УДК 004.056.53
DOI: 10.18469/ikt.2025.23.1.08

ПРОТОТИП СИСТЕМЫ ЗАЩИТЫ ОТ УТЕЧЕК И НЕЙРОСЕТЕВАЯ МОДЕЛЬ ДЛЯ ДЕТЕКЦИИ КОНФИДЕНЦИАЛЬНЫХ ТЕКСТОВЫХ СООБЩЕНИЙ

Сулавко А.Е.[1], Панфилова И.Е.[2], Варкентин Ю.А.[1], Клиновенко С.А.[1]
[1]Омский государственный технический университет, Омск, РФ
[2]Самарский государственный технический университет, Самара, РФ
E-mail: sulavich@mail.ru, panfilova_2015@bk.ru, varkentinyuri@gmail.com, sergey.klinovenko@gmail.com

В работе рассматривается проблема утечек конфиденциальной информации, возникающих вследствие действий внутренних нарушителей. В качестве методов противодействия таким атакам рассмотрены современные программные продукты, в том числе, нацеленные на защиту конфиденциальных данных посредствам анализа поведения пользователя. В качестве решения проблемы предлагается прототип расширенной DLP-системы, использующей методы искусственного интеллекта для анализа активности пользователей и предотвращения таких утечек. Разработанный прототип реализует механизм мониторинга контекста поведения пользователя, а также использует в основе модели машинного обучения для детекции конфиденциальной информации в коротких текстовых сообщениях. В частности, для извлечения признаков используется предварительно обученная нейронная сеть на базе архитектуры E5. Для детекции конфиденциальной информации применяется сверточная нейронная сеть типа автокодировщик, которая обучается исключительно на данных, содержащих конфиденциальные документы. Экспериментальные результаты показали, что предложенная модель успешно выполняет бинарную классификацию сообщений, с ошибкой второго рода на уровне 5,1% и отсутствием ошибок первого рода (в рамках проведенного эксперимента на собственном наборе данных). Предлагаемый комплекс может послужить основой для разработки более сложных решений в сфере информационной безопасности.

Ключевые слова: *утечка конфиденциальной информации, хищение данных, инсайдерские угрозы, конфиденциальные документы, искусственный интеллект, автокодировщик, машинное обучение, поведенческая аналитика*

Введение

Сегодня наиболее остро стоит проблема утечек конфиденциальной информации (представленной в виде документов, данных или знаний искусственного интеллекта (ИИ)). Прежде всего, данная проблема связана с действиями внутренних нарушителей, подрывающих систему безопасности предприятий изнутри. Оценить масштабы потенциальных потерь от утечек численно сложно. В соответствии с результатами исследований компании InfoWatch в 2021 году средняя стоимость одной утечки конфиденциальных данных в мире составила $4,24 млн. [1].

На рынке систем защиты информации имеется множество интеллектуальных программных продуктов, которые позволяют эффективно отражать внешние угрозы. Однако по-прежнему возникают большие сложности в борьбе с внутренними угрозами, а именно: хищением конфиденциальной информации, выявлением инсайдеров и внутренних нарушителей, их мотивов, определением всей цепочки взаимодействия внутренних нарушителей. Наконец, важнейшей задачей является не просто выявить противоправные действия, а блокировать их. На сегодняшний день на рынке нет полнофункциональных решений, способных в полной мере предотвратить утечки информации и обеспечить превентивную защиту от внутренних угроз, основанную на анализе данных и событий безопасности с применением всего возможного потенциала технологий ИИ.

Сегодня уже существует техническая возможность для осуществления непрерывной идентификации и мониторинга пользователей в компьютерных системах, оценки их психофизиологического и эмоционального состояния в процессе профессиональной деятельности, определения уровня конфиденциальности информации, с которой работает пользователь в реальном времени, потенциально опасных действий с этой информацией, а также отслеживания всех взаимодействий пользователя с внешними и внутренними источниками. Объединение технологий регистрации и анализа больших данных с технологиями машинного обучения гипотетически позволяет реализовать указанный функционал и получить достаточный эффект для обеспечения защиты от внутренних угроз.

Настоящее исследование посвящено разработке прототипа программного комплекса для анализа активности пользователей компьютерных систем с целью обеспечения превентивной

защиты от внутренних угроз на основе методов ИИ, а также модели для детекции конфиденциальной информации в потоке сообщений. В работе также проведено исследование существующих программных продуктов.

Научная новизна заключается в построении бинарного классификатора конфиденциальных сообщений на основе автокодировщика. Мы использовали автокодировщик для сжатия и восстановления векторов признаков, которые извлекаются с помощью большой языковой модели. Данный подход применим для детекции аномалий во временных рядах, однако ранее его не применяли для анализа текстовых данных, обработанных языковыми моделями.

Типы программных продуктов для противодействия внутренним угрозам

Существующие системы защиты построены на мониторинге действий пользователя (трудового процесса) и его поведения в корпоративной среде. Чаще всего такие системы реализуются в виде UAM (User Activity Monitoring – мониторинг действий пользователей) решений, позволяющих организациям отслеживать и анализировать активность пользователей в информационных системах и сетях. В наши дни также можно встретить термин EM (Employee Monitoring – мониторинг сотрудников), однако специалисты отмечают, что для Employee Monitoring важно понимание среды, где работает компьютер, какой состав имеет программное обеспечение и т.д. Главная цель UAM заключается в обнаружении аномального поведения пользователей, угроз безопасности и неправомерных действий в сети [2]. UAM обеспечивает возможность регистрации и анализа действий пользователей, таких как входы в систему, выходы из системы, доступ к файлам и ресурсам, выполнение конкретных операций и другую активность. Это позволяет организациям выявлять несанкционированные действия, нарушения политик безопасности, утечки информации и другие потенциальные угрозы для безопасности информации.

Альтернативным вариантом исполнения решения для трудового мониторинга являются системы аналитики поведения пользователей: UEBA (User and Entity Behavior Analytics) и UBA (User Behavior Analytics) [3]. Оба термина используются в области кибербезопасности для обозначения аналитических методов и технологий, которые направлены на обнаружение аномального поведения пользователей в информационных системах. По своей сути указанные технологии очень схожи, и часто используются взаимозаменяемо. Однако UEBA расширяет концепцию UBA, включая в анализ не только поведение пользователей, но и других сущностей в информационной системе, таких как устройства, приложения, сервисы и т.д. Это позволяет более полно оценивать безопасность системы, учитывая не только действия пользователей, но и их взаимодействие с другими элементами инфраструктуры. UEBA часто использует методы машинного обучения, статистики и аналитики данных для выявления потенциальных угроз безопасности, которые могут быть не замечены с помощью традиционных методов обнаружения угроз.

Безусловно, для обеспечения комплексной защиты от инсайдерских угроз необходимо использовать дополнительные решения, позволяющие анализировать не только модель поведения пользователя, но и периметры, к которым он имеет доступ. Среди таких систем можно выделить давно зарекомендовавшие себя DLP (Data Loss Prevention – предотвращение утечки данных) [4] решения, направленные на предотвращение утечек конфиденциальной информации из организации. DLP-системы производят непрерывный мониторинг потоков данных в реальном времени, сканируют и анализируют информацию, проходящую через сеть и хранилища данных, с целью обнаружения конфиденциальной информации. Как правило, DLP-системы позволяют устанавливать и применять политики безопасности, определяющие, какие действия допустимы с конкретными данными. Например, блокирование попыток отправки конфиденциальных данных по электронной почте или на внешних накопителях.

Стоит отметить, что в последнее время DLP, UAM, EM, а также UEBA решения стали взаимодополняющими системами, которые значительно усиливают защиту от инсайдерских угроз [5]. Эти решения предоставляют возможности для записи видео с экрана, изображений с веб-камеры, а также аудиозаписей с микрофона и динамиков. В ряде продуктов поддерживается возможность распознавания лиц с целью контроля присутствия, идентификации и аутентификации сотрудников, а также сбор данных с систем контроля доступа (СКУД) [6]. Помимо этого, программы осуществляют мониторинг посещаемых веб-ресурсов, запускаемых приложений, анализ поисковых запросов, лингвистический анализ текстов, работу с файлами и папками, активность сотрудников, а также анализ снимков экрана и другие важные функции.

Анализ современных и расширенных DLP-систем

DLP-системы нацелены на защиту конфиденциальных данных компании от утечек [7]. Они помогают контролировать передачу данных внутри организации и за ее пределами, обнаруживать и предотвращать утечку информации, а также соблюдать законодательные требования к защите данных и конфиденциальности. Базовый принцип работы DLP-систем – это фильтрация контента: на рабочих станциях и файловых хранилищах, при отправке за периметр организации или в облако.

DLP-системы работают с тремя классами корпоративных данных: data-in-use, data-in-motion и data-at-rest [8]. Data-in-use – это данные, к которым пользователи активно обращаются, обрабатывают и обновляют. DLP предотвращает утечку таких данных благодаря функции мониторинга действий на рабочих станциях (например, съемные USB-носители, буфер обмена, приложения) и анализа поведения пользователей. Data-in-motion называют информацию, которая перемещается из одной точки в другую. DLP предотвращает утечку подобных данных через сетевые коммуникации (например, электронная почта, инструменты совместной работы, программы мгновенного обмена сообщениями и практически любой публичный канал связи). И наконец, data-at-rest, данные, которые не перемещаются между устройствами или сетями. Чаще всего DLP-системы обнаруживают конфиденциальный контент в данных, хранящихся на корпоративных ИТ-ресурсах (например, сетевые файловые ресурсы, файловые системы конечных устройств, базы данных, хранилища документов и облачные хранилища), и устраняет нарушения политики хранения данных.

Почти каждая организация работает со всеми тремя потоками рассмотренных данных, в связи с чем внедрение DLP-системы на предприятии может иметь несколько основных преимуществ:

1. Защита конфиденциальности данных. DLP-система позволяет предотвращать несанкционированный доступ к чувствительным данным и утечку информации. Это особенно важно для компаний, работающих с конфиденциальными клиентскими или корпоративными данными.

2. Соблюдение законодательства. DLP-системы помогают организациям соблюдать законы и нормативные требования относительно защиты данных и конфиденциальности. Нарушение законодательства может привести к штрафам и угрозам репутации предприятия.

3. Предотвращение утечки данных. DLP-системы помогают предотвращать случайную или намеренную утечку данных вследствие действий сотрудников предприятия. Они могут контролировать и отслеживать передачу и использование данных на устройствах компании и за ее пределами.

4. Повышение безопасности информации. DLP-системы помогают предотвращать угрозы безопасности, такие как уязвимости в данных, фишинговые атаки, вредоносное программное обеспечение и утечки информации.

5. Сокращение финансовых потерь. Внедрение DLP-системы помогает предприятию сократить риски финансовых потерь, связанные с утечкой или утратой конфиденциальной информации, а также снизить затраты на восстановление после инцидентов безопасности.

В этой связи компаниям стоит внимательно изучить реальные потребности своего бизнеса, описать информационные активы, оценить риски в случае потенциальных утечек критичных данных, а уже после этого запускать полноценный DLP-проект и выбирать решение, максимально учитывающее его актуальные задачи. В системах обеспечения информационной безопасности рекомендуется использовать продукты, сертифицированные регуляторами, например, ФСТЭК [9]. Для государственных учреждений использование несертифицированных DLP-систем недопустимо.

Для оценки полного спектра возможностей современных DLP решений проведен сравнительный анализ российских (таблица 1) и зарубежных (таблица 2) поставщиков. В качестве критериев сравнения использовались наиболее распространенные функции DLP-систем, включая контроль действий пользователя с файлами, мониторинг посещаемых сайтов, загруженных данных и использования социальных сетей, контроль выполнения исполняемых файлов, возможности лингвистического анализа и др. Отметим, что такие функции, как распознавание лиц для контроля присутствия и идентификации сотрудников, запись с микрофона или контроль рабочего времени сотрудника, скорее можно отнести к поведенческой аналитике, то есть к классу решений UBA или UAM. Российские решения DLP, как правило, либо интегрируют возможности UBA непосредственно в систему, либо предлагают дополнительные модули с функциями UBA для обогащения данных, получаемых от стандартной DLP-системы. Такой подход не характерен для зарубежных решений, в которых

проводится четкая граница между DLP и UBA: DLP занимаются контролем информационных потоков, выявлением инцидентов информационной безопасности, связанных с передачей данных и блокировкой утечек конфиденциальных данных, UBA (иногда UAM) же контролируют действия сотрудников, выявляют инциденты информационной безопасности (ИБ) на основе поведенческой аналитики сотрудников, а также осуществляют сбор доказательной базы при расследованиях.

Таблица 1. Сравнительная таблица типовых характеристик DLP решений, представленных российскими вендорами

Продукт / Функции DLP	InfoWatch Traffic Monitor	Гарда технологии DLP	«Кибер Протего»	Solar Dozor	«СерчИнформ КИБ»	Falcongaze SecureTower	Zecurion	«Стахановец»
Работа с файлами системы	+	+	+	+		+	+	+
Контроль веб-трафика	+	+	+	+	+	+	+	+
Контроль и блокировка запуска приложений	+	+	+		—	+	+	+
Работа с электронной почтой, вложенные файлы	+	+	+	+	+	+	+	+
Контроль периферийного оборудования рабочих мест	+	+	+	+	+	+	+	+
Контроль работы с документами	+	+	+	+	+	+	+	+
Лингвистический анализ	+	+	+	—		+	—	+
Технологии OCR			—	—		+		+
Перехват клавиатурного ввода	+	+	+	+	+	—	—	+
Контроль мессенджеров	+	+	+	+	+	+	+	+
Контроль буфера обмена	+	+	+	+		+	+	+
Интеграция с СКУД	+	—	—	—	—	+		+
Контроль облачных хранилищ	+	+	—	+	+	+		+
Анализ поведения пользователя								
Детектирование факта съемки крана мобильным телефоном	+	—	—	—		—	+	+
Детекция скриншотов с рабочего стола	+		—	+	+	+		+
Запись экрана рабочей станции	+	+	+	+	+	+		+
Запись происходящего в пределах веб-камеры	+	—	—	—	+	+	+	+

(Продолжение) Таблица 1. Сравнительная таблица типовых характеристик DLP решений, представленных российскими вендорами

Распознавание лиц с веб-камер для контроля присутствия и идентификации сотрудников	+	—	—	—	—	—	—	+
Запись с микрофона	+	—	—	+	+	+		+
Контроль рабочего времени сотрудника	+		—	+		+		+
Клавиатурный почерк	—	—	—	—	—	—	—	+

Таблица 2. Сравнительная таблица типовых характеристик DLP решений, представленных зарубежными вендорами

Продукт / Функции DLP	Symantec DLP	ForcePoint	McAfee	Eset Safetica	Sophos Endpoint Protection	Digital Guardian
Работа с файлами системы	+	+	+	+	+	+
Контроль веб-трафика	+	+	+	+	+	+
Контроль и блокировка запуска приложений	+	+	+	+	+	+
Работа с электронной почтой, вложенные файлы	+	+	—	+	—	
Контроль периферийного оборудования рабочих мест	+		+	+	—	
Контроль работы с документами	+	+	+	+	—	
Лингвистический анализ			+		—	
Технологии OCR			+		—	
Контроль мессенджеров			+		—	
Контроль буфера обмена			+		—	
Интеграция с СКУД	—	—	—	—	—	
Контроль облачных хранилищ	+	+	+	—	—	
Анализ поведения пользователя (UBA)	—		—	—	—	—

Отметим, что о некоторых функциях производители не заявляют.

Представленные критерии оценки продуктов отечественного и зарубежного рынка отражают основные тенденции развития DLP-систем. Судя по всему, отечественные решения совершили качественный переход от классических реализаций, исключающих возможность утечки конфиденциальных данных по стандартным каналам коммутации, к своим расширенным представлениям, включающим анализ поведения пользователя (UBA) и его биометрии.

Прототип расширенной системы защиты от утечек на основе анализа активности пользователей компьютерных систем основан на применении следующих ключевых механизмов:

1. Механизм непрерывного мониторинга и понимания контекста поведения пользователя, позволяющий детектировать внутренние нарушения информационной безопасности.

2. Модели искусственного интеллекта для детекции конфиденциальной информации.

3. Дополнительные (опциональные) модели искусственного интеллекта, применяемые для реализации иных действий (непрерывная аутентификация, оценка функционального состояния пользователя и др.). На данный момент они не реализованы.

Механизм непрерывного мониторинга и понимания контекста поведения пользователя

Мониторинг контекста поведения пользователя является первым рубежом для определения атаки или факта хищения конфиденциальной информации в информационно-коммуникационной сети организации. Контекст поведения пользователя – это последовательность действий пользователя в интервале времени. Каждый пользователь имеет свой уникальный типовой портрет работы в системе – контекст поведения, соответствующий ожидаемому.

Прежде всего, на данном уровне определяются потенциально опасные действия пользователя или его потенциально опасное поведение, которое может гипотетически привести к атаке (а в некоторых случаях уже указывает на попытку реализации атаки или нарушение).

Механизм непрерывного мониторинга и понимания контекста поведения пользователя осуществляет мониторинг действий пользователей рабочих станций в локальной сети предприятия. Действие определяется следующими ключевыми аспектами:

1. Кто работает в системе в данный момент (источник действия)? Осуществление непрерывной идентификации и/или верификации личности пользователя в реальном времени по типовому портрету работы в компьютерной системе (в процессе его профессиональной деятельности). Является ли субъект тем, за кого себя выдал при аутентификации (верификация личности), если нет, то кем он является (идентификация личности среди других зарегистрированных пользователей сети организации), либо он является неизвестным субъектом (потенциальным злоумышленником, не зарегистрированным в системе).

2. В каком состоянии находится субъект и адекватно ли его поведение? Субъект может быть в состоянии стресса, усталости, перевозбуждения, опьянения под действием алкоголя и т.д. [10]. Если субъект находится в измененном и потенциально опасном состоянии, это повод обратить пристальное внимание на его действия.

3. Какое конкретно действие совершает пользователь (какой метод воздействия на информацию применяет)? Запускает приложение (какое?), отправляет сообщение в мессенджере, отправляет электронную почту, открывает/копирует/перезаписывает/удаляет файл, фотографирует экран и т.д. Любое значимое действие может нести потенциальную угрозу, в зависимости от контекста.

4. С помощью какого инструмента? Каким приложением или веб-сервисом он пользуется либо это внешнее по отношению к системе событие (например, он фотографирует экран). Данный аспект может быть важен, так как действие может быть инициировано не пользователем, а приложением (например, в случае заражения компьютерным вирусом).

5. По отношению к какому объекту направлено действие? Например, документ (файл), текстовое сообщение, изображение, информация на экране и т.д. Чтобы определить, содержит ли объект чувствительную информацию, следует проводить контентный анализ (для текста – лингвистический, изображение следует сначала преобразовать в текст, схемы на изображении могут быть проанализированы методами распознавания графических образов и т.д.).

6. Кто является получателем информации? Кому или куда отправляется сообщение или документ? Место назначения может определяться сетевым диском, флеш-накопителем, интернет-адресом (например, IP), электронной почтой получателя, именем или логином контакта в мессенджере, а также другим приложением (если информация передается из одного приложения в

другое по какому-то протоколу) и т.д. Получателя также может не быть, например, если информация удаляется или скопирована в буфер обмена.

Для определения почти каждого аспекта может быть применен ИИ – идентификация пользователя, его состояния, регистрация внешних действий пользователя (фотографирование экрана, запись разговора с передачей конфиденциальной информации), классификация объектов по уровню конфиденциальности, в некоторых случаях определение личности получателя путем анализа сообщений, приходящих от него. Исключением являются регистрация внутренних событий (обращения к файлам, отправка данных) и регистрация приложения, от имени которого осуществляется действие.

Действие пользователя можно определить как составное событие – структура с параметрами, которые определяют ответы на шесть вопросов, указанных выше.

Контекст поведения пользователя – это последовательность действий пользователя в интервале времени, т.е. процесс (в терминах математической статистики – случайный процесс), который ограничен промежутком времени и может быть для наглядности представлен в виде матрицы (таблица 3).

Как можно видеть из таблицы 3, каждый аспект характеризуется набором параметров. Некоторые параметры могут являться вероятностными характеристиками. Например, классифицировать сообщение по уровню конфиденциальности можно только на основе ИИ. Также, при фотографировании экрана нельзя быть точно уверенным в том, что на фотографию попал конфиденциальный документ. Однако это можно вероятностно оценить либо с помощью анализа скриншота на основе ИИ, либо просто исходя из контекста (документ имя_файла_3, был открыт, но не был закрыт).

Введем обозначения для каждого параметра события:

U – пользователь (user);

A – приложение (application);

S – состояние (state);

A – метод воздействия (method);

Таблица 3. Контекст поведения пользователя за период времени

Момент времени t	Дата-час: минута: секунда: миллисекунда	Пользователь	Приложение	Состояние	Метод воздействия	Объект	Получатель
1	01.01.2024-12:00:00.001	Иванов (уровень доступа 3)	explorer.exe	Норма	Удаление	Не конф. (имя_файла_1)	—
2	01.01.2024-12:01:10.123	Иванов (уровень доступа 3)	explorer.exe	Норма	Чтение	Не конф. (имя_файла_2)	Иванов (рабочая станция 1)
3	01.01.2024-12:01:46.100	Иванов (уровень доступа 3)	explorer.exe	Опьянение	Чтение	Конф. уровень 1 (имя_файла_3)	Иванов (рабочая станция 1)
4	01.01.2024-12:10:10.011	Иванов (уровень доступа 3)	firefox.exe (веб-сервис telegram)	Опьянение	Отправка сообщения	Конф. уровень 1, вероятность 0.95	Петров (рабочая станция 3)
5	01.01.2024-12:10:30.111	Иванов (уровень доступа 3)	explorer.exe	Опьянение	Копирование	Конф. уровень 1 (имя_файла_3)	Иванов, флеш-накопитель (рабочая станция 1)
6	01.01.2024-12:10:45.001	Иванов (уровень доступа 3)	—	Опьянение	Фотографирование	Предположительно конф. уровень 1 (имя_файла_3, открыт в MSWord)	Иванов
…	…	…		…	…	…	…
1010	01.01.2024-19:00.001	Иванов (уровень доступа 3)	explorer.exe	Норма	Запуск приложения (firefox.exe)	—	—

O – объект (object);

R – получатель (recipient).

Как можно видеть из таблицы 3, представленная в ней ситуация является неоднозначной. С одной стороны, оба пользователя (Иванов и Петров) имеют доступ к файлу имя_файла_3 (их уровень доступа превышает уровень конфиденциальности файла). Однако с момента $t = 3$ состояние пользователя «Иванов» становится измененным (предположительно, так как об этом можно сделать вывод только на основе анализа данных с помощью методов ИИ). Возможно, его поведение неадекватно. Также настораживает, что в этот момент пользователь осуществляет действия с конфиденциальной информацией и передает ее другому пользователю. В данном случае с момента $t = 3$ следует осуществлять архивирование (сохранение) всех данных, которые регистрируются на рабочей станции 1 (связанных с пользователем Иванов). С момента $t = 5$ пользователь Иванов копирует конфиденциальные данные на флеш-накопитель, а это уже может являться нарушением (в зависимости от установленных правил в организации, накопитель также может быть служебным). В момент $t = 6$ скорее всего происходит нарушение. В определенный момент ($t = 1010$), когда пользователь снова находится в адекватном состоянии, архивирование абсолютно всех данных можно прекратить, но только если подозрения были ранее сняты с пользователя Иванов.

Многие нарушения информационной безопасности имеют сложный контекст, в который вовлекается множество пользователей или рабочих станций. Кроме того, развитие событий, предшествующих инциденту, не всегда имеет линейный сценарий, инцидент может не иметь четкого критерия, который бы однозначно указывал на попытку хищения информации. Таким образом, контекст поведения пользователей, являющийся, с одной стороны, результатом анализа первичных данных и событий с помощью специализированных моделей искусственного интеллекта, с другой стороны, должен рассматриваться в качестве входных данных для других моделей искусственного интеллекта (моделей более высокого уровня или метамоделей). Метамодель ИИ (модель машинного обучения) – модель ИИ, созданная для анализа контекста поведения пользователя с целью формирования дальнейших более сложных предсказаний, позволяющих детектировать аномальное поведение пользователей или прогнозировать инциденты определенного типа, в том числе, новые, ранее неизвестные типы инцидентов. Регрессионная или классификационная метамодель ИИ анализирует поток событий контекста поведения пользователя, учитывая определенные параметры этих событий, и распознает или предсказывает инцидент (реализацию атаки в будущем, попытку совершения нарушений информационной безопасности, попытку хищения информации). Такой двухуровневый подход позволяет создать систему превентивного реагирования на инциденты и хищение информации.

Метамоделей может быть множество. Каждая метамодель позволяет распознавать или предсказывать определенный тип нарушения. Кроме того, должен существовать определенный тип метамоделей, нацеленных на идентификацию новых, неизвестных ранее угроз и инцидентов. Данный тип метамоделей должен обладать способностью к онлайн-обучению в процессе функционирования (возможно в автоматизированном режиме, т.е. частично с помощью сотрудников службы информационной безопасности).

Этот отдельный тип непрерывно обучаемых метамоделей ИИ («threat hunter») нацелен по поиск аномалий в контексте поведения пользователей. Онлайн-обучение метамоделей такого рода позволяет постоянно повышать точность предсказаний возникновения новых угроз.

На данный момент метамодели не реализованы. Разработано и реализовано программное ядро системы мониторинга, позволяющее подключать сторонние модели и метамодели для распознавания действий пользователя.

Режимы работы монитора обращений

Следует предусмотреть несколько режимов работы монитора обращений, определяющих, в каком случае имеет смысл усилить контроль, когда следует блокировать действие, а когда – сообщать в службу безопасности. Должна быть предусмотрена возможность настройки правил активации этих режимов.

Предлагается ввести следующие режимы:

1. Режим спокойствия. Осуществляется мониторинг, без сохранения какой-либо информации.
2. Штатный режим. Осуществляется мониторинг с сохранением основных параметров контекста поведения пользователя, необходимых для ведения статистики активности пользователя, процессов и т.д. (идентификаторы/имена пользователя, процессов, состояния, файлов, методов, получателей).
3. Режим внимания (осуществляется мониторинг с сохранением всех параметров контекста поведения пользователя, в том числе, тексты сообщений, но без сохранения подробной инфор-

мации – записей с микрофона, видеозаписей с камеры и т.д.).

4. Режим повышенного внимания (осуществляется мониторинг с сохранением всех данных).

5. Режим тревоги (блокировка действий пользователя или рабочей станции).

Для гибкости любой режим может активироваться как для всего предприятия, так и только для определенных пользователей или рабочих станций. Чтобы настроить этот механизм для конкретного предприятия, предлагается сформировать матрицу переходов из одного состояния в другое, например, как показано в таблице 4. Правила перехода действуют для всей организации, однако могут применяться отдельно для конкретных пользователей и рабочих станций.

Модель для детекции конфиденциальной информации

Основная сложность задачи классификации текстов по уровню конфиденциальности заключается в том, что каждая организация имеет собственную совокупность конфиденциальных данных и документов. Поэтому на практике необходимо применять методы автоматического машинного обучения [11] (AutoML), чтобы настроить языковые модели с учетом специфики чувствительной информации на каждом предприятии. Другим сложным аспектом является возможность распознавания конфиденциальной информации в коротких текстовых сообщениях, перехватываемых DLP-системой.

Прежде всего, текстовый образ необходимо преобразовать в вектор признаков [12]. Каждый образ (текстовое сообщение) состоит из двух полей: тема (состоит из одного предложения) и содержание сообщения (состоит из 1–10 предложений). Для этого поля образа (тема и содержание) обрабатываются большой языковой моделью, которая извлекает вектор признаков из каждого поля отдельно. Мы отдали предпочтение одной из моделей на базе новой архитектуры E5 (название обученной версии модели – goldenrooster/multilingual-e5-large [13]), предложенной как вариант расширения существующих моделей трансформеров (таких как BERT и производные от нее модели) для улучшения задач извлечения информации, многозадачного обучения и поиска. Важнейшей характеристикой модели является ее оптимизация под задачи поиска и классификации,

Таблица 4. Матрица переходов из одного режима в другой

Режим был / Режим стал	Режим спокойствия	Штатный режим	Режим внимания	Режим повышенного внимания	Режим тревоги
Режим спокойствия	—	Активируется службой безопасности	Активируется службой безопасности & S = «Норма»	Активируется службой безопасности & S = «Норма» & О.уровень_конф < 1	—
Штатный режим	Активируется службой безопасности	—	S = «Норма»	S = «Норма» & О.уровень_конф < 1	Служба безопасности сняла ограничения S = «Норма» & О.уровень_конф < 1
Режим внимания	—	S ≠ «Норма»	—	S ≠ «Норма» & О.уровень_конф < 1	Служба безопасности сняла ограничения & S ≠ «Норма» & О.уровень_конф < 1
Режим повышенного внимания	—	S ≠ «Норма» & О.уровень_конф > 1	S ≠ «Норма» & О.уровень_конф > 1	—	Служба безопасности сняла ограничения
Режим тревоги	—	Детекция инцидента	Детекция инцидента	Детекция инцидента	—

реализуемая через использование для обучения смешанного набора данных, включающего задачи ранжирования и тематической классификации.

Основу E5 составляет стек трансформеров с многоголовым механизмом внимания, который позволяет модели выделять наиболее значимые текстовые фрагменты. В рамках трансформеров E5 использует специализированные слои (например, контекстуальные слои внимания) и адаптивные весовые коэффициенты, что позволяет изменять представление текста в зависимости от специфики задачи. Для извлечения признаков E5 задействует модульные уровни обработки, разделяющие текст на семантические блоки, за счет чего обеспечивается эффективное сжатие информации, а также формирование компактных векторных представлений.

Далее два вектора признаков объединяются в один. Вектор признаков нормируется таким образом, чтобы все значения принадлежали интервалу [0; 1]. Нормированный вектор признаков отправляется на вход нейросетевой модели, которая определяет, являются ли данные конфиденциальными.

Конфиденциальные данные – это, как правило, тематически узкая категория, которая ограничивается определенной тематикой или несколькими тематиками. А все что не относится к данным тематикам – это практически любой текст. Поэтому подобрать репрезентативную выборку не конфиденциальных данных достаточно сложно, а обучить классификатор на них проблематично. Любой текст, который стилистически или тематически выбивается из обучающей выборки, может фактически быть отнесен случайным образом к любому из классов.

По указанным причинам сначала требуется отделить конфиденциальные сообщения от любых других. В данном случае придется проводить обучение только на конфиденциальных данных. В этом отношении интересные свойства наблюдаются у автокодировщиков. Это архитектура нейронной сети, которая состоит из двух сегментов – кодировщика и декодировщика. Кодировщик сжимает данные до компактного представления, а декодировщик их восстанавливает. Обучается сеть как единое целое, при этом на вход и выход подаются одни и те же данные – вектор признаков. Часто автокодировшики используются для извлечения признаков [14].

Наша гипотеза заключается в том, что если переобучить автокодировщик, то он будет хорошо сжимать и восстанавливать данные, которые «видел» и очень близкие к ним (конфиденциальная информация). При этом он будет плохо восстанавливать данные, которые он «не видел» (любая другая информация). Среднее арифметическое модулей отклонений (СМО) выходов нейронов последнего слоя автокодировщика от соответствующих входов можно рассматривать как меру близости, позволяющую отличать обычные данные от конфиденциальных. При этом в ходе вычислительного эксперимента необходимо найти оптимальное пороговое значение СМО, балансирующее вероятности ошибок 1-го и 2-го рода (Er1 – ложное принятие обычных данных за конфиденциальные, Er2 – ложный пропуск конфиденциальных данных, соответственно). Критерием оптимальности при выставлении порога может служить:

– Equal Error Rate (EER) – когда вероятности ошибок 1-го и 2-го рода примерно равны;

– минимальная сумма вероятностей ошибок 1-го и 2-го рода;

– достаточно низкий уровень ошибок 2-го рода, стремящийся к нулю (порог, при котором не зафиксировано ложных пропусков конфиденциальных данных по результатам эксперимента на рассматриваемом наборе достаточно большого объема данных).

В настоящей работе решено использовать третий критерий, так как цена ошибки 2-го рода гораздо выше, чем цена ошибки 1-го рода.

Рисунок 1. Архитектура автокодировщика

В основе автокодировщика решено использовать сверточные слои на базе одномерных и двумерных сверток. Сверточные сети успешно применяются не только для анализа изображений, но и временных рядов [15].

Мы предложили архитектуру автокодировщика, представленную на рисунке 1.

Оценка надежности детекции конфиденциальной информации

Проведен эксперимент по оценке надежности распознавания конфиденциальной информации. Для проведения эксперимента мы сформировали набор данных 10 организаций на русском языке с общим объемом 310 тысяч слов (примерно в равном соотношении на каждую организацию и уровень конфиденциальности). Из этих данных было сгенерировано по 100000 образов с конфиденциальной информацией для каждой организации (всего 1 млн. образов). Также мы подготовили 100000 образов (общих для всех организаций), не относящихся к конфиденциальной информации, сформированных из текстов на русском языке по тематике – финансы, музыка, фильмы, информационные технологии (по 25000 для каждой тематики). Сформированный корпус данных использовался в дальнейших экспериментах и делился на обучающую, валидационную и тестовую выборки в соответствии с задачами эксперимента.

Тестирование выполнено для данных 10 организаций, для каждой из которых обучался отдельный автокодировщик. В качестве обучающей выборки каждого автоэнкодера использовано 80000 конфиденциальных образов соответствующей организации. В качестве тестовых данных использовано:

– 20 000 других конфиденциальных образов организации;

– 20 000 случайных образов, не относящихся к конфиденциальной информации (идентичных для всех организаций).

Таким образом, всего проведено 400 000 опытов.

Исходя из практики работы автокодировщика, отметим, что для его устойчивой работы требуется обучающая выборка объемом не менее 20 000 размеченных конфиденциальных образов на организацию. При меньшем количестве данных наблюдается снижение точности классификации и рост числа ложных срабатываний, особенно при наличии тематического разнообразия. Это обусловлено тем, что модель обучается только на положительном классе и требует репрезентативного охвата различных форм представления конфиденциальной информации. В реальных условиях с ограниченным количеством данных рекомендуется либо использовать предварительно обученные модели, либо применять подходы активного и полуавтоматического обучения.

Для обучения использовался оптимизатор Adam. В соответствии с критерием оптимальности определялся порог, при котором для всех организаций ошибок второго рода зафиксировано не было ($Er2 \approx 0$). Значение ошибки 1-го рода составило $Er1=5{,}1\%$ (рисунок 2). Экспериментально установлено, что 300 эпох достаточно для обучения автокодировщика.

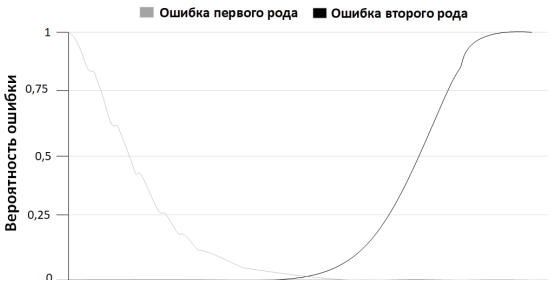

Рисунок 2. Результаты обучения автокодировщика

Несмотря на положительные результаты, следует учитывать потенциальные ограничения предложенного подхода, вытекающие из его специфики. В частности, возможны ложные срабатывания на тексты, близкие по тематике и структуре к конфиденциальным, а также ограниченная способность модели к обобщению на ранее не встречавшиеся типы данных. Хотя подобные случаи не были зафиксированы в ходе проведенных экспериментов, вероятность их возникновения сохраняется и требует дополнительного изучения при внедрении системы в реальных условиях.

Заключение

Стоит отметить, что в последнее время DLP, UAM (и EM) и UEBA решения выступают как комплементарные системы, дополняющие и расширяющие возможности друг друга, а также осуществляющие гораздо более надежную защиту от инсайдерских угроз. UEBA позволяют организациям обнаруживать в своей сети сложные угрозы нулевого дня и внутренние угрозы, которые, скорее всего, не смогли бы обойти традиционные инструменты безопасности. Чтобы обнаружить эти угрозы, решения UEBA не только создают базовые показатели «нормального» поведения пользователей, но и дополнительно анализируют поведение так называемых сущностей: конечных точек, приложений, серверов, маршрутизаторов, узлов и хранилищ данных. В таком случае и люди, и сущности превращаются в объекты ИБ,

обладающие некоторым эталонным поведением, отклонение от которого можно расценивать как потенциально опасное. Для описанного анализа UEBA используют комбинацию алгоритмов машинного и глубокого обучения, а также статистический анализ.

Ключевое отличие UEBA от рассмотренных расширенных DLP-систем состоит в том, что DLP-системы, а также их расширенные реализации, направлены на предотвращение утечки конфиденциальных данных за пределы организации, в то время как UEBA анализирует поведение пользователей и сущностей с целью выявления потенциальных, зачастую заранее не обозначенных, инцидентов безопасности. Таким образом, с одной стороны, UEBA направлены на предотвращение более широкого спектра угроз, а с другой, являются узконаправленными в отношении специфической поведенческой аналитики объектов ИБ.

В настоящем исследовании разработан прототип расширенной DLP-системы и модель для детекции конфиденциальной информации в коротких текстовых сообщениях. Для извлечения признаков используется предварительно обученная нейронная сеть на базе архитектуры E5, поддерживающая множество языков. Для детекции конфиденциальной информации применяется сверточнвая нейронная сеть типа автокодировщик. Система обучается автоматически на конфиденциальных документах каждой организации. Документы, не являющиеся конфиденциальными, не используются при обучении. Установлено, что применение автокодировщика успешно позволяет производить бинарную классификацию сообщений с вероятностью ошибок 1-го рода 5,1% при отсутствии ошибок 2-го рода в рамках эксперимента с использованным нами набором данных (другой набор данных не гарантирует отсутствия ошибок).

Дальнейшие исследования будут направлены на снижение объемов обучающей выборки при сохранении или повышении точности классификации.

Работа выполнена федеральным государственным автономным образовательным учреждением высшего образования «Омский государственный технический университет» в рамках государственного задания Минобрнауки России на 2023-2025 годы (FSGF-2023-0004).

Литература

1. Средняя стоимость ущерба от утечек данных // Infowatch. URL: https://www.infowatch.ru/analytics/daydzhesty-i-obzory/srednyaya-stoimost-uscherba-ot-utechek-dannykh (дата обращения: 11.09.2024).
2. Pervasive runtime monitoring for detection and assessment of emerging hazards for advanced UAM Systems / C. Elks [et al.] // AIAA AVIATION 2022 Forum. Chicago, 2022. P. 3541. DOI: 10.2514/6.2022-3541
3. Khan M.Z.A., Khan M.M., Arshad J. Anomaly detection and enterprise security using user and entity behavior analytics (UEBA) // 2022 3rd International Conference on Innovations in Computer Science & Software Engineering (ICONICS). Karachi, Pakistan, 2022. P. 1–9. DOI: 10.1109/ICONICS56716.2022.10100596
4. Аусилова Н.М., Зарынбеков А.Б., Ахмет Г.Б. Применение DLP-систем как инструмента обеспечения информационной безопасности // Наука и реальность. 2023. №. 1 (13). С. 93–96.
5. Manoharan P. Supervised Learning for Insider Threat Detection: PhD Thesis. Australia, Victoria University, 2024. 184 p.
6. Сагидова М.Л. Современные системы контроля и управления доступом // Международный журнал гуманитарных и естественных наук. 2022. № 9-1 (72). С. 64–68. DOI: 10.24412/2500-1000-2022-9-1-64-68
7. Александров А.С. Преимущества практической реализации альтернативной концепции развития DLP-систем // Достижения науки и технологий-ДНиТ-11-2023: материалы II Всероссийской научной конференции. Красноярск: Общественное учреждение «Красно-ярский краевой Дом науки и техники Российского союза научных и инженерных обще-ственных объединений», 2023. С. 389–394. DOI: 10.47813/dmt-n.2023.7.389—394
8. Hussain M.E., Hussain R. Cloud security as a service using data loss prevention: Challenges and solution // International Conference on Internet of Things and Connected Technologies. Cham: Springer International Publishing, 2021. P. 98–106.
9. Лабазин М.А., Фатеев А.Г. Применение средств контроля и предотвращения утечки информации // Инжиниринг и технологии. 2020. Т. 5, № 1. С. 7–11. DOI: 10.21685/2587-7704-2020-5-1-2
10. Жумажанова С.С., Сулавко А.Е., Ложников П.С. Распознавание психофизиологиче-ского состояния субъектов-операторов на основе анализа термографических изображений лица с применением сверточных нейронных сетей // Системная инженерия и информационные технологии. 2023. Т. 5, № 2 (11). С. 41–55.

DOI: 10.54708/2658-5014-SIIT-2023-no2-p41
11. AutoML to date and beyond: Challenges and opportunities / S.K. Karmaker [et al.] //ACM Computing Surveys (CSUR). 2021. Vol. 54, no. 8. P. 1–36. DOI: 10.1145/3470918
12. Grohe M. word2vec, node2vec, graph2vec, X2vec: Towards a theory of vector embeddings of structured data // Proceedings of the 39th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems. 2020. P. 1–16. DOI: 10.1145/3375395.3387641
13. Goldenrooster_Multilingual-E5-large // Hugging Face. URL: https://huggingface.co/goldenrooster/multilingual-e5-large (дата обращения: 11.09.2024).
14. Sulavko A. Biometric-based key generation and user authentication using acoustic characteristics of the outer ear and a network of correlation neurons // Sensors. 2022. Vol. 22, no. 23. P. 9551. DOI: 10.3390/s22239551
15. Evaluation of EEG identification potential using statistical approach and convolutional neural networks / A.E. Sulavko [et al.] // Информационно-управляющие системы. 2020. № 6. С. 37–49. DOI: 10.31799/1684-8853-2020-6-37-49.

Дата поступления 24.02.2025
Дата принятия 25.03.2025

Сулавко Алексей Евгеньевич, д.т.н., профессор кафедры комплексной защиты информации (КЗИ) Омского государственного технического университета (ОмГТУ). 644050, Российская Федерация, г. Омск, ул. Мира, 11. Тел. +7 953 394-90-54. E-mail: sulavich@mail.ru

Панфилова Ирина Евгеньевна, старший преподаватель кафедры электронных систем и информационной безопасности Самарского государственного технического университета. 443100, Российская Федерация, г. Самара, ул. Молодогвардейская, 244. Тел. +7 917 953-83-87. E-mail: panfilova_2015@bk.ru

Варкентин Юрий Андреевич, инженер кафедры КЗИ ОмГТУ. 644050, Российская Федерация, г. Омск, ул. Мира, 11. Тел. +7 996 122-03-89. E-mail: varkentinyuri@gmail.com

Клиновенко Сергей Александрович, аспирант кафедры КЗИ ОмГТУ. 644050, Российская Федерация, г. Омск, ул. Мира, 11. Тел. +7 3812 95-79-17. E-mail: sergey.klinovenko@gmail.com

UDC 004.056.53
DOI: 10.18469/ikt.2025.23.1.08

PROTOTYPE OF LEAKAGE PROTECTION SYSTEM AND NEURAL NETWORK MODEL FOR DETECTION OF CONFIDENTIAL TEXT MESSAGES

Sulavko A.E.[1], Panfilova I.E.[2], Varkentin Yu.A.[1], Klinovenko S.A.[1]
[1]Omsk State Technical University, Omsk, Russian Federation
[2]Samara State Technical University, Samara, Russian Federation
E-mail: sulavich@mail.ru, panfilova_2015@bk.ru, varkentinyuri@gmail.com, sergey.klinovenko@gmail.com

The study raises the problem of confidential information leaks arising from the actions of internal intruders. Modern software products, including those aimed at protecting confidential data by analyzing user behavior, are considered as methods of countering such attacks. To solve the problem, we propose a prototype of an advanced DLP-system that uses artificial intelligence methods to analyze user activity and prevent such leaks. The developed prototype implements a mechanism for monitoring the context of user behavior and uses machine learning models to detect sensitive information in short text messages. Specifically, a pre-trained neural network based on E5 architecture is used for feature extraction. For confidential information detection, a convolutional neural network of autoencoder type is applied, which is trained exclusively on data containing confidential documents. Experimental results have shown that the proposed model successfully performs binary classification of messages, with the error of the second kind at the level of 5.1% and the absence of errors of the first kind (within the framework of the conducted experiment on its own data set). The proposed complex can serve as a basis for the creation of more complex solutions in the field of information security.

Keywords: *confidential information leakage, data theft, insider threats, confidential documents, artificial intelligence, autoencoder, machine learning, behavioral analytics*

Sulavko Alexey Evgenyevich, Omsk State Technical University, 11, Mira Street, Omsk, 644050, Russian Federation; Professor of Complex Information Protection Department, Doctor of Technical Science. Tel. +7 953 394-90-54. E-mail: sulavich@mail.ru

Panfilova Irina Evgenyevna, Samara State Technical University, 244, Molodogvardeyskaya Street, 443100, Russian Federation; Senior Teacher of Electronic Systems and Information Security Department. Tel. +7 917 953-83-87. E-mail: panfilova_2015@bk.ru

Varkentin Yuriy Andreevich, Omsk State Technical University, 11, Mira Street, Omsk, 644050, Russian Federation; Engineer of Complex Information Protection Department. Tel. +7 996 122-03-89. E-mail: varkentinyuri@gmail.com

Klinovenko Sergey Aleksandrovich, Omsk State Technical University, 11, Mira Street, Omsk, 644050, Russian Federation; PhD Student of Complex Information Protection Department. Tel. +7 3812 95-79-17. E-mail: sergey.klinovenko@gmail.com

References

1. The average cost of damage from data leakage. *Infowatch*. URL: https://www.infowatch.ru/analytics/daydzhesty-i-obzory/srednyaya-stoimost-uscherba-ot-utechek-dannykh (accessed: 11.09.2024). (In Russ.)
2. Elks C. et al. Pervasive runtime monitoring for detection and assessment of emerging hazards for advanced UAM systems. *AIAA AVIATION 2022 Forum*. Chicago, 2022, pp. 3541. DOI: 10.2514/6.2022-3541
3. Khan M.Z.A., Khan M.M., Arshad J. Anomaly detection and enterprise security using user and entity behavior analytics (UEBA). *2022 3rd International Conference on Innovations in Computer Science & Software Engineering* (*ICONICS*). Karachi, Pakistan, 2022, pp. 1–9. DOI: 10.1109/ICONICS56716.2022.10100596
4. Ausilova N.M., Zarynbekov A.B., Akhmet G.B. The use of DLP systems as an information security tool. *Nauka i real'nost'*, 2023, no. 1 (13), pp. 93–96. (In Russ.)
5. Manoharan P. Supervised Learning for Insider Threat Detection: PhD Thesis. Australia, Victoria University, 2024, 184 p.
6. Sagidova M.L. Modern access control and management systems. *Mezhdunarodnyj zhurnal gumanitarnyh i estestvennyh nauk*, 2022, no. 9–1 (72), pp. 64–68. DOI: 10.24412/2500-1000-2022-9-1-64-68 (In Russ.)
7. Aleksandrov A.S. Advantages of the practical implementation of the alternative concept for the development of DLP systems. *Dostizheniya nauki i tekhnologij-DNiT-11-2023: materialy II Vserossijskoj nauchnoj konferencii*. Krasnoyarsk: Obshchestvennoe uchrezhdenie «Krasnoyarskij kraevoj Dom nauki i tekhniki Rossijskogo soyuza nauchnyh i inzhenernyh obshchestvennyh ob"edinenij», 2023, pp. 389–394. DOI: 10.47813/dmt-n.2023.7.389–394 (In Russ.)
8. Hussain M.E., Hussain R. Cloud security as a service using data loss prevention: Challenges and solution. *International Conference on Internet of Things and Connected Technologies*. Cham: Springer International Publishing, 2021, pp. 98–106.
9. Labazin M.A., Fateev A.G. The use of data loss prevention systems. *Inzhiniring i tekhnologii*, 2020, vol. 5, no. 1, pp. 7–11. DOI: 10.21685/2587-7704-2020-5-1-2 (In Russ.)
10. ZHumazhanova S.S., Sulavko A.E., Lozhnikov P.S. Recognition of the psychophysiological state of subject-operators based on the analysis of thermographic images of the face using convolutional neural networks. *Sistemnaya inzheneriya i informacionnye tekhnologii*, 2023, vol. 5, no. 2 (11), pp. 41–55. DOI: 10.54708/2658-5014-SIIT-2023-no2-p41 (In Russ.)
11. Karmaker S.K. et al. AutoML to date and beyond: Challenges and opportunities. *ACM Computing Surveys* (*CSUR*), 2021, vol. 54, no. 8, pp. 1–36. DOI: 10.1145/3470918

12. Grohe M. word2vec, node2vec, graph2vec, X2vec: Towards a theory of vector embeddings of structured data. *Proceedings of the 39th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, 2020, pp. 1–16. DOI: 10.1145/3375395.3387641
13. Goldenrooster_Multilingual-E5-large. *Hugging Face*. URL: https://huggingface.co/goldenrooster/multilingual-e5-large (accessed: 11.09.2024).
14. Sulavko A. Biometric-based key generation and user authentication using acoustic characteristics of the outer ear and a network of correlation neurons. *Sensors*, 2022, vol. 22, no. 23, pp. 9551. DOI: 10.3390/s22239551
15. Sulavko A.E. et al. Evaluation of EEG identification potential using statistical approach and convolutional neural networks. *Informatsionno-upravliaiushchie sistemy*, 2020, no. 6, pp. 37–49. DOI: 10.31799/1684-8853-2020-6-37-49

Received 24.02.2025
Accepted 25.03.2025

Реклама