УДК 004.852 **DOI** 10.15622/ia.23.1.3

С.А. Давыденко, Е.Ю. Костюченко, С.Н. Новиков

ОЦЕНИВАНИЕ ИНФОРМАТИВНОСТИ ПРИЗНАКОВ В НАБОРАХ ДАННЫХ ДЛЯ ПРОВЕДЕНИЯ ПРОДЛЁННОЙ АУТЕНТИФИКАЦИИ

Давыденко С.А., Костюченко Е.Ю., Новиков С.Н. **Оценивание информативности признаков в наборах данных для проведения продлённой аутентификации.**

Аннотация. Продлённая аутентификация позволяет избавиться от недостатков, присущих статической аутентификации, например, идентификаторы могут быть потеряны или забыты, пользователь совершает только первоначальный вход в систему, что может быть опасно не только для областей, требующих обеспечения высокого уровня безопасности, но и для обычного офиса. Динамическая проверка пользователя во время всего сеанса работы может повысить безопасность системы, поскольку во время работы пользователь может подвергнуться воздействию со стороны злоумышленника (например, быть атакованным) или намеренно передать ему права. В таком случае оперировать машиной будет не пользователь, который выполнил первоначальный вход. Классификация пользователей во время работы системы позволит ограничить доступ к важным данным, которые могут быть получены злоумышленником. Во время исследования были изучены методы и наборы данных, использующихся для продлённой аутентификации. Затем был сделан выбор наборов данных, которые использовались в дальнейшем исследовании: данные о движении смартфона и смарт-часов (WISDM) и динамике активности мыши (Chao Shen's, DFL, Balabit). Помочь улучшить результаты работы моделей при классификации может предварительный отбор признаков, например, через оценивание их информативности. Уменьшение размерности признаков позволяет снизить требования к устройствам, которые будут использоваться при их обработке, повысить объём перебора значений параметров классификаторов при одинаковых временных затратах, тем самым потенциально повысить долю правильных ответов при классификации за счёт более полного перебора параметров значений. Для оценивания информативности использовались метод Шеннона, а также алгоритмы, встроенные в программы для анализа данных и машинного обучения (WEKA: Machine Learning Software и RapidMiner). В ходе исследования были выполнены расчёты информативности каждого признака в выбранных для исследования наборах данных, затем с помощью RapidMiner были проведены эксперименты по классификации пользователей с последовательным уменьшением количества используемых при классификации признаков с шагом в 20%. В результате была сформирована таблица с рекомендуемыми наборами признаков для каждого набора данных, а также построены графики зависимостей точности и времени работы различных моделей от количества используемых при классификации признаков.

Ключевые слова: информативность, классификация, продлённая аутентификация, машинное обучение, отбор признаков, информационная безопасность.

1. Введение. Аутентификация пользователей критически важна для компьютерных систем. В настоящее время наиболее популярные подходы к аутентификации – это методы, основанные на знании (пароли) и методы, основанные на владении (например, смарт-картой, токеном). При этом данные средства могут быть легко украдены, потеряны или забыты [1], пользователи зачастую используют простые

пароли, такие как «1234», свою фамилию или же используют один пароль для большого числа различных ресурсов. Для решения данных проблем существуют различные подходы, например биометрия [2]. Несмотря на это, всё ещё существует один критический недостаток: пользователь аутентифицируется только при первоначальном входе в систему и не аутентифицируется повторно до тех пор, пока не выйдет из системы или не пройдёт значительный интервал между его действиями на рабочей станции. Это представляет большую проблему не только для областей, в которых требуется высокая безопасность, но и в обычном офисе: ведь кто угодно может получить доступ к данным, с которыми работает пользователь, если он не вышел из системы на время перерыва. Уже установлено, что 29% атак на организации происходят по вине инсайдеров [3]. При этом сети, использующие концепцию интернета вещей для обмена информацией между устройствами (например, в «умном доме»), тоже требуют защиты от атак [4, 5, 6]. Чтобы уменьшить вероятность того, что злоумышленник сможет совершить какие-либо действия с устройства пользователя, используют продлённую аутентификацию. Под этим термином подразумевается, что личность человека, работающего на устройстве, постоянно проверяется. В настоящее время способы проведения продлённой аутентификации можно разделить на два вида: использующие поведенческие характеристики и использующие физические характеристики [7].

Исследования, направленные на изучение данной области, ставят перед собой следующие цели:

– поиск данных, при использовании которых будет достигнута высокая точность определения легитимности пользователя;

– упрощение процесса прохождения аутентификации пользователем, с сохранением точности определения.

Основной целью данной работы является сокращение времени работы моделей при решении задачи продлённой аутентификации путём поиска наиболее информативных признаков в нескольких наборах данных и изучение зависимостей доли правильных ответов при классификации пользователей и времени работы различных моделей от выбранного алгоритма и количества используемых наиболее информативных признаков. Объект исследования – процедура проведения продлённой аутентификации. Предмет исследования – изучение зависимости времени и точности работы различных моделей для классификации пользователя от количества

используемых информативных признаков для задачи продлённой аутентификации.

Основные задачи, поставленные в данном исследовании, включают в себя:

– изучение существующих методов и наборов данных для выполнения продлённой аутентификации;

– изучение существующих методов оценивания информативности признака;

– оценивание информативности признаков в выбранных наборах данных;

– изучение влияния количества используемых информативных признаков на точность и время работы различных моделей классификации.

Под информативностью в общем случае понимают совокупное количество информации, получаемое потребителем по пространственным, спектрально-энергетическим, временным и иным признакам при их восприятии и анализе [8]. Соответственно под оцениванием информативности понимается расчёт этого значения, а само это значение называется оценкой информативности.

Для измерения точности работы системы исследователи используют следующие метрики: FAR, FRR, EER [9], TPR, FPR [10], для которых можно выделить следующие ситуации: FP (false positive) – самозванец определён неверно, FN (false negative) – легитимный пользователь определён неверно, TN (true negative) – самозванец определён верно, TP (true positive) – легитимный пользователь определён верно. FAR (false acceptance rate) – показывает отношение FP к общему числу попыток нелегитимного пользователя попасть в систему, FRR (false rejection rate) – показывает отношение FN к общему числу попыток легитимного пользователя попасть в систему. На рисунке 1 показан способ расчёта метрики EER (equal error rate) – на оси абсцисс находится пороговое значение выходной метрики классификатора, а на оси ординат – значения FAR и FRR. Понимать данный график необходимо следующим образом: по мере увеличения порогового значения FAR уменьшается (при минимальном значении порога система впускает всех, в том числе всех нелегитимных пользователей, при максимальном не впускает никого), а FRR увеличивается (при минимальном значении порога система впускает всех легитимных пользователей, при максимальном не впускает никого). EER показывает в каком значении величина FAR совпадает с FRR. TPR и FPR вычисляются по формулам 1 – 2.

$$TPR = \frac{TP}{TP + FN}, \quad (1)$$

$$FPR = \frac{FP}{FP + TN}. \quad (2)$$

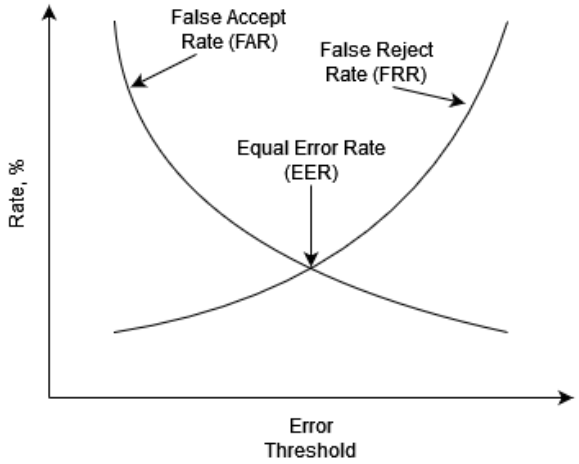

Рис. 1. Equal Error Rate

2. Подходы к выполнению продлённой аутентификации. На данный момент можно выделить следующие подходы: аутентификация по поведенческим признакам и аутентификация по физическим характеристикам, которые, в свою очередь, делятся на внешние признаки и признаки жизнедеятельности.

В статье [11] авторы рассматривают использование данных со смарт-часов пользователя для продлённой аутентификации. В плюсы они приводят то, что данные возникают естественным путём при взаимодействии с устройствами ввода (мышью, клавиатурой, сенсорным экраном), для их сбора не требуется активного вмешательства пользователя, а устройства, собирающие данные, по умолчанию оснащены необходимыми датчиками (гироскопом, акселерометром) и имеют подключение к интернету. Авторы использовали общедоступный набор данных WISDM, собранный во время использования смарт-часов и смартфонов для различных действий в течение трёх минут, но рассматривали они только активность при печатании. Всего в наборе данных содержится 90

высокоуровневых признаков, но исследователями также были проведены эксперименты, в которых использовались только 60 признаков. Авторы использовали 6 разных классификаторов и получили Avg. TP Rate в 77,3-87,2%, Avg. FP Rate в 0,2-0,5% и Avg. Precision в 78,2-87,9%, после чего пришли к следующим выводам: классификатор MLP справляется с задачей лучше всего, имея высокий TPR и низкий FPR, уменьшение признаков до 60 не сильно влияет на точность, при этом увеличивая скорость работы системы, данные акселерометра показывают более высокую точность, FPR остаётся ниже 1% в большинстве случаев, что говорит о высокой защищённости.

В статье [12] авторы тоже используют набор данных WISDM, но в этот раз исследование проводилось для всех типов деятельности, а не только печатания. Исследователи подтвердили, что поведенческие характеристики, получаемые с помощью акселерометра в телефоне или смарт-часах, хорошо справляются с задачей обеспечения продлённой аутентификации, такая система подходит для многофакторной аутентификации. При этом использованы три разных комбинации входных данных: только данные с акселерометра телефона (Ph Acell), данные с акселерометра часов и акселерометра телефона (All Acell) и данные с гироскопа и акселерометра как телефона, так и часов (All Sensors). Средняя точность идентификации пользователя при выполнении различных действий с помощью разных классификаторов составила от 59,3% (SVM по данным All Sensors) до 95,7% (Random Forest по данным Ph Accel). Расширив количество действий для классификации пользователя и используя только данные акселерометра для аутентификации, исследователи не только смогли добиться точностей выше, чем в предыдущей работе, но и расширить возможности выполнения продлённой аутентификации. При работе за компьютером человек может не только печатать, но и принимать пищу, активно шевелить руками, или просто сидеть спокойно; увеличение количества рассматриваемых активностей позволяет приблизить эксперименты по продлённой аутентификации с помощью данных со смарт-часов к реальности.

В то же время данные, получаемые с помощью мобильных устройств, предлагаются к использованию не только для продлённой аутентификации при работе за компьютером, но и при работе с самим смартфоном [13], поскольку в нём также могут содержаться очень важные данные, которые требуют защиты (например, банковские приложения). В качестве используемых данных для верификации

пользователя исследователи уже рассмотрели особенности походки пользователя [14, 15], клавиатурный почерк [14, 16, 17], данные с акселерометра [17 – 25], гироскопа [17 – 25], данные GPS [17, 21], магнитометра [17 – 20, 25], данные с сенсорного экрана [19, 24, 26], данные об использовании приложений [21, 27, 28], данные о нажатиях на экран (сила давления, место нажатия, длина нажатий) [21, 29] и датчики вращения [17]. Большое количество исследований по применению акселерометра и гироскопа различных устройств говорит о достаточно частом применении таких наборов данных, а также о том, что их получение не является нетривиальной задачей и может быть использовано не только на специфических предприятиях с высокими требованиями к безопасности, но быть доступным для обычных пользователей.

Одна из поведенческих характеристик человека при работе с компьютером – это динамика нажатия клавиш. Её плюсами является то, что для её измерения нет необходимости в использовании дополнительного оборудования, а пользователи используют клавиатуру как для работы, так и во время отдыха, поэтому у них не возникает дополнительного раздражения [30]. Подходы, использующие динамику нажатия клавиш, используются в статьях [31 – 38]. В частности, в статье [39] исследователи утверждают, что при нажатии клавиши генерируется три события: key_up (клавиша отпущена), key down (клавиша нажата), key_press (фактический символ вносится в текст); также они делят признаки на общие: частота появления ошибок при наборе текста (использования backspace, delete), использования клавиш управления (ctrl, alt), и общую скорость печатания; и временные: Interval (время между нажатием клавиш), Dwell time (продолжительность нажатия), а также метрики для двух последовательных нажатий клавиш: Latency (время между нажатием первой клавиши и моментом когда отпущена вторая), Flight time (время между нажатием первой клавиши и нажатием второй) и Up to Up (время между моментом, когда отпущена первая клавиша и когда отпущена вторая). В основе работы с динамикой нажатия клавиш лежит работа с N-граммами, в данном исследовании «схожие» N-граммы объединялись в кластеры, поиск оптимального количества кластеров (k) стоял как одна из основных задач перед авторами. При классификации входных данных X генерируется вектор вероятностей принадлежностей к классам. Обычно X определяется как принадлежащий к классу с наибольшей вероятностью, но исследователи приняли решение об использовании порогового значения (Pt), малое значение Pt увеличивает FAR и уменьшает FRR,

высокое – оказывает обратный эффект. Использованный набор данных был взят из исследования, проводившегося ранее [40], при этом авторами было принято решение о вынесении вердикта не после полной сессии, а после одной, двух и трёх четвертей сессий. FAR составил от 0,3% до 0,61%, а EER от 0,15% до 0,3%.

В статье [41] автор также рассматривает динамику нажатия клавиш. Он приходит к выводу, что такие системы можно рассматривать как кейлоггер, поэтому для того, чтобы их использовать необходимо чтобы пользователь полностью доверял разработчикам системы, ведь они могут отслеживать его сообщения, пароли иные важные и секретные данные. Другая часто используемая для продлённой аутентификации поведенческая характеристика – активность мыши. Такой подход рассматривается в статьях [42 – 46]. В частности, в статье [47] авторы собрали информацию об использовании мыши студентами с конца ноября по начало декабря 2014 года. Для сбора данных о курсоре использовался встроенный в ОС Windows GetCursorPos() из заголовочного файла winuser.h, отслеживающий его положение в определённые моменты времени. Помимо данных о курсоре также собиралась информация о состоянии клавиш мыши (нажаты или нет), данные записывались каждые 15,6 мс. Всего было собрано 76500 образцов данных для 45 пользователей. Для составления профиля пользователя использовались следующие признаки: длина правого и левого клика, длина двойного левого клика, скорость, ускорение и резкость движения (изменение ускорения с течением времени). Исследователи разбили полученные данные каждого пользователя на 2 набора, один используется для того, чтобы получить профиль пользователя, другой используется для тестирования. Оценка соответствия профиля пользователя и тестовых данных – это отношение принятых системой признаков к общему числу полученных признаков. По результатам исследования авторы получили значения EER, самое низкое из которых: 6,7%, а среднее: 13,19%. Следует отметить, что под R values исследователи подразумевали пороговое значение для того, чтобы уменьшить количество шума в полученных данных, но при этом сохранить уникальность показателей, а под M values – размах диапазона в стандартных отклонениях, чем меньше M, тем сложнее успешно пройти аутентификацию.

Большое количество исследований, посвящённых динамике нажатия на клавиши и динамике движения мыши, говорит о высокой востребованности и применимости такого типа данных для продлённой аутентификации. Широкое распространение устройств,

используемых для получения информации о пользователе для составления его профиля, позволяет упростить сбор данных, а также применять модели для классификации пользователей почти на любом компьютере.

В статье [48] авторами был рассмотрен подход по использованию лингвистического стиля текста. Для своего исследования авторы использовали образцы текста из CASIS (Center for Advanced Studies in Identity Science): 4000 образцов текста от 1000 пользователей, не являющихся носителями английского языка. В каждом образце в среднем около 1600 символов, 300 слов и 13 предложений, оценивание производилось для блоков с размерами в 50 и 100 символов. Текст характеризуется признаками, происходящими из шести категорий: символьные (N-граммы), лексические (отличия на уровне слова), синтаксические (структура на уровне предложения), семантические (смысловое значение), структурные (организация документа, например отступы) и предметно-специфические. В качестве классификатора авторы используют алгоритм изоляционного леса с 50 деревьями. По графику TPR, приведённому исследователями видно, что с увеличением числа образцов TPR увеличивается, при этом уже на 5 образцах достигаются значения в 99,35% и 98,5% соответственно. Авторы делают вывод, что масштабируемость такого подхода, а также малый шанс ошибочной верификации (меньше 1%) говорят о высокой актуальности такого решения. При этом они отмечают, что такая система имеет смысл в ситуациях с одним пользователем, но может быть непрактичной, если в системе несколько участников (например, несколько авторов, работающих над одной статьёй), также масштабируемость ограничивается другими языками. Следует отметить, что TPR является не самой лучшей метрикой для измерения качества классификации, поскольку добиться 100% TPR можно путём простого трактования всех пользователей как легитимных. Для более объективной оценки статьи не хватает дополнительных метрик, показывающих иные параметры работы построенной модели. Профиль применения такого набора данных для продлённой аутентификации весьма ограничен, а данные, необходимые для построения профиля пользователя, сложны для сбора. Данные недостатки не позволяют говорить о возможности широкого использования моделей, использующих такие наборы данных.

В статье [49] авторы предлагают использовать «мягкую» биометрию (цвет одежды и цвет кожи пользователя) для выполнения

продлённой аутентификации. В плюсы данного метода выдвигается следующее: пользователю не требуется вводить никаких данных, в том числе биометрических, также нет необходимости в предварительной регистрации признаков, они регистрируются автоматически при каждом входе пользователя в систему. Ключевым отличием подхода авторов статьи от исследований, использующих схожие данные (лицо пользователя) является то, что перед системой ставится не задача идентификации пользователя, а проверка того, является ли он тем же человеком, что совершил первоначальный вход в систему. На рисунке 5 показана начальная регистрация данных пользователя, для работы с «мягкой» биометрией используются цветовые гистограммы, а для работы с лицом пользователя – каскады Хаара. При этом система учитывает изменение освещения в помещении, в такой ситуации данные сессии пользователя будут обновлены. Используя «мягкую» биометрию в дополнение к фотографии лица пользователя, авторы смогли упростить процесс продлённой аутентификации. Система показала свою работоспособность при различных действиях пользователя перед монитором. При сборе данных для модели не используется сложных устройств, многие ноутбуки оснащены веб-камерой по умолчанию поэтому применимость такого способа выполнения аутентификации достаточно высока.

В статье [50] авторы используют устройства отслеживания направления взгляда для анализа радужки глаза. По итогу исследования они получили EER в 9% и пришли к выводу, что для единственного средства аутентификации данный способ не подходит, но он может быть использован как дополнительное средство.

В статье [51] авторы предлагают использовать систему продлённой аутентификации на основе сердцебиения. Для получения данных используется доплеровский радиолокационный датчик. Плюсы данного подхода следующие: эта характеристика уникальна (различима для разных субъектов), измерима (тяжело скрыть), непроизвольна (неизвестна пользователю), безопасна (трудно подделать) и есть у любого живого человека. Также на её результаты не влияют шумы, которые присутствуют в методах, использующих камеру. Исследователи сообщают, что добились BAC (balanced accuracy) в 98,61% и EER в 4,42%. Авторы использовали данные 78 человек, для эксперимента выбирались люди, не обладающие сердечными заболеваниями, для каждого было собрано 20 образцов по 8-10 сердечных циклов. За признаки были взяты внутренние

геометрические дескрипторы (амплитуда, площадь, угол), полученные из реперных точек.

В статье [52] исследователи использовали электрокардиограммы для идентификации и верификации пользователя, в работе использовалось 5 открытых наборов данных и один, созданный исследователями. Данные получают с помощью сенсоров, прикрепляемых к телу человека для снятия сигнала, или с применением специального устройства, позволяющего получать данные ЭКГ без предварительной подготовки человека к снятию данных. Авторы смогли добиться 100% точности распознавания с EER, варьирующимся от 0,5% до 1,8%. Для извлечения данных об активности сердца также можно использовать фитнесс-трекеры, что было продемонстрировано исследователями в статье [53]. Поскольку в статьях [50 – 52] для сбора данных о пользователе используются специальные устройства, которые не применяются при ежедневной работе за компьютером, применимость такого способа выполнения продлённой аутентификации в реальных условиях достаточно мала.

В статье [54] авторы рассматривают использование голоса для продлённой аутентификации во время выполнения звонков. Исследователи получали запись звуков через VoxGuard, в дальнейшем извлекались признаки с помощью скрытых Марковских моделей. Каждая модель представляет собой фонему, всего их было получено 41, для работы с ними использовался распознаватель фонем, который показал не очень высокую точность (около 30-50%). Верификация производится следующим образом: запись голоса делится на сегменты, затем сегмент разбивается на фонемы. Фонемы и первоначальная запись голоса передаются системе VoxGuard, она выносит оценку схожести. Такая операция выполняется для каждого сегмента записи после чего выносится общий вердикт. В итоге исследователями была получена EER в 15%. Микрофоны – достаточно широко распространённое устройство, зачастую встроенное в ноутбуки, но при этом для получения данных о голосе необходимо говорить, в то время как при ежедневной работе за компьютером пользователь, как правило, соблюдает тишину и ни с кем не разговаривает.

В статье [55] рассматривается использование паттернов дыхания для выполнения продлённой аутентификации. В плюсы приводятся следующие аргументы: не требуется каких-либо дополнительных действий от пользователя, поскольку дышит человек автоматически, данная система может быть использована

для любых устройств, работающих с Wi-Fi, например в умном доме. Wi-Fi используется во многих корпоративных и домашних сетях, поэтому применение такого подхода может быть весьма практичным, главной проблемой будет сбор данных о дыхании пользователя через данные каналы обмена информацией. Дыхание каждого человека уникально, оно содержит следующие признаки: глубину, ритм вдохов/выдохов, частоту. Для сбора данных о дыхании пользователей используется информация о состоянии канала Wi-Fi, которая позволяет получать информацию о жизнедеятельности человека [56], на основании эти данных строится профиль пользователя. Также авторы используют нейронную сеть для определения личности пользователя и собирают обратную связь от пользователей (в случае изменения паттерна дыхания, связанного, например, с болезнью).

Исследователи использовали данные 20 человек (14 мужчин и 6 женщин), для каждого из них было собрано около 200-300 образцов дыхания во время сидения на кресле на протяжении четырёх месяцев. Для оценивания использовались следующие метрики: Authentication Success Rate (процент успешной аутентификации), FPR, Spoofing Detection Rate (процент атак, которые отмечены как атака), Receiver Operating Characteristic (показывает компромисс между частотой FPR и SDR при различных пороговых значениях). В результате экспериментов были получены следующие результаты: средний ASR около 90%, SDR – 92,14% с FPR около 5%. Авторы отмечают влияние размеров обучающей выборки и скорости передачи данных на конечную точность системы.

В настоящей работе были изучены существующие подходы к проведению продлённой аутентификации для компьютеров и смартфонов, а также рассмотрены используемые исследователями входные данные для классификации. В таблице 1 приведены краткие итоги исследования.

Рассмотренные исследования не содержат информации об оценивании информативности признаков для классификации, поэтому можно предположить, что авторы не использовали этот метод для отбора признаков. Для изучения зависимости доли правильных ответов и времени работы классификаторов от количества используемых признаков, было принято решение предварительно оценить информативность каждого признака на примере наборов данных для продлённой аутентификации.

Таблица 1. Итоги исследования существующих типов данных для продлённой аутентификации

Тип данных	Устройство	Используемый канал связи
Движения во время набора текста, полученные через смартфон/смарт-часы	Смартфон/смарт-часы	Электронная почта, бухгалтерский учёт
Клавиатурный почерк	Клавиатура, смартфон	Электронная почта, бухгалтерский учёт
Динамика активности мыши	Мышь	Любая работа за ПЭВМ
Лицо	Камера	Сеансы видеосвязи
Отпечаток пальца	Сканер отпечатка пальца	Любая работа за ПЭВМ
Дыхание	Wi-Fi роутеры	Любая работа за ПЭВМ
Радужка глаза	Отслеживатель глаз, сканер радужки, камера	Любая работа за ПЭВМ
Сердечная активность	Доплеровский радиолокационный датчик, устройства для получения электрокардиограмм	Любая работа за ПЭВМ
Голос	Микрофон	Сеансы видеосвязи, голосовой связи
Мягкая биометрия	Камера	Сеансы видеосвязи
Лингвистический анализ	Специальных устройств не требуется	Электронная почта, бухгалтерский учёт
Движения при использовании мобильного устройства	Смартфон	Работа с приложениями на смартфоне
Динамика использования приложений	Смартфон	Работа с приложениями на смартфоне
Динамика использования сенсорного экрана	Мобильный телефон с сенсорным экраном	Работа с приложениями на смартфоне

2. Оценивание информативности. После изучения типов наборов данных, представленных в разделе 1, было принято решение использовать поведенческие характеристики для дальнейших исследований. Обосновано это следующими факторами: по поведенческим признакам больше исследований, а следовательно, и больше источников данных, в то время как для внешних признаков и признаков жизнедеятельности исследователи, как правило, собирают данные самостоятельно и в открытый доступ их потом не выкладывают. Следует отметить, что при поиске следует ориентироваться на следующие параметры: открытость набора данных и наличие в нём выделенных признаков. В первую очередь был выполнен поиск наборов данных, которые работают с динамикой нажатия клавиш. Исключая те

исследования, в которых авторы собирали данные самостоятельно, в основном используют следующие наборы данных:

– Clarkson University Dataset (требует запроса) [57 – 60];
– Buffalo Dataset (требует запроса) [58, 60, 61];
– University of Victoria (требует запроса) [62];
– Keystroke and Mouse Dynamics for UEBA Dataset (нет признаков) [63];
– The WOLF of SUDT (требует запроса) [64];
– BB-MAS (нет признаков) [65].

Также помимо этих наборов данных несколько исследователей упоминают об использовании в своих работах собранные ими наборы данных:

– Free vs. transcribed text for keystroke-dynamics evaluations (нет признаков) [66];
– On continuous user authentication via typing behavior (рассматривают видеозаписи набора текста) [67];
– Shared Data Set for Free-Text Keystroke Dynamics Authentication Algorithms. (нет признаков) [68];
– Identity verification through dynamic keystroke analysis (нет в открытом доступе) [40, 69];
– Keystroke patterns as prosody in digital writings: A case study with deceptive reviews and essays (нет признаков) [70];
– Shared Dataset on Natural Human-Computer Interaction to Support Continuous Authentication Research (нет в открытом доступе) [71].

Поскольку ни один из перечисленных ранее наборов данных не удовлетворяет вышеобозначенным условиям, был выполнен поиск наборов данных, осуществляющих продлённую аутентификацию с использованием активности мыши. Найдены были исследования, использующие следующие наборы данных:

– Bogazici mouse dynamics dataset (нет в открытом доступе) [46, 72];
– Chao Shen’s [43, 73, 74];
– Balabit [75 – 77];
– DFL [78].

Из данного списка Balabit, DFL и Chao Shen подходят для дальнейшего исследования, поскольку они открыты и в них уже выделены признаки. Также, помимо этих наборов данных, было принято решение использовать набор данных WISDM. Этот набор данных тоже соответствует обозначенным критериям, но требует предварительной обработки, поскольку рассматривает не только

работу за компьютером, но и другие действия, которые может выполнять пользователь. Данные из WISDM [79] были обработаны следующим образом: для одного устройства (телефона или смарт-часов) выбирался один тип датчика (гироскоп или акселерометр), после чего данные всех пользователей, измеренные этими устройством и датчиком, объединялись в один файл. Затем из данного файла удалялись все активности помимо набора текста.

Для оценивания информативности [80, 81] использовались следующие средства: WEKA machine learning tool [82], RapidMiner Studio [83] и собственноручно написанная программа на языке python для расчёта информативности признака по методу Шеннона (формулы 6 – 8). Такой выбор программ для расчётов связан в первую очередь с тем, что в отличие от аналогов они предоставляют не только алгоритмы, позволяющие проводить предварительную обработку признаков перед классификацией, но и данные об информативности, рассчитанные во время этой обработки. Это позволяет выполнять сортировку признаков по данному параметру (по возрастанию или убыванию). Из рассмотренных аналогов в Orange доступен инструмент предобработки, в том числе с отбором признаков по информативности, но получить список признаков и соответствующей им информативности нельзя, а в инструменте Knime плагин InformativityGainCalculator – платный. В связи с данными ограничениями, а также с тем, что WEKA и RapidMiner бесплатны, они были выбраны в качестве основных инструментов, используемых в исследовании. При этом RapidMiner поддерживает как собственный алгоритм расчёта информативности признака, так и алгоритм из библиотеки WEKA (функция weka.information_gain_ratio), также это средство позволяет выполнять классификацию объектов с заданием определённых признаков, поэтому дальнейшие эксперименты, не связанные с методом Шеннона, будут выполняться в этой программе.

Для расчёта информативности признака с помощью алгоритмов WEKA используются следующие формулы:

$$I = \frac{((H(class) - H(class \mid attribute))}{H(attribute)}, \tag{3}$$

$$H(x) = -\sum(P_i * \ln P_i), \tag{4}$$

где I – информативность признака, а $H(x)$ – энтропия, P – вероятность появления примера, принадлежащего к классу i.

Для расчёта H(class | attribute) разделим набор данных на части с одинаковыми значениями признаков (или принадлежащих одинаковым интервалам значений – для непрерывных признаков, не поддающихся прямой группировке по отдельным значениям). Для каждого значения или интервала значений признака проводим расчёт условной энтропии, а итоговую энтропию находим как сумму соответствующих условных энтропий. В RapidMiner для расчёта информативности используются схожая формула, но вместо натурального логарифма используется логарифм по основанию 2.

Метод Шеннона позволяет оценивать информативность как средневзвешенное количество информации, приходящееся на различные градации признака. Под количеством информации в теории информации понимают величину устраненной энтропии [84]. Следует отметить, что метод Шеннона даёт оценку информативности в виде нормированной величины, варьирующейся от 0 до 1. Об информативности в таком случае говорят следующее: чем *I(x)* ближе к 1 тем выше информативность x, и наоборот, чем ближе *I(x)* к нулю, тем информативность ниже [85]. Рассчитывается информативность признака *x* по следующей формуле:

$$I(x) = 1 + \sum_{i=1}^{G} (P_i * \sum_{k=1}^{K} P_{i,k} * \log_K P_{i,k}) , \quad (5)$$

где G – число градаций признака, K – количество классов, P_i – вероятность попадания значения признака в i-ю градацию (формула 2), $P_{i,k}$ – вероятность появления i-ой градации признака в k-ом классе (формула 3).

Для нахождения P_i используется следующая формула:

$$P_i = \frac{\sum_{k=1}^{K} m_{i,k}}{N} , \quad (6)$$

где N – общее число наблюдений признака, $m_{i,k}$ – частота появления i-ой градации признака в k-ом классе.

Для нахождения $P_{i,k}$ используется следующая формула:

$$P_{i,k} = \frac{m_{i,k}}{\sum_{k=1}^{K} m_{i,k}} . \quad (7)$$

Пример данных об информативности признаков в наборе данных Balabit, рассчитанных с использованием трёх методов представлены в таблице 2.

Таблица 2. Сравнение информативности признаков в наборе данных Balabit

Признак	Способ расчёта		
	shannon	weka	rapid miner
a_beg_time	0,000623	0,25805	0,32903
direction_of_movement	0,00543	0,00000	0,00455
dist_end_to_end_line	0,003989	0,03265	0,22556
elapsed_time	0,001628	0,04777	0,24204
largest_deviation	0,012828	0,06026	0,25286
max_a	0,008029	0,04548	0,31813
max_curv	0,011305	0,05192	0,22556
max_jerk	0,009352	0,07502	0,31320
max_omega	0,022478	0,13271	0,27680
max_v	0,006949	0,04698	0,31320
max_vx	0,00344	0,06239	0,27384
max_vy	0,001489	0,07896	0,24204
mean_a	0,012486	0,08390	0,26115
mean_curv	0,001805	0,02117	0,22556
mean_jerk	0,007407	0,07900	0,22556
mean_omega	0,002935	0,02644	0,22556
mean_v	0,001375	0,03230	0,22556
mean_vx	0,00085	0,01892	0,26798
mean_vy	0,000339	0,01419	0,22556
min_a	0,015152	0,06077	0,25286
min_curv	0,001036	0,06454	0,22556
min_jerk	0,001937	0,08974	0,23720
min_omega	0,034644	0,12979	0,29641
min_v	0,005781	0,04389	0,25286
min_vx	0,00282	0,06531	0,31059
min_vy	0,00523	0,09105	0,22556
num_critical_points	0,004077	0,08430	0,22556
num_points	0,00206	0,07257	0,22556
sd_a	0,005084	0,04127	0,27384
sd_curv	0,001067	0,03506	0,25286
sd_jerk	0,002005	0,08441	0,27384
sd_omega	0,005857	0,07404	0,24204
sd_v	0,005983	0,04235	0,24204
sd_vx	0,001944	0,03298	0,26637
sd_vy	0,003355	0,03670	0,22556
straightness	0,00699	0,03250	0,16915
sum_of_angles	0,00332	0,14158	0,33866
traveled_distance_pixel	0,007664	0,03891	0,24204
type_of_action	0,008792	0,02382	0,02184

Для данной таблицы был проведён корреляционный анализ, который показал наличие значимой корреляции между способами оценки информативности WEKA и RapidMiner (коэффициент корреляции Спирмена 0,494 статистически значим при уровне значимости 0,01) и отсутствие статистически значимой корреляции для метода Шеннона (наибольший коэффициент корреляции Спирмена 0,241 статистически не значим при уровне значимости 0,01). Это говорит об отсутствии полной согласованности между рассматриваемыми критериями информативности и необходимости дальнейшего рассмотрения каждого из них.

3. Оценка влияния отобранных информативных признаков на качество работы системы. Следующий этап исследования – проверка зависимости доли правильных ответов при классификации и времени работы различных моделей классификации от количества используемых признаков. Для этого с помощью RapidMiner будет поочерёдно выполняться классификация с разным количеством используемых признаков: сперва будут использоваться 100%, затем 80% самых информативных, после этого 60% и так далее. Для классификации с помощью RapidMiner использовался инструмент Auto Model, он упрощает расчёты долей правильных ответов при работе разных классификаторов на определённом наборе данных, при этом происходит автоматическое разделение данных на обучающую (60%) и валидационную (40%) выборки. В процессе выполнения работы были задействованы все модели, доступные через Auto Model: Naïve Bayes, Generalized Linear Model, Logistic Regression, Fast Large Margin, Deep Learning, Decision Tree, Random Forest, Gradient Boosted Trees, Support Vector Machine. Каждый классификатор выполнял многоклассовую классификацию, рассматривая каждого пользователя как отдельный класс.

После измерения показателей моделей с разным числом признаков полученные ими данные были экспортированы и по ним были построены графики. Примеры графиков зависимости доли правильных ответов моделей от количества признаков показаны на рисунках 2 – 4.

На каждом графике указан инструмент, использованный для расчёта информативности, на оси абсцисс показано количество признаков, использованное при распознавании, а на оси ординат – итоговая доля правильных ответов для модели. Каждая линия на графике обозначает соответствующую модель и показывает зависимость доли правильных ответов при работе модели от количества наиболее информативных признаков, использованных при распознавании.

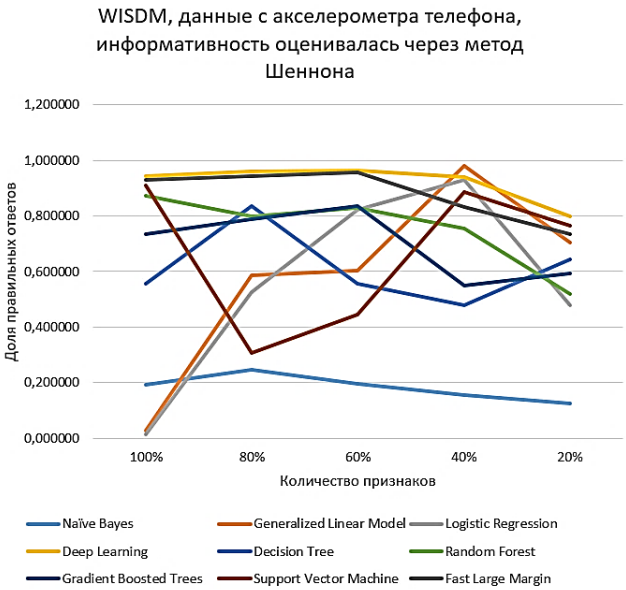

Рис. 2. Оценивание информативности через метод Шеннона, график точности

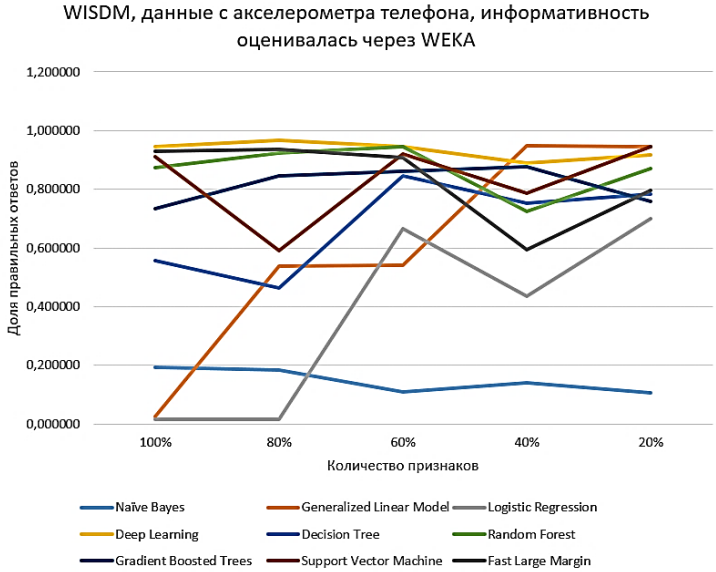

Рис. 3. Оценивание информативности через расчёты в WEKA, график точности

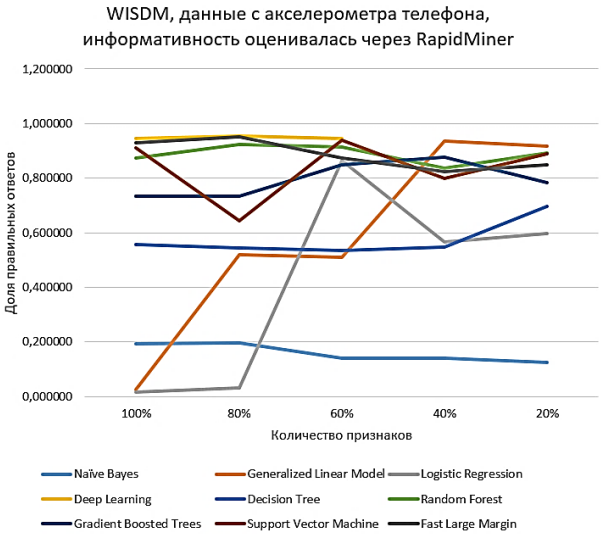

Рис. 4. Оценивание информативности через расчёты в RapidMiner, график точности

Примеры графиков зависимости времени работы моделей от количества признаков показаны на рисунках 5 – 7.

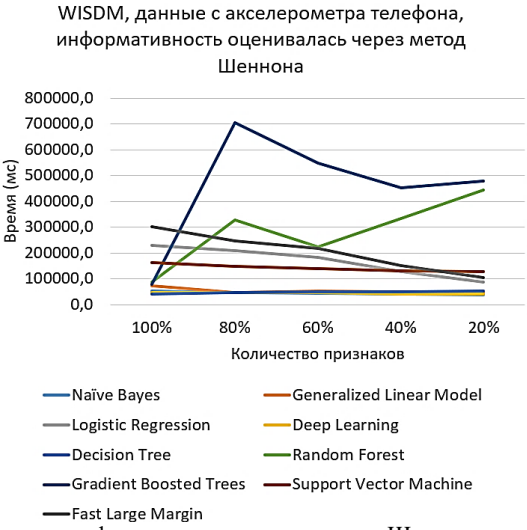

Рис. 5. Оценивание информативности через метод Шеннона, график времени обработки

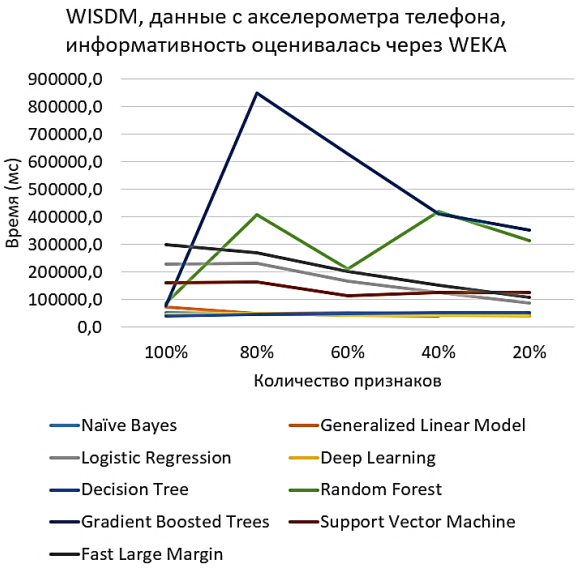

Рис. 6. Оценивание информативности через расчёты в WEKA, график времени обработки

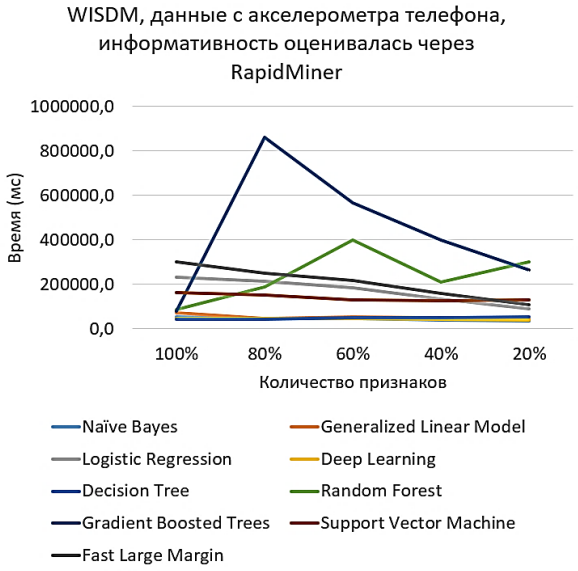

Рис. 7. Оценивание информативности через расчёты в RapidMiner, график времени обработки

Исходя из полученных графиков, можно сделать следующие выводы:

– некоторые модели обладают стабильно высокой точностью, на которую в большинстве случаев не сильно влияет изменение количества признаков (Deep Learning, Generalized Linear Model), а некоторые, наоборот, стабильно низкой (Naive Bayes, Decision Tree);

– при уменьшении числа используемых для классификации признаков у большинства моделей уменьшается время обработки, исключением являются Gradient Boosted Trees и Random Forest при обработке признаков набора данных WISDM;

– при классификации на основе данных с акселерометра смарт часов и телефона, а также гироскопа телефона не имеет смысла отбрасывать слишком много неинформативных признаков, поскольку у большинства моделей понижается точность. Особенно это заметно при понижении с 40% признаков до 20% (рисунок 2);

– уменьшение количества признаков для наборов данных, выполняющих классификацию пользователей по активности мыши, в целом сказывается положительно и в большинстве случаев позволяет увеличить низкую изначальную точность.

Исходя из полученных в результате экспериментов данных, были выбраны наборы признаков, которые рекомендуются для достижения наибольшей точности классификации. Данные рекомендации приведены в таблице 3.

Таблица 3. Рекомендуемые наборы признаков для наборов данных

Набор данных	Рекомендуемый набор признаков
WISDM, акселерометр телефона	40% самых информативных признаков, информативность рассчитывать через метод Шеннона
WISDM, акселерометр смарт-часов	40% самых информативных признаков, информативность рассчитывать через метод Шеннона
WISDM, гироскоп телефона	80% самых информативных признаков, информативность рассчитывать через RapidMiner
WISDM, гироскоп смарт-часов	80% самых информативных признаков, информативность рассчитывать через WEKA
Balabit	80% самых информативных признаков, информативность рассчитывать через RapidMiner
Chao Shen	60% самых информативных признаков, информативность рассчитывать через метод Шеннона
DFL	20% самых информативных признаков, информативность рассчитывать через метод Шеннона

В таблице 4 приведено сравнение метрик, полученных в ходе исследования и метрик из анализируемых источников. Следует отметить, что низкие точности у наборов данных Balabit, Chao Shen и DFL могут быть связаны с тем, что в исследованиях, посвящённых им, рассматривалась бинарная классификация, при которой и были получены такие результаты, в то время как в данной работе производилась многоклассовая классификация (каждый пользователь является отдельным классом). Данный подход также применяется в исследованиях, посвящённых продлённой аутентификации как для сравнения с бинарной [86 – 89] так и как отдельный способ идентификации пользователя [11, 90, 91].

В [86] при использовании бинарной аутентификации средний EER составил 14,7%, в то время как при использовании многоклассовой классификации он достиг 24,9%. В [87] при использовании многоклассовой аутентификации достигли F меры в 82%, а при использовании бинарной классификации 92%. В [88] с помощью нейронной сети ANN, разработанной авторами удалось добиться максимальной точности на тестовой выборке в 92,4% на 40 классах. В [89] бинарная аутентификация показала EER на 1,01% меньше чем многоклассовая. При этом авторы отмечают, что хотя продуктивность моделей и будет выше, сбор данных для расширения моделей, использующих бинарную аутентификацию в разы сложнее чем для многоклассовой классификации, поскольку требует индивидуальной модели для каждого пользователя. В [11] при использовании многоклассового классификатора в системе с 49 пользователями, исследователи достигли TPR в 80,7-87,2% для разных типов считывающих устройств. В [90] при использовании многоклассовой аутентификации авторы достигли точности в 91,72-98%. В [91] с использованием многоклассовой классификации авторы достигли среднего значения EER в 3,96%. Рассмотренные исследования показывают, что многоклассовая аутентификация, хотя и обладает недостатками по сравнению с бинарной классификацией, может быть успешно применима для задачи продлённой аутентификации (особенно на наборе данных с малым количеством пользователей). Таким образом, использование многоклассовой классификации позволяет увеличить количество исследований, с которыми могут быть сравнены результаты работы.

Таблица 4. Сравнение метрик

Набор данных	Данные, полученные в результате данного исследования	Данные из анализируемых источников
WISDM (акселерометр телефона)	Accuracy: Random forest: 51,7-94,3% SVM: 30,6-94,6%	Accuracy: Random forest: 98,8% SVM: 94,5%
Balabit	Accuracy: Random forest: 46,12%%	Accuracy: Avg. Random forest: 72,29% [75] Avg. Random Forest: 79,7% [77]
DFL	AUC: 0.928	AUC: 0.99
Chao Shen	FAR: 70,6% FRR: 70,6%	FAR: 0,37-44,65% FRR: 1,12-34,78%

4. Заключение. В ходе проведённого исследования было выполнено оценивание информативности в доступных для исследования наборах данных для выполнения продлённой аутентификации, экспериментально были проверены зависимость точности и времени работы различных моделей для классификации пользователя от количества используемых признаков, построены графики этих зависимостей. Получены рекомендации по выбору информативных признаков, позволяющих сократить время построения и работы моделей без существенного снижения метрики качества их работы. Все задачи, поставленные в начале работы были успешно выполнены. В результате изучения построенных графиков была составлена таблица рекомендованных наборов признаков для каждого набора данных, рассмотренного в ходе исследования, а также выполнено сравнение результатов без отбора признаков и с отбором признаков с помощью оценивания информативности (таблица 4). Уменьшение количества признаков за счёт использования наиболее значимых позволит уменьшить время обучения моделей (в зависимости от алгоритма классификации и набора данных удалось сократить время работы от 5 до 82%, в среднем максимальный прирост скорости работы модели составил 40%), затраты на расширение наборов данных за счёт сокращения снимаемых признаков, а также потенциально повысить точность работы моделей при выполнении продлённой аутентификации и классификации. Перспективы дальнейшего исследования лежат в применении рассмотренных методов и инструментов для сокращения количества используемых признаков в исследованиях, также посвящённых

продлённой аутентификации для других факторов, например, аутентификации по динамике движения зрачков.

Литература

1. Jain A.K., Ross A., Pankanti S. Biometrics: a tool for information security // IEEE transactions on information forensics and security. 2006. vol. 1. no. 2. pp. 125–143.
2. Jain A.K., Ross A., Prabhakar S. An introduction to biometric recognition // IEEE Transactions on circuits and systems for video technology. 2004. vol. 14. no. 1. pp. 4–20.
3. Zhang N., Yu W., Fu X., Das S.K. Maintaining defender's reputation in anomaly detection against insider attacks // IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics). 2009. vol. 40. no. 3. pp. 597–611.
4. Liang Y., Samtani S., Guo B., Yu Z. Behavioral biometrics for continuous authentication in the internet-of-things era: An artificial intelligence perspective // IEEE Internet of Things Journal. 2020. vol. 7. no. 9. pp. 9128–9143.
5. Al-Naji F.H., Zagrouba R. CAB-IoT: Continuous authentication architecture based on Blockchain for internet of things // Journal of King Saud University-Computer and Information Sciences. 2022. vol. 34. no. 6. pp. 2497–2514.
6. Ashibani Y., Kauling D., Mahmoud Q.H. Design and implementation of a contextual-based continuous authentication framework for smart homes // Applied System Innovation. 2019. vol. 2(1). DOI: 10.3390/asi2010004.
7. Oak R. A literature survey on authentication using Behavioural biometric techniques // Intelligent Computing and Information and Communication: Proceedings of 2nd International Conference, ICICC. 2018. pp. 173–181.
8. Бондаренко М.А., Дрынкин В.Н. Оценка информативности комбинированных изображений в мультиспектральных системах технического зрения // Программные системы и вычислительные методы. 2016. Т. 1. С. 64–79.
9. Soltane M., Bakhti M. Multi-modal biometric authentications: concept issues and applications strategies // International Journal of Advanced Science and Technology. 2012. vol. 48. pp. 23–60.
10. Hong C.S., Oh T.G. TPR-TNR plot for confusion matrix // Communications for Statistical Applications and Methods. 2021. vol. 28. no. 2. pp. 161–169.
11. Rahman K.A., Alam N., Musarrat J., Madarapu A., Hossain M.S. Smartwatch Dynamics: A Novel Modality and Solution to Attacks on Cyber-behavioral Biometrics for Continuous Verification? // International Symposium on Networks, Computers and Communications (ISNCC). 2020. pp. 1–5.
12. Verma A., Moghaddam V., Anwar A. Data-driven behavioural biometrics for continuous and adaptive user verification using Smartphone and Smartwatch // Sustainability. 2022. vol. 14(12). DOI: 10.3390/su14127362.
13. Abuhamad M., Abusnaina A., Nyang D., Mohaisen D. Sensor-based continuous authentication of smartphones' users using behavioral biometrics: A contemporary survey // IEEE Internet of Things Journal. 2020. vol. 8. no. 1. pp. 65–84.
14. Lamiche I., Bin G., Jing Y., Yu Z., Hadid A. A continuous smartphone authentication method based on gait patterns and keystroke dynamics // Journal of Ambient Intelligence and Humanized Computing. 2019. vol. 10. pp. 4417–4430.
15. Giorgi G., Saracino A., Martinelli F. Using recurrent neural networks for continuous authentication through gait analysis // Pattern Recognition Letters. 2021. vol. 147.pp. 157–163.
16. Kim D.I., Lee S., Shin J.S. A new feature scoring method in keystroke dynamics-based user authentications // IEEE Access. 2020. vol. 8. pp. 27901–27914.

17. Deb D., Ross A., Jain A.K., Prakah-Asante K., Prasad K.V. Actions speak louder than (pass) words: Passive authentication of smartphone* users via deep temporal features // International conference on biometrics (ICB). 2019. pp. 1–8.
18. Abuhamad M., Abuhmed T., Mohaisen D., Nyang D. AUToSen: Deep-learning-based implicit continuous authentication using smartphone sensors // IEEE Internet of Things Journal. 2020. vol. 7. no. 6. pp. 5008–5020.
19. Barlas Y., Basar O.E., Akan Y., Isbilen M., Alptekin G.I., Incel O.D. DAKOTA: Continuous authentication with behavioral biometrics in a mobile banking application // 5th International Conference on Computer Science and Engineering (UBMK). 2020. pp. 1–6.
20. Mekruksavanich S., Jitpattanakul A. Deep learning approaches for continuous authentication based on activity patterns using mobile sensing // Sensors. 2021. vol. 21. no. 22. DOI: 10.3390/s21227519.
21. Acien A., Morales A., Vera-Rodriguez R., Fierrez J., Tolosana R.. Multilock: Mobile active authentication based on multiple biometric and behavioral patterns // 1st International Workshop on Multimodal Understanding and Learning for Embodied Applications. 2019. pp. 53–59.
22. Li Y., Hu H., Zhu Z., Zhou G. SCANet: sensor-based continuous authentication with two-stream convolutional neural networks // ACM Transactions on Sensor Networks (TOSN). 2020. vol. 16. no. 3. pp. 1–27.
23. Mekruksavanich S., Jitpattanakul A. Deep convolutional neural network with rnns for complex activity recognition using wrist-worn wearable sensor data // Electronics. 2021. vol. 10. no. 14. DOI: 10.3390/electronics10141685.
24. Volaka H.C., Alptekin G., Basar O.E., Isbilen M., Incel O.D. Towards continuous authentication on mobile phones using deep learning models // Procedia Computer Science. 2019. vol. 155. pp. 177–184.
25. Li Y., Zou B., Deng S., Zhou G. Using feature fusion strategies in continuous authentication on smartphones // IEEE Internet Computing. 2020. vol. 24. no. 2. pp. 49–56.
26. Incel O.D., Gunay S., Akan Y., Barlas Y., Basar O.E., Alptekin G.I., Isbilen M. Dakota: sensor and touch screen-based continuous authentication on a mobile banking application // IEEE Access. 2021. vol. 9. pp. 38943–38960.
27. Alotaibi S., Alruban A., Furnell S., Clarke N. A Novel Behaviour Profiling Approach to Continuous Authentication for Mobile Applications // ICISSP. 2019. pp. 246–251.
28. Mahbub U., Komulainen J., Ferreira D., Chellappa R. Continuous authentication of smartphones based on application usage // IEEE Transactions on Biometrics, Behavior, and Identity Science. 2019. vol. 1. no. 3. pp. 165–180.
29. Dee T., Richardson I., Tyagi A. Continuous transparent mobile device touchscreen soft keyboard biometric authentication // 32nd international conference on VLSI design and 18th international conference on embedded systems (VLSID). 2019. pp. 539–540.
30. Rahman K.A., Balagani K.S., Phoha V.V. Making impostor pass rates meaningless: A case of snoop-forge-replay attack on continuous cyber-behavioral verification with keystrokes // CVPR 2011 workshops. 2011. pp. 31–38.
31. Messerman A., Mustafic T., Camtepe S.A., Albayrak S. Continuous and non-intrusive identity verification in real-time environments based on free-text keystroke dynamics // International Joint Conference on Biometrics (IJCB). 2011. pp. 1–8.
32. Zack R.S., Tappert C.C., Cha S.H. Performance of a long-text-input keystroke biometric authentication system using an improved k-nearest-neighbor classification method // Fourth IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS). 2010. pp. 1–6.

33. Traore I., Woungang I., Obaidat M.S., Nakkabi Y., Lai I. Combining mouse and keystroke dynamics biometrics for risk-based authentication in web environments // Fourth international conference on digital home. 2012. pp. 138–145.
34. Quraishi S.J., Bedi S.S. On keystrokes as continuous user biometric authentication // International Journal of Engineering and Advanced Technology. 2019. vol. 8. no. 6. pp. 4149–4153.
35. Ayotte B., Huang J., Banavar M.K., Hou D., Schuckers S. Fast continuous user authentication using distance metric fusion of free-text keystroke data // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2019. pp. 1–9.
36. Mhenni A., Cherrier E., Rosenberger C., Amara N.E.B. Double serial adaptation mechanism for keystroke dynamics authentication based on a single password // Computers & Security. 2019. vol. 83. pp. 151–166.
37. Lu X., Zhang S., Hui P., Lio P. Continuous authentication by free-text keystroke based on CNN and RNN // Computers and Security. 2020. vol. 96. no. 101861.
38. Kiyani A.T., Lasebae A., Ali K., Rehman M.U., Haq B. Continuous user authentication featuring keystroke dynamics based on robust recurrent confidence model and ensemble learning approach // IEEE Access. 2020. vol. 8. pp. 156177–156189.
39. Shimshon T., Moskovitch R., Rokach L., Elovici Y. Continuous verification using keystroke dynamics // International conference on computational intelligence and security. IEEE Computer Society. 2010. pp. 411–415.
40. Gunetti D., Picardi C. Keystroke analysis of free text // ACM Transactions on Information and System Security (TISSEC). 2005. vol. 8. no. 3. pp. 312–347.
41. Stanic M. Continuous user verification based on behavioral biometrics using mouse dynamics // Proceedings of the ITI 2013 35th International Conference on Information Technology Interfaces. IEEE, 2013. C. 251–256.
42. Feher C., Elovici Y., Moskovitch R., Rokach L., Schclar A. User identity verification via mouse dynamics // Information Sciences. 2012. vol. 201. pp. 19–36.
43. Shen C., Cai Z., Guan X., Du Y., Maxion R.A. User authentication through mouse dynamics // IEEE Transactions on Information Forensics and Security. 2012. vol. 8. no. 1. pp. 16–30.
44. Zheng N., Paloski A., Wang H. An efficient user verification system using angle-based mouse movement biometrics // ACM Transactions on Information and System Security (TISSEC). 2016. vol. 18. no. 3. pp. 1–27.
45. Ahmed A.A.E., Traore I. A new biometric technology based on mouse dynamics // IEEE Transactions on dependable and secure computing. 2007. vol. 4. no. 3. pp. 165–179.
46. Siddiqui N., Dave R., Vanamala M., Seliya N. Machine and deep learning applications to mouse dynamics for continuous user authentication // Machine Learning and Knowledge Extraction. 2022. vol. 4. no. 2. pp. 502–518.
47. Rahman K.A., Moormann R., Dierich D., Hossain M.S. Continuous User Verification via Mouse Activities. Multimedia Communications, Services and Security: 8th International Conference, MCSS. 2015. pp. 170–181.
48. Neal T., Sundararajan K., Woodard D. Exploiting linguistic style as a cognitive biometric for continuous verification // International Conference on Biometrics (ICB). IEEE, 2018. C. 270–276.
49. Niinuma K., Park U., Jain A.K. Soft biometric traits for continuous user authentication // IEEE Transactions on information forensics and security. 2010. vol. 5. no. 4. pp. 771–780.

50. Mock K., Hoanca B., Weaver J., Milton M. Real-time continuous iris recognition for authentication using an eye tracker // Proceedings of the 2012 ACM conference on Computer and communications security. 2012. pp. 1007–1009.
51. Lin F., Song C., Zhuang Y., Xu W., Li C., Ren K.. Cardiac scan: A non-contact and continuous heart-based user authentication system // Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking. 2017. pp. 315–328.
52. Ingale M., Cordeiro R., Thentu S., Park Y., Karimian N. Ecg biometric authentication: A comparative analysis // IEEE Access. 2020. vol. 8. pp. 117853–117866.
53. Ekiz D., Can Y.S., Dardagan Y.C., Ersoy C. Can a smartband be used for continuous implicit authentication in real life // IEEE Access. 2020. vol. 8. pp. 59402–59411.
54. Kunz M., Kasper K., Reininger H., Mobius M., Ohms J. Continuous speaker verification in realtime // BIOSIG 2011–Proceedings of the Biometrics Special Interest Group. 2011. pp. 79–87.
55. Liu J., Chen Y., Dong Y., Wang Y., Zhao T., Yao Y.-Do. Continuous User Verification via Respiratory Biometrics. IEEE INFOCOM 2020 – IEEE Conference on Computer Communications. 2020. pp. 1–10.
56. Zhuravchak A., Kapshii O., Pournaras E. Human Activity Recognition based on Wi-Fi CSI Data-A Deep Neural Network Approach // Procedia Computer Science. 2022. vol. 198. pp. 59–66.
57. Ceker H., Upadhyaya S. User authentication with keystroke dynamics in long-text data // IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS). IEEE, 2016. pp. 1–6.
58. Wang X., Shi Y., Zheng K., Zhang Y., Hong W., Cao S. User Authentication Method Based on Keystroke Dynamics and Mouse Dynamics with Scene-Irrelated Features in Hybrid Scenes // Sensors. 2022. vol. 22. no. 17. pp. 6627.
59. Vural E., Huang J., Hou D., Schuckers S. Shared research dataset to support development of keystroke authentication // IEEE International joint conference on biometrics. IEEE, 2014. pp. 1–8.
60. Li J., Chang H.C., Stamp M. Free-text keystroke dynamics for user authentication // Artificial Intelligence for Cybersecurity. Cham: Springer International Publishing, 2022. pp. 357–380.
61. Sun Y., Ceker H., Upadhyaya S. Shared keystroke dataset for continuous authentication // IEEE International Workshop on Information Forensics and Security (WIFS). IEEE, 2016. pp. 1–6.
62. Ahmed A.A., Traore I. Biometric recognition based on free-text keystroke dynamics // IEEE transactions on cybernetics. 2013. vol. 44. no. 4. pp. 458–472.
63. Martin A.G., de Diego I.M., Fernandez-Isabel A., Beltran M., Fernandez R.R. Combining user behavioural information at the feature level to enhance continuous authentication systems // Knowledge-Based Systems. 2022. vol. 244. no. 108544. DOI: 10.1016/j.knosys.2022.108544.
64. Harilal A., Toffalini F., Castellanos J., Guarnizo J., Homoliak I., Ochoa M.. Twos: A dataset of malicious insider threat behavior based on a gamified competition // Proceedings of the International Workshop on Managing Insider Security Threats. 2017. pp. 45–56.
65. Belman A.K. et al. Insights from BB-MAS--A Large Dataset for Typing, Gait and Swipes of the Same Person on Desktop, Tablet and Phone // arXiv preprint arXiv:1912.02736. 2019. 2019.
66. Killourhy K.S., Maxion R.A. Free vs. transcribed text for keystroke-dynamics evaluations // Proceedings of the Workshop on Learning from Authoritative Security Experiment Results. 2012. pp. 1–8.

67. Roth J., Liu X., Metaxas D. On continuous user authentication via typing behavior // IEEE Transactions on Image Processing. 2014. vol. 23. no. 10. pp. 4611–4624.
68. Iapa A.C., Cretu V.I. Shared Data Set for Free-Text Keystroke Dynamics Authentication Algorithms. 2021. DOI: 10.20944/preprints202105.0255.v1.
69. Bergadano F., Gunetti D., Picardi C. Identity verification through dynamic keystroke analysis // Intelligent Data Analysis. 2003. vol. 7. no. 5. pp. 469–496.
70. Banerjee R., Feng S., Kang J.S., Choi Y. Keystroke patterns as prosody in digital writings: A case study with deceptive reviews and essays // Proceedings of the Conference on empirical methods in natural language processing (EMNLP). 2014. pp. 1469–1473.
71. Murphy C., Huang J., Hou D., Schuckers S. Shared dataset on natural human-computer interaction to support continuous authentication research // IEEE International Joint Conference on Biometrics (IJCB). IEEE, 2017. pp. 525–530.
72. Kilic A.A., Yildirim M., Anarim E. Bogazici mouse dynamics dataset // Data in Brief. 2021. vol. 36. no. 107094.
73. Shen C., Cai Z., Guan X. Continuous authentication for mouse dynamics: A pattern-growth approach // IEEE/IFIP International Conference on Dependable Systems and Networks (DSN 2012). IEEE, 2012. pp. 1–12.
74. Shen C., Cai Z., Guan X., Maxion R. Performance evaluation of anomaly-detection algorithms for mouse dynamics // Computers and security. 2014. vol. 45. pp. 156–171.
75. Antal M., Egyed-Zsigmond E. Intrusion detection using mouse dynamics // IET Biometrics. 2019. vol. 8. no. 5. pp. 285–294.
76. Fulop A., Kovacs L., Kurics T., Windhager-Pokol E. Balabit Mouse Dynamics Challenge data set. 2017. Available at: https://github.com/balabit/Mouse-Dynamics-Challenge (accessed: 16.10.2023).
77. Almalki S., Chatterjee P., Roy K. Continuous authentication using mouse clickstream data analysis // Security, Privacy, and Anonymity in Computation, Communication, and Storage: SpaCCS Proceedings 12. Springer International Publishing, 2019. pp. 76–85.
78. Antal M., Denes-Fazakas L. User verification based on mouse dynamics: a comparison of public data sets // IEEE 13th International Symposium on Applied Computational Intelligence and Informatics (SACI). IEEE, 2019. pp. 143–148.
79. Weiss G.M. Wisdm smartphone and smartwatch activity and biometrics dataset // UCI Machine Learning Repository: WISDM Smartphone and Smartwatch Activity and Biometrics Dataset Data Set. 2019. vol. 7. pp. 133190–133202.
80. Матвеев Ю.Н. Исследование информативности признаков речи для систем автоматической идентификации дикторов // Известия высших учебных заведений. Приборостроение. 2013. Т. 56. № 2. С. 47–51.
81. Стародубов Д.Н. Методика определения информативности признаков объектов // Алгоритмы, методы и системы обработки данных. 2008. № 13. С. 140–146.
82. Frank E., Hall M., Holmes G., Kirkby R., Pfahringer B., Witten I.H., Trigg L. Weka-a machine learning workbench for data mining // Data mining and knowledge discovery handbook. 2010. pp. 1269–1277.
83. Sharma P., Singh D., Singh A. Classification algorithms on a large continuous random dataset using rapid miner tool // 2nd International Conference on Electronics and Communication Systems (ICECS). IEEE, 2015. pp. 704–709.
84. Жигулин П.В., Мальцев А.В., Мельников М.А., Подворчан Д.Э. Анализ сетевого трафика на основе нейронных сетей // Электронные средства и системы управления. Материалы докладов Международной научно-практической конференции. 2013. № 2. С. 44–48.

85. Быкова В.В., Катаева А.В. Методы и средства анализа информативности признаков при обработке медицинских данных // Программные продукты и системы. 2016. № 2(114). С. 172–178.
86. Burton A, Parikh T., Mascarenhas S., Zhang J., Voris J., Artan N.S., Li W. Driver identification and authentication with active behavior modeling // 12th International Conference on Network and Service Management (CNSM). IEEE, 2016. pp. 388–393.
87. Milton L.C., Memon A. Intruder detector: A continuous authentication tool to model user behavior // IEEE Conference on Intelligence and Security Informatics (ISI). IEEE, 2016. pp. 286–291.
88. Siddiqui N., Dave R., Vanamala M., Seliya N. Machine and deep learning applications to mouse dynamics for continuous user authentication // Machine Learning and Knowledge Extraction. 2022. vol. 4. no. 2. pp. 502–518.
89. Feher C., Elovici Y., Moskovitch R., Rokach L., Schclar A. User identity verification via mouse dynamics // Information Sciences. 2012. vol. 201. pp. 19–36.
90. Kuzminykh I., Mathur S., Ghita B. Performance Analysis of Free Text Keystroke Authentication using XGBoost // Proc. of 6th International Conference on Computer Science, Engineering and Education Applications. (ICCSEEA). 2023. pp. 429–439.
91. Agrafioti F., Bui F.M., Hatzinakos D. Secure telemedicine: Biometrics for remote and continuous patient verification // Journal of Computer Networks and Communications. 2012. vol. 2012. DOI: 10.1155/2012/924791.

Давыденко Сергей Андреевич — младший научный сотрудник, центр компетенций Национальной технологической инициативы «технологии доверенного взаимодействия», Федеральное государственное бюджетное образовательное учреждение высшего образования «Томский государственный университет систем управления и радиоэлектроники». Область научных интересов: искусственный интеллект и машинное обучение. Число научных публикаций — 122. sergun_dav@mail.ru; проспект Ленина, 40, 634050, Томск, Россия; р.т.: +7(3822)51-3262.

Костюченко Евгений Юрьевич — канд. техн. наук, доцент, заведующий лабораторией, лаборатория съема, анализа и управления биологическими сигналами; доцент кафедры, кафедра комплексной информационной безопасности электронно-вычислительных систем, Федеральное государственное бюджетное образовательное учреждение высшего образования «Томский государственный университет систем управления и радиоэлектроники». Область научных интересов: искусственный интеллект и машинное обучение, обработка речи, биометрия, анализ данных. Число научных публикаций — 145. key@fb.tusur.ru; проспект Ленина, 40, 634050, Томск, Россия; р.т.: +7(3822)70-1529.

Новиков Сергей Николаевич — д-р техн. наук, доцент, заведующий кафедрой, кафедра безопасности и управления в телекоммуникациях, Сибирский государственный университет телекоммуникаций и информатики. Область научных интересов: искусственный интеллект и машинное обучение. Число научных публикаций — 122. snovikov@ngs.ru; улица Кирова, 86, 630102, Новосибирск, Россия; р.т.: +7(3832)69-2245.

Поддержка исследований. Работа выполнена при финансовой поддержке Министерства науки и высшего образования РФ в рамках базовой части государственного задания ТУСУРа на 2023–2025 гг. (проект № FEWM-2023-0015).

DOI 10.15622/ia.23.1.3

S. DAVYDENKO, E. KOSTYUCHENKO, S. NOVIKOV

EVALUATION OF THE INFORMATIVENESS OF FEATURES IN DATASETS FOR CONTINUOUS VERIFICATION

Davydenko S., Kostyuchenko E., Novikov S. **Evaluation of the Informativeness of Features in Datasets for Continuous Verification.**

Abstract. Continuous verification eliminates the flaws of existing static authentication, e.g. identifiers can be lost or forgotten, and the user logs in the system only once, which may be dangerous not only for areas requiring a high level of security but also for a regular office. Checking the user dynamically during the whole session of work can improve the security of the system, since while working with the system, the user may be exposed to an attacker (to be assaulted for example) or intentionally transfer rights to him. In this case, the machine will not be operated by the user who performed the initial login. Classifying users continuously will limit access to sensitive data that can be obtained by an attacker. During the study, the methods and datasets used for continuous verification were checked, then some datasets were chosen, which were used in further research: smartphone and smart watch movement data (WISDM) and mouse activity (Chao Shen's, DFL, Balabit). In order to improve the performance of models in the classification task it is necessary to perform a preliminary selection of features, to evaluate their informativeness. Reducing the number of features makes it possible to reduce the requirements for devices that will be used for their processing, and to increase the volume of enumeration of classifier parameter values at the same time, thereby potentially increasing the proportion of correct answers during classification due to a more complete enumeration of value parameters. For the informativeness evaluation, the Shannon method was used, as well as the algorithms built into programs for data analysis and machine learning (WEKA: Machine Learning Software and RapidMiner). In the course of the study, the informativeness of each feature in the selected datasets was evaluated, and then users were classified with RapidMiner. The used in classifying features selection was decreased gradually with a 20% step. As a result, a table was formed with recommended sets of features for each dataset, as well as dependency graphs of the accuracy and operating time of various models.

Keywords: informativeness, classification, continuous verification, machine learning, feature selection, information security.

References

1. Jain A.K., Ross A., Pankanti S. Biometrics: a tool for information security. IEEE transactions on information forensics and security. 2006. vol. 1. no. 2. pp. 125–143.
2. Jain A.K., Ross A., Prabhakar S. An introduction to biometric recognition. IEEE Transactions on circuits and systems for video technology. 2004. vol. 14. no. 1. pp. 4–20.
3. Zhang N., Yu W., Fu X., Das S.K. Maintaining defender's reputation in anomaly detection against insider attacks. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics). 2009. vol. 40. no. 3. pp. 597–611.
4. Liang Y., Samtani S., Guo B., Yu Z. Behavioral biometrics for continuous authentication in the internet-of-things era: An artificial intelligence perspective. IEEE Internet of Things Journal. 2020. vol. 7. no. 9. pp. 9128–9143.
5. Al-Naji F.H., Zagrouba R. CAB-IoT: Continuous authentication architecture based on Blockchain for internet of things. Journal of King Saud University-Computer and Information Sciences. 2022. vol. 34. no. 6. pp. 2497–2514.

6. Ashibani Y., Kauling D., Mahmoud Q.H. Design and implementation of a contextual-based continuous authentication framework for smart homes. Applied System Innovation. 2019. vol. 2(1). DOI: 10.3390/asi2010004.
7. Oak R. A literature survey on authentication using Behavioural biometric techniques. Intelligent Computing and Information and Communication: Proceedings of 2nd International Conference, ICICC. 2018. pp. 173–181.
8. Bondarenko M.A., Drynkin V.N. [Evaluation of the informativeness of combined images in multispectral complex technical studies]. Programmnye sistemy i vychislitel'nye metody – Software systems and computational methods. 2016. vol. 1. pp. 64–79. (In Russ.).
9. Soltane M., Bakhti M. Multi-modal biometric authentications: concept issues and applications strategies. International Journal of Advanced Science and Technology. 2012. vol. 48. pp. 23–60.
10. Hong C.S., Oh T.G. TPR-TNR plot for confusion matrix. Communications for Statistical Applications and Methods. 2021. vol. 28. no. 2. pp. 161–169.
11. Rahman K.A., Alam N., Musarrat J., Madarapu A., Hossain M.S. Smartwatch Dynamics: A Novel Modality and Solution to Attacks on Cyber-behavioral Biometrics for Continuous Verification?. International Symposium on Networks, Computers and Communications (ISNCC). 2020. pp. 1–5.
12. Verma A., Moghaddam V., Anwar A. Data-driven behavioural biometrics for continuous and adaptive user verification using Smartphone and Smartwatch. Sustainability. 2022. vol. 14(12). DOI: 10.3390/su14127362.
13. Abuhamad M., Abusnaina A., Nyang D., Mohaisen D. Sensor-based continuous authentication of smartphones' users using behavioral biometrics: A contemporary survey. IEEE Internet of Things Journal. 2020. vol. 8. no. 1. pp. 65–84.
14. Lamiche I., Bin G., Jing Y., Yu Z., Hadid A. A continuous smartphone authentication method based on gait patterns and keystroke dynamics. Journal of Ambient Intelligence and Humanized Computing. 2019. vol. 10. pp. 4417–4430.
15. Giorgi G., Saracino A., Martinelli F. Using recurrent neural networks for continuous authentication through gait analysis. Pattern Recognition Letters. 2021. vol. 147.pp. 157–163.
16. Kim D.I., Lee S., Shin J.S. A new feature scoring method in keystroke dynamics-based user authentications. IEEE Access. 2020. vol. 8. pp. 27901–27914.
17. Deb D., Ross A., Jain A.K., Prakah-Asante K., Prasad K.V. Actions speak louder than (pass) words: Passive authentication of smartphone* users via deep temporal features. International conference on biometrics (ICB). 2019. pp. 1–8.
18. Abuhamad M., Abuhmed T., Mohaisen D., Nyang D. AUToSen: Deep-learning-based implicit continuous authentication using smartphone sensors. IEEE Internet of Things Journal. 2020. vol. 7. no. 6. pp. 5008–5020.
19. Barlas Y., Basar O.E., Akan Y., Isbilen M., Alptekin G.I., Incel O.D. DAKOTA: Continuous authentication with behavioral biometrics in a mobile banking application. 5th International Conference on Computer Science and Engineering (UBMK). 2020. pp. 1–6.
20. Mekruksavanich S., Jitpattanakul A. Deep learning approaches for continuous authentication based on activity patterns using mobile sensing. Sensors. 2021. vol. 21. no. 22. DOI: 10.3390/s21227519.
21. Acien A., Morales A., Vera-Rodriguez R., Fierrez J., Tolosana R.. Multilock: Mobile active authentication based on multiple biometric and behavioral patterns. 1st International Workshop on Multimodal Understanding and Learning for Embodied Applications. 2019. pp. 53–59.

22. Li Y., Hu H., Zhu Z., Zhou G. SCANet: sensor-based continuous authentication with two-stream convolutional neural networks. ACM Transactions on Sensor Networks (TOSN). 2020. vol. 16. no. 3. pp. 1–27.
23. Mekruksavanich S., Jitpattanakul A. Deep convolutional neural network with rnns for complex activity recognition using wrist-worn wearable sensor data. Electronics. 2021. vol. 10. no. 14. DOI: 10.3390/electronics10141685.
24. Volaka H.C., Alptekin G., Basar O.E., Isbilen M., Incel O.D. Towards continuous authentication on mobile phones using deep learning models. Procedia Computer Science. 2019. vol. 155. pp. 177–184.
25. Li Y., Zou B., Deng S., Zhou G. Using feature fusion strategies in continuous authentication on smartphones. IEEE Internet Computing. 2020. vol. 24. no. 2. pp. 49–56.
26. Incel O.D., Gunay S., Akan Y., Barlas Y., Basar O.E., Alptekin G.I., Isbilen M. Dakota: sensor and touch screen-based continuous authentication on a mobile banking application. IEEE Access. 2021. vol. 9. pp. 38943–38960.
27. Alotaibi S., Alruban A., Furnell S., Clarke N. A Novel Behaviour Profiling Approach to Continuous Authentication for Mobile Applications. ICISSP. 2019. pp. 246–251.
28. Mahbub U., Komulainen J., Ferreira D., Chellappa R. Continuous authentication of smartphones based on application usage. IEEE Transactions on Biometrics, Behavior, and Identity Science. 2019. vol. 1. no. 3. pp. 165–180.
29. Dee T., Richardson I., Tyagi A. Continuous transparent mobile device touchscreen soft keyboard biometric authentication. 32nd international conference on VLSI design and 18th international conference on embedded systems (VLSID). 2019. pp. 539–540.
30. Rahman K.A., Balagani K.S., Phoha V.V. Making impostor pass rates meaningless: A case of snoop-forge-replay attack on continuous cyber-behavioral verification with keystrokes. CVPR 2011 workshops. 2011. pp. 31–38.
31. Messerman A., Mustafic T., Camtepe S.A., Albayrak S. Continuous and non-intrusive identity verification in real-time environments based on free-text keystroke dynamics. International Joint Conference on Biometrics (IJCB). 2011. pp. 1–8.
32. Zack R.S., Tappert C.C., Cha S.H. Performance of a long-text-input keystroke biometric authentication system using an improved k-nearest-neighbor classification method. Fourth IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS). 2010. pp. 1–6.
33. Traore I., Woungang I., Obaidat M.S., Nakkabi Y., Lai I. Combining mouse and keystroke dynamics biometrics for risk-based authentication in web environments. Fourth international conference on digital home. 2012. pp. 138–145.
34. Quraishi S.J., Bedi S.S. On keystrokes as continuous user biometric authentication. International Journal of Engineering and Advanced Technology. 2019. vol. 8. no. 6. pp. 4149–4153.
35. Ayotte B., Huang J., Banavar M.K., Hou D., Schuckers S. Fast continuous user authentication using distance metric fusion of free-text keystroke data. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2019. pp. 1–9.
36. Mhenni A., Cherrier E., Rosenberger C., Amara N.E.B. Double serial adaptation mechanism for keystroke dynamics authentication based on a single password. Computers & Security. 2019. vol. 83. pp. 151–166.
37. Lu X., Zhang S., Hui P., Lio P. Continuous authentication by free-text keystroke based on CNN and RNN. Computers and Security. 2020. vol. 96. no. 101861.
38. Kiyani A.T., Lasebae A., Ali K., Rehman M.U., Haq B. Continuous user authentication featuring keystroke dynamics based on robust recurrent confidence model and ensemble learning approach. IEEE Access. 2020. vol. 8. pp. 156177–156189.

39. Shimshon T., Moskovitch R., Rokach L., Elovici Y. Continuous verification using keystroke dynamics. International conference on computational intelligence and security. IEEE Computer Society. 2010. pp. 411–415.
40. Gunetti D., Picardi C. Keystroke analysis of free text. ACM Transactions on Information and System Security (TISSEC). 2005. vol. 8. no. 3. pp. 312–347.
41. Stanic M. Continuous user verification based on behavioral biometrics using mouse dynamics. Proceedings of the ITI 2013 35th International Conference on Information Technology Interfaces. IEEE, 2013. C. 251–256.
42. Feher C., Elovici Y., Moskovitch R., Rokach L., Schclar A. User identity verification via mouse dynamics. Information Sciences. 2012. vol. 201. pp. 19–36.
43. Shen C., Cai Z., Guan X., Du Y., Maxion R.A. User authentication through mouse dynamics. IEEE Transactions on Information Forensics and Security. 2012. vol. 8. no. 1. pp. 16–30.
44. Zheng N., Paloski A., Wang H. An efficient user verification system using angle-based mouse movement biometrics. ACM Transactions on Information and System Security (TISSEC). 2016. vol. 18. no. 3. pp. 1–27.
45. Ahmed A.A.E., Traore I. A new biometric technology based on mouse dynamics. IEEE Transactions on dependable and secure computing. 2007. vol. 4. no. 3. pp. 165–179.
46. Siddiqui N., Dave R., Vanamala M., Seliya N. Machine and deep learning applications to mouse dynamics for continuous user authentication. Machine Learning and Knowledge Extraction. 2022. vol. 4. no. 2. pp. 502–518.
47. Rahman K.A., Moormann R., Dierich D., Hossain M.S. Continuous User Verification via Mouse Activities. Multimedia Communications, Services and Security: 8th International Conference, MCSS. 2015. pp. 170–181.
48. Neal T., Sundararajan K., Woodard D. Exploiting linguistic style as a cognitive biometric for continuous verification. International Conference on Biometrics (ICB). IEEE, 2018. C. 270–276.
49. Niinuma K., Park U., Jain A.K. Soft biometric traits for continuous user authentication. IEEE Transactions on information forensics and security. 2010. vol. 5. no. 4. pp. 771–780.
50. Mock K., Hoanca B., Weaver J., Milton M. Real-time continuous iris recognition for authentication using an eye tracker. Proceedings of the 2012 ACM conference on Computer and communications security. 2012. pp. 1007–1009.
51. Lin F., Song C., Zhuang Y., Xu W., Li C., Ren K.. Cardiac scan: A non-contact and continuous heart-based user authentication system. Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking. 2017. pp. 315–328.
52. Ingale M., Cordeiro R., Thentu S., Park Y., Karimian N. Ecg biometric authentication: A comparative analysis. IEEE Access. 2020. vol. 8. pp. 117853–117866.
53. Ekiz D., Can Y.S., Dardagan Y.C., Ersoy C. Can a smartband be used for continuous implicit authentication in real life. IEEE Access. 2020. vol. 8. pp. 59402–59411.
54. Kunz M., Kasper K., Reininger H., Mobius M., Ohms J. Continuous speaker verification in realtime. BIOSIG 2011–Proceedings of the Biometrics Special Interest Group. 2011. pp. 79–87.
55. Liu J., Chen Y., Dong Y., Wang Y., Zhao T., Yao Y.-Do. Continuous User Verification via Respiratory Biometrics. IEEE INFOCOM 2020 – IEEE Conference on Computer Communications. 2020. pp. 1–10.
56. Zhuravchak A., Kapshii O., Pournaras E. Human Activity Recognition based on Wi-Fi CSI Data-A Deep Neural Network Approach. Procedia Computer Science. 2022. vol. 198. pp. 59–66.

57. Ceker H., Upadhyaya S. User authentication with keystroke dynamics in long-text data. IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS). IEEE, 2016. pp. 1–6.
58. Wang X., Shi Y., Zheng K., Zhang Y., Hong W., Cao S. User Authentication Method Based on Keystroke Dynamics and Mouse Dynamics with Scene-Irrelated Features in Hybrid Scenes. Sensors. 2022. vol. 22. no. 17. pp. 6627.
59. Vural E., Huang J., Hou D., Schuckers S. Shared research dataset to support development of keystroke authentication. IEEE International joint conference on biometrics. IEEE, 2014. pp. 1–8.
60. Li J., Chang H.C., Stamp M. Free-text keystroke dynamics for user authentication. Artificial Intelligence for Cybersecurity. Cham: Springer International Publishing, 2022. pp. 357–380.
61. Sun Y., Ceker H., Upadhyaya S. Shared keystroke dataset for continuous authentication. IEEE International Workshop on Information Forensics and Security (WIFS). IEEE, 2016. pp. 1–6.
62. Ahmed A.A., Traore I. Biometric recognition based on free-text keystroke dynamics. IEEE transactions on cybernetics. 2013. vol. 44. no. 4. pp. 458–472.
63. Martin A.G., de Diego I.M., Fernandez-Isabel A., Beltran M., Fernandez R.R. Combining user behavioural information at the feature level to enhance continuous authentication systems. Knowledge-Based Systems. 2022. vol. 244. no. 108544. DOI: 10.1016/j.knosys.2022.108544.
64. Harilal A., Toffalini F., Castellanos J., Guarnizo J., Homoliak I., Ochoa M.. Twos: A dataset of malicious insider threat behavior based on a gamified competition. Proceedings of the International Workshop on Managing Insider Security Threats. 2017. pp. 45–56.
65. Belman A.K. et al. Insights from BB-MAS--A Large Dataset for Typing, Gait and Swipes of the Same Person on Desktop, Tablet and Phone. arXiv preprint arXiv:1912.02736. 2019. 2019.
66. Killourhy K.S., Maxion R.A. Free vs. transcribed text for keystroke-dynamics evaluations. Proceedings of the Workshop on Learning from Authoritative Security Experiment Results. 2012. pp. 1–8.
67. Roth J., Liu X., Metaxas D. On continuous user authentication via typing behavior. IEEE Transactions on Image Processing. 2014. vol. 23. no. 10. pp. 4611–4624.
68. Iapa A.C., Cretu V.I. Shared Data Set for Free-Text Keystroke Dynamics Authentication Algorithms. 2021. DOI: 10.20944/preprints202105.0255.v1.
69. Bergadano F., Gunetti D., Picardi C. Identity verification through dynamic keystroke analysis. Intelligent Data Analysis. 2003. vol. 7. no. 5. pp. 469–496.
70. Banerjee R., Feng S., Kang J.S., Choi Y. Keystroke patterns as prosody in digital writings: A case study with deceptive reviews and essays. Proceedings of the Conference on empirical methods in natural language processing (EMNLP). 2014. pp. 1469–1473.
71. Murphy C., Huang J., Hou D., Schuckers S. Shared dataset on natural human-computer interaction to support continuous authentication research. IEEE International Joint Conference on Biometrics (IJCB). IEEE, 2017. pp. 525–530.
72. Kilic A.A., Yildirim M., Anarim E. Bogazici mouse dynamics dataset. Data in Brief. 2021. vol. 36. no. 107094.
73. Shen C., Cai Z., Guan X. Continuous authentication for mouse dynamics: A pattern-growth approach. IEEE/IFIP International Conference on Dependable Systems and Networks (DSN 2012). IEEE, 2012. pp. 1–12.
74. Shen C., Cai Z., Guan X., Maxion R. Performance evaluation of anomaly-detection algorithms for mouse dynamics. Computers and security. 2014. vol. 45. pp. 156–171.

75. Antal M., Egyed-Zsigmond E. Intrusion detection using mouse dynamics. IET Biometrics. 2019. vol. 8. no. 5. pp. 285–294.
76. Fulop A., Kovacs L., Kurics T., Windhager-Pokol E. Balabit Mouse Dynamics Challenge data set. 2017. Available at: https://github.com/balabit/Mouse-Dynamics-Challenge (accessed: 16.10.2023).
77. Almalki S., Chatterjee P., Roy K. Continuous authentication using mouse clickstream data analysis. Security, Privacy, and Anonymity in Computation, Communication, and Storage: SpaCCS Proceedings 12. Springer International Publishing, 2019. pp. 76–85.
78. Antal M., Denes-Fazakas L. User verification based on mouse dynamics: a comparison of public data sets. IEEE 13th International Symposium on Applied Computational Intelligence and Informatics (SACI). IEEE, 2019. pp. 143–148.
79. Weiss G.M. Wisdm smartphone and smartwatch activity and biometrics dataset. UCI Machine Learning Repository: WISDM Smartphone and Smartwatch Activity and Biometrics Dataset Data Set. 2019. vol. 7. pp. 133190–133202.
80. Matveev Ju.N. [Study of the informativeness of speech features for automatic speaker identification systems]. Izvestija vysshih uchebnyh zavedenij. Priborostroenie – News of higher educational institutions. Instrumentation. 2013. vol. 56. no. 2. pp. 47–51. (In Russ.).
81. Starodubov D.N. [Methods for determining the informativeness of features of objects]. Algoritmy, metody i sistemy obrabotki dannyh – Algorithms, methods and data processing systems. 2008. no. 13. pp. 140–146. (In Russ.).
82. Frank E., Hall M., Holmes G., Kirkby R., Pfahringer B., Witten I.H., Trigg L. Weka-a machine learning workbench for data mining. Data mining and knowledge discovery handbook. 2010. pp. 1269–1277.
83. Sharma P., Singh D., Singh A. Classification algorithms on a large continuous random dataset using rapid miner tool. 2nd International Conference on Electronics and Communication Systems (ICECS). IEEE, 2015. pp. 704–709.
84. Zhigulin P.V., Mal'cev A.V., Mel'nikov M.A., Podvorchan D.E. [Network traffic analysis based on neural networks]. Jelektronnye sredstva i sistemy upravlenija. Materialy dokladov Mezhdunarodnoj nauchno-prakticheskoj konferencii – Electronic means and control systems. Materials of reports of the International Scientific and Practical Conference. 2013. no. 2. pp. 44–48. (In Russ.).
85. Bykova V.V., Kataeva A.V. [Methods and tools for analyzing the informativeness of features in the processing of medical data]. Programmnye produkty i sistemy – Software products and systems. 2016. no. 2(114). pp. 172–178.
86. Burton A, Parikh T., Mascarenhas S., Zhang J., Voris J., Artan N.S., Li W. Driver identification and authentication with active behavior modeling. 12th International Conference on Network and Service Management (CNSM). IEEE, 2016. pp. 388–393.
87. Milton L.C., Memon A. Intruder detector: A continuous authentication tool to model user behavior. IEEE Conference on Intelligence and Security Informatics (ISI). IEEE, 2016. pp. 286–291.
88. Siddiqui N., Dave R., Vanamala M., Seliya N. Machine and deep learning applications to mouse dynamics for continuous user authentication. Machine Learning and Knowledge Extraction. 2022. vol. 4. no. 2. pp. 502–518.
89. Feher C., Elovici Y., Moskovitch R., Rokach L., Schclar A. User identity verification via mouse dynamics. Information Sciences. 2012. vol. 201. pp. 19–36.
90. Kuzminykh I., Mathur S., Ghita B. Performance Analysis of Free Text Keystroke Authentication using XGBoost. Proc. of 6th International Conference on Computer Science, Engineering and Education Applications. (ICCSEEA). 2023. pp. 429–439.

91. Agrafioti F., Bui F.M., Hatzinakos D. Secure telemedicine: Biometrics for remote and continuous patient verification. Journal of Computer Networks and Communications. 2012. vol. 2012. DOI: 10.1155/2012/924791.

Davydenko Sergey — Junior researcher, NTI competence center "trusted interaction technologies", Tomsk State University of Control Systems and Radioelectronics. Research interests: artificial intelligence and machine learning. The number of publications — 122. sergun_dav@mail.ru; 40, Lenina Av., 634050, Tomsk, Russia; office phone: +7(3822)51-3262.

Kostyuchenko Evgeny — Ph.D., Associate Professor, Head of the laboratory, Laboratory for reception, analysis and control of biological signals; Associate professor of the department, Department of integrated information security of electronic computing systems, Tomsk State University of Control Systems and Radioelectronics. Research interests: artificial intelligence and machine learning, speech processing, biometry, data analysis. The number of publications — 145. key@fb.tusur.ru; 40, Lenina Av., 634050, Tomsk, Russia; office phone: +7(3822)70-1529.

Novikov Sergey — Ph.D., Dr.Sci., Associate Professor, Head of the department, Department of security and management in telecommunications, Siberian State University of Telecommunications and Informatics. Research interests: artificial intelligence and machine learning. The number of publications — 122. snovikov@ngs.ru; 86, Kirova St., 630102, Novosibirsk, Russia; office phone: +7(3832)69-2245.

Acknowledgements. This research was funded by the Ministry of Science and Higher Education of Russia, Government Order for 2023–2025, project no. FEWM-2023-0015 (TUSUR).