to the spherical coordinate system when representing the parameters of the geomagnetic field. For complex visualization of tensor field parameters, an approach based on the use of tensor glyphs is proposed. The latter are superellipses with axes corresponding to the rank of the tensor. In this case, the attribute values themselves are proposed to be visualized relative to the corresponding axes of the graphic primitive in such a way that the data distribution can be specified by varying the gradient of the corresponding monochrome representation of the parameter along the corresponding axis. The performance of the proposed concept was investigated during a comparative analysis of the tensor approach with known solutions based on the scalar decomposition of the corresponding complex values with their subsequent representation in the form of one or many spatial layers. The analysis showed that the use of the proposed approach will significantly increase the visibility of the generated geospatial image without the need for complex overlapping of spatial layers.

**Keywords:** tensor fields, tensor calculus, geographic information technologies, glyphs, superellipses.

## References

1. Vorobev A.V., Pilipenko V.A., Sakharov Ya.A., Selivanov V.N. Statistical relationships between variations of the geomagnetic field, auroral electrojet, and geomagnetically induced currents. Solar-Terrestrial Physics. 2019. vol. 5. no. 1. pp. 35–42.
2. Vorobev A.V., Pilipenko V.A., Enikeev T.A., Vorobeva G.R. Geoinformation system for analyzing the dynamics of extreme geomagnetic disturbances from observations of ground stations. Computer Optics. 2020. vol. 44. no. 5. pp. 782–790.

3. Fleming J., Marvel S., Supak S., Motsinger-Reif A., Reif D. ToxPi*GIS Toolkit: creating, viewing, and sharing integrative visualizations for geospatial data using ArcGIS. Journal of Exposure Science & Environmental Epidemiology. 2022. vol. 32. no. 6. pp. 900–907. DOI: 10.1038/s41370-022-00433-w.
4. Simonyan A., Ohanyan M. Refined Spatio-Temporal Model of Accelerations of the Main Geomagnetic Field on the Earth's Surface and Geomagnetic Jerks. Geomagnetism and Aeronomy. 2023. vol. 63. no. 3. pp. 325–348. DOI: 10.1134/S0016793223600078.
5. Boyarchuk M.A., Zhurkin I.G., Nepoklonov V.B. Concept of a visualization method for Earth's gravity field on plain maps. Scientific Visualization. 2019. vol. 11. no. 1. pp. 70–79. DOI: 10.26583/sv.11.1.06.
6. Peng Z, Laramee S. Higher Dimensional Vector Field Visualization. A Survey. Theory and Practice of Computer Graphics (TPCG '09). 2009. pp. 149–163.
7. Meuschke M., Vob S., Gaidzik F., Preim B., Lawonn K. Skyscraper Visualization of Multiple Time-Dependent Scalar Fields on Surfaces. Computers & Graphics. 2021. vol. 99. pp. 22–42. DOI: 10.1016/j.cag.2021.05.005.
8. Lobo M.-J., Telea A., Hurter C. Feature Driven Combination of Animated Vector Field Visualizations. Computer Graphics Forum. 2020. vol. 39. no. 3. pp. 429–441. DOI: 10.1111/cgf.13992.
9. Hergl C., Blecha C., Kretzschmar V., Raith F., Gunther F., Stommel M., Jankowai J., Hotz I., Nagel T., Scheuermann G. Visualization of Tensor Fields in Mechanics. Computer Graphics Forum. 2021. vol. 40. no. 6. pp. 135–161. DOI: 10.1111/cgf.14209.
10. He Z., Hu X., Teng Yu., Zhang X., Shen X. Data agreement analysis and correction of comparative geomagnetic vector observations. Earth, Planets and Space. 2022. vol. 74. DOI: 10.1186/s40623-022-01583-9.
11. Huang Y., Wu L., Li D. Theoretical Research on Full Attitude Determination Using Geomagnetic Gradient Tensor. The Journal of Navigation. 2015. no. 68(5). pp. 951–961. DOI: 10.1017/S0373463315000259.
12. Vorobev A.V., Vorobeva G.R., Yusupova N.I. Conception of geomagnetic data integrated space. SPIIRAS Proceedings. 2019. vol. 18. no. 2. pp. 390–415. DOI: 10.15622/sp.18.2.390-415.
13. Reddy B., Bommala H., Bhyrapuneni S. Strategies and Approaches for Generating Identical Extensive XML Tree Instances. International Journal on Recent and Innovation Trends in Computing and Communication. 2023. vol. 11. pp. 559–564. DOI: 10.17762/ijritcc.v11i8s.7238.
14. Yu Q., Zhang X., Huang Zh.-H. Tensor Factorization-Based Method for Tensor Completion with Spatio-temporal Characterization. Journal of Optimization Theory and Applications. 2023. vol. 119. pp. 337–362. DOI: 10.1007/s10957-023-02287-0.
15. Xia S., Qiu D., Zhang X. Tensor factorization via transformed tensor-tensor product for image alignment. Numerical Algorithms. 2023. vol. 22. pp. 1251–1289. DOI: 10.1007/s11075-023-01607-9.
16. Tomasevic D., Peer P., Solina F., Jaklic A., Struc V. Reconstructing Superquadrics from Intensity and Color Images. Sensors. 2022. vol. 22(14). no. 5332. DOI: 10.3390/s22145332.
17. Mamieva I. Ruled algebraic surfaces with a main frame from three superellipses. Structural Mechanics of Engineering Constructions and Buildings. 2022. vol. 18. no. 4. pp. 387–395. DOI: 10.22363/1815-5235-2022-18-4-387-395.
18. Borisenko V., Ustenko S., Ustenko I. Constructing a method for the geometrical modeling of the lame superellipses in the oblique coordinate systems. Eastern-European Journal of Enterprise Technologies. 2020. vol. 2. no. 4. pp. 51–59. DOI: 10.15587/1729-4061.2020.201760.

19. Olayiwola T., Choi S.-J. Superellipse model: An accurate and easy-to-fit empirical model for photovoltaic panels. Solar Energy. 2023. vol. 262. DOI: 10.1016/j.solener.2023.05.026.
20. Vorobev A.V., Pilipenko V.A., Enikeev T.A., Vorobeva G.R. Geoinformation system for analyzing the dynamics of extreme geomagnetic disturbances from observations of ground stations. Computer Optics. 2020. vol. 44. no. 5. pp. 782–790.
21. Vorobev A.V., Pilipenko V.A., Enikeev T.A., Vorobeva G.R., Khristodulo O.I. System for dynamic visualization of geomagnetic disturbances according to the data of ground magnetic stations. Scientific Visualization. 2021. vol. 13. no. 1. pp. 162–176. DOI: 10.26583/sv.13.1.11.

**Vorobeva Gulnara** — Ph.D., Dr.Sci., Professor of the department, Computational mathematics and cybernetics institute of mathematics, informatics and robotics, Ufa University of Science and Technology. Research interests: geoinformation and web technologies, systems of information storing and processing. The number of publications — 158. gulnara.vorobeva@gmail.com; 12, Karl Marx St., 450077, Ufa, Russia; office phone: +7(917)417-4111.

**Vorobev Andrei** — Ph.D., Dr.Sci., Associate Professor, Professor of the department, informatics department of institute of mathematics, informatics and robotics, Ufa University of Science and Technology; Researcher, Geophysical Center of RAS. Research interests: geoinformation technologies, digital signal processing. The number of publications — 172. geomagnet@list.ru; 12, Karl Marx St., 450077, Ufa, Russia; office phone: +7(917)345-2299.

**Orlov Gleb** — Graduate student, Department of computational mathematics and cybernetics, institute of mathematics, informatics and robotics, Ufa University of Science and Technology. Research interests: geoinformation and web technologies, systems for protecting software from unauthorized copying. The number of publications — 2. orlovgleb99@mail.ru; 12, Karl Marx St., 450077, Ufa, Russia; office phone: +7(919)145-5147.

**Acknowledgements.** The reported study was funded by RSF, project number 21-77-30010.

**Руководство для авторов**

Взаимодействие автора с редакцией осуществляется через личный кабинет на сайте журнала «Информатика и автоматизация» http://ia.spcras.ru/. При регистрации 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 pesoal 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 27.03.2024. Passed for print 01.04.2024.
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 27.03.2024. Дата выхода в свет 01.04.2024.
Формат 60×90 1/16. Усл. печ. л. 16,6. Заказ № 86. 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»