DOI: 10.18137/RNU.V9187.25.01.P.112 УДК 004.056

Ма Анастасия Алексеевна
магистрант факультета программной инженерии и компьютерной техники, Научно-исследовательский университет ИТМО, Санкт-Петербург. ORCID: 0009-0009-4942-4211
Электронный адрес: kryukovskayaanastacia@yandex.ru
Anastasia A. Ma
Master's students at the Faculty of software engineering and computer systems, ITMO University, Saint Petersburg. ORCID: 0009-0009-4942-4211
E-mail address: kryukovskayaanastacia@yandex.ru

Авксентьева Елена Юрьевна
кандидат педагогических наук, доцент, доцент факультета программной инженерии и компьютерной техники, Научно-исследовательский университет ИТМО, Санкт-Петербург. ORCID: 0000-00015000-4868, Author ID: 559672, SPIN-код: 2688-1540.
Электронный адрес: eavksenteva@itmo.ru
Elena Yu. Avksentieva
Ph.D. of Pedagogical Sciences, Docent, Associate Professor at the Faculty of software engineering and computer systems, ITMO University, Saint Petersburg. ORCID: 0000-00015000-4868, AuthorID: 559672, SPIN-code: 2688-1540.
E-mail address: eavksenteva@itmo.ru

ВЛИЯНИЕ ОБЪЕМА ДАННЫХ НА ТОЧНОСТЬ ОБНАРУЖЕНИЯ АНОМАЛИЙ В СЕТЕВОМ ТРАФИКЕ: ИССЛЕДОВАНИЕ ЭФФЕКТА БОЛЬШИХ ДАННЫХ

Аннотация. В статье рассматривается использование алгоритмов машинного обучения для обнаружения аномалий на основе набора данных CICIDS2017, который был специально разработан для имитации реальных сценариев сетевых атак. Особое внимание уделено трем популярным алгоритмам: логистической регрессии, случайному лесу и нейронным сетям. Эти алгоритмы были выбраны благодаря своей способности эффективно обрабатывать большие объемы данных и выявлять сложные паттерны. В рамках статьи проведена серия экспериментов, в которых будут варьироваться объем обучающих данных и оцениваться производительность моделей как на чистых, так и на зашумленных данных. Результаты данного исследования помогут понять, как различные алгоритмы реагируют на изменения в объеме данных и качество входной информации, что является важным аспектом для разработки эффективных систем кибербезопасности.

Ключевые слова: аномалии сетевого трафика, машинное обучение, эффект больших данных, нейронные сети, случайный лес, логистическая регрессия.

Для цитирования: *Ма А.А., Авксентьева Е.Ю.* Влияние объема данных на точность обнаружения аномалий в сетевом трафике: исследование эффекта больших данных // Вестник Российского нового университета. Серия: Сложные системы: модели, анализ и управление. 2025. № 1. С. 112 – 126. DOI: 10.18137/RNU.V9187.25.01.P.112

THE EFFECT OF DATA VOLUME ON THE ACCURACY OF DETECTING ANOMALIES IN NETWORK TRAFFIC: EXPLORING THE BIG DATA EFFECT

Abstract. The article discusses the use of machine learning algorithms to detect anomalies based on the CICIDS2017 dataset, which was specifically designed to simulate real-world network attack scenarios. Special attention is paid to three popular algorithms: logistic regression, random forest and neural networks. These algorithms were chosen due to their ability to efficiently process large amounts of data and identify complex patterns. Within the framework of this article, a series of experiments has been conducted in which the amount of training data will vary and the performance of models will be evaluated, both on pure and noisy data. The results of this study will help to better understand how different algorithms respond to changes in the amount of data and the quality of input information, which is an important aspect for developing effective cybersecurity systems.

Keywords: network traffic anomalies, machine learning, big data effect, neural networks, random forest, logistic regression.

For citation: Ma A.A., Avksentieva E.Yu. (2025) The effect of data volume on the accuracy of detecting anomalies in network traffic: Exploring the big data effect. *Vestnik of Russian New University. Series: Complex Systems: Models, Analysis and Management.* No. 1. Pp. 112 – 126. DOI: 10.18137/RNU.V9187.25.01.P.112 (In Russian).

Введение

В современном мире обнаружение аномалий в сетевом трафике становится одной из ключевых задач для защиты информационных систем от различных угроз. С увеличением объема и сложности сетевых данных возрастает необходимость в эффективных методах анализа и классификации, способных выявлять подозрительные активности и предотвращать атаки.

Актуальность темы обнаружения аномалий в сетевом трафике также подчеркивается необходимостью разработки адаптивных систем, которые могут эффективно реагировать на изменения в сетевом трафике и новые виды атак. Исследование влияния объема данных на производительность алгоритмов является важным шагом к созданию более надежных и эффективных систем защиты [1–3].

Обзор литературы

Тема обнаружения аномалий в сетевом трафике с учетом объема данных остается чрезвычайно актуальной в свете роста информационных потоков и новых требований к точности и производительности анализа [4]. Современные исследования подчеркивают необходимость адаптации методов к условиям больших данных, что позволяет улучшить качество анализа и своевременно выявлять угрозы. В данном обзоре рассмотрены работы отечественных и зарубежных авторов, опубликованные в последние годы, которые освещают различные аспекты влияния объема данных на эффективность обнаружения аномалий.

Отечественные исследования. Российские исследования показывают, что анализ больших объемов сетевого трафика требует новых подходов к обработке данных и оптимизации алгоритмов. Так, в работе И.А. Ушакова и соавторов [5] показано, что методы машинного обучения адаптируются к растущим объемам данных, повышая точность обнаружения вторжений и снижая ложные срабатывания.

Исследование В.Н. Труфанова и соавторов [6] показывает, что использование методов машинного обучения для анализа больших объемов сетевого трафика позволяет эффек-

Вестник Российского нового университета
Серия «Сложные системы: модели, анализ и управление», выпуск 1 за 2025 год

тивно выявлять аномалии с высокой точностью, а также описаны сильные и слабые стороны применения этих решений.

Зарубежные исследования. Зарубежные исследования также подчеркивают важность больших данных в повышении точности обнаружения аномалий. Так, в работе Song Shijun и Fan Min [7] показано, что случайный лес повышает точность обнаружения аномалий в больших данных и снижает ложные срабатывания.

Исследование 2025 года, выполненное S. Ness и соавторами [8], посвящено методам машинного обучения, которые эффективно выявляют аномалии в сетевом трафике, что важно для сетевой безопасности.

N. Abinaya, A.V. Senthil Kumar и соавторы [9] в 2024 году рассмотрели применение методов анализа больших данных для обнаружения аномалий в сетевом трафике. Они показали, что современные методы машинного обучения, адаптированные для больших объемов данных, эффективно выявляют сложные аномалии, что критично для кибербезопасности.

C. Cavallaro и соавторы [10] обсуждают подходы к обнаружению аномалий в больших данных с использованием машинного обучения и метаэвристик. Авторы отмечают, что автоматизированная классификация повышает точность прогнозирования и снижает затраты, что важно для предотвращения кибератак.

Таким образом, современные исследования подтверждают, что увеличение объема данных является одним из решающих факторов в повышении точности алгоритмов обнаружения аномалий. Важно отметить, что при работе с большими данными необходима соответствующая адаптация алгоритмов и инфраструктуры.

Понимание того, как объем и качество данных влияют на производительность алгоритмов машинного обучения, является ключевым аспектом для разработки эффективных систем обнаружения аномалий.

Выбор данных

Для выбора подходящего набора данных исследования влияния объема данных на точность обнаружения аномалий были рассмотрены несколько известных наборов данных, используемых в области кибербезопасности и анализа сетевого трафика. Ниже приведены критерии, по которым сравнивались эти наборы данных, а также таблица с результатами сравнения. Критерии сравнения наборов данных:

- *объем данных* – количество записей в наборе данных, что позволяет оценить влияние объема на производительность моделей;
- *разнообразие аномалий* – наличие различных типов атак и нормального трафика, что необходимо для обучения моделей на разнообразных сценариях;
- *реалистичность* – насколько хорошо набор данных отражает реальные условия сетевого трафика и атак;
- *структурированность* – наличие хорошо организованных и структурированных данных с четкими признаками, что упрощает процесс предобработки;
- *доступность* – наличие набора данных для общественного использования, что позволяет другим исследователям воспроизводить результаты.

Выбор набора данных CICIDS2017 для исследования обусловлен его уникальными характеристиками, которые делают его особенно подходящим для анализа влияния объема данных на точность обнаружения аномалий в сетевом трафике (см. Таблицу).

Таблица

Сравнение датасетов по заданным критериям

Набор данных	Объем данных (количество записей, млн)	Разнообразие аномалий	Реалистичность	Структурированность	Доступность
CICIDS2017	2,8	DDoS, Brute Force, SQL Injection и др.	Высокая	Хорошо структурированные данные	Открытый доступ
KDD Cup 1999	4,9	22 типа атак	Умеренная	Структурированные данные	Открытый доступ
NSL-KDD	0,13	22 типа атак	Умеренная	Хорошо структурированные данные	Открытый доступ
UNSW- NB15	2,5	9 типов атак	Высокая	Хорошо структурированные данные	Открытый доступ
CICIDS 2018	1,2	DDoS, DoS, Brute Force и др.	Высокая	Хорошо структурированные данные	Открытый доступ

Источник: таблица составлена авторами на основе данных из [6].

CICIDS2017 был создан Канадским институтом кибербезопасности и включает в себя более 2,8 млн записей, что предоставляет обширный объем данных для анализа. Набор данных содержит разнообразные типы атак, такие как DDoS, BruteForce и SQL Injection, что позволяет моделям обучаться на множестве сценариев, отражающих реальные угрозы. Это разнообразие аномалий важно для повышения устойчивости и точности алгоритмов обнаружения. Кроме того, CICIDS2017 отличается высокой реалистичностью, так как данные были собраны в условиях, максимально приближенных к реальному сетевому трафику. Это позволяет точнее оценить эффективность алгоритмов в реальных условиях. Структурированность данных также играет важную роль: набор содержит множество четко определенных признаков, что упрощает процесс предобработки и анализа данных. Наконец, доступность набора данных для общественного использования делает его идеальным выбором для исследователей и практиков, позволяя воспроизводить результаты и делиться методологиями.

Таким образом, набор данных CICIDS2017 был выбран из-за его большого объема, разнообразия аномалий, реалистичности, структурированности и доступности, что делает его оптимальным для исследования влияния объема данных на точность обнаружения аномалий в сетевом трафике.

Предобработка данных

Предобработка данных является критически важным этапом в любом исследовании, связанном с анализом и машинным обучением. В контексте обнаружения аномалий в сетевом трафике качество и подготовленность данных напрямую влияют на точность и эффективность обучаемых моделей. Набор данных CICIDS2017, используемый в данном исследовании, содержит большое количество записей, что делает его подходящим для анализа. Однако, как и любой другой набор данных, он требует внимательной предобработки.

Вестник Российского нового университета
Серия «Сложные системы: модели, анализ и управление», выпуск 1 за 2025 год

На данном этапе необходимо решить несколько ключевых задач. Во-первых, необходимо устранить бесконечные значения и заполнить пропуски, чтобы обеспечить корректность анализа. Во-вторых, потребуется кодировать категориальные переменные, чтобы алгоритмы машинного обучения могли корректно интерпретировать данные. Наконец, нормализация признаков поможет привести данные к единому масштабу, что особенно важно для алгоритмов, чувствительных к масштабу, таких как логистическая регрессия и нейронные сети.

Экспериментальная установка

Подробно опишем экспериментальную установку, использованную для исследования влияния объема данных на точность обнаружения аномалий в сетевом трафике с использованием набора данных CICIDS2017. Экспериментальная установка включает в себя этапы предобработки данных, создание различных наборов данных, выбор алгоритмов машинного обучения, а также методы оценки производительности моделей.

Загрузка и предобработка данных

Первым шагом в эксперименте является загрузка набора данных CICIDS2017. Данные содержат как нормальный, так и аномальный трафик, записанный в условиях, приближенных к реальным. После загрузки данных проводится предобработка, которая включает несколько ключевых этапов:

Замена бесконечных значений и пропусков. В процессе анализа данных могут встречаться бесконечные значения или пропуски. Эти значения заменяются на NaN, а затем заполняются нулями или средними значениями соответствующих признаков. Это гарантирует, что алгоритмы машинного обучения не столкнутся с ошибками во время обучения.

Кодирование меток классов. Поскольку набор данных содержит категориальные переменные, такие как метки классов (например, BENIGN, DDoS), они кодируются в числовой формат с использованием метода LabelEncoder. Это позволяет алгоритмам машинного обучения корректно интерпретировать данные.

Нормализация данных. Признаки нормализуются с помощью Standard Scaler, что приводит их к единому масштабу. Нормализация особенно важна для алгоритмов, чувствительных к масштабу признаков, таких как логистическая регрессия и нейронные сети.

Создание наборов данных различного объема

После предобработки данных создаются несколько наборов данных с различным объемом. Для этого используется метод train_test_split, который позволяет разделить данные на обучающую и тестовую выборки. Подобраны размеры обучающих выборок: 10, 25, 50 и 75 % от общего объема данных. Это позволяет исследовать, как изменение объема данных влияет на точность моделей.

Введение искусственного шума

Для оценки устойчивости моделей к искажениям в данных добавляется искусственный шум в часть обучающих данных. Это достигается с помощью функции, которая генерирует случайные значения, добавляемые к признакам данных. Этот шаг важен, поскольку в реальных условиях данные могут содержать ошибки или шумы, и понимание того, как это влияет на производительность алгоритмов, критично для разработки надежных систем обнаружения аномалий.

Оценка производительности моделей

Для оценки производительности моделей используются стандартные метрики, такие как точность, полнота и F1-мера. Каждая модель обучается на чистых и зашумленных данных, после чего производится оценка их точности на тестовой выборке. Результаты сравниваются для выявления влияния объема данных и качества данных на производительность моделей.

Экспериментальная установка, описанная выше, обеспечивает структурированный подход к исследованию влияния объема данных на точность обнаружения аномалий в сетевом трафике. Каждый этап, начиная от предобработки данных и заканчивая оценкой производительности, играет важную роль в получении надежных и воспроизводимых результатов, что критично для дальнейшего развития методов обнаружения аномалий.

Алгоритмы

Рассмотрим три основных алгоритма машинного обучения, используемых для обнаружения аномалий в сетевом трафике: логистическая регрессия, случайный лес и нейронные сети.

Логистическая регрессия – простой и эффективный алгоритм, который используется для бинарной классификации. Он основан на модели, которая предсказывает вероятность принадлежности объекта к определенному классу. В отличие от линейной регрессии логистическая регрессия использует логистическую функцию (сигмоидную), чтобы ограничить предсказанные значения в диапазоне от 0 до 1.

Случайный лес – ансамблевый метод, который использует множество деревьев решений для улучшения точности предсказаний. Каждое дерево в лесу обучается на случайной подвыборке данных, и итоговое предсказание формируется путем голосования между всеми деревьями.

Нейронные сети – мощные инструменты для решения задач классификации, которые имитируют работу человеческого мозга. Они состоят из слоев взаимосвязанных нейронов, которые обрабатывают входные данные и обучаются на основе примеров. Нейронные сети могут быть как простыми (один скрытый слой), так и глубокими (много скрытых слоев).

Каждый из этих алгоритмов имеет свои сильные и слабые стороны, что делает их подходящими для различных сценариев. В зависимости от характеристик данных и требований к модели выбор конкретного алгоритма может существенно повлиять на точность и эффективность обнаружения аномалий.

Обоснование выбора алгоритмов

В данном исследовании были выбраны три алгоритма машинного обучения: логистическая регрессия, случайный лес и нейронные сети. Каждый из этих алгоритмов был выбран на основе своих уникальных характеристик, которые делают их подходящими для задачи обнаружения аномалий в сетевом трафике.

Логистическая регрессия была выбрана благодаря своей простоте и интерпретируемости. Этот алгоритм позволяет быстро обучать модели и получать ясные результаты, что особенно важно в контексте кибербезопасности, где необходимо объяснять, почему система приняла то или иное решение. Логистическая регрессия хорошо работает на линейно разделимых данных, что делает ее эффективной для базового анализа и как отправной точки для более сложных моделей [11; 12].

Вестник Российского нового университета
Серия «Сложные системы: модели, анализ и управление», выпуск 1 за 2025 год

Случайный лес был выбран за его высокую точность и устойчивость к переобучению. Этот ансамблевый метод использует множество деревьев решений, что позволяет ему обрабатывать сложные зависимости в данных и улучшать производительность по сравнению с одиночными деревьями. Случайный лес также хорошо справляется с большими объемами данных и множеством признаков, что делает его идеальным для работы с набором данных CICIDS2017, который содержит много различных характеристик сетевого трафика [13; 14].

Нейронные сети были выбраны из-за их способности моделировать сложные паттерны в данных. Они могут эффективно обрабатывать большие объемы информации и выявлять скрытые зависимости, что делает их особенно полезными для задач, связанных с обнаружением аномалий. Нейронные сети могут адаптироваться к различным типам данных и хорошо работают в ситуациях, когда другие алгоритмы могут не справляться [15].

Таким образом, выбор логистической регрессии, случайного леса и нейронных сетей основан на их уникальных преимуществах и способности решать задачи классификации в контексте обнаружения аномалий в сетевом трафике [16]. Эти алгоритмы обеспечивают разнообразие подходов, что позволяет более полно оценить влияние объема данных на точность моделей и их устойчивость к шумам.

Random Forest

График (см. Рисунок 1) иллюстрирует влияние объема данных на точность модели Random Forest для двух типов данных: чистых (Clean Data) и шумных (Noisy Data). По оси абсцисс (горизонтальной) отображается объем данных, используемых для обучения модели, выраженный в процентах от исходного набора данных, а по оси ординат (вертикальной) – точность модели.

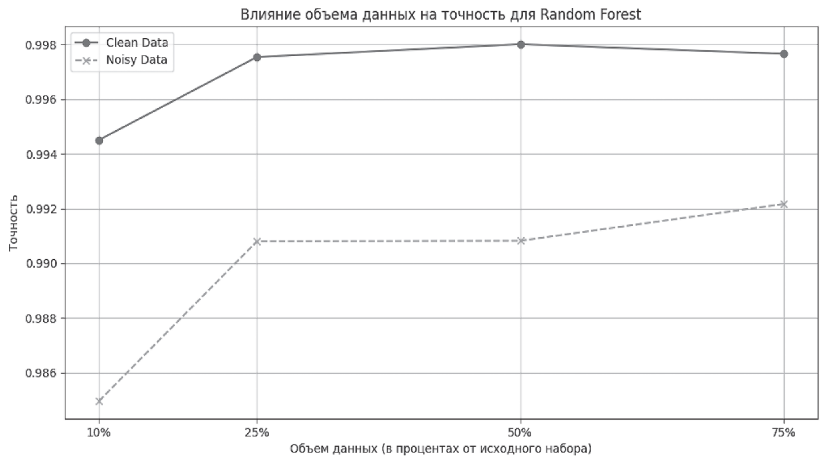

Рисунок 1. Влияние объема данных на точность модели Random Forest
Источник: здесь и далее рисунки выполнены авторами.

Анализ графика на чистых данных (Clean Data)

Начальный этап (10 % данных). При использовании 10 % от исходного набора данных точность модели достигает примерно 99,4 %. Это высокий показатель, указывающий на то, что даже небольшой объем данных может дать хорошие результаты на чистом наборе данных.

Рост объема данных до 25 %. Точность модели значительно возрастает до уровня около 99,8 %. Это подчеркивает, что увеличение объема данных улучшает обучаемость модели, позволяя ей более точно идентифицировать закономерности в данных.

Дальнейшее увеличение объема данных: при использовании 50 % данных наблюдается незначительное улучшение точности, которая достигает максимума (около 99,8 %). Однако при увеличении объема данных до 75 % точность немного снижается. Это может свидетельствовать о том, что модель достигает насыщения, и дальнейшее увеличение объема данных не оказывает значительного влияния на ее производительность, а в некоторых случаях может даже приводить к снижению точности из-за избыточного обучения.

Шумные данные (Noisy Data)

Начальный этап (10 % данных). На этом этапе точность модели составляет около 98,6%, что ниже, чем у чистых данных. Это ожидаемо, поскольку шумные данные усложняют процесс обучения модели.

Рост объема данных до 25 %. Точность возрастает до уровня около 99,0 %, что указывает на позитивное влияние увеличения объема данных на производительность модели.

Дальнейшее увеличение объема данных. Интересно отметить, что начиная с 25 % и до 50 % данных точность модели остается почти неизменной. Это говорит о том, что увеличение объема данных на этом этапе не оказывает существенного влияния на улучшение модели. Лишь при использовании 75 % данных наблюдается незначительное увеличение точности.

График демонстрирует, что увеличение объема данных, используемых для обучения модели, положительно влияет на точность модели, особенно в случаях с чистыми данными. Однако для шумных данных эффект от увеличения объема данных менее выражен. Важно отметить, что после определенного порога точность модели может стабилизироваться или даже начать снижаться, что требует тщательного подхода к выбору объема данных и качеству их предварительной обработки.

Neural Network

График (см. Рисунок 2) иллюстрирует влияние объема данных на точность модели Neural Network для двух типов данных: чистых (Clean Data) и шумных (Noisy Data).

Анализ графика на чистых данных (Clean Data)

Начальный этап (10 % данных). На данном этапе точность модели составляет около 99,25 %. Это хороший результат, однако он ниже, чем у модели Random Forest на аналогичном этапе. Это может говорить о том, что нейронной сети требуется больше данных для достижения высокой точности.

Рост объема данных до 25 %. Точность модели значительно возрастает до уровня 99,75 %, что свидетельствует о существенном улучшении производительности модели с увеличением объема данных.

Дальнейшее увеличение объема данных. При увеличении объема данных до 50 и 75 % точность модели стабилизируется на уровне около 99,75…99,8 %. Это означает, что на этом этапе модель уже почти полностью обучена и дальнейшее увеличение объема данных не приводит к значительному улучшению точности.

Шумные данные (Noisy Data)

Начальный этап (10 % данных). На этом этапе точность модели составляет около 97,75 %, что заметно ниже, чем у модели Random Forest на аналогичном этапе. Это подтверждает чувствительность нейронной сети к шуму в данных.

Вестник Российского нового университета
Серия «Сложные системы: модели, анализ и управление», выпуск 1 за 2025 год

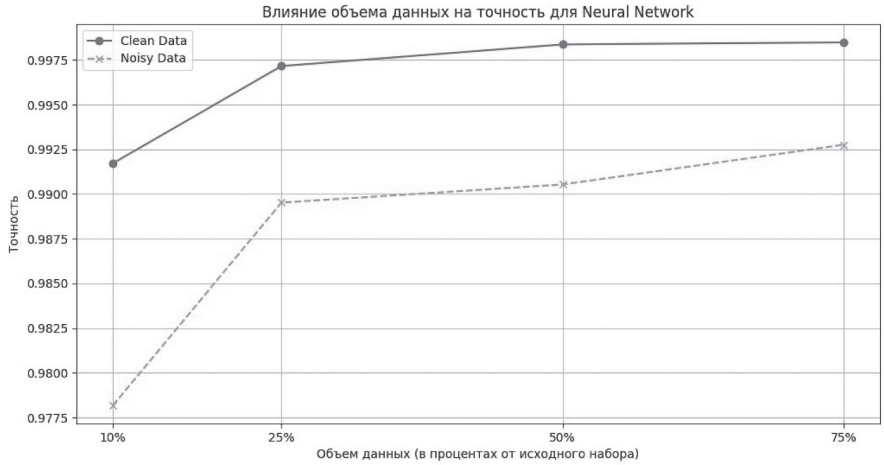

Рисунок 2. Влияние объема данных на точность модели Neural Network

Рост объема данных до 25 %. Точность значительно возрастает и достигает уровня около 99,0 %, что говорит о позитивном влиянии увеличения объема данных.

Дальнейшее увеличение объема данных. Точность модели продолжает постепенно повышаться, достигая примерно 99,2 % при использовании 75 % данных. Несмотря на наличие шума, увеличение объема данных помогает нейронной сети компенсировать влияние шумных данных и улучшить свою точность.

График демонстрирует, что увеличение объема данных положительно сказывается на точности модели нейронной сети, особенно в случае с чистыми данными. Однако модель нейронной сети более чувствительна к шумным данным, чем Random Forest, и требует большего объема данных для достижения схожей точности. Как и в случае с Random Forest, точность модели стабилизируется на определенном уровне при увеличении объема данных, что говорит о необходимости балансирования между объемом данных и качеством их обработки.

Logistic Regression

График (см. Рисунок 3) иллюстрирует влияние объема данных на точность модели Logistic Regression для двух типов данных: чистых (Clean Data) и шумных (Noisy Data).

Анализ графика на чистых данных (Clean Data)

Начальный этап (10 % данных). На этом этапе точность модели составляет около 98,75 %. Это неплохой показатель, учитывая простоту модели логистической регрессии.

Рост объема данных до 25 %. Точность модели возрастает до уровня около 99,0 %. Это показывает, что увеличение объема данных положительно влияет на обучаемость модели.

Дальнейшее увеличение объема данных. При увеличении объема данных до 50 и 75 % наблюдается незначительное улучшение точности, которая достигает максимума около 99,25 %. Это указывает на то, что после определенного объема данных модель практически достигает своего максимума точности, и дальнейшее увеличение объема данных приводит лишь к небольшим изменениям.

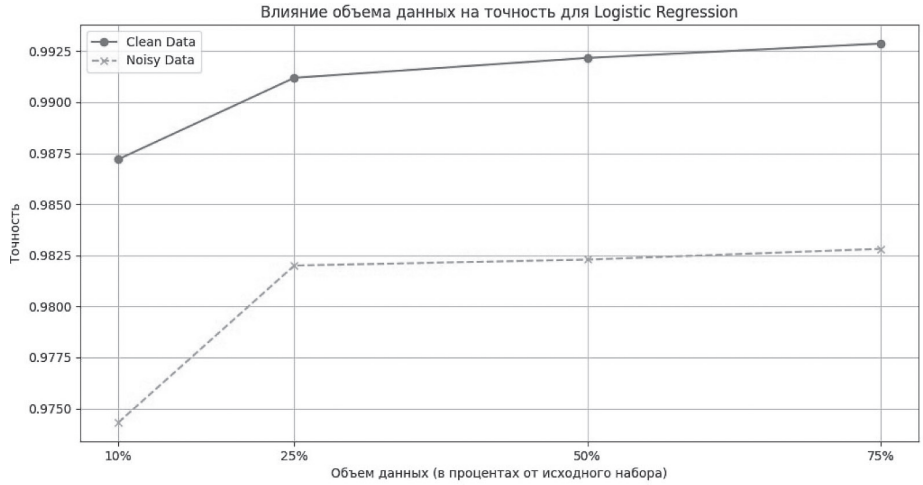

Рисунок 3. Влияние объема данных на точность модели Logistic Regression

Шумные данные (Noisy Data)

Начальный этап (10 % данных). Точность модели составляет около 97,5 %, что ниже по сравнению с чистыми данными, но не так значительно, как в случае с более сложными моделями (например, нейронной сетью). Это может говорить о некоторой устойчивости логистической регрессии к шуму.

Рост объема данных до 25 %. Точность модели увеличивается до 98,25 %, что показывает положительное влияние увеличения объема данных, но эффект выражен не так сильно, как у чистых данных.

Дальнейшее увеличение объема данных. Точность модели стабилизируется на уровне около 98,3 % при использовании 50 и 75 % данных. Это говорит о том, что увеличение объема данных в условиях шумного набора данных почти не влияет на точность модели.

График демонстрирует, что логистическая регрессия чувствительна к увеличению объема данных, особенно в случае чистыми данными. Однако модель достигает своего предела точности довольно быстро, после чего увеличение объема данных не приводит к значительному улучшению. В случае шумных данных эффект от увеличения объема данных выражен слабо, что может свидетельствовать о некоторой устойчивости логистической регрессии к шуму, но также и о ее ограниченных возможностях в обработке сложных данных.

Рассмотрим общий график (см. Рисунок 4) влияния объема данных на точность обнаружения аномалий с использованием различных алгоритмов машинного обучения. Он показывает, как изменяется точность при обработке чистых данных (Clean) и шумных данных (Noisy) в зависимости от процента использованного набора данных (10, 25, 50, 75 %).

Проанализируем полученный общий график.

Общие тенденции. Все алгоритмы показывают рост точности при увеличении объема данных. Наибольшей точности достигает нейронная сеть на чистых данных, причем даже при небольших объемах данных (10 %) ее точность уже близка к максимальной (около 0,995).

Вестник Российского нового университета
Серия «Сложные системы: модели, анализ и управление», выпуск 1 за 2025 год

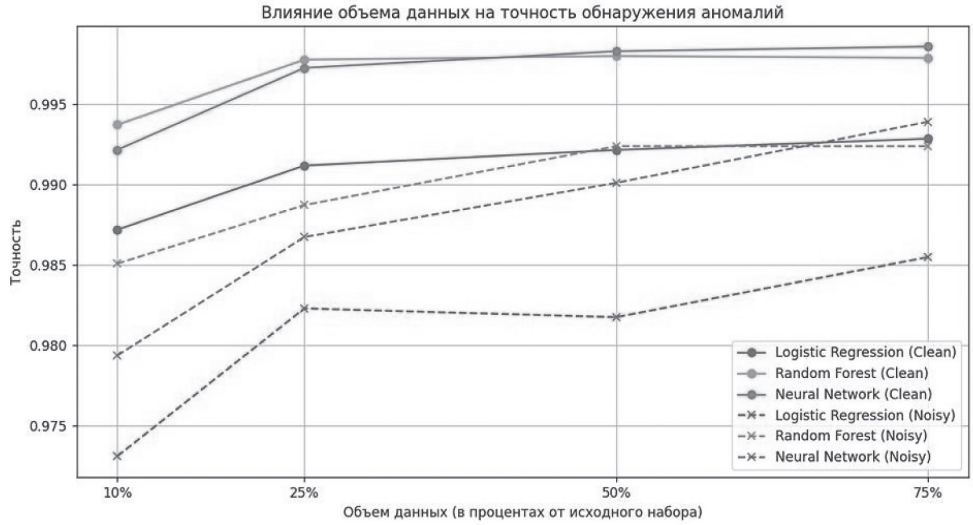

Рисунок 4. Общий график влияния объема данных на точность модели

Чистые данные (Clean). Нейронная сеть (Neural Network) показывает стабильную и высокую точность с самого начала (10 % данных) и практически не изменяет ее при увеличении объема данных.

Случайный лес (Random Forest) на втором месте по точности. Начиная с 25 % данных, точность достигает 0,995 и далее остается практически на том же уровне.

Логистическая регрессия (Logistic Regression) демонстрирует наименьшую точность среди всех методов на чистых данных, но всё же увеличивается с ростом объема данных, достигая около 0,992 при 75 % объема данных.

Шумные данные (Noisy). Нейронная сеть (Neural Network). сохраняет свои лидирующие позиции и при работе с шумными данными, хотя ее точность несколько ниже, чем на чистых данных. Точность возрастает с увеличением объема данных, но не достигает уровня чистых данных, останавливаясь около 0,990 при 75 %.

Случайный лес (Random Forest) показывает устойчивую динамику роста точности по мере увеличения объема данных, однако разрыв с чистыми данными заметен, особенно при 10 % объема данных.

Логистическая регрессия (Logistic Regression) наиболее чувствительна к шуму в данных. На 10 % объема данных ее точность наименее высокая (около 0,975), но с увеличением данных она значительно улучшает свои результаты, достигая около 0,985 при 75 %.

Обсуждение и выводы

На основе проведенного анализа графиков можно сделать несколько ключевых выводов о производительности алгоритмов обнаружения аномалий в сетевом трафике на основе набора данных CICIDS2017.

Во-первых, все три алгоритма показывают общую тенденцию к увеличению точности с ростом объема данных. Нейронные сети демонстрируют наивысшую точность как на чистых, так и на шумных данных, что подтверждает их эффективность в условиях, когда объем данных даже минимален. Это делает нейронные сети предпочтительным выбором

для задач обнаружения аномалий, особенно когда доступно ограниченное количество данных.

Во-вторых, на чистых данных случайный лес показывает высокую устойчивость и точность, что делает его надежным инструментом для анализа. Логистическая регрессия, хотя и демонстрирует наименьшую точность среди всех алгоритмов, всё же улучшает свои результаты с увеличением объема данных, что указывает на ее потенциал в условиях, когда данные имеют хорошее качество.

На шумных данных нейронные сети сохраняют свои лидирующие позиции, хотя и с некоторым снижением точности. Случайный лес также показывает хорошую динамику, но его эффективность заметно снижается по сравнению с чистыми данными. Логистическая регрессия, будучи наиболее чувствительной к шуму, демонстрирует значительное улучшение с увеличением объема данных, что подчеркивает важность качественной предобработки данных.

Таким образом, в ходе исследования получены следующие результаты.

1. Нейронные сети наиболее эффективны как для чистых, так и для шумных данных, обеспечивая высокую точность обнаружения аномалий даже при небольшом объеме данных.

2. Случайный лес также демонстрирует высокую устойчивость, особенно на чистых данных, но его эффективность снижается в условиях шума.

3. Логистическая регрессия наименее устойчива к шуму, но она значительно улучшает свою точность с увеличением объема данных.

В целом нейронные сети и случайный лес представляют собой более надежные алгоритмы для обнаружения аномалий в условиях как чистых, так и шумных данных, особенно при наличии достаточного объема данных.

Результаты проведенного исследования подтверждают, что нейронные сети и случайный лес являются более надежными алгоритмами для обнаружения аномалий как в чистых, так и в шумных данных, особенно при наличии достаточного объема информации. Эти выводы могут служить основой для дальнейших исследований и улучшений в области кибербезопасности, способствуя разработке более эффективных систем защиты от кибератак.

Литература

1. *Wang S., Balarezo J.F., Sithamparanathan K., Al-Hourani A., Chavez K.G., Rubinstein B.* Machine Learning in Network Anomaly Detection: A Survey // IEEE Access. 2021. Vol. 9. P. 152379–152396. DOI: 10.1109/ACCESS.2021.3126834

2. *Rzym G., Masny A., Chołda P.* Dynamic Telemetry and Deep Neural Networks for Anomaly Detection in 6G Software-Defined Networks // Electronics. 2024. Vol. 13. No. 2. Article no. 382. DOI: 10.3390/electronics13020382.

3. *Перов Р.А., Лаута О.С., Крибель А.М., Федулов Ю.В.* Метод выявления аномалий в сетевом трафике // Наукоемкие технологии в космических исследованиях Земли. 2022. Т. 14. № 3. С. 25–31. EDN QSBJKM. DOI: 10.36724/2409-5419-2022-14-3-25-31

4. *Кравцова В.А., Ушаков И.А.* Обнаружение аномалий сетевого трафика с использованием больших данных // Актуальные проблемы инфотелекоммуникаций в науке и образовании (АПИНО 2023) : Сборник научных статей. XII Международная научно-техническая и научно-методическая конференция : В 4 т. Санкт-Петербург, 28 февраля – 01 марта 2023 г. Т. 1. Санкт-Петербург : Санкт-

Петербургский государственный университет телекоммуникаций им. проф. М.А. Бонч-Бруевича, 2023. С. 697–701. EDN IHZQZK.
5. *Ушаков И.А., Исмоилов Ф.Х., Фёдорова А.Э., Манкаев Р.М., Деркач А.Ю.* Обнаружение аномалий в сетевом трафике, используя методы машинного обучения // Актуальные вопросы современной науки и образования: сборник статей XX Международной научно-практической конференции. В 2-х ч. Пенза, 20 июня 2022 г. Ч. 1. Пенза : Наука и Просвещение, 2022. С. 96–98. EDN UGMWZZ.
6. *Труфанов В.Н., Огарык А.Л., Нестеров С.Г.* Исследование сетевых систем обнаружения вторжений, использующих методы машинного обучения // Информатизация и связь. 2023. Т. 14. № 4. С. 59–72. EDN LKGOXL. DOI: 10.34219/2078-8320-2023-14-4-59-72
7. *Shijun S., Min F.* Design of big data anomaly detection model based on random forest algorithm // Scientific Insights and Discoveries Review. 2024. Vol. 1. P. 166–172. DOI: 10.59782/sidr.v1i1.40.
8. *Ness S., Eswarakrishnan V., Sridharan H., Shinde V., Janapareddy N.V.P., Dhanawat V.* Anomaly Detection in Network Traffic Using Advanced Machine Learning Techniques // IEEE Access. 2025. Vol. 13. P. 16133–16149. DOI: 10.1109/ACCESS.2025.3526988
9. *Abinaya N., Senthil Kumar A.V., Chaturvedi A., Bin Musirin I., Rao M., Kaur G., Kaur S., Saleh O.S., Malladi R., Arya N.* Big Data in Real Time to Detect Anomalies // Darwish D. (Ed.) Big Data Analytics Techniques for Market Intelligence. 2024. Hershey, PA: IGI Global. P. 372–396. DOI: 10.4018/979-8-3693-0413-6.ch015
10. *Cavallaro C., Cutello V., Pavone M., Zito F.* Discovering anomalies in big data: A review focused on the application of metaheuristics and machine learning techniques // Frontiers in Big Data. 2023. Vol. 6. DOI: 10.3389/fdata.2023.1179625
11. *Kolukisa B., Dedeturk B.K., Hacılar H., Gungor V.C.* An efficient network intrusion detection approach based on logistic regression model and parallel artificial bee colony algorithm // Computer Standards & Interfaces. 2023. Vol. 89. No. 4. Article no. 103808. DOI: 10.1016/j.csi.2023.103808
12. *Inuwa M., Das R.* A comparative analysis of various machine learning methods for anomaly detection in cyber attacks on IoT networks // Internet of Things. 2024. Vol. 26. Article no. 101162. DOI: 10.1016/j.iot.2024.101162
13. *Safaei M., Driss M., Boulila W., Sundararajan E.* Global outliers detection in wireless sensor networks: A novel approach integrating time-series analysis, entropy, and random forest-based classification // Software Practice and Experience. 2021. Vol 52. No. 1. P. 277–295. DOI: 10.1002/spe.3020
14. *Rafique S. H., Abdallah A., Musa N. S., Murugan T.* Machine Learning and Deep Learning Techniques for Internet of Things Network Anomaly Detection – Current Research Trends // Sensors. 2024. Vol. 24. No. 6. Article no. 1968. DOI: 10.3390/s24061968
15. *Albuquerque Filho J.E., Brandão L.C.P., Fernandes B.J.T., Maciel A.M.A.* A Review of Neural Networks for Anomaly Detection // IEEE Access. 2022. Vol. 10. Pp. 112342–112367. DOI: 10.1109/ACCESS.2022.3216007
16. *Yang Z., Liu X., Li T., Wu D.* A systematic literature review of methods and datasets for anomaly-based network intrusion detection // Computers & Security. 2022. Vol. 116. Article no. 102675. DOI: 10.1016/j.cose.2022.102675

References

1. Wang S., Balarezo J.F., Sithamparanathan K., Al-Hourani A., Chavez K.G., Rubinstein B. (2021) Machine Learning in Network Anomaly Detection: A Survey. *IEEE Access.* Vol. 9. Pp. 152379–152396. DOI: 10.1109/ACCESS.2021.3126834.

2. Rzym G., Masny A., Chołda P. (2024) Dynamic Telemetry and Deep Neural Networks for Anomaly Detection in 6G Software-Defined Networks. *Electronics.* Vol. 13. No. 2. Article no. 382. DOI: 10.3390/electronics13020382

3. Perov R.A., Lauta O.S., Kribel A.M., Fedulov Yu.V. (2022) A method for detecting anomalies in network traffic. *H&ES Research.* Vol. 14. No. 3. Pp. 25–31. DOI: 10.36724/2409-5419-2022-14-3-25-31 (In Russian).

4. Kravtsova V.A., Ushakov I.A. (2023) Detection of network traffic anomalies using big data. In: Makarenko S.I. (Ed) *Aktual'nye problemy infotelekommunikatsii v nauke i obrazovanii (APINO 2023)* [Current problems of infocommunications in science and education (APINO 2023)] : Proceedings of the XII International Conference on Science, Technology and Methods : In 4 vols. St. Petersburg, 28 February – 01 March 2023. Vol. 1. St. Petersburg : Bonch-Bruevich St Petersburg State University of Telecommunications Publ. Pp. 697–701. (In Russian).

5. Ushakov I.A., Ismoilov F.Kh., Fedorova A.E., Mankaev R.M., Derkach A.Yu. (2022) Detection of network traffic anomalies using machine learning methods. In: Gulyaev G.Yu. (Ed) *Aktual'nye voprosy sovremennoi nauki i obrazovaniya* [Current issues of modern science and education] : Proceedings of the XX International Scientific and Practical Conference. Penza, June 20, 2022. Part 1. Penza : Nauka i Prosveshchenie Publ. Pp. 96–98. (In Russian).

6. Trufanov V.N., Ogarok A., Nesterov S. (2023) A research on network intrusion detection systems using machine learning techniques. *Informatization and Communication.* Vol. 14. No. 4. Pp. 59–72. DOI: 10.34219/2078-8320-2023-14-4-59-72 (In Russian).

7. *Shijun S., Min F.* (2024) Design of big data anomaly detection model based on random forest algorithm. Scientific Insights and Discoveries Review. Vol. 1. Pp. 166–172. DOI: 10.59782/sidr.v1i1.40

8. Ness S., Eswarakrishnan V., Sridharan H., Shinde V., Janapareddy N.V.P., Dhanawat V. (2025) Anomaly Detection in Network Traffic Using Advanced Machine Learning Techniques. *IEEE Access.* Vol. 13. Pp. 16133–16149. DOI: 10.1109/ACCESS.2025.3526988

9. Abinaya N., Senthil Kumar A.V., Chaturvedi A., Bin Musirin I., Rao M., Kaur G., Kaur S., Saleh O.S., Malladi R., Arya N. (2024) Big Data in Real Time to Detect Anomalies. In: Darwish D. (Ed.) *Big Data Analytics Techniques for Market Intelligence.* Hershey, PA : IGI Global. Pp. 372–396. DOI: 10.4018/979-8-3693-0413-6.ch015

10. Cavallaro C., Cutello V., Pavone M., Zito F. (2023) Discovering anomalies in big data: A review focused on the application of metaheuristics and machine learning techniques. *Frontiers in Big Data.* Vol. 6. DOI: 10.3389/fdata.2023.1179625

11. Kolukisa B., Dedeturk B.K., Hacılar H., Gungor V.C. (2023) An efficient network intrusion detection approach based on logistic regression model and parallel artificial bee colony algorithm. *Computer Standards & Interfaces.* Vol. 89. No. 4. Article no. 103808. DOI: 10.1016/j.csi.2023.103808

12. Inuwa M., Das R. (2024) A comparative analysis of various machine learning methods for anomaly detection in cyber attacks on IoT networks. *Internet of Things.* Vol. 26. Article no. 101162. DOI: 10.1016/j.iot.2024.101162.

13. Safaei M., Driss M., Boulila W., Sundararajan E. (2021) Global outliers detection in wireless sensor networks: A novel approach integrating time-series analysis, entropy, and random forest-based classification. *Software Practice and Experience*. Vol 52. No. 1. Pp. 277–295. DOI: 10.1002/spe.3020

14. Rafique S. H., Abdallah A., Musa N. S., Murugan T. (2024) Machine Learning and Deep Learning Techniques for Internet of Things Network Anomaly Detection – Current Research Trends. *Sensors*. Vol. 24. No. 6. Article no. 1968. DOI: 10.3390/s24061968

Вестник Российского нового университета
Серия «Сложные системы: модели, анализ и управление», выпуск 1 за 2025 год

15. Albuquerque Filho J.E., Brandão L.C.P., Fernandes B.J.T., Maciel A.M.A. (2022) A Review of Neural Networks for Anomaly Detection. *IEEE Access.* Vol. 10. Pp. 112342–112367. DOI: 10.1109/ACCESS.2022.3216007
16. Yang Z., Liu X., Li T., Wu D. (2022) A systematic literature review of methods and datasets for anomaly-based network intrusion detection. *Computers & Security*. Vol. 116. Article no. 102675. DOI: 10.1016/j.cose.2022.102675

Поступила в редакцию: 16.01.2025	Received: 16.01.2025
Поступила после рецензирования: 10.02.2025	Revised: 10.02.2025
Принята к публикации: 21.02.2025	Accepted: 21.02.2025

DOI: 10.18137/RNU.V9187.25.01.P.127 УДК 004.032

Сергиенко Михаил Владимирович
аспирант, ассистент кафедры практической и прикладной информатики, МИРЭА – Российский технологический университет, Москва. SPIN-код: 6664-9116, AuthorID: 1230038.
Электронный адрес: Sergienko.m@inbox.ru
Mikhail V. Sergienko
Postgraduate, Assistant at the Department of practical and applied informatics, MIREA – Russian Technological University, Moscow. SPIN-code: 6664-9116, AuthorID: 1230038.
E-mail address: Sergienko.m@inbox.ru

Алпатов Алексей Николаевич
кандидат технических наук, доцент кафедры инструментального и прикладного программного обеспечения, МИРЭА – Российский технологический университет, Москва. SPIN-код: 9012-0246, AuthorID: 1064377.
Электронный адрес: aleksej01-91@mail.ru
Aleksey N. Alpatov
Ph.D. of Technical Sciences, Associate Professor at the Department of instrumental and applied software, MIREA – Russian Technological University, Moscow. SPIN-code: 9012-0246, AuthorID: 1064377.
E-mail address: aleksej01-91@mail.ru

АНАЛИЗ ПРИМЕНИМОСТИ БЕСПРОВОДНЫХ СЕТЕЙ СТАНДАРТА IEEE 802.11AX В СЦЕНАРИЯХ ПРОМЫШЛЕННОГО ИНТЕРНЕТА ВЕЩЕЙ И УМНЫХ СКЛАДОВ

Аннотация. Индустрия 4.0 стремится к полной цифровизации производственных процессов, и беспроводные технологии играют в этом важную роль. Однако для различных областей применения, таких, например, как умные склады, где роботы контролируют и выполняют все задачи, современные стандарты и запатентованные решения должны соответствовать строгим отраслевым требованиям. Одной из ключевых задач исследований в этой области является разработка беспроводных систем, которые могут передавать короткие пакеты данных в многопользовательских средах. Однако существующие предложения и модели промышленных каналов не всегда могут удовлетворить эти требования. В качестве возможного решения авторы предлагают оптимизировать стандарт IEEE 802.11ax. Благодаря использованию OFDMA (Orthogonal Frequency Division Multiple Access) эта оптимизация может создать высокопроизводительную систему для умных складов. В данной статье представлена беспроводная система для интеллектуальных складов, разработанная на основе стандарта IEEE 802.11ax. Система включает в себя оптимизацию на двух уровнях: MAC и PHY.

Ключевые слова: интернет вещей, промышленность, умный склад, связь, беспроводные технологии, IEEE 802.11ax.

Для цитирования: *Сергиенко М.В., Алпатов А.Н.* Анализ применимости беспроводных сетей стандарта IEEE 802.11ax в сценариях промышленного интернета вещей и умных складов // Вестник Российского нового университета. Серия: Сложные системы: модели, анализ, управление. 2025. № 1. С. 127 – 140. DOI: 10.18137/RNU.V9187.25.01.P.127