Разработка подходов к увеличению устойчивости моделей машинного обучения для обнаружения распределенных атак отказа обслуживания

Д.В. Бербер, О.Р. Лапонина

Аннотация **– Данная статья посвящена анализу и разработке подходов для повышения устойчивости к состязательным атакам моделей машинного обучения, используемым для обнаружения распределенных атак отказа обслуживания. Для повышения устойчивости моделей машинного обучения выполнено увеличение обучающей выборки релевантными примерами, полученными с помощью генератора GAN модели. В качестве набора данных используется выборка DDoS Evaluation Dataset (CIC-DDoS2019) 2019г. Датасет разработан компанией Canadian Institute for Cybersecurity (CIC) и предназначен для использования в исследованиях и разработках в области обнаружения и предотвращения DDoS-атак. Данные содержат 128 027 наборов атак и 97 718 обычных запросов. Аномальные запросы содержат различные типы DDoS-атак, такие как: SYN-флуд, UDP-флуд, ICMP-флуд и HTTP-флуд. Используется обучение с учителем. Реализовано преобразование значений признаков выборки. В качестве модели выбран градиентный бустинг XGBoost, показывающий хорошие результаты на стандартных метриках. Произведен анализ состязательных атак на модель, обученную на исходном тренировочном датасете, и на модель, обученную на расширенном наборе данных, дополненном релевантными примерами, сгенерированными с помощью GAN модели над тренировочными данными. Генератором выступала G-часть сети Wasserstein GAN с градиентным штрафом (WGAN-GP). Для выполнения состязательных black box атак использовалась модифицированная библиотека ART класса ZooAttack. Для сохранения семантики вредоносных данных выполнена модификация библиотеки от IBM Adversarial Robustness Toolbox (ART). Использовались стандартные метрики качества для моделей. Получены результаты, показывающие, что добавление релевантных примеров в обучающую выборку повышает устойчивость модели к состязательным атакам. Однако на 200 итерациях даже более устойчивая модель не смогла показать качество сопоставимое с исходной тестовой выборкой, что говорит о том, что при достаточном количестве времени и неограниченном доступе к входам и выходам модели возможно подобрать состязательные примеры, на которых классификатор будет ошибаться.**

Ключевые слова **– обучение с учителем, состязательное обучение, состязательные атаки, распределенные атаки отказа обслуживания DDoS, устойчивость моделей машинного обучения.**

I. Введение

DDoS-атака (Distributed Denial of Service) — это тип кибератаки, при которой злоумышленники используют множество скомпрометированных компьютеров или устройств для отправки большого количества запросов или любого другого трафика на целевой сервер или сеть. Это приводит к перегрузке системы, делая её недоступной для легитимных пользователей.

Цель DDoS-атаки — нарушить нормальное функционирование системы, сделав её недоступной или замедлив её работу. Это может привести к финансовым потерям, потере репутации и другим негативным последствиям.

За последние годы системы обнаружения вторжений (NIDS) значительно улучшились благодаря использованию моделей машинного обучения. Современные системы NIDS состоят из экстрактора признаков и модели, основанной на машинном обучении. Экстрактор обрабатывает неструктурированный сетевой трафик, чтобы выделить ключевые характеристики, из которых формируются данные для анализа. Модель обучается на примерах вредоносных и безвредных данных, после чего она способна классифицировать новые случаи атак в реальном времени. Было предложено множество таких систем, которые показывают отличные результаты в обнаружении DDoS-атак [1], [2].

Однако и злоумышленники могут изменять свои атаки, чтобы оставаться незаметными для систем обнаружения, работающих по порогу или сигнатурам. Модели машинного обучения тоже могут быть подвержены атакам [3]. Цель состязательной атаки —

обойти модель классификации, чтобы успешно провести DDoS-атаку.

Существует два подхода к реализации состязательных атак против NIDS: возмущение данных и возмущение сетевого трафика. В первом подходе злоумышленник вносит возмущения непосредственно в предварительно обработанные образцы перед подачей их в ML модель. Во втором подходе злоумышленник генерирует состязательно возмущенные потоки DDoS, преобразуемые в признаковое пространство данных, которые будут классифицироваться моделью. При реализации первого подхода при генерации состязательных атак важно сохранить семантику входных данных. Так, например в [4] авторы статьи разделили признаки в наборе данных на неизменяемые, которые существенны для поддержания вредоносной функции, и изменяемые, которыми можно манипулировать, не затрагивая эту функцию.

Для предотвращения обхода моделей машинного обучения используются различные методы, в том числе метода состязательного обучения [5]. В данной работе мы используем генеративные состязательные сети (GAN). Генеративные состязательные сети (GAN) — это модели глубокого обучения, которые могут изучать распределение входных данных и создавать новые, похожие на них примеры. GAN состоит из двух моделей: дискриминатора и генератора, которые обучаются в процессе соревнования друг с другом. Генератор учится на распределении обучающих данных и создает новые образцы, которые напоминают оригинальные.

В работе с помощью GAN мы генерируем состязательные примеры на основе примеров из обучающей выборки, и дополняем ими новую обучающую выборку для второй модели. В работе показано, что такое расширение обучающей выборки делает модель более устойчивой к состязательным атакам и повышает метрики качества.

II. Постановка задачи

A. *Цели работы*

1. Разработать подходы к увеличению устойчивости к состязательным атакам модели машинного обучения для обнаружения распределенных атак отказа обслуживания. Устойчивость повышается за счет дополнения обучающей выборки релевантными примерами, полученными с помощью генератора GAN модели.
2. Сравнить результаты состязательных атак на модели.

B. *Задачи работы*

1. Собрать обучающий набор данных для построения модели бинарной классификации DDoS-атак и выполнить предварительную обработку данных. Разделить выборку на тренировочную и тестовую части.
2. Обучить модель на исходном тренировочном наборе данных и протестировать на тестовом наборе данных.
3. Дополнить исходную тренировочную выборку релевантными примерами на основе состязательного обучения GAN модели.
4. Обучить вторую модель на обновленном наборе данных и протестировать на тестовом наборе данных.
5. Провести состязательные атаки на первую и вторую модели и сравнить результаты классификации с тестовыми данными.
6. Проанализировать вклад состязательного обучения на качество классификации атак.

III. Описание и генерация данных

A. *Описание и предварительная обработка набора данных*

В данной статье рассматривается обучение с учителем, поэтому необходимо определить обучающую выборку, которая содержит как обычный трафик, так и аномальный трафик. Для обучения был выбран датасет CIC-DDoS2019, который можно найти в открытом доступе [6]. Он используется в исследованиях и разработках в области обнаружения и предотвращения DDoS-атак. CIC-DDoS2019 включает в себя данные о сетевых атаках типа «отказ в обслуживании» (DDoS). В датасете представлены различные типы DDoS-атак, такие как:

- SYN-флуд — атака, при которой злоумышленник отправляет большое количество SYN-запросов на сервер, пытаясь переполнить очередь TCP-соединений.
- UDP-флуд — атака, при которой злоумышленник отправляет большое количество UDP-пакетов на сервер, пытаясь перегрузить его обработкой запросов.
- ICMP-флуд — атака, при которой злоумышленник отправляет большое количество ICMP-пакетов на сервер, пытаясь вызвать его перегрузку.
- HTTP-флуд — атака, при которой злоумышленник отправляет большое количество HTTP-запросов на веб-сервер, пытаясь вызвать его перегрузку.

Датасет содержит 225 745 строк для обучения и 85 столбцов, которые характеризуют поступающие данные. В том числе столбец label, который сигнализирует наличие или отсутствие DDoS атаки ('DDoS' наличие атаки и 'Benign' отсутствие). Колонка label выбирается как целевая переменная. С помощью алгоритмов машинного обучения мы хотим по входным признакам предсказать значение колонки label - наличие или отсутствие атаки.

Из выборки убирались служебные колонки 'Flow ID', 'Sourse IP', 'Destination IP ', 'Timestamp', убирались константные и полностью уникальные признаки. Далее производилась нормировка данных с помощью

MinMaxScaler из библиотеки sklearn. Значения столбца label менялись с 'DDoS' и 'Benign' на 1 и 0 соответственно.
Для обучения и тестирования моделей датасет был разбит на обучающую и тестовую выборки в отношении размеров выборок 6 к 1.

Таблица 1. Размеры обучающей и тестовой выборок для исходной модели

	Датасет	#samples	#DDoS	#Benign
train	CIC-DDoS2019	188120	106689	81431
test	CIC-DDoS2019	37625	21338	16287

B. Расширение выборки релевантными примерами

Для тестирования гипотез об увеличении устойчивости моделей к состязательным атакам был сформирован расширенный вариант датасета с помощью генератора GAN моделей. Генератором выступала G-часть сети Wasserstein GAN с градиентным штрафом (WGAN-GP) [7], которая достигает высочайшего уровня производительности с точки зрения стабильности обучения и разнообразия сгенерированных образцов. Интуиция заключается в том, что, генерируя примеры похожие на стандартные запросы, а также генерируя примеры с DDoS атаками мы увеличиваем устойчивость модели к состязательным атакам, которые незначительно изменяют признаковое пространство с целью обмануть модель классификации.
На данном этапе тренировочная выборка для обучения второй модели была увеличена в 3 раза. Отдельно отметим, что тестовый набор данных оставался без изменений.

IV. МОДЕЛЬ МАШИННОГО ОБУЧЕНИЯ И МЕТРИКИ КАЧЕСТВА

A. Модели машинного обучения

В данной работе решается задача бинарной классификации, признаковое пространство сформировано из различных характеристик запросов. В качестве алгоритма машинного обучения в работе используется одна из реализаций градиентного бустинга - XGBoost. Среди алгоритмов классического машинного обучения именно градиентный бустинг показывает наилучший результат и считается SOTA (state-of-the-art) алгоритмом для табличного рода данных.

Градиентный бустинг — это последовательный метод обучения ансамбля, используемый для задач классификации и регрессии [8]. Базовые улучшения генерируются последовательно для достижения лучшего качества по сравнению с предыдущей итерацией, т. е. последовательного улучшения общей модели для каждой итерации.

Основная идея заключается в минимизации функции потерь, вызванной прогнозами предыдущего ансамбля. Алгоритм градиентного бустинга работает следующим образом:

1. Зафиксировать входные данные для обучающего набора $(a_i, b_i)_{i=1}^{n}$, функцию потерь $L(b, F(a))$ и количество итерации M.
2. Инициализировать модель с постоянным значением.
   $$F_0(a) = \underset{\varphi}{\operatorname{argmin}} \sum_{i=1}^{m} L(b_i, \varphi) \quad (1)$$
   где φ – предсказанное значение, $L(b_i, \varphi)$ – функция потерь.
3. Повторять для каждого нового дерева m = 1 до M.
   a. Рассчитать остатки или ошибку.
   $$e_{im} = \frac{-\vartheta L(b_i, F(a_i))}{\vartheta F(a_i)} \Big|_{F(a)=F_{m-1}(a)} \quad for i=1,n \quad (2)$$
   b. Обучить новый слабый предсказатель $h_m(a)$ на остатках предыдущего ансамбля на тренировочном датасете
   c. Вычислить φ_m решив следующую задачу.
   $$\varphi_m = \underset{\varphi}{\operatorname{argmin}} \sum_{i=1}^{m} L(b_i, F_{m-1}(a_i)) + \varphi h_m(a_i) \quad (3)$$
4. Обновить ансамбль
   $$F_m(a_i) = F_{m-1}(a_i) + \varphi_m h_m(\text{a}) \quad (4)$$
5. Вывести результат $F_m(a)$

B. Метрики качества

Для оценки эффективности работы моделей будем использовать стандартный набор метрик качества моделей бинарной классификации, а также auc_roc, так как алгоритм градиентного бустинга на выходе может выдавать вероятности принадлежности данному классу. В качестве метрик качества каждой модели используются следующие показатели.
True Positive (TP) – количество правильно классифицированного нормального трафика.
True Negative (TN) - количество правильно классифицированного аномального трафика.
False Negative (FN) - количество неправильно классифицированного нормального трафика (нормальный трафик классифицирован как аномальный).
False Positive (FP) - количество неправильно классифицированного аномального трафика (аномальный трафик классифицирован как нормальный).
Рассматриваемые метрики качества:
Accuracy (доля верных ответов). Вычисляется по формуле:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

Precision (точность). Определяет, на сколько можно доверять классификатору. Определяется формулой:

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

Recall (полнота). Показывает, как много объектов класса «аномальный трафик» распознает классификатор. Для вычисления используется формула:

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

F1-metrics - среднее гармоническое точности и полноты, которое вычисляется по формуле:

$$F1 = 2\frac{Precision*Recall}{Precision+Recall} \quad (8)$$

AUC ROC. Для расчета метрики **auc roc** – площади под графиком в осях FPR и TPR. Физический смысл метрики: AUC ROC равен доле пар объектов вида (объект класса 1, объект класса 0), которые алгоритм упорядочил верно.

С помощью метрик accuracy, precision, recall, F1, AUC ROC далее будем оценивать качество моделей.

V. Состязательные атаки

A. Классификация состязательных атак

При атаке на модель цель злоумышленника состоит в том, чтобы сгенерировать примеры, классификация которых может быть изменена во время применения модели на произвольный класс по выбору злоумышленника с минимальным изменением данных [9].

Атаки на модели машинного обучения можно разбить на 2 группы:

1) White-box атаки с полным доступом к весам и градиентам модели.
2) Black-box атаки с доступом только к входным и выходным данным модели.

В 2013 году Szedegy [10] и Biggio [11] независимо друг от друга обнаружили эффективный метод генерации примеров состязательных атак с линейными моделями и нейронными сетями путем применения градиентной оптимизации к состязательной целевой функции. Оба этих метода требуют доступа к модели (white-box атаки).

Даже в условиях black-box атаки, когда злоумышленники получают предсказанные моделью метки или вероятности классов, нейронные сети и классические модели машинного обучения по-прежнему уязвимы. Методы создания состязательных примеров при black-box атаках включают оптимизацию нулевого порядка [12], дискретную оптимизацию [13] и байесовскую оптимизацию [14], а также возможность переноса, которая предполагает генерацию состязательных примеров в white-box на другой архитектуре модели перед их переносом в целевую модель [15]. Наиболее многообещающими направлениями для снижения критической угрозы атак на модели являются состязательное обучение, рандомизированное сглаживание [16] (оценка прогноза ML в условиях шума); и формальные методы проверки [17] (применение формальных методов для проверки выходных данных модели). Всегда существует компромисс между устойчивостью модели и точностью классификации.

B. Применение состязательных атак

В данной работе к датасетам (начальном и дополненном генеративными примерами) применялись атаки типа black-box. Популярным методом является оптимизация нулевого порядка, которая оценивает градиенты модели без явного вычисления производных [12]. Также существуют методы, включающие дискретную оптимизацию [13], генетические алгоритмы [18] и случайные блуждания [19].

В текущей работе применялись атаки с оптимизацией нулевого порядка без доступа к градиентам модели (Zeroth order optimization based black-box attacks). За основу была взята реализация атак в библиотеке от IBM Adversarial Robustness Toolbox (ART) с помощью класса ZooAttack. Отметим, что при возмущении признаков трафика необходимо оставлять семантику запроса. Некоторые поля датасета после генерации состязательных примеров должны оставаться в исходном виде, например поле адрес получателя. Однако реализация класса ZooAttack в библиотеке ART не позволяет оставлять конкретные поля без изменений. В данной работе была произведена модификация алгоритма, для возможности сохранения выбранных полей в исходном виде, чтобы оптимизация нулевого порядка происходила в оставшемся наборе признаков.

Для экспериментов данный алгоритм запускался в двух вариантах: со значениями параметра max_iter 10 и 200. Этот параметр отвечает за максимальное число итераций при итеративном изменении входных данных для поиска состязательного примера. Чем больше это число, тем больше вероятность найти состязательный пример и тем дольше работа алгоритма.

VI. Итоги экспериментов

A. Схема экспериментов

Для исследования модели на устойчивость к состязательным атакам было сформировано две обучающие выборки:

1) Исходная обучающая выборка, после деления на train/test части.
2) Обучающая выборка, дополненная сгенерированными данными GAN моделью.

Тестовая выборка оставалась неизменной. Схема эксперимента показана на Рис. 1.

Для того, чтобы сохранять семантику запросов при генерации состязательных атак мы сохраняли неизменными следующие поля датасета:

- Destination Port *Порт назначения трафика*
- Source Port *Порт источника трафика*

- Fwd Header Length *Общая длина заголовков пакетов, отправленных в прямом направлении*
- Bwd Header Length *Общая длина заголовков пакетов, отправленных в обратном направлении*
- Protocol *Протокол, используемый для передачи данных (TCP, UDP, ICMP и т.д.)*

Также из датасета исключались служебные поля, добавление которых в модель ведет к переобучению:

- Source IP *IP-адрес источника трафика*
- Destination IP *IP-адрес назначения трафика*
- Flow ID *Уникальный идентификатор потока трафика*
- Timestamp *Временная метка начала потока*

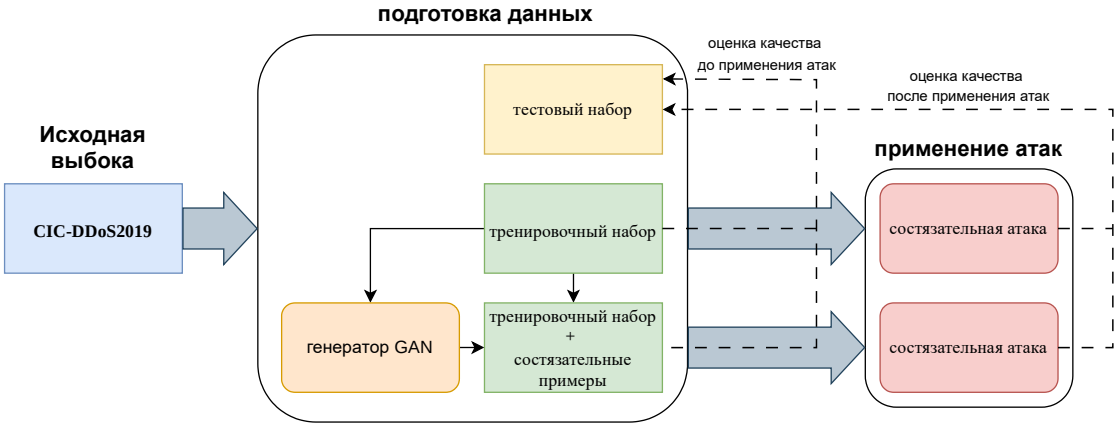

Рисунок 1. Схема проведения экспериментов

На каждой из обучающих выборок была обучена модель градиентного бустинга XGBoostClassifier. Эксперимент проводился в среде Google Collab со следующими версиями библиотек:
adversarial-robustness-toolbox==1.19.1
matplotlib==3.10.0
numpy==2.0.2
pandas==2.2.2
scikit-learn==1.6.1
scipy==1.14.1
tqdm==4.67.1
xgboost==2.1.4

B. Результат экспериментов

Обе модели (далее будем называть их default и augmented) на тестовом наборе данных показали качество, близкое к идеальному, почти полностью разделив классы (Таблица 2). Из эксперимента видно, что качество классификации за счет добавления генеративных примеров на выбранных метриках дополненной моделью уменьшилось несущественно.

Таблица 2. Качество классификации исходной (Default Model) и дополненной (Augm. Model) моделей на тестовых данных.

	Accuracy	Precision	Recall	F1	roc_auc
Default Model	0.9999	1.0000	0.9999	0.9999	0.9999
Augm. Model	0.9998	0.9999	0.9998	0.9999	0.9999

Теперь посмотрим воздействие атак на две модели. На каждую модель было сгенерировано 1000 примеров состязательных атак с параметрами **max_iter** равным 10 и 200 для более качественной генерации состязательных атак. Качество моделей классификации на состязательных примерах можно видеть в Таблице 3.

Таблица 3. Качество классификации исходной (Default Model) и дополненной (Augm. Model) моделей на состязательных примерах данных, сгенерированными с параметрами max_iter 10 и 200.

	Accuracy	Precision	Recall	F1	roc_auc
Default Model 10 iters	0.605	0.735	0.446	0.555	0.886
Default Model 200 iters	0.420	0.437	0.1225	0.194	0.810
Augm. Model 10 iters	0.999	1.000	0.982	0.990	0.996
Augm. Model 200 iters	0.950	0.999	0.910	0.952	0.970

Из таблицы видно, что для исходной модели даже на небольшом количестве итераций модифицированная ZooAttack смогла сгенерировать состязательные атаки, которые существенно понизили метрики классификации.

Для модели, обученной на выборке, дополненной генеративными состязательными примерами, 10 итераций не хватило, чтобы создать большое

количество неправильных срабатываний. Однако на 200 итерациях даже более устойчивая модель не смогла показать качество сопоставимое с исходной тестовой выборкой, что говорит о том, что при достаточном количестве времени и неограниченном доступе к входам и выходам модели возможно подобрать состязательные примеры, на которых классификатор будет ошибаться.

VII. Заключение

В данной работе показано, что модель детектирования DDoS атак, обучающая выборка которой не подразумевает расширение состязательными примерами, является неустойчивой даже к black-box типу атак. Расширение обучающей выборки генеративными примерами повышает устойчивость к состязательным атакам, однако в данном процессе необходимо находить компромисс между устойчивостью и точностью, так как излишнее увеличение выборки состязательными примерами может вести к падению качества классификации.

В работе также показано, что дополнение обучающих данных состязательными примерами повышает устойчивость модели к состязательным атакам типа black-box, но не полностью защищает от них.

Благодарности

Тема статьи, написанной по результатам магистерской диссертации, соответствует направлению “Кибербезопасность” на факультете ВМК МГУ [20]. Как примеры похожих работ см. публикации [21,22].

Как обычно, отмечаем работы В.П. Куприяновского и его многочисленных соавторов, положивших начало цифровой повестке в журнале [23, 24].

Библиография

[1] M. S. Elsayed, N.-A. Le-Khac, S. Dev, and A. D. Jurcut, “DDoSNet: A deep-learning model for detecting network attacks,” in 2020 IEEE 21st International Symposium on” A World of Wireless, Mobile and Multimedia Networks”(WoWMoM). IEEE, 2020, pp. 391–396.

[2] T. Fardusy, S. Afrin, I. J. Sraboni and U. K. Dey, "An Autoencoder-Based Approach for DDoS Attack Detection Using Semi-Supervised Learning," 2023 International Conference on Next-Generation Computing, IoT and Machine Learning (NCIM), Gazipur, Bangladesh, 2023, pp. 1-7, doi: 10.1109/NCIM59001.2023.10212626.

[3] Costa, Joana C., et al. "How deep learning sees the world: A survey on adversarial attacks & defenses." IEEE Access 12 (2024): 61113-61136.

[4] Q. Yan, M. Wang, W. Huang, X. Luo, and F. R. Yu, “Automatically synthesizing dos attack traces using generative adversarial networks,” International Journal of Machine Learning and Cybernetics, vol. 10, no. 12, pp. 3387–3396, 2019.

[5] Abdelaty, Maged, et al. "Gadot: Gan-based adversarial training for robust ddos attack detection." 2021 IEEE Conference on Communications and Network Security (CNS). IEEE, 2021.

[6] Kaggle https://www.kaggle.com/datasets/aymenabb/ddos-evaluation-dataset-cic-ddos2019 Retrieved: 05.05.2025

[7] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, “Improved training of wasserstein gans,” in Advances in neural information processing systems, 2017, pp. 5767–5777.

[8] A. Alsirhani, S. Sampalli, P. Bodorik, “DDoS detection system: utilizing gradient boosting algorithm and apache spark”, in: Proceedings of the”IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), 2018 (pp. 1-6). IEEE.

[9] Apostol Vassilev, Alina Oprea, Alie Fordyce, Hyrum Anderson “Adversarial Machine Learning A Taxonomy and Terminology of Attacks and Mitigations” NIST Trustworthy and Responsible AI NIST AI 100-2e2023 https://doi.org/10.6028/NIST.AI.100-2e2023

[10] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In International Conference on Learning Representations, 2014.

[11] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Srndic´, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. “Evasion attacks against machine learning at test time.” In Joint European conference on machine learning and knowledge discovery in databases, pages 387–402. Springer, 2013.

[12] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. “Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training” substitute models. In Proceedings of the 10th ACM Workshop on Artif- cial Intelligence and Security, AISec ’17, page 15–26, New York, NY, USA, 2017. Association for Computing Machinery.

[13] Seungyong Moon, Gaon An, and Hyun Oh Song. “Parsimonious black-box adversarial attacks via effcient combinatorial optimization”. In International Conference on Machine Learning (ICML), 2019.

[14] Satya Narayan Shukla, Anit Kumar Sahu, Devin Willmott, and Zico Kolter. “Simple and effcient hard label black-box adversarial attacks in low query budget regimes.” In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, KDD ’21, page 1461–1469, New York, NY, USA, 2021. Association for Computing Machinery.

[15] Papernot, Nicolas, Patrick McDaniel, and Ian Goodfellow. "Transferability in machine learning: from phenomena to black-box attacks using adversarial samples." arXiv preprint arXiv:1605.07277 (2016).

[16] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. “Certifed adversarial robustness via randomized smoothing.” In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, Proceedings of the 36th International Conference on Machine Learning, volume 97 of Proceedings of Machine Learning Research, pages 1310–1320. PMLR, 09–15 Jun 2019.

[17] Timon Gehr, Matthew Mirman, Dana Drachsler-Cohen, Petar Tsankov, Swarat Chaudhuri, and Martin Vechev. “AI2: Safety and robustness certifcation of neural networks with abstract interpretation.” In 2018 IEEE Symposium on Security and Privacy (S&P), pages 3–18, 2018.

[18] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. “Black-box adver- sarial attacks with limited queries and information. In Jennifer G. Dy and An- dreas Krause, editors”, Proceedings of the 35th International Conference on Ma- chine Learning, ICML 2018, Stockholmsma¨ssan, Stockholm, Sweden, July 10-15, 2018, volume 80 of Proceedings of Machine Learning Research, pages 2142–2151. PMLR, 2018.

[19] Nina Narodytska and Shiva Kasiviswanathan. “Simple black-box adversarial attacks on deep neural networks.” In 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pages 1310–1318, 2017.

[20] Сухомлин, Владимир Александрович, et al. "Модель цифровых навыков кибербезопасности 2020." Современные информационные технологии и ИТ-образование 16.3 (2020): 695-710.

[21] Yudova, E. A., and Olga R. Laponina. "Analysis of the possibilities of using machine learning technologies to detect attacks on web applications." International Journal of Open Information Technologies 10.1 (2021): 61-68.

[22] Korniukhina, Sofia P., and Olga R. Laponina. "Research of the Capabilities of Deep Learning Algorithms to Protection Against Phishing Attacks." International Journal of Open Information Technologies 11.6 (2023): 163-174.

[23] Искусственный интеллект как стратегический инструмент экономического развития страны и совершенствования ее государственного управления. Часть 2. Перспективы применения искусственного интеллекта в России для государственного управления / И. А. Соколов, В. И. Дрожжинов, А. Н. Райков [и др.] // International Journal of Open Information Technologies. – 2017. – Т. 5, № 9. – С. 76-101. – EDN ZEQDMT.

[24] Развитие транспортно-логистических отраслей Европейского Союза: открытый BIM, Интернет Вещей и кибер-физические системы / В. П. Куприяновский, В. В. Аленьков, А. В. Степаненко [и др.] // International Journal of Open Information Technologies. – 2018. – Т. 6, № 2. – С. 54-100. – EDN YNIRFG.

Статья получена 07 мая 2025.

Д.В. Бербер – МГУ имени М.В. Ломоносова (email: s_dasha-00@mail.ru).

О.Р. Лапонина – МГУ имени М.В. Ломоносова (email: laponina@oit.cmc.msu.ru).

Developing approaches to increase the robustness of machine learning models for detecting distributed denial of service attacks

D.V. Berber, O.R. Laponina

***Abstract* – This paper analyzes and develops approaches to improve the resilience of machine learning models used to detect distributed denial of service attacks to adversarial attacks. To improve the resilience of machine learning models, the training set was augmented with relevant examples obtained using the GAN model generator. The 2019 DDoS Evaluation Dataset (CIC-DDoS2019) is used as a dataset. The dataset was developed by the Canadian Institute for Cybersecurity (CIC) and is intended for use in research and development in the field of DDoS attack detection and prevention. The data contains 128,027 attack sets and 97,718 normal requests. The abnormal requests contain various types of DDoS attacks, such as SYN flood, UDP flood, ICMP flood, and HTTP flood. Supervised learning is used. The transformation of sample feature values is implemented. XGBoost gradient boosting, which shows good results on standard metrics, is chosen as a model. An analysis of adversarial attacks was performed on the model trained on the original training dataset and on the model trained on an extended dataset supplemented with relevant examples generated using the GAN model over the training data. The generator was the G-part of the Wasserstein GAN network with a gradient penalty (WGAN-GP). To perform adversarial black box attacks, a modified ART library of the ZooAttack class was used. To preserve the semantics of malicious data, a modification of the library from IBM Adversarial Robustness Toolbox (ART) was performed. Standard quality metrics for models were used. The results obtained show that adding relevant examples to the training set increases the model's resistance to adversarial attacks. However, at 200 iterations, even a more robust model could not show quality comparable to the original test set, which suggests that with sufficient time and unlimited access to the model's inputs and outputs, it is possible to select adversarial examples on which the classifier will make mistakes.**

***Keywords* – supervised learning, adversarial learning, adversarial attacks, distributed denial of service attacks (DDoS), resilience of machine learning models**

REFERENCES

[1] M. S. Elsayed, N.-A. Le-Khac, S. Dev, and A. D. Jurcut, "DDoSNet: A deep-learning model for detecting network attacks," in 2020 IEEE 21st International Symposium on" A World of Wireless, Mobile and Multimedia Networks"(WoWMoM). IEEE, 2020, pp. 391–396.

[2] T. Fardusy, S. Afrin, I. J. Sraboni and U. K. Dey, "An Autoencoder-Based Approach for DDoS Attack Detection Using Semi-Supervised Learning," 2023 International Conference on Next-Generation Computing, IoT and Machine Learning (NCIM), Gazipur, Bangladesh, 2023, pp. 1-7, doi: 10.1109/NCIM59001.2023.10212626.

[3] Costa, Joana C., et al. "How deep learning sees the world: A survey on adversarial attacks & defenses." IEEE Access 12 (2024): 61113-61136.

[4] Q. Yan, M. Wang, W. Huang, X. Luo, and F. R. Yu, "Automatically synthesizing dos attack traces using generative adversarial networks," International Journal of Machine Learning and Cybernetics, vol. 10, no. 12, pp. 3387–3396, 2019.

[5] Abdelaty, Maged, et al. "Gadot: Gan-based adversarial training for robust ddos attack detection." 2021 IEEE Conference on Communications and Network Security (CNS). IEEE, 2021.

[6] Kaggle https://www.kaggle.com/datasets/aymenabb/ddos-evaluation-dataset-cic-ddos2019 Retrieved: 05.05.2025

[7] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," in Advances in neural information processing systems, 2017, pp. 5767–5777.

[8] A. Alsirhani, S. Sampalli, P. Bodorik, "DDoS detection system: utilizing gradient boosting algorithm and apache spark", in: Proceedings of the"IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), 2018 (pp. 1-6). IEEE.

[9] Apostol Vassilev, Alina Oprea, Alie Fordyce, Hyrum Anderson "Adversarial Machine Learning A Taxonomy and Terminology of Attacks and Mitigations" NIST Trustworthy and Responsible AI NIST AI 100-2e2023 https://doi.org/10.6028/NIST.AI.100-2e2023

[10] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In International Conference on Learning Representations, 2014.

[11] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Srndic´, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. "Evasion attacks against machine learning at test time." In Joint European conference on machine learning and knowledge discovery in databases, pages 387–402. Springer, 2013.

[12] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. "Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training" substitute models. In Proceedings of the 10th ACM Workshop on Artif- cial Intelligence and Security, AISec '17, page 15–26, New York, NY, USA, 2017. Association for Computing Machinery.

[13] Seungyong Moon, Gaon An, and Hyun Oh Song. "Parsimonious black-box adversarial attacks via effcient combinatorial optimization". In International Conference on Machine Learning (ICML), 2019.

[14] Satya Narayan Shukla, Anit Kumar Sahu, Devin Willmott, and Zico Kolter. "Simple and effcient hard label black-box adversarial attacks in low query budget regimes." In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, KDD '21, page 1461–1469, New York, NY, USA, 2021. Association for Computing Machinery.

[15] Papernot, Nicolas, Patrick McDaniel, and Ian Goodfellow. "Transferability in machine learning: from phenomena to black-box attacks using adversarial samples." arXiv preprint arXiv:1605.07277 (2016).

[16] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. "Certifed adversarial robustness via randomized smoothing." In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, Proceedings of the 36th International Conference on Machine Learning, volume 97 of Proceedings of Machine Learning Research, pages 1310–1320. PMLR, 09–15 Jun 2019.

[17] Timon Gehr, Matthew Mirman, Dana Drachsler-Cohen, Petar Tsankov, Swarat Chaudhuri, and Martin Vechev. "AI2: Safety and

robustness certifcation of neural networks with abstract interpretation." In 2018 IEEE Symposium on Security and Privacy (S&P), pages 3–18, 2018.

[18] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. "Black-box adver- sarial attacks with limited queries and information. In Jennifer G. Dy and An- dreas Krause, editors", Proceedings of the 35th International Conference on Ma- chine Learning, ICML 2018, Stockholmsma¨ssan, Stockholm, Sweden, July 10-15, 2018, volume 80 of Proceedings of Machine Learning Research, pages 2142–2151. PMLR, 2018.

[19] Nina Narodytska and Shiva Kasiviswanathan. "Simple black-box adversarial attacks on deep neural networks." In 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pages 1310–1318, 2017.

[20] Suhomlin, Vladimir Aleksandrovich, et al. "Model' cifrovyh navykov kiberbezopasnosti 2020." Sovremennye informacionnye tehnologii i IT-obrazovanie 16.3 (2020): 695-710.

[21] Yudova, E. A., and Olga R. Laponina. "Analysis of the possibilities of using machine learning technologies to detect attacks on web applications." International Journal of Open Information Technologies 10.1 (2021): 61-68.

[22] Korniukhina, Sofia P., and Olga R. Laponina. "Research of the Capabilities of Deep Learning Algorithms to Protection Against Phishing Attacks." International Journal of Open Information Technologies 11.6 (2023): 163-174.

[23] Iskusstvennyj intellekt kak strategicheskij instrument jekonomicheskogo razvitija strany i sovershenstvovanija ee gosudarstvennogo upravlenija. Chast' 2. Perspektivy primenenija iskusstvennogo intellekta v Rossii dlja gosudarstvennogo upravlenija / I. A. Sokolov, V. I. Drozhzhinov, A. N. Rajkov [i dr.] // International Journal of Open Information Technologies. – 2017. – T. 5, # 9. – S. 76-101. – EDN ZEQDMT.

[24] Razvitie transportno-logisticheskih otraslej Evropejskogo Sojuza: otkrytyj BIM, Internet Veshhej i kiber-fizicheskie sistemy / V. P. Kuprijanovskij, V. V. Alen'kov, A. V. Stepanenko [i dr.] // International Journal of Open Information Technologies. – 2018. – T. 6, # 2. – S. 54-100. – EDN YNIRFG