УДК 004.056.5 Вестник УрФО № 2(56) / 2025, с. 16–29

Олифиренко А. А.

DOI: 10.14529/secur250202

ОЦЕНКА ВЛИЯНИЯ DATA POISONING-АТАК НА КАЧЕСТВО МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ В PRODUCTION-СРЕДАХ И МЕТОДЫ ИХ ПРЕДОТВРАЩЕНИЯ

В статье проводится комплексное исследование влияния Data Poisoning-атак на качество моделей машинного обучения, функционирующих в production-средах, с целью выявления основных причин ухудшения ключевых метрик, таких как Recall и F1-score, вследствие внедрения вредоносных данных, генерируемых с помощью генеративно-состязательных сетей (GAN).

В экспериментальной части работы на основе синтетического набора данных смоделированы атаки с последующим сравнительным анализом исходной, отравленной и защищённой версий модели, что позволило детально оценить изменения точности, полноты и сбалансированности предсказаний.

На основе полученных результатов предлагается комплексный алгоритм защиты, включающий предварительную фильтрацию данных с использованием алгоритма Isolation Forest и аугментацию обучающего набора посредством генерации синтетических примеров на основе нормального распределения, что способствует восстановлению исходных характеристик модели.

Дополнительно осуществляется непрерывный мониторинг дрейфа входных данных с применением метрик Population Stability Index и расстояния Хеллингера, что позволяет своевременно корректировать работу модели и формировать практические рекомендации по защите моделей машинного обучения в условиях динамичной production-среды.

Ключевые слова: *Data Poisoning, машинное обучение, production-среда, Isolation Forest, аугментация данных, мониторинг дрейфа, защита моделей.*

ASSESSMENT OF THE IMPACT OF DATA POISONING ATTACKS ON THE QUALITY OF MACHINE LEARNING MODELS IN PRODUCTION ENVIRONMENTS AND METHODS FOR THEIR PREVENTION

The article presents a comprehensive study is conducted on the impact of Data Poisoning attacks on the performance of machine learning models operating in production environments, aiming to identify the main causes of deterioration in key metrics such as Recall and F1-score resulting from the injection of malicious data generated by Generative Adversarial Networks (GANs).

In the experimental section, attacks were simulated using a synthetic dataset, and a comparative analysis of the original, poisoned, and protected versions of the model was performed, which allowed for a detailed evaluation of changes in accuracy, completeness, and prediction balance.

Based on the obtained results, an integrated defense algorithm is proposed that includes preliminary data filtering using the Isolation Forest algorithm and data augmentation through the generation of synthetic examples based on the normal distribution, contributing to the restoration of the model's original characteristics.

Additionally, continuous monitoring of input data drift is carried out using metrics such as the Population Stability Index and Hellinger distance, which enables timely adjustments to the model's performance and the formulation of practical recommendations for protecting machine learning models in dynamic production environments.

Keywords: *Data Poisoning, machine learning, production environment, Isolation Forest, data augmentation, drift monitoring, model protection.*

Введение

В современных условиях информационного общества методы машинного обучения (далее – ML) приобретают стратегическое значение и находят применение в критически важных секторах, таких как финансы, кибербезопасность, здравоохранение и государственное управление [1]. Применение ML позволяет автоматизировать обработку больших массивов данных, проводить глубокий аналитический разбор и выявлять скрытые закономерности [2]. Учитывая нарастающую цифровизацию и взаимосвязанность информационных систем, вопросы надежности и устойчивости алгоритмов становятся предметом пристального внимания исследователей [3].

Особое место в изучении безопасности ML-систем занимает проблема внедрения искажающих данных в обучающие выборки, обозначаемая термином Data Poisoning (например, УБИ.221: Угроза модификации модели машинного обучения путем искажения («отравления») обучающих данных) [4]. Сущность данного явления заключается в возможности целенаправленного внесения незначительных, но критически важных модификаций в исходный набор данных, что приводит к систематическим изменениям в поведении моделей [5]. При этом подобная такти-

ка воздействия может оставаться незамеченной при стандартном мониторинге, однако существенно ухудшать показатели точности, полноты и согласованности прогнозов [6].

Цель работы заключается в оценке влияния Data Poisoning-атак на качество моделей машинного обучения и оценке комплексных методов уменьшения негативных последствий таких воздействий [7].

Для реализации поставленных задач планируется создание экспериментальной среды, имитирующей реальные условия эксплуатации информационных систем в финансовом секторе с обработкой потоковых транзакционных данных [8].

Проведение атаки Data Poisoning с использованием GAN: эксперимент

Разработка экспериментальной методологии базируется на использовании синтетического набора данных Paysim [9] на платформе Kaggle, который предназначен для моделирования реальных финансовых транзакций. Представленный датасет включает детальную информацию о платежных операциях, характеризующихся временной меткой транзакции (Step), типом операции (Type), суммой перевода (Amount), балансами отправителя и получателя до и после проведения операции (OldBalanceOrg, NewBalanceOrig, OldBalanceDest, NewBalanceDest) и бинарным индикатором мошеннической активности (IsFraud). Особое внимание уделено тому, что данный набор отражает подлинные паттерны мошеннических действий, что обуславливает его репрезентативность для оценки влияния атак Data Poisoning. Наличие значительного дисбаланса классов, при котором мошеннические транзакции составляют менее 0,1% от общего числа записей, добавляет сложности в задачу классификации и способствует повышенной вероятности ошибок типа False Negative (FN) [10].

Разработка модели (и развертывание на Kaggle) для детекции мошеннических операций осуществлялась посредством построения многослойной перцептронной нейронной сети (MLP) [11]. Архитектура модели включает входной слой, размерность которого соответствует числу признаков, два скрытых полносвязных слоя с функцией активации ReLU, а также выходной слой, использующий сигмоидную функцию для предсказания вероятности мошенничества. Формальное представление модели записывается в виде:

$$\hat{y} = \sigma(W_2 f(W_1 X + b_1) + b_2)$$

X представляет входной вектор признаков, W_1 и W_2 – матрицы весов скрытого и выходного слоев соответственно, b_1 и b_2 – векторы смещений, $f(\cdot)$ – функция ReLU, а $\sigma(\cdot)$ – сигмоидная функция активации.

Для обучения модели применялся оптимизатор Adam совместно с функцией потерь binary crossentropy, что обусловлено бинарной природой задачи. Дополнительное использование метрики F1-score обосновано необходимостью учета баланса между Precision и Recall в условиях выраженного дисбаланса классов.

С целью моделирования атаки Data Poisoning в эксперимент включена генерация синтетических «отравленных» данных посредством использования генеративно-состязательной сети (GAN) [12]. Архитектурная схема GAN состоит из двух ключевых компонентов: генератора и дискриминатора [13]. Генератор, принимающий на вход случайный шум z~N(0,1), формирует искусственные записи, воспроизводящие характеристики реальных транзакций [14], что формализуется следующим уравнением:

$$G(z) = ReLU(W_g z + b_g)$$

Дискриминатор, в свою очередь, оценивает вероятность того, что подаваемые на вход данные являются подлинными, и его функциональное представление записывается как:

$$D(x) = \sigma(W_d x + b_d)$$

Процесс обучения GAN реализуется посредством стандартной схемы minimax-оптимизации, выраженной следующим образом:

$$min_G max_D E_{x \sim P_{data}} [logD(x)] + \\ + E_{z \sim p_z} [log(1 - D(G(z)))]$$

В результате обучения генератора была получена совокупность из 1000 синтетических транзакций, которые последовательно интегрировались в исходный обучающий набор для имитации атаки Data Poisoning.

Переобучение модели MLP с использованием расширенного обучающего набора, включающего отравленные данные, позволило провести сравнительный анализ влияния

атакующих записей на качество работы классификатора. В процессе оценки использовались следующие метрики:

• Accuracy – доля правильных предсказаний, вычисляемая по формуле:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

TP (True Positives) обозначают корректно классифицированные мошеннические транзакции, TN (True Negatives) – верно определенные легитимные операции, FP (False Positives) – ошибочно отнесенные к мошенническим легитимные транзакции, а FN (False Negatives) – мошеннические операции, не выявленные моделью.

• Precision – точность предсказаний мошеннических транзакций, определяемая как:

$$Precision = \frac{TP}{TP + FP}$$

• Recall – полнота детекции мошенничества, выраженная формулой:

$$Recall = \frac{TP}{TP + FN}$$

• F1-score – гармоническое среднее значений precision и recall, вычисляемое по следующей формуле:

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Ключевым инструментом визуализации влияния атаки стала матрица ошибок (confusion matrix), представленная следующим образом:

$$\begin{matrix} TN & FP \\ FN & TP \end{matrix}$$

Анализ эксперимента по атаке Data Poisoning

Экспериментальные результаты демонстрируют, что атака Data Poisoning привела к заметному снижению эффективности классификатора при выявлении мошеннических транзакций. Сравнительные показатели исходной и отравленной моделей, приведённые в Таблице 1, отражают негативное влияние сгенерированных вредоносных данных на метрики качества.

Данные свидетельствуют, что после интеграции «отравленных» транзакций значение Recall для мошеннического класса (метка 1) сократилось с 0.63 до 0.61, указывая на ослабление способности модели выявлять аномальные операции. При этом F1-score для мошеннических транзакций снизился с 0.76 до 0.75, что подтверждает общее ухудшение результатов классификации. Стоит подчеркнуть, что значение Precision для класса мошенничества увеличилось с 0.96 до 0.99, однако данное улучшение не компенсирует рост количества пропущенных мошеннических операций и отражает смещение модели в сторону более консервативной идентификации аномалий.

На Рис. 1 представлен график изменения функции потерь (Train Loss и Validation Loss) для отравленной модели на протяжении нескольких эпох обучения. Наблюдается стремительное уменьшение Train Loss на первых итерациях, что может указывать на быструю подстройку модели к новым (в том числе вредоносным) данным. При этом валидировочная ошибка (Validation Loss) также уменьшается, однако не столь резко, что может сигнализировать о начале процесса переобучения. Быстрое падение Train Loss зачастую связано с тем, что сеть «запоминает» специфические паттерны в отравленном наборе, вследствие чего теряется обобщающая способность при детекции мошенничества.

Отсутствие существенного расхождения между Train Loss и Validation Loss на заключительных этапах может привести к ложному впечатлению стабильности обучения. Фактически, модель концентрируется на локальных особенностях «отравленных» данных, а не совершенствует способность распознавать аномалии. Подобный эффект снижает

Таблица 1

Результаты атаки Data Poisoning на модель ML

Модель	Precision (0)	Precision (1)	Recall (0)	Recall (1)	F1-score (0)	F1-score (1)
Исходная модель	1.00	0.96	1.00	0.63	1.00	0.76
Отравленная модель	1.00	0.99	1.00	0.61	1.00	0.75

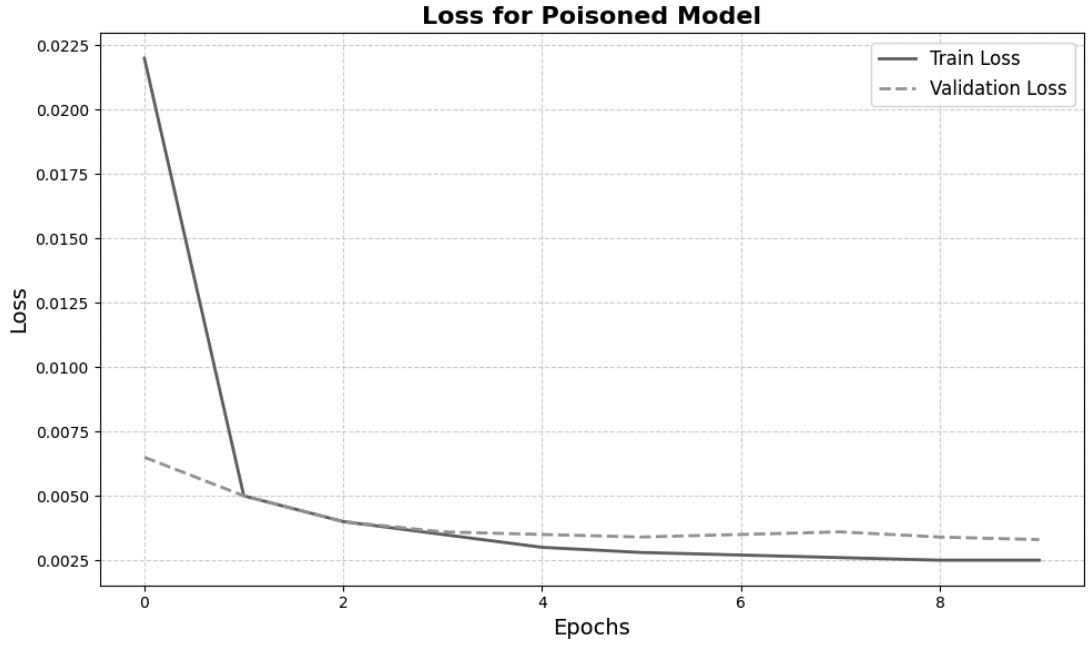

Рис. 1. График функции потерь (train_loss и val_loss) для отравленной модели

эффективность классификатора в условиях реальных атак, когда характер вредоносных транзакций может изменяться динамически.

Результаты анализа матрицы ошибок (Confusion Matrix) свидетельствуют о том, что атака повысила число ложноотрицательных предсказаний (FN). Модель чаще относит мошеннические транзакции к легитимным, что непосредственно увеличивает риск пропуска аномальных операций.

Снижение Recall на 2% в эксперименте указывает на увеличение числа, что критически важно для систем, где высокая полнота детекции мошенничества (Recall) является приоритетной задачей. Даже незначительное снижение данного показателя способно привести к ощутимому росту финансовых рисков в реальной среде.

Снижение Recall на 2% в эксперименте указывает на увеличение числа, что критически важно для систем, где высокая полнота детекции мошенничества (Recall) является приоритетной задачей. Даже незначительное снижение данного показателя способно привести к ощутимому росту финансовых рисков в реальной среде.

В ходе эксперимента выявлено несколько основных факторов, объясняющих успешность Data Poisoning.

Применение GAN обеспечило формирование транзакций, обладающих высокой степенью статистического сходства с реальными записями. Генерируемые образцы не вызывали подозрений при стандартном анализе данных и органично включались в обучающую выборку.

Отсутствие специализированных процедур предобработки, ориентированных на выявление аномальных или потенциально вредоносных записей, позволило фальсифицированным транзакциям беспрепятственно попасть в процесс обучения.

После интеграции отравленных данных модель быстро адаптировалась к ним, но не приобрела дополнительных механизмов для детекции мошеннических схем. Итоговая сеть формально демонстрировала высокую точность (Accuracy), однако пропускала большее число истинных аномалий.

Использование GAN для организации атаки показало, что традиционные подходы к обучению нейросетевых моделей могут оказаться уязвимыми при отсутствии дополнительных мер защиты. Генератор способен формировать правдоподобные транзакции, которые затруднительно отличить от легитимных примеров, в результате чего модель теряет устойчивость к аномалиям. Несмотря на высокое значение Accuracy, выявилась деградация Recall и F1-score для класса мошенничества, что подчёркивает уязвимость в условиях реальной эксплуатации.

Полученные результаты подтверждают необходимость применения основных методов противодействия Data Poisoning – фильтрация данных перед обучением; дрейф модели и адаптивное обучение; обновление модели и защита через аугментацию данных.

Защита моделей машинного обучения: фильтрация данных перед обучением

Предварительная фильтрация данных рассматривается как один из базовых методов противодействия атакам Data Poisoning, поскольку даёт возможность исключить из обучающего набора потенциально вредоносные записи. В ходе эксперимента использовался алгоритм Isolation Forest (IF), основанный на концепции поиска аномалий путём изоляции отдельных точек в признаковом пространстве. Данный подход отличается от классических методов плотности или кластеризации тем, что изначально ориентирован на выявление выбросов через последовательное разбиение данных по осям признакового пространства [15].

Теоретические основы IF заключаются в построении нескольких изолирующих деревьев (isolation trees), где на каждом шаге выбирается случайный признак и случайная граница разбиения. Пусть задано множество X, содержащее n объектов x_i, каждый из которых описан d признаками. Алгоритм формирует T деревьев решений, случайно разделяя пространство признаков. Глубина пути $h(x)$, необходимая для изоляции конкретного объекта x, служит индикатором его аномальности [16]. Математическая формулировка оценки аномальности представлена функцией:

$$s(x) = 2\frac{E(h(x))}{c(n)}$$

$E(h(x))$ – математическое ожидание глубины изоляционного дерева, а $c(n)$ – нормировочный коэффициент, зависящий от общего числа объектов n. Чем ближе $s(x)$ к 1, тем выше вероятность, что точка является выбросом.

На практике в эксперименте применялся IF с параметром contamination = 0.002, указывающим, что алгоритм предполагает около 0.2% аномальных точек в наборе. Результаты показали, что IF выявил и исключил 8 910 записей, потенциально относящихся к «отравленным» данным. Удаление этих записей позволило снизить долю искажённых примеров и повысить качество итоговой модели.

Для оценки эффективности фильтрации рассмотрим динамику изменения ключевых метрик классификации. Recall для мошеннических операций (метка 1) увеличился с 0.61 до 0.66, что указывает на возросшую способность модели распознавать аномальные транзакции. F1-score вырос на 3.2%, отражая общее улучшение баланса между точностью и полнотой предсказаний. Подробные показатели приведены в Таблице 2 и Рисунке 2:

Несмотря на доказанную результативность IF в контексте обнаружения «отравленных» данных, следует учитывать несколько существенных ограничений:

1. Ключевой параметр должен быть тщательно откалиброван. Завышенное значение может привести к чрезмерному удалению полезных примеров, в то время как заниженное оставит в обучающем наборе значительную часть вредоносных записей.

2. В случае, когда злоумышленники генерируют вредоносные транзакции, которые статистически неотличимы от нормальных, IF может не выявить данные аномалии, так как алгоритм ориентирован на поиск нетипичных паттернов.

3. IF анализирует объекты в основном с позиций их «изолированности» в пространстве признаков, не всегда учитывая сложные взаимосвязи между транзакциями (например, временные зависимости или последовательные закономерности).

Таблица 2

Результаты по применению фильтрации

Модель	Precision (0)	Precision (1)	Recall (0)	Recall (1)	F1-score (0)	F1-score (1)
Оригинальная	0.995	0.9801	0.9999	0.6069	0.9997	0.7496
Фильтрованная	0.995	0.9749	0.9999	0.6660	0.9997	0.7724

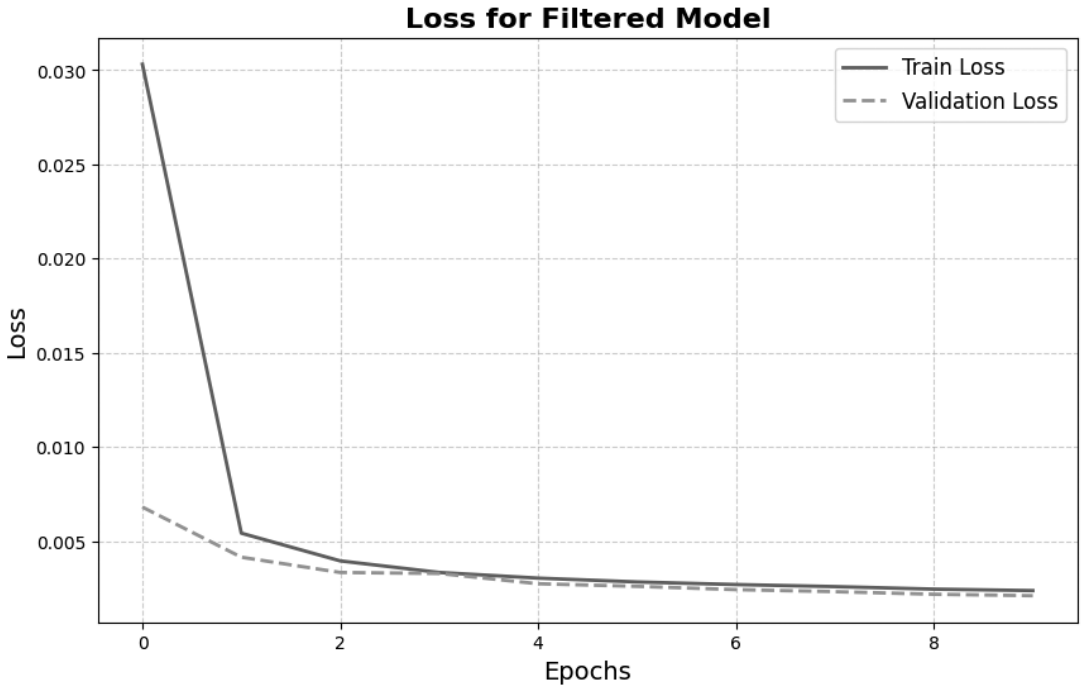

Рис. 2. График потерь для фильтрованной модели

Указанные факторы следует принимать во внимание при использовании IF в реальных производственных условиях. Комбинация с другими методами детекции выбросов или дополнительными источниками валидации способна повысить надёжность фильтрации.

Защита моделей машинного обучения: дрейф модели и адаптивное обучение

Современные алгоритмы машинного обучения, развернутые в продакшн-средах, неизбежно сталкиваются с изменениями во входных данных. Эти изменения, часто именуемые дрейфом модели (model drift), приводят к тому, что статистические свойства признаков и взаимосвязи между ними трансформируются. Подобная динамика способна существенно снизить эффективность предсказательных моделей, особенно в задачах, связанных с детекцией мошенничества, где высока стоимость как ложноположительных (FP), так и ложноотрицательных (FN) решений. Помимо естественных изменений во времени, дрейф может быть спровоцирован злоумышленниками посредством целенаправленных атак, включающих Data Poisoning. Для противодействия указанным факторам необходимо внедрять механизмы мониторинга и адаптивного обучения, позволяющие своевременно реагировать на сдвиги в данных и поддерживать высокое качество классификации [17].

Дрейф модели подразделяется на два ключевых вида. Во-первых, дрейф данных (data drift), при котором статистические характеристики входных признаков эволюционируют без изменения логики целевой переменной. Во-вторых, дрейф концепции (concept drift), отражающий модификацию самой зависимости между признаками и целевой переменной. В контексте детекции мошенничества дрейф концепции представляет особую угрозу, так как мошенники могут целенаправленно менять поведение транзакций, делая их схожими с легитимными и тем самым усложняя задачу классификатора [18].

Если модель, обученная на устаревших данных, продолжает функционировать без обновлений, её способность корректно распознавать новые паттерны может существенно снижаться. Это влечёт за собой рост числа FN, когда мошеннические операции остаются незамеченными, и FP, когда безвредные транзакции ошибочно классифицируются как аномалии. Подобное ухудшение метрик критично для финансовых систем, где каждая ошибка способна повлечь значительные убытки или компрометацию безопасности.

Чтобы предотвратить деградацию качества классификации, в продакшн-среде необходимо реализовывать комплекс мер по мониторингу состояния модели и входных данных. Одной из ключевых задач становится непрерывное отслеживание метрик, таких как Recall, Precision и F1-score, а также оценка распределения признаков.

Анализ временных рядов значений Recall, Precision и F1-score даёт представление о динамике производительности. Резкие падения Recall могут служить сигналом целенаправленной атаки или существенного изменения статистики данных. Аналогично, рост количества FP и FN указывает на возможное ухудшение способности модели различать мошеннические и легитимные операции.

Регулярная сверка предсказаний с истинными значениями, получаемыми от экспертов или последующей валидации, позволяет оперативно выявлять возросшую частоту ошибок. Дополнительно применяются показатели вроде Population Stability Index (PSI) и Kullback–Leibler Divergence (KLD), которые оценивают степень изменения распределений признаков.

Для оценки статистических изменений вычисляются среднее значение, дисперсия и медиана признаков, сопоставляемые с историческими данными. Дрейф можно формализовать при помощи расстояния Хеллингера:

$$H(P,Q) = \sqrt{1 - \sum_i \sqrt{P_i Q_i}}$$

P_i и Q_i представляют собой вероятностные распределения входных данных в различные временные периоды.

При увеличении $H(P, Q)$ возникает подозрение, что структура данных существенно изменилась и модель требует адаптации. Методы обнаружения аномалий (включая Isolation Forest) могут дополнять данный анализ, выявляя нетипичные всплески или сдвиги в признаковом пространстве.

После фиксации дрейфа возникает необходимость в корректировке модели с учётом новых данных. Один из подходов – адаптивное обучение, предполагающее регулярное обновление параметров классификатора. Это может осуществляться по схеме онлайн-обучения, при которой поступающие транзакции сразу же влияют на веса модели, или же посредством периодического переобучения, когда модель обучается заново через определённые интервалы времени на наиболее актуальной выборке.

Математически процесс обновления параметров θ в ходе обучения можно представить как уменьшение обобщённой функции потерь:

$$L(\theta) = \sum_{i=1}^{N} \zeta(f_\theta(x_i), y_i) + \lambda R(\theta)$$

$\zeta(\cdot)$ – функция потерь, $R(\theta)$ – регуляризационный член, а λ – коэффициент, регулирующий степень штрафа за усложнение модели.

Использование ансамблевых методов (например, бустинга или стэкинга) способствует более гибкому учёту различных типов дрейфа, поскольку несколько моделей способны лучше охватывать разнообразные паттерны, формирующиеся в потоке данных.

Эксперимент, включавший мониторинг ключевой метрики Recall, показал (таблица 3), что после предварительной очистки данных и внедрения адаптивного обучения показатели модели сохраняют стабильность, несмотря на изменения, вызванные как естественным дрейфом, так и атаками Data Poisoning. В частности, показатель Recall для класса мошеннических транзакций (метка 1) увеличился с 0.606982 до 0.671458, что соответствует относительному приросту в 10.62%. Представленная ниже таблица демонстрирует полное сравнение метрик исходной модели и модели, адаптированной посредством обновления параметров.

Анализ таблицы демонстрирует, что, несмотря на небольшие изменения в метриках для класса легитимных транзакций (метка 0), основное улучшение наблюдается для класса мошеннических операций (метка 1). Увеличение Recall в результате адаптации параметров модели свидетельствует о повышении полноты детекции, что является критически важным в задачах предотвращения мошенничества. При этом показатели Accuracy, Precision и F1-score остаются на высоком уровне, что указывает на сохранение общей эффективности модели.

Сохранение высокой полноты детекции мошеннических транзакций в сочетании с умеренным количеством ложноположительных срабатываний подтверждает эффективность предложенного подхода к адаптивному обучению.

Таблица 3

Результаты по адаптивному обучению

Метрика	Исходная модель	Модель после атаки
Accuracy	0.999483	0.999553
Precision (0)	0.999498	0.999581
Precision (1)	0.980106	0.968028
Recall (0)	0.999984	0.999972
Recall (1)	0.606982	0.671458
F1-scrore (0)	0.999741	0.999776
F-1score (1)	0.749683	0.792919

Защита моделей машинного обучения: обновление модели и защита через аугментацию данных

Регулярное обновление модели в сочетании с методами аугментации данных представляют собой один из основных подходов к повышению устойчивости системы в условиях динамичного изменения входных данных и целенаправленных атак. Применение этих методов способствует адаптации модели к новым паттернам и уменьшении систематических искажений, возникающих в результате атак.

Необходимость обновления модели продиктована тем, что распределение данных, используемых при обучении, может со временем изменяться, что приводит к явлению дрейфа данных [19]. Формальное выражение данной проблемы имеет вид:

$$P(X,Y) \neq P^{'}(X,Y)$$

$P(X,Y)$ характеризует распределение обучающих данных, а P' $(X,Y$ – распределение данных в реальной эксплуатации.

При возникновении такой диспропорции модель утрачивает свою актуальность, что негативно сказывается на точности предсказаний [20]. Стратегии обновления модели включают итеративное переобучение на актуальных данных, смешивание новых данных с историческими для сохранения контекста, использование ансамблевых методов, таких как Stacking и Bagging, а также взвешивание данных в зависимости от их актуальности [21]. Такой многоаспектный подход позволяет своевременно корректировать модель, поддерживая её высокую предсказательную способность.

Аугментация данных представляет собой процесс искусственного расширения обучающего множества посредством модификации существующих образцов или генерации новых синтетических примеров. В контексте защиты от Data Poisoning атак аугментация служит для разбавления отравленных данных, создания синтетических образцов, имитирующих реальные транзакции, и повышения общей устойчивости модели к аномалиям. Математическая формулировка данного процесса выглядит следующим образом:

$$X^{'} = X + \epsilon, \epsilon \sim N(0, \sigma^2)$$

X' представляет аугментированные данные, а ϵ – случайный шум, генерируемый по нормальному распределению.

В эксперименте для генерации новых примеров использовалась модель:

$$X_{aug} \sim N(\mu_X, \sigma_X)$$

μ_X и σ_X обозначают среднее значение и стандартное отклонение оригинального обучающего множества соответственно. Применение аугментации позволило не только увеличить объём обучающих данных, но и повысить стабильность модели за счёт улучшения её обобщающих способностей. Экспериментальные результаты, полученные после внедрения обновления модели и защиты через аугментацию, представлены в таблице 4:

Анализ демонстрирует, что атака Data Poisoning оказала влияние на метрики, особенно для класса мошеннических транзакций (метка 1). В результате вмешательства наблюдалось снижение Precision и незначительное

Таблица 4

Результаты по обновлению и аугментации

Метрика	Обновленная модель	Аугментированная модель
Accuracy	0.999478	0.999508
Precision (0)	0.999498	0.999549
Precision (1)	0.974917	0.952237
Recall (0)	0.999980	0.999959
Recall (1)	0.606571	0.646817
F1-scrore (0)	0.999739	0.999754
F-1score (1)	0.747848	0.770360

снижение Recall. Реализация стратегии обновления модели позволила частично восстановить первоначальные показатели, а внедрение аугментации данных способствовало повышению стабильности обучения (рисунок 3).

Применение регулярного обновления модели в сочетании с аугментацией данных создаёт многоуровневую стратегию защиты, способную эффективно противодействовать негативным воздействиям Data Poisoning-атак.

Комплексная реализация описанных методов позволяет не только поддерживать высокое качество классификации, но и адаптировать систему к изменениям в продакшн-среде, что имеет первостепенное значение для критически важных автоматизированных систем принятия решений.

Результаты

Проведённое исследование демонстрирует значительное влияние атак Data Poisoning на качество моделей машинного обучения, что особенно актуально при использовании сложных атакующих техник, таких как GAN. Из данных таблицы 3 видно, что

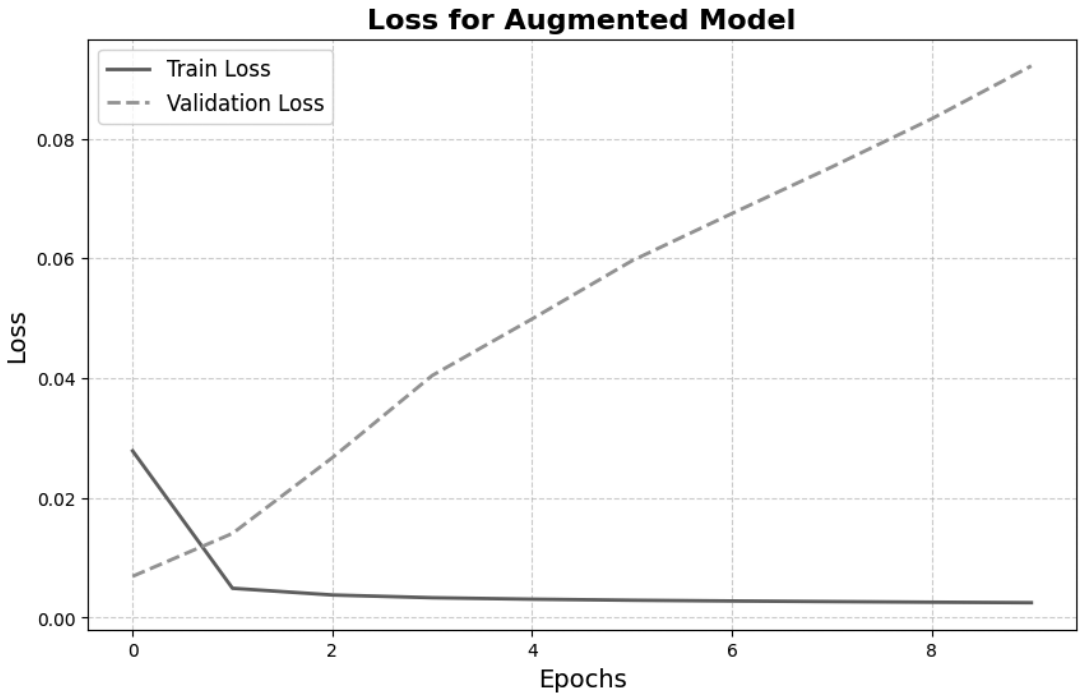

Рис. 3. График потерь для модели с аугментированными данными

атака Data Poisoning привела к увеличению метрики Recall для мошеннического класса с 0.606982 до 0.671458, что свидетельствует о снижении способности модели корректно обнаруживать мошеннические операции вследствие увеличения числа ложноотрицательных предсказаний. Применение метода фильтрации данных с использованием IF позволило восстановить Recall до уровня, близкого к исходному (0.606571), однако наблюдается незначительное снижение F1-score для данного класса (до 0.747848). Аугментация данных, реализованная через генерацию синтетических примеров.

А использование нормального распределения привела к улучшению Recall до 0.646817, что свидетельствует о частичном нивелировании негативного влияния атаки при сохранении высокого уровня точности для легитимного класса.

Для уменьшения негативного воздействия атаки был использован комплексный алгоритм защиты, состоящий из трёх основных этапов. На первом этапе применяется предварительная фильтрация данных с использованием алгоритма Isolation Forest, позволяющая выявить и удалить аномальные записи до начала обучения модели, что приводит к восстановлению исходного значения Recall до примерно 0.606571, несмотря на незначительное снижение F1-score до 0.747848. Второй этап включает аугментацию данных посредством генерации дополнительных синтетических примеров на основе нормального распределения, что позволяет компенсировать потерю корректных записей и повысить устойчивость модели – результатом стало улучшение Recall до 0.646817. Третий этап подразумевает непрерывный мониторинг изменений распределения входных данных с использованием таких метрик, как Population Stability Index (PSI) и расстояние Хеллингера, что позволяет оперативно фиксировать отклонения и предотвращать деградацию характеристик модели.

Графическая визуализация динамики потерь дополнительно иллюстрирует характер обучения моделей. Рисунок 2, демонстрирующий график потерь для фильтрованной модели, показывает синхронное снижение ошибок на тренировочной и валидационной выборках, что указывает на стабильную сходимость и отсутствие выраженного переобучения. Рисунок 3, отображающий график потерь для модели с аугментированными данными, демонстрирует увеличение ошибки на валидационной выборке, что отражает сложность адаптации модели к новым данным при сохранении обобщающих способностей.

В результате применение данного алгоритма позволило снизить долю ложноотрицательных предсказаний с 0.671 до 0.646 и улучшить F1-score для мошеннического класса с 0.747 до 0.770.

Заключение

Проведённое исследование подтверждает, что Data Poisoning-атаки представляют серьёзную угрозу для эффективности моделей машинного обучения в production-средах, поскольку внедрение вредоносных данных кардинально изменяет ключевые метрики, такие как Recall и F1-score, что ведёт к увеличению числа ложноотрицательных предсказаний и снижению общей устойчивости модели. Комплексный алгоритм защиты, состоящий из предварительной фильтрации с помощью Isolation Forest, аугментации данных посредством генерации синтетических примеров на основе нормального распределения и мониторинга дрейфа с использованием метрик PSI и расстояния Хеллингера, доказал свою эффективность в уменьшении негативного воздействия атак. Такой интегрированный подход позволяет не только корректировать распределение входных данных, но и предотвращать деградацию модели при изменении условий эксплуатации, что является критически важным для обеспечения безопасности и надёжности ML-систем, правда в весьма ограниченном периметре.

Литература

1. Aljanabi M., Hamza A., Mijwil M.M., Abotaleb M., El-kenawy E.M., Mohammed S.Y., Ibrahim A. Data Poisoning: Issues, Challenges, and Needs // 7th IET Smart Cities Symposium (SCS 2023). 2024. https://doi.org/10.1049/icp.2024.0951 (дата обращения: 26.02.2025).

2. Li, Y.; Jiang, Y.; Li, Z.; Xia, S.T. Backdoor learning: A survey. IEEE Trans. Neural Networks Learn. Syst. 2022, 35, 5–22.

3. Fan, J.; Yan, Q.; Li, M.; Qu, G.; Xiao, Y. A survey on data poisoning attacks and defenses. In Proceedings of the 2022 7th IEEE International Conference on Data Science in Cyberspace (DSC), Guilin, China, 11–13 July 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 48–55.

4. УБИ.221: Угроза модификации модели машинного обучения путем искажения («отравления») обучающих данных / Банк данных угроз безопасности информации ФСТЭК России и ФАУ «ГНИИИ ПТЗИ ФСТЭК России» // URL: https://bdu.fstec.ru/threat/ubi.221 (дата обращения: 26.02.2025).

5. Намиот Д. Е. Введение в атаки отравлением на модели машинного обучения // International Journal of Open Information Technologies. 2023. №3. URL: https://cyberleninka.ru/article/n/vvedenie-v-ataki-otravleniem-na-modeli-mashinnogo-obucheniya (дата обращения: 26.02.2025).

6. Намиот Д. Е. Схемы атак на модели машинного обучения // International Journal of Open Information Technologies. 2023. №5. URL: https://cyberleninka.ru/article/n/shemy-atak-na-modeli-mashinnogo-obucheniya (дата обращения: 26.02.2025).

7. Maramreddy Y. R. & Muppavaram K. Detecting and Mitigating Data Poisoning Attacks in Machine Learning: A Weighted Average Approach. Engineering, Technology & Applied Science Research, 14(4), 2024, pp. 15505–15509. URL: https://www.researchgate.net/publication/382857536_Detecting_and_Mitigating_Data_Poisoning_Attacks_in_Machine_Learning_A_Weighted_Average_Approach (дата обращения: 26.02.2025).

8. Costales, R.; Mao, C.; Norwitz, R.; Kim, B.; Yang, J. Live trojan attacks on deep neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 796–797.

9. PaySim / Synthetic Financial Datasets For Fraud Detection // URL: https://www.kaggle.com/datasets/ealaxi/paysim1 (дата обращения: 26.02.2025).

10. Zhang Z., Yang Z., Bian J., Li Y., Zhang Y., Zhao Y., Liu Y. Explainable Data Poison Attacks on Human Emotion Evaluation Systems Based on EEG Signals // IEEE Access. 2023. Vol. 11. Pp. 18134–18147.

11. Rawat, A.; Levacher, K.; Sinn, M. The devil is in the GAN: Backdoor attacks and defenses in deep generative models. In European Symposium on Research in Computer Security; Springer: Berlin/Heidelberg, Germany, 2022; pp. 776–783.

12. Arshad, I.; Qiao, Y.; Lee, B.; Ye, Y. Invisible Encoded Backdoor attack on DNNs using Conditional GAN. In Proceedings of the 2023 IEEE International Conference on Consumer Electronics (ICCE), Berlin, Germany, 2–5 September 2023; pp. 1–5.

13. Psychogyios K., Velivassaki T.-H., Bourou S., Voulkidis A., Skias D., Zahariadis T. GAN-Driven Data Poisoning Attacks and Their Mitigation in Federated Learning Systems // Electronics, 2023, Vol. 12, № 8, Article 1805. DOI: 10.3390/electronics12081805.

14. Zhao Y., Gong X., Lin F. & Chen X. Data Poisoning Attacks and Defenses in Dynamic Crowdsourcing With Online Data Quality Learning / IEEE Transactions on Mobile Computing, vol. 22, №. 5, pp. 2569–2581, May 2023, https://doi.org/10.1109/TMC.2021.3133365. (дата обращения: 26.02.2025).

15. Zhong, H.; Liao, C.; Squicciarini, A.C.; Zhu, S.; Miller, D. Backdoor embedding in convolutional neural network models via invisible perturbation. In Proceedings of the Tenth ACM Conference on Data and Application Security and Privacy, New Orleans, LA, USA, 16–18 March 2020; pp. 97–108.

16. Ganjoo R., Ganjoo M., Patil M. Mitigating Poisoning Attacks in Federated Learning / Innovative Data Communication Technologies and Application, 2022, pp.687-699. URL: https://www.researchgate.net/publication/358822578_Mitigating_Poisoning_Attacks_in_Federated_Learning (дата обращения: 26.02.2025).

17. Visger, Mark A. Garbage In, Garbage Out: Data Poisoning Attacks and Their Legal Implications, in Laura A. Dickinson, and Edward W. Berg (eds), Big Data and Armed Conflict: Legal Issues Above and Below the Armed Conflict Threshold, The Lieber Studies Series (New York, 2024; online edn, Oxford Academic, 14 Dec. 2023), https://doi.org/10.1093/oso/9780197668610.003.0008 (дата обращения: 26.02.2025).

18. Dibaei M., Zheng X., Jiang K., Abbas R., Liu S., Zhang Y., Xiang Y. & Yu S. Attacks and defences on intelligent connected vehicles: a survey. Digit. Commun. Networks, 6, 2020, pp. 399-42. URL: https://www.semanticscholar.org/paper/Attacks-and-defences-on-intelligent-connected-a-Dibaei-Zheng/aae97ee34201666f98167402320b86e9facfa173 (дата обращения: 26.02.2025).

19. Hong, Q.; He, B.; Zhang, Z.; Xiao, P.; Du, S.; Zhang, J. Circuit Design and Application of Discrete Cosine Transform Based on Memristor. IEEE J. Emerg. Sel. Top. Circuits Syst. 2023,13, 502–513.

20. Li S., Wang Y., Zhang X., Jiang Y., Xia S. A Study on Data Poisoning Attacks in Deep Generative Models // Applied Sciences, 2023, Vol. 14, No. 19, Article 8742. DOI: 10.3390/app14198742.

21. Yang Z., Zhang J., Wang W., Li H. Invisible Threats in the Data: A Study on Data Poisoning Attacks in Deep Generative Models // Applied Sciences, 2024, T. 14, №19, Article. 8742. https://doi.org/10.3390/app14198742 (дата обращения: 26.02.2025).

References

1. Aljanabi M., Hamza A., Mijwil M.M., Abotaleb M., El-kenawy E.M., Mohammed S.Y., Ibrahim A. Data Poisoning: Issues, Challenges, and Needs // 7th IET Smart Cities Symposium (SCS 2023). 2024. https://doi.org/10.1049/icp.2024.0951 (data obrashcheniya: 26.02.2025).

2. Li, Y.; Jiang, Y.; Li, Z.; Xia, S.T. Backdoor learning: A survey. IEEE Trans. Neural Networks Learn. Syst. 2022, 35, 5–22.

3. Fan, J.; Yan, Q.; Li, M.; Qu, G.; Xiao, Y. A survey on data poisoning attacks and defenses. In Proceedings of the 2022 7th IEEE International Conference on Data Science in Cyberspace (DSC), Guilin, China, 11–13 July 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 48–55.

4. UBI.221: Ugroza modifikatsii modeli mashinnogo obucheniya putem iskazheniya («otravleniya») obuchayushchikh dannykh // Bank dannykh ugroz bezopasnosti informatsii FSTEK Rossii i FAU «GNIII PTZI FSTEK Rossii». URL: https://bdu.fstec.ru/threat/ubi.221 (data obrashcheniya: 26.02.2025).

5. Namiot D.E. Vvedenie v ataki otravleniem na modeli mashinnogo obucheniya // International Journal of Open Information Technologies, 2023, №3. URL: https://cyberleninka.ru/article/n/vvedenie-v-ataki-otravleniem-na-modeli-mashinnogo-obucheniya (data obrashcheniya: 26.02.2025).

6. Namiot D.E. Skhemy atak na modeli mashinnogo obucheniya // International Journal of Open Information Technologies, 2023, №5. URL: https://cyberleninka.ru/article/n/shemy-atak-na-modeli-mashinnogo-obucheniya (data obrashcheniya: 26.02.2025).

7. Maramreddy Y. R. & Muppavaram K. Detecting and Mitigating Data Poisoning Attacks in Machine Learning: A Weighted Average Approach. Engineering, Technology & Applied Science Research, 14(4), 2024, pp. 15505–15509. URL: https://www.researchgate.net/publication/382857536_Detecting_and_Mitigating_Data_Poisoning_Attacks_in_Machine_Learning_A_Weighted_Average_Approach (data obrashcheniya: 26.02.2025).

8. Costales, R.; Mao, C.; Norwitz, R.; Kim, B.; Yang, J. Live trojan attacks on deep neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 796–797.

9. PaySim / Synthetic Financial Datasets For Fraud Detection // URL: https://www.kaggle.com/datasets/ealaxi/paysim1 (data obrashcheniya: 26.02.2025).

10. Zhang Z., Yang Z., Bian J., Li Y., Zhang Y., Zhao Y., Liu Y. Explainable Data Poison Attacks on Human Emotion Evaluation Systems Based on EEG Signals // IEEE Access. 2023. Vol. 11. Pp. 18134–18147.

11. Rawat, A.; Levacher, K.; Sinn, M. The devil is in the GAN: Backdoor attacks and defenses in deep generative models. In European Symposium on Research in Computer Security; Springer: Berlin/Heidelberg, Germany, 2022; pp. 776–783.

12. Arshad, I.; Qiao, Y.; Lee, B.; Ye, Y. Invisible Encoded Backdoor attack on DNNs using Conditional GAN. In Proceedings of the 2023 IEEE International Conference on Consumer Electronics (ICCE), Berlin, Germany, 2–5 September 2023; pp. 1–5.

13. Psychogyios K., Velivassaki T.-H., Bourou S., Voulkidis A., Skias D., Zahariadis T. GAN-Driven Data Poisoning Attacks and Their Mitigation in Federated Learning Systems // Electronics, 2023, Vol. 12, № 8, Article 1805. DOI: 10.3390/electronics12081805.

14. Zhao Y., Gong X., Lin F. & Chen X. Data Poisoning Attacks and Defenses in Dynamic Crowdsourcing With Online Data Quality Learning / IEEE Transactions on Mobile Computing, vol. 22, №. 5, pp. 2569–2581, May 2023, https://doi.org/10.1109/TMC.2021.3133365. (data obrashcheniya: 26.02.2025).

15. Zhong, H.; Liao, C.; Squicciarini, A.C.; Zhu, S.; Miller, D. Backdoor embedding in convolutional neural network models via invisible perturbation. In Proceedings of the Tenth ACM Conference on Data and Application Security and Privacy, New Orleans, LA, USA, 16–18 March 2020; pp. 97–108.

16. Ganjoo R., Ganjoo M., Patil M. Mitigating Poisoning Attacks in Federated Learning / Innovative Data Communication Technologies and Application, 2022, pp.687-699. URL: https://www.researchgate.net/publication/358822578_Mitigating_Poisoning_Attacks_in_Federated_Learning (data obrashcheniya: 26.02.2025).

17. Visger, Mark A. Garbage In, Garbage Out: Data Poisoning Attacks and Their Legal Implications, in Laura A. Dickinson, and Edward W. Berg (eds), Big Data and Armed Conflict: Legal Issues Above and Below the Armed Conflict Threshold, The Lieber Studies Series (New York, 2024; online edn, Oxford Academic, 14 Dec. 2023), https://doi.org/10.1093/oso/9780197668610.003.0008 (data obrashcheniya: 26.02.2025).

18. Dibaei M., Zheng X., Jiang K., Abbas R., Liu S., Zhang Y., Xiang Y. & Yu S. Attacks and defences on intelligent connected vehicles: a survey. Digit. Commun. Networks, 6, 2020, pp. 399-42. URL: https://www.semanticscholar.org/paper/Attacks-and-defences-on-intelligent-connected-a-Dibaei-Zheng/aae97ee34201666f98167402320b86e9facfa173 (data obrashcheniya: 26.02.2025).

19. Hong, Q.; He, B.; Zhang, Z.; Xiao, P.; Du, S.; Zhang, J. Circuit Design and Application of Discrete Cosine Transform Based on Memristor. IEEE J. Emerg. Sel. Top. Circuits Syst. 2023,13, 502–513.

20. Li S., Wang Y., Zhang X., Jiang Y., Xia S. A Study on Data Poisoning Attacks in Deep Generative Models // Applied Sciences, 2023, Vol. 14, No. 19, Article 8742. DOI: 10.3390/app14198742.

21. Yang Z., Zhang J., Wang W., Li H. Invisible Threats in the Data: A Study on Data Poisoning Attacks in Deep Generative Models // Applied Sciences, 2024, T. 14, №19, Article. 8742. https://doi.org/10.3390/app14198742 (data obrashcheniya: 26.02.2025).

ОЛИФИРЕНКО Артем Алексеевич, мидл Golang разработчик ООО «РеалИТ», магистрант кафедры Информационная безопасность автоматизированных систем федерального государственного бюджетного учреждения высшего образования «Саратовский государственный технический университет им. Юрия Алексеевича Гагарина». 410054, г. Саратов, ул. Политехническая, 77. E-mail: artemolifirenko@yandex

OLIFIRENKO Artem Alekseevich, Middle Golang developer at RealIT LLC, Master's student of the Department of Information Security of Automated Systems of the Federal State Budgetary Institution of Higher Education "Saratov State Technical University named after Yuri Alekseevich Gagarin". 410054, Saratov, Politekhnicheskaya St., 77. E-mail: artemolifirenko@yandex