МЕТОДИКА ПОСТРОЕНИЯ УСТОЙЧИВОЙ СИСТЕМЫ ЗАЩИТЫ НА ОСНОВЕ СОСТЯЗАТЕЛЬНОГО МАШИННОГО ОБУЧЕНИЯ В БЕСПРОВОДНЫХ СЕТЯХ 6G

Легашев Л.В.[1], Гришина Л.С.[2]

Аннотация

Цель исследования: *разработка методики аналитической обработки больших массивов данных сервисов и приложений в сетях последнего поколения для обнаружения инцидентов кибербезопасности и построения устойчивых систем защиты на основе состязательного машинного обучения.*

Метод исследования: *анализ современных методов машинного обучения и нейросетевых технологий, синтез и формализация алгоритмов состязательных атак на модели машинного обучения.*

Результат исследования: *в статье предложена методика построения устойчивой системы защиты от состязательных атак в беспроводных самоорганизующихся сетях последнего поколения. Формализованы основные виды состязательных атак, в том числе отравляющие атаки и атаки уклонения, а также описаны методы генерации состязательных примеров на табличные, текстовые и визуальные данные. Проведена генерация нескольких сценариев и исследовательский анализ наборов данных с помощью эмулятора DeepMIMO. Выделены потенциальные прикладные задачи бинарной классификации и прогнозирования затухания сигнала между пользователем и базовой станцией для проведения состязательных атак. Представлена алгоритмизация процессов построения и обучения устойчивой системы от состязательных атак в беспроводных сетях последнего поколения на примере эмулируемых данных.*

Научная новизна: *представлена методика аналитической обработки больших массивов эмулируемых данных сервисов и приложений для обнаружения инцидентов кибербезопасности, которая обеспечивает задел в области исследования вопросов безопасности сложных интеллектуальных сервисов и приложений в инфраструктуре беспроводных сетей последнего поколения.*

Ключевые слова: *состязательные атаки, беспроводные самоорганизующиеся сети, машинное обучение, MIMO.*

DOI:10.21681/2311-3456-2023-2-99-108

Введение и обзор современного состояния исследований

Повсеместное распространение беспроводных сетей последнего поколения и, как следствие, возросший уровень передаваемых по сети данных влечет за собой проблемы обеспечения безопасности. Современные методы машинного обучения активно используются для анализа трафика и выявления злонамеренной сетевой активности, но при этом сами модели глубокого обучения уязвимы к состязательным атакам, цель которых заключается в компрометации эффективности таких моделей. Как отмечается в докладе сотрудников лаборатории Касперского[3] существуют два основных сценария атак на модели машинного обучения: атаки вида «белый ящик» (white box) характерны для случая, когда злоумышленник имеет прямой доступ к модели машинного обучения с возможностью исследования исходного кода, архитектуры, маскировки и усиления отдельных признаков; атаки вида «черный ящик» (black box) характерны для случая, когда злоумышленник не имеет прямого доступа к модели машинного обучения, но имеет

3 Атаки на искусственный интеллект URL: https://media.kaspersky.com/ru/business-security/attacks-on-artificial-intelligence-whitepaper.pdf

1 Легашев Леонид Вячеславович, кандидат технических наук, ведущий научный сотрудник лаборатории цифровых решений и аналитики больших данных Оренбургского государственного университета, г. Оренбург, Россия. E-mail: silentgir@gmail.com. ORCID: 0000-0001-6351-404X

2 Гришина Любовь Сергеевна, младший научный сотрудник лаборатории цифровых решений и аналитики больших данных Оренбургского государственного университета, г. Оренбург, Россия. E-mail: grishina_ls@inbox.ru ORCID: 0000-0003-2752-7198

возможность тестировать модель на специально подготовленных вредоносных файлах для выявления ее слабых сторон. Как правило, у злоумышленника отсутствует прямой доступ к модели машинного обучения, поэтому атаки вида black box имеют наибольшее распространение и в текущем исследовании наше внимание будет сфокусировано именно на них.

Атака на нейронную сеть и/или модель машинного обучения называется состязательной атакой (adversarial attack) и ее цель состоит в намеренном искажении выходных данных при подаче на вход специально подготовленных данных (adversarial samples). Цель злоумышленника в этом случае заключается в снижении эффективности обученных моделей машинного обучения. Авторы исследования [1] успешно выполняют состязательную атаку по отношению к системе распознавания лиц, добавляя невидимый человеческому глазу шум в исходные данные, в результате чего лица на фотографиях не распознаются в большинстве случаев. В работах [2] и [3] также описывается успешное построение состязательных изображений с целью снижения качества классификации. В работе [4] исследователи используют универсальные состязательные триггеры (universal adversarial triggers) – специально подготовленные токены, которые после операции конкатенации с любыми входными данными модели обработки естественных языков приводят к получению специфического предсказания. Авторы отмечают возможность изменения сентиментного анализа текста на противоположный, а также успешную генерацию расистских высказываний для языковой модели GPT-2. В статье [5] представлена библиотека TextAttack для генерации состязательных примеров атак на языковые модели для некорректной классификации и логического вывода заданного текста. Авторы исследования [6] используют алгоритмы на основе эмбеддингов для некорректной классификации жанра текста на русском языке для модели XLM-RoBERTa. Публикация [7] посвящена исследованию состязательных атак по отношению к финансовым транзакциям. Авторы отмечают, что добавление токенов в конец записи приводит к снижению качества модели. Наибольшее негативное влияние состязательные атаки могут оказывать на различные экспертные медицинские системы прогнозирования на основе исходных данных пациентов, как отмечается в публикации [8]. В исследовании [9] успешно реализованы атаки вида black box и white box в интеллектуальных системах здравоохранения с целью компрометации медицинских данных пациентов и некорректной классификации диагнозов.

Для повышения эффективности сетей беспроводной связи поколения 6G используются концепции сверхплотной сети, связи миллиметрового диапазона и метод пространственного кодирования сигнала с использованием множества антенн (massive Multiple Input Multiple Output, massive MIMO). В исследовании [10] представлена реализация состязательной атаки вида «белый ящик» на информацию о состоянии канала системы massive MIMO. Анализ производительности нормализованной среднеквадратической ошибки показал ярко выраженный деструктивный эффект проводимой атаки. В работе [11] описывается безопасный автопрекодер на основе состязательного обучения (ASAP) для каналов прослушивания MIMO. В статье [12] предлагается метод смягчения состязательных атак на предложенные модели машинного обучения 6G для прогнозирования луча миллиметрового диапазона mmWave методом быстрого градиентного знака. Похожая задача решается в исследовании [13], при этом авторы используют состязательное обучение и защитную дистилляцию для смягчения последствий проводимой атаки.

Одним из наиболее перспективных направлений развития беспроводных сетей являются интеллектуальные транспортные системы и связанные с ними автомобильные самоорганизующиеся сети (Vehicular ad hoc networks, VANETs). В исследовании [14] авторами представлен новый протокол на основе MIMO-OFDM, который повышает пропускную способность и уменьшает задержку в VANETs. Авторы статьи [15] обсуждают вопросы развития интеллектуальных транспортных 6G сетей, а также потенциал MIMO коммуникаций. Авторы работы [16] фокусируют исследование на том, что моделирование каналов с использованием технологий MIMO является одним из ключевых факторов проектирования и улучшения автомобильной коммуникации для сетей B5G и 6G.

В этой статье будет предложена методика построения устойчивой системы защиты от состязательных атак на основе состязательного машинного обучения и выделены примеры прикладных задач для беспроводных самоорганизующихся сетей и различных моделируемых сценариях распространения сигнала MIMO антенн.

1. Формализация видов состязательных атак

Рассмотрим три основные области искусственного интеллекта, в рамках которых будут проводиться исследования защиты от состязательных атак на модели машинного обучения в беспроводных сетях: визуальные данные $X_{samples}$ из области компьютерного зрения, ко-

торую обозначим через *cv*, текстовые данные $X_{triggers}$ из области обработки естественного языка, которую обозначим через *nlp*, табличные данные X_{arrays} из области машинного обучения, которые обозначим через *tbl*.

Рассмотрим два основных типа состязательных атак, на которые будем акцентировать внимание: отравляющие атаки (poisoning attacks) – вид атак, выполняемых в момент обучения моделей искусственного интеллекта, связанных с подмешиванием «отравленных» данных в тренировочный набор данных ($data_{train}$). Обозначим такой вид атак через *poi*. Атаки уклонения (evasion attacks) – вид атак, выполняемых на готовые модели машинного обучения с целью компрометации их надёжности при работе с тестовым набором данных ($data_{test}$). В этом случае злоумышленник работает только с вводимыми данными, обозначим такой вид атак через *eva*.

Рассмотрим два основных типа атак с точки зрения цели злоумышленника: целенаправленные атаки (targeted attacks) – вид атак, при котором у злоумышленника есть целевой вектор признаков, который должен быть ошибочно классифицирован как нормальный или принадлежащий определенному классу. Атаки на надёжность моделей (reliability attacks) – вид атак, выполняемых на готовые модели машинного обучения с целью ухудшения основных метрик качества моделей. В текущем исследовании мы сфокусируемся именно на атаках на надёжность моделей.

Для исходного множества данных X и множества возможных классов (меток) Y в общем виде представим модель классификации как $f(x) \to y$, где каждому элементу x из X в соответствие присваивается метка y из Y. Доля правильных ответов (accuracy) является одной из множества метрик для оценки моделей классификации машинного обучения. В общем виде доля правильных ответов рассчитывается как отношение количества корректных предсказаний к общему числу предсказаний. Для случая бинарной классификации доля правильных ответов традиционно задаётся в виде:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

где: *TP* – истинно-положительный результат классификации, *TN* – истинно-отрицательный результат классификации, *FP* – ложно-положительный результат классификации и *FN* – ложно-отрицательный результат классификации.

Тогда определим состязательные атаки следующим образом:

1. $aa\langle cv, poi\rangle$ – состязательная атака в области компьютерного зрения, связанная с отравлением исходных изображений $X_{samples}$ таким образом, чтобы на выходе классификатора минимизировать долю правильных ответов $accuracy(data_{train}) \to \min$.

2. $aa\langle cv, eva\rangle$ – состязательная атака в области компьютерного зрения, связанная с вводом таких изображений $X_{samples}$, чтобы максимизировать ложно-отрицательность $FN(data_{test}) \to \max$ и минимизировать долю правильных ответов $accuracy(data_{test}) \to \min$.

3. $aa\langle nlp, poi\rangle$ – состязательная атака в области обработки естественного языка, связанная с отравлением исходных текстовых данных $X_{triggers}$ таким образом, чтобы на выходе классификатора минимизировать долю правильных ответов $accuracy(data_{train}) \to \min$.

4. $aa\langle nlp, eva\rangle$ – состязательная атака в области обработки естественного языка, связанная с вводом таких текстовых данных $X_{triggers}$, чтобы максимизировать ложно-отрицательность $FN(data_{test}) \to \max$ и минимизировать долю правильных ответов $accuracy(data_{test}) \to \min$.

5. $aa\langle tbl, poi\rangle$ – состязательная атака машинного обучения, связанная с отравлением исходных табличных данных X_{arrays} таким образом, чтобы на выходе классификатора минимизировать долю правильных ответов $accuracy(data_{train}) \to \min$.

6. $aa\langle tbl, eva\rangle$ – состязательная атака машинного обучения, связанная с вводом таких табличных данных X_{arrays}, чтобы максимизировать ложно-отрицательность $FN(data_{test}) \to \max$ и минимизировать долю правильных ответов $accuracy(data_{test}) \to \min$.

В интеллектуальных транспортных системах атаки вида $aa\langle cv, poi\rangle$ и $aa\langle cv, eva\rangle$ могут использоваться для того, чтобы некорректно классифицировать виды дорожных знаков бортовыми камерами транспортных средств. Атаки вида $aa\langle nlp, poi\rangle$ и $aa\langle nlp, eva\rangle$ могут использоваться для некорректной классификации сообщений, рассылаемых транспортными средствами друг другу (например, информация об аварии, пробке, состоянии дорожного полотна и др.). Атаки вида $aa\langle tbl, poi\rangle$ и $aa\langle tbl, eva\rangle$ могут использоваться для компрометации числовых показателей транспортного средства (скорость, направление движения, координаты).

2. Формализация методов генерации состязательных примеров

Рассмотрим базовые подходы для генерации состязательных образцов, которые могут быть применены для атаки моделей машинного обучения, построенных на основе визуальных, текстовых или табличных данных.

2.1. Генерация состязательных примеров для атак на визуальные данные

2.1.1 Метод быстрого градиентного знака (Fast Gradient Sign Method, FGSM).

Идея данного метода заключается в том, что он вычисляет градиенты функции потерь по отношению к исходным данным (в частности, изображению), а затем использует знак градиентов для создания нового «отравленного» изображения, которое максимизирует потери J модели машинного обучения: $x' = \varepsilon \cdot sign(\nabla_x J(\theta, x, y))$, где ε – минимальный уровень шума, практически незаметный для человеческого глаза, θ – модель нейронной сети, $sign(\nabla_x J(\theta, x, y))$ – знак градиента, ∇_x – градиент, x – исходное изображение, y – корректная метка для x.

2.1.2 Метод Бройдена-Флетчера-Гольдфарба-Шанно с ограниченной памятью (Limited-memory Broyden–Fletcher–Goldfarb–Shanno, L-BFGS).

Алгоритм минимизации функции потерь в направлении целевой метки, который формулируется следующим образом: $\min_{x'} \|x - x'\|_p$ при условии $f(x') \neq y'$, где $\|x - x'\|_p$ – L_p норма состязательных возмущений; y' – состязательная метка цели ($y' \neq y$).

2.1.3 Базовый итеративный метод (Basic Iterative Method, BIM).

Данный метод используется для повышения производительности метода FGSM путем запуска более точного итеративного оптимизатора. BIM выполняется с меньшим размером шага и обрезает обновленный состязательный сэмпл в допустимый диапазон для T итераций; то есть в t-й итерации правило обновления можно задать следующим образом:

$x'_{t+1} = Clip\{x'_t + \alpha \cdot sign[\nabla_x J(\theta, x'_t, y]\}$, где α – величина возмущения в каждой итерации, $\alpha T = \varepsilon$.

2.2. Генерация состязательных примеров для атак на текстовые данные

Пусть дан текст $w = (w_1, ..., w_n)$ состоящий из слов словаря W, а модель МО для данного текста имеет выход $y = f(w) \in Y$. Тогда построение состязательного образца можно рассматривать как решение следующей оптимизационной задачи:

$w'_{i+1} = \arg\min[Loss(w, w'_i) - \|y - f(w'_i)\|_2], i = 1, ..., N$

где функция $Loss(w, w'_i)$ может представлять:

1) долю измененных слов в состязательном примере w'_i по сравнению с исходной записью w:

$$Loss(w, w'_i) = \frac{\sum_{j=1,n} sgn(|w_j - w'_{i_j}|)}{n};$$

2) формальное сходство между состязательным примером w'_i и исходной записью w, которое измеряется расстоянием Левенштейна:

$$Loss(w, w'_i) = lev(w, w'_i) =$$

$$= \begin{cases} |w|, & \text{если } |w'_i| = 0, \\ |w'_i|, & \text{если } |w| = 0, \\ lev(tail(w), tail(w'_i)), & \text{если } w_0 = w'_{i_0}, \\ 1 + \min \begin{cases} lev(tail(w), w'_i) \\ lev(w, tail(w'_i)) \\ lev(tail(w), tail(w'_i)) \end{cases}, & \text{в остальных случаях} \end{cases}$$

где функция $tail(x)$ это строка, состоящая из всех символов, кроме первого.

3) формальное сходство между состязательным примером w'_i и исходной записью w, которое измеряется сходством Жаккарда на уровне символов и слов:

$$Loss(w, w'_i) = J(w, w'_i) = \frac{|w \cap w'_i|}{|w \cup w'_i|}.$$

2.3. Генерация состязательных примеров для атак на табличные данные

2.3.1 Алгоритм атаки на расстоянии (Distance-based attack, Advdcr).

Данный метод состоит в том, чтобы минимизировать расстояние между объектом и синтетической записью с разными выходными метками. Особенность данного подхода состоит в предварительной группировке состязательных образцов в соответствии с квазиидентификаторами и выставлении соответствующего секретного признака как наиболее распространенное значение (моду). Для алгоритма Advdcr правило обновления можно задать следующим образом:

$y' = \arg\max_{|y'_i - y|} \min_{r \in R} \|(x'_i | t) - r\|_2$

где r – вектор возмущений значений признаков.

2.3.2 Алгоритм низкого профиля (Low Profile Algorithm, LowProFool).

Данный метод [17] состоит в том, чтобы минимизировать взвешенную норму вектора возмущений на признаках табличных данных при максимизации доли примеров x ∈ X, с ложными ответами на выходе. Для алгоритма LowProFool правило обновления можно задать следующим образом:

$$x'_{i+1} = Clip\left\{x' + (r'_i + \alpha \cdot \left[-\nabla_r J(x'_i, t) + \lambda \|v \odot r\|_p\right]\right\},$$

$$i = 0, ..., N-1, \; x' = \arg\min_{x'_i} d_v(x_i),$$

где λ – коэффициент компромисса, v – вектор важности признаков, N – максимальное количество итераций, α – коэффициент масштабирования.

3. Генерация и исследование наборов данных массивных MIMO сетей

3.1 Генерация набора данных сценария “O1_drone scenario”

Для генерации наборов данных массивных MIMO сетей на основе точной 3D-трассировки лучей Remcom мы использовали фреймворк DeepMIMO [18]. Рассмотрен сценарий “O1_drone scenario” – сценарий на открытом пространстве, с двумя улицами, зданиями варьируемой высоты и одним перекрестком. На улице зафиксирована одна базовая станция на высоте 6 м и одна летающая реконфигурируемая интеллектуальная поверхность (Reconfigurable Intelligent Surface), находящаяся на высоте 80 м. В качестве массивов пользователей выступают четыре слоя дронов с общим количеством дронов около 270 тыс. на высоте от 40 м. до 42,4 м, расстояние между дронами составляет 81 см. Стандартная рабочая частота эмуляции – 200 ГГц. Каждый пользователь (дрон) состоит из одной антенны. Задано вращение антенн базовой станции по осям x, y и z в виде (30 °, 30 °, 90 °), задано вращение антенны пользователей по осям x, y и z в виде (35 °, 20 °, 60 °). Общая схема расположения пользователей и базовой станции представлена на рисунке 1.

Для сгенерированного набора данных доступны координаты отправителя и получателей, матрица каналов отправителя и получателей, а также различные характеристики путей при распространении сигнала. Нами выделены следующие признаки:

1. Distance – расстояние между базовой станцией и каждым пользователем, в метрах.
2. Pathloss – комбинированные потери на пути канала между отправителем и получателем («затухание» сигнала антенны), в децибелах относительно 1 милливатта.
3. DoA_phi – азимутальный угол прибытия, в градусах.
4. DoA_theta – зенитный угол прибытия, в градусах.
5. DoD_phi – азимутальный угол отправления, в градусах.
6. DoD_theta – зенитный угол отправления, в градусах.
7. Phase – фаза пути распространения сигнала, в градусах.
8. Power – сила сигнала при получении, в мегаватт.
9. Time of arrival – время получения сигнала, в секундах.

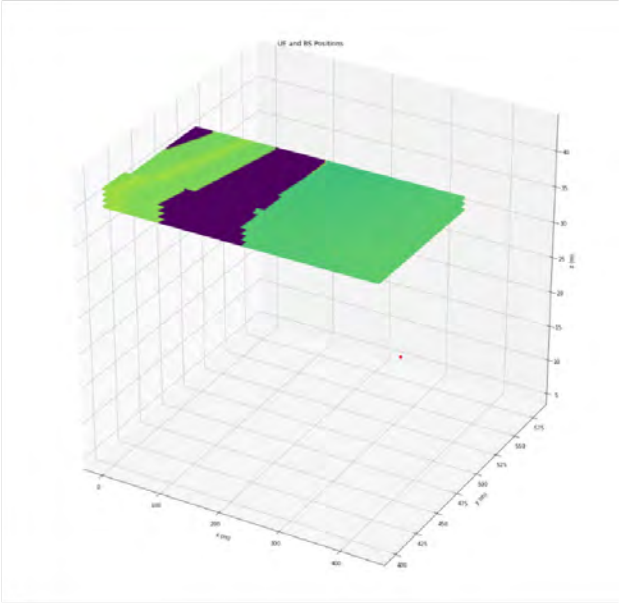

Рис.1. Расположение базовой станции и групп пользователей для сценария “O1_drone scenario”. Базовая станция отмечена красной точкой. Четыре слоя дронов размещены на высоте ~ 41 метра. Цветовая схема соответствует потерям сигнала на пути канала между пользователями и базовой станцией. Зеленый цвет – высокие потери, темно-синий цвет – низкие.

10. Line of Sight (LoS) – статус сигнала, принимаемый одно из трёх значений из {-1, 0, 1}. (LoS = 1): Путь прямой видимости существует. (LoS = 0): существуют только пути вне прямой видимости, при этом путь прямой видимости заблокирован. (LoS = -1): Между передатчиком и приемником нет путей (полная блокировка).

Итоговый набор данных содержит 180 999 записей. Из набора данных удалены записи с нулевыми признаками (для которых LoS = -1), получено следующее распределение по двум классам: 109 017 записей с LoS = 1 и 55 342 записи с LoS = 0. Для такого набора данных возможна реализация бинарной задачи классификации определения нахождения пути между источником и получателем в прямой или непрямой видимости и выполнение состязательной атаки вида $aa\langle tbl, poi\rangle$.

3.2. Исследование сгенерированного набора данных

На рисунке 2 представлены гистограммы распределения нескольких признаков по двум классам метрики LoS. Из рисунка 2(a) мы можем видеть, что комбинированные потери на пути канала увеличиваются

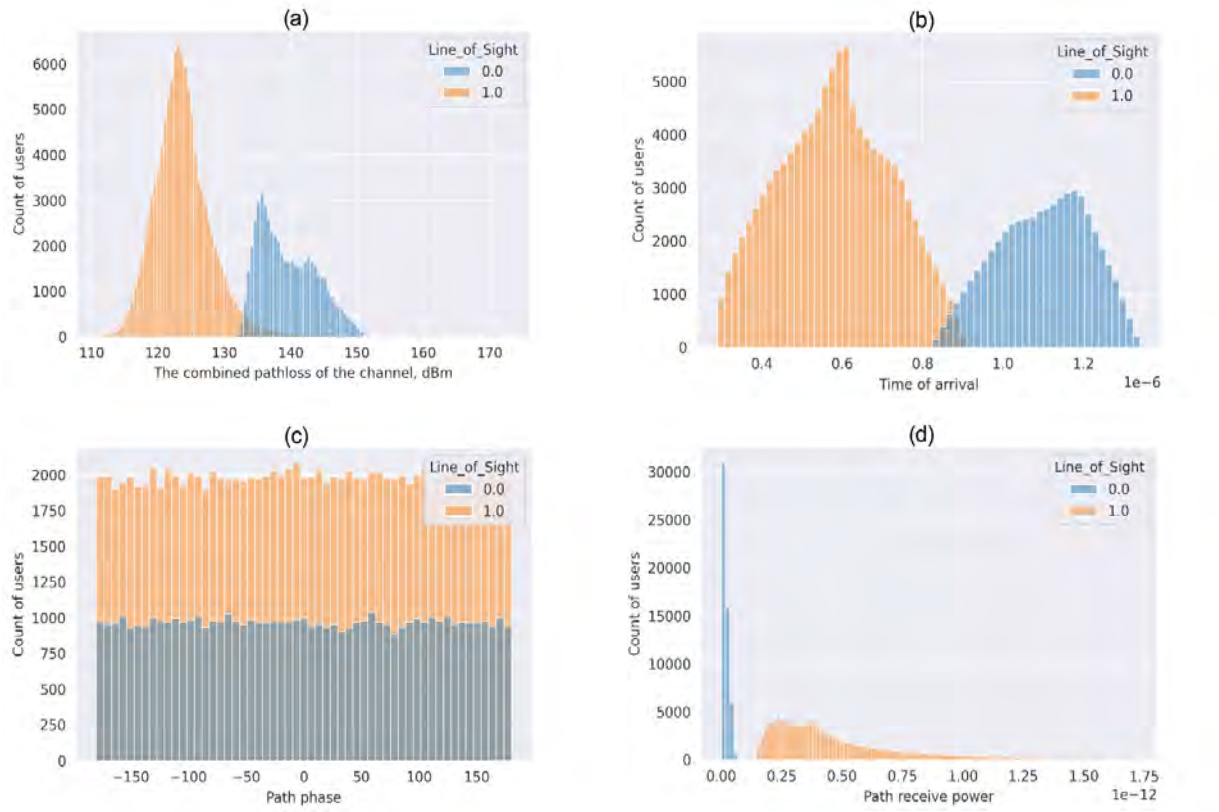

Рис. 2. Гистограммы распределения признаков по двум классам метрики Line of sight

для пользователей-дронов, находящихся вне прямой видимости базовой станции. При этом время прибытия сигнала, исходя из рисунка 2(b) также увеличивается. Из рисунка 2(c) мы можем видеть, что фаза пути распространения сигнала равномерно распределена на интервале от -150° до 150° вне зависимости от состояния статуса сигнала. Из рисунка 2(d) мы можем видеть, что сила сигнала при получении очень слабая для пользователей-дронов, находящихся вне прямой видимости базовой станции, и усиливается для пользователей-дронов, находящихся в прямой видимости.

На рисунках 3(a) и 3(b) представлены графики рассеяния и гистограммы для азимутальных и зенитных углов отправления и прибытия сигнала. Из рисунка 3(a) можно визуально выделить два кластера азимутальных углов отправления и прибытия сигнала, из рисунка 3(b) мы можем наблюдать линейную зависимость зенитных углов отправления и прибытия сигнала. Из рисунка 3(c) комбинированных потерь на пути канала и расстояний от пользователей до базовой станции визуально можно выделить два крупных кластера в соответствии со статусом сигнала.

3.3 Генерация набора данных сценария “O2_dyn_3p5”

Также рассмотрен динамический сценарий “O2_dyn_3p5”, в котором реализовано 1000 записанных сцен с движением автомобильного транспорта по дорогам моделируемого сегмента карты городской инфраструктуры. Общая схема расположения пользователей и базовой станции представлена на рисунке 4. На улице зафиксирована одна базовая станция на высоте 6 м и три группы пользователей с общим количеством 116 303 записей. Стандартная рабочая частота эмуляции – 3.5 ГГц. Каждый пользователь состоит из одной антенны. Временной интервал эмуляции – 100 секунд с шагом изменения состояния сцены в 100 мс.

В этом случае для подготовки набора данных будет зафиксирован произвольный пользователь и одна произвольная базовая станция, при это записи в наборе данных будут соответствовать временным отрезкам той или иной сцены. Будут зафиксированы изменения значений десяти метрик из раздела 3.1, а также добавлен новый признак с временной меткой. Для такого набора данных возможна реализация задачи прогнозирования “затухания” сигнала LoS по косвен-

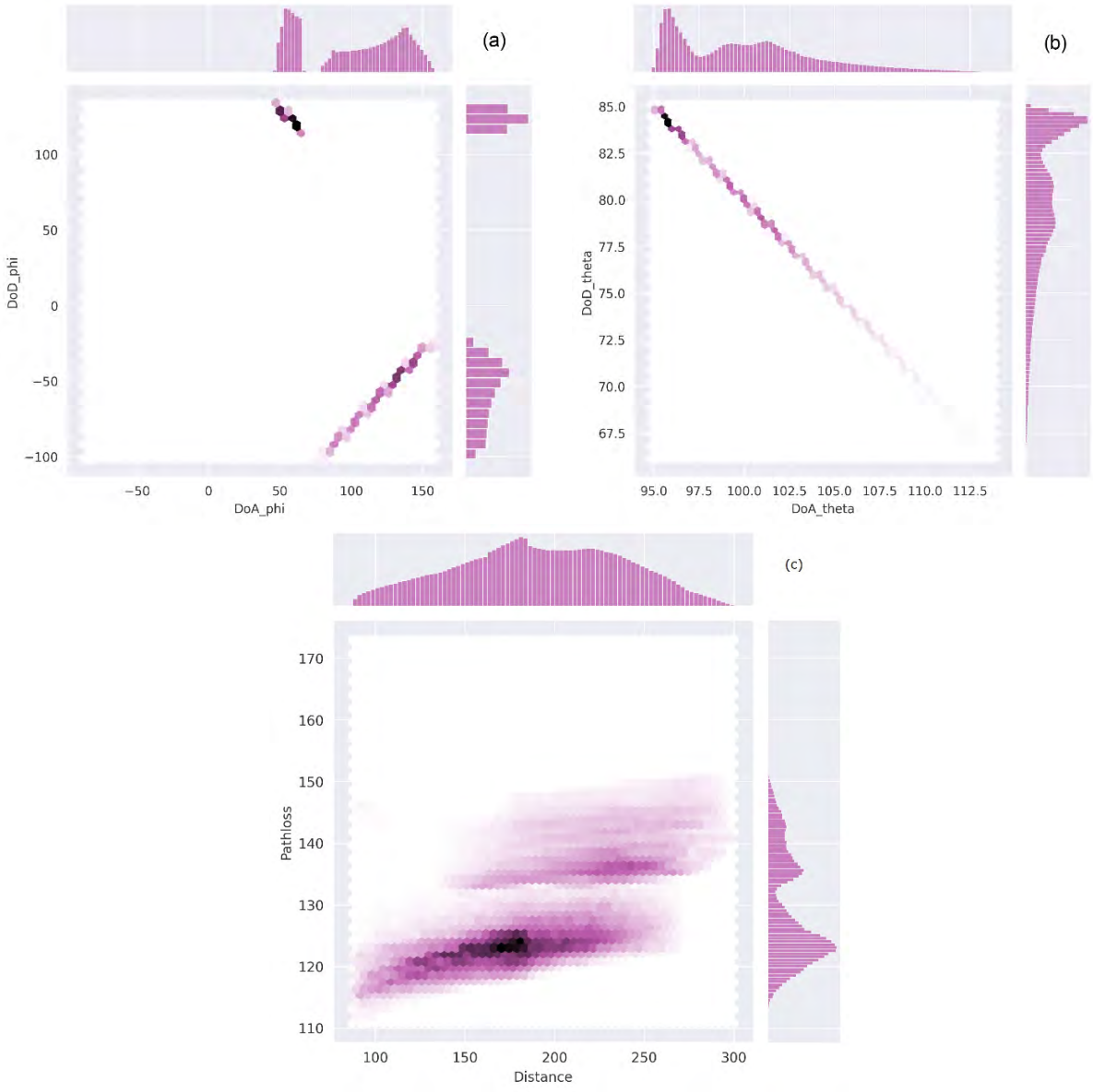

Рис. 3. Графики рассеяния и гистограммы 3(a) и 3(b) для пар признаков DoA_phi-DoD_phi и DoA_theta-DoD_theta, 3(c) для пар признаков Pathloss и Distance

ным признакам и выполнение состязательной атаки вида $aa\langle tbl, eva\rangle$. На рисунке 5 приведен фрагмент графика статуса сигнала относительно потерь сигнала вдоль временной оси моделирования сценария.

4. Обсуждение и выводы

Основные особенности методики построения устойчивой системы от состязательных атак в беспроводных сетях состоят в следующем:

1. Для генерации наборов данных массивных MIMO сетей на основе точной 3D-трассировки лучей можно использовать фреймворк DeepMIMO. Указанное программное решение позволяет наиболее точно и достоверно эмулировать данные MIMO сетей для разработки и оценивания различных приложений машинного обучения в беспроводных сетях.

2. Генерируемые данные представлены в табличном виде, поэтому при построении состязательных примеров для атак можно использовать методы, описанные в разделе 2.3. В частности, предлагается использовать генеративно-состязательные сети для генерации синтетических состязательных примеров на основе подстановки реальных значений выходного признака.

3. Для противодействия состязательным атакам планируется использовать вероятностные методы (метод стохастической защиты, иерархическое случайное переключение), алгоритмы перекрёстной проверки моделей, а также методы состязательного обучения.

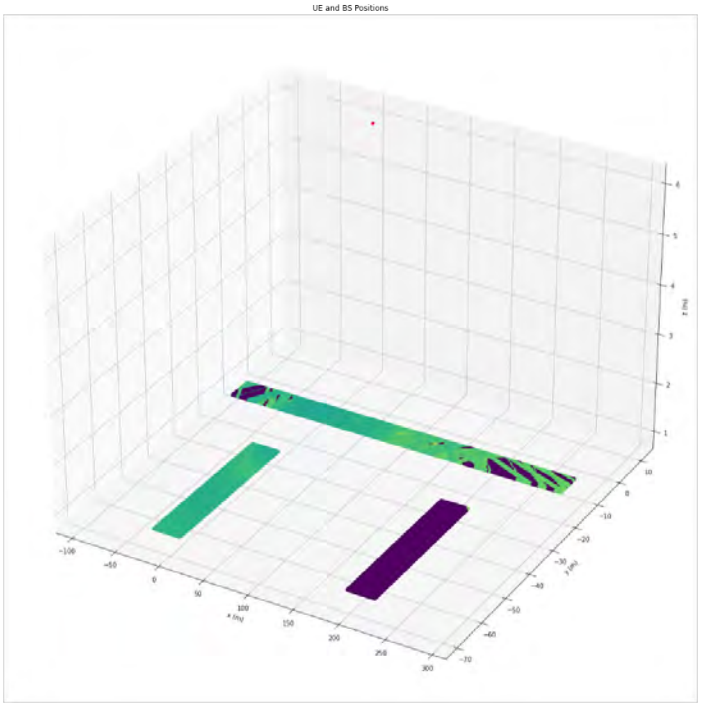

Рис. 4. Расположение базовой станции и групп пользователей для динамического сценария “O2_dyn_3p5”. Базовая станция отмечена красной точкой. Три группы пользователей размещены на высоте 1 метра. Между пользователями и базовой станцией расположена дорога, на которой моделируется движение транспорта. Цветовая схема соответствует потерям сигнала на пути канала между пользователями и базовой станцией. Зеленый цвет – высокие потери, темно-синий цвет – низкие.

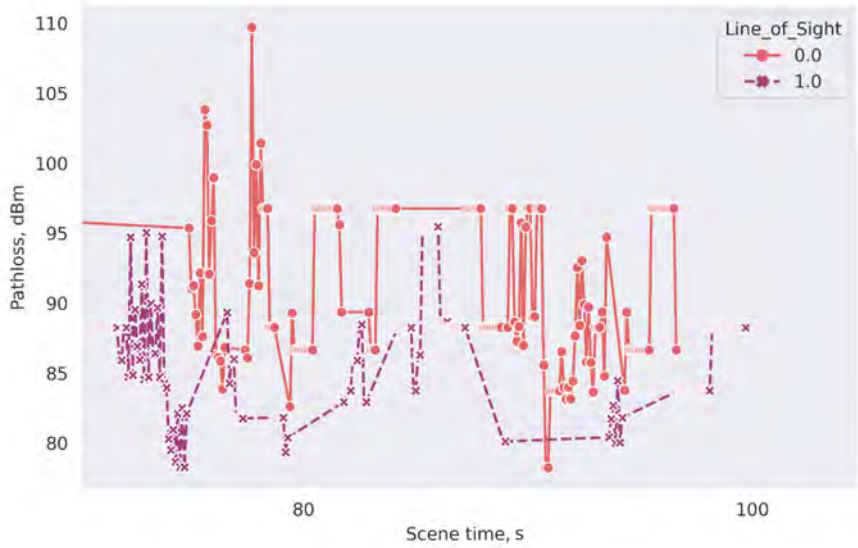

Рис. 5. Фрагмент линейного графика статуса сигнала относительно потерь сигнала во временном интервале от 70 до 100 секунды

Приоритетом будет максимальное восстановление исходных значений метрик оценки качества моделей машинного обучения.

4. Отдельно планируется выполнить построение состязательных примеров для атак на визуальные и текстовые данные при решении различных прикладных задач машинного обучения в области интеллектуальных транспортных систем. Комплексная система защиты будет анализировать табличные, текстовые и визуальные данные на предмет обнаружения состязательных атак, осуществляя меры по противодействию.

В рамках исследования предложена методика аналитической обработки больших массивов данных сервисов и приложений в сетях последнего поколения для построения устойчивых систем защиты на основе состязательного машинного обучения. Проведена генерация и исследовательский анализ наборов данных с помощью эмулятора DeepMIMO. Полученные наборы

данных будут использованы в дальнейших исследованиях для построения состязательных примеров и проведения состязательных атак, а также для отыскания способов выявления отравленных данных в приложениях беспроводных сетей поколения 6G.

Исследование выполнено за счет гранта Российского научного фонда (проект № 22-71-10124).

Литература

1. Bose A. J., Aarabi P. Adversarial attacks on face detectors using neural net based constrained optimization // 2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP). – IEEE, 2018. – P. 1-6. DOI: 10.1109/MMSP.2018.8547128
2. Laidlaw C., Feizi S. Functional adversarial attacks // arXiv preprint arXiv:1906.00001. – 2019. DOI: https://doi.org/10.48550/arXiv.1906.00001
3. Guo C. et al. Simple black-box adversarial attacks // International Conference on Machine Learning. – PMLR, 2019. – P. 2484-2493. DOI: https://doi.org/10.48550/arXiv.1905.07121
4. Wallace E. et al. Universal adversarial triggers for attacking and analyzing NLP // arXiv preprint arXiv:1908.07125. – 2019. DOI: https://doi.org/10.48550/arXiv.1908.07125
5. Morris J. X. et al. Textattack: A framework for adversarial attacks in natural language processing. – 2020. DOI: https://doi.org/10.48550/arXiv.2005.05909
6. Lepekhin M., Sharoff S. Experiments with adversarial attacks on text genres // arXiv preprint arXiv:2107.02246. – 2021. DOI: https://doi.org/10.48550/arXiv.2107.02246
7. Fursov I. et al. Adversarial Attacks on Deep Models for Financial Transaction Records // arXiv preprint arXiv:2106.08361. – 2021. DOI: https://doi.org/10.1145/3447548.3467145
8. Finlayson S. G. et al. Adversarial attacks on medical machine learning // Science. – 2019. – V. 363. – №. 6433. – P. 1287-1289. DOI: 10.1126/science.aaw4399
9. Newaz A. K. M. I. et al. Adversarial attacks to machine learning-based smart healthcare systems // GLOBECOM 2020-2020 IEEE Global Communications Conference. – IEEE, 2020. – P. 1-6. DOI: 10.1109/GLOBECOM42002.2020.9322472
10. Liu Q. et al. Adversarial attack on DL-based massive MIMO CSI feedback // Journal of Communications and Networks. – 2020. – V. 22. – №. 3. – P. 230-235. DOI: 10.1109/JCN.2020.000016
11. Wang X., Zheng Z., Fei Z. ASAP: Adversarial Learning Based Secure Autoprecoder Design for MIMO Wiretap Channels // IEEE Wireless Communications Letters. – 2022. – V. 11. – №. 9. – P. 1915-1919. DOI: 10.1109/LWC.2022.3187089
12. Catak E., Catak F. O., Moldsvor A. Adversarial machine learning security problems for 6G: mmWave beam prediction use-case // 2021 IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom). – IEEE, 2021. – P. 1-6. DOI: 10.1109/BlackSeaCom52164.2021.9527756
13. Kuzlu M. et al. The Adversarial Security Mitigations of mmWave Beamforming Prediction Models using Defensive Distillation and Adversarial Retraining // arXiv preprint arXiv:2202.08185. – 2022. DOI: https://doi.org/10.1007/s10207-022-00644-0
14. Karabulut M. A., Shah A. F. M. S., Ilhan H. A Novel MIMO-OFDM Based MAC Protocol for VANETs // IEEE Transactions on Intelligent Transportation Systems. – 2022. DOI: 10.1109/TITS.2022.3180697
15. Guo H. et al. Vehicular intelligence in 6G: Networking, communications, and computing // Vehicular Communications. – 2022. – V. 33. – P. 1-19. DOI: https://doi.org/10.1016/j.vehcom.2021.100399
16. Cheng X., Huang Z., Chen S. Vehicular communication channel measurement, modelling, and application for beyond 5G and 6G // IET Communications. – 2020. – V. 14. – №. 19. – P. 3303-3311. DOI: https://doi.org/10.1049/iet-com.2020.0531
17. Ballet V. et al. Imperceptible adversarial attacks on tabular data // arXiv preprint arXiv:1911.03274. – 2019. DOI: https://doi.org/10.48550/arXiv.1911.03274
18. Alkhateeb A. DeepMIMO: A generic deep learning dataset for millimeter wave and massive MIMO applications // arXiv preprint arXiv:1902.06435. – 2019. DOI: https://doi.org/10.48550/arXiv.1902.06435

THE TECHNIQUE OF BUILDING A SUSTAINABLE PROTECTION SYSTEM BASED ON ADVERSARIAL MACHINE LEARNING IN 6G WIRELESS NETWORKS

Legashev L.V.[4], Grishina L.S.[5]

4 Leonid V. Legashev, Ph.D., Leading Researcher at the Laboratory of Digital Solutions and Big Data Analytics, Orenburg State University, Orenburg, Russia. Email: silentgir@gmail.com. CODE: 0000-0001-6351-404X

5 Lyubov S. Grishina, Junior Researcher, Laboratory of Digital Solutions and Big Data Analytics Orenburg State University, Orenburg, Russia. E-mail: grishina_ls@inbox.ru ORCID: 0000-0003-2752-7198

Abstract

The purpose of research *is to develop the technique of analytical processing of big data of services and applications in the new generation communication networks to detect cybersecurity incidents and build sustainable protection systems based on adversarial machine learning.*

The methods of research: *Analysis of modern methods of machine learning and neural network technologies, synthesis and formalization of algorithms for adversarial attacks on machine learning models.*

Scientific novelty: *a technique for analytical processing of emulated data of services and applications for detecting cybersecurity incidents is presented, which provides a groundwork in the field of research into the security issues of complex intelligent services and applications in the infrastructure of wireless networks of the next generation.*

The result of research: *The article proposes a technique of building a sustainable protection system against adversarial attacks in wireless ad hoc networks of the next generation. The main types of adversarial attacks, including poisoning attacks and evasion attacks, are formalized, and methods for generating adversarial examples on tabular, textual, and visual data are described. Several scenarios were generated and exploratory analysis of datasets was carried out using the DeepMIMO emulator. Potential application problems of binary classification and prediction of signal attenuation between a user and a base station for adversarial attacks are presented. The algorithmization of the processes of building and training a sustainable protection system against adversarial attacks in wireless networks of the next generation is presented on the example of emulated data.*

Keywords: *adversarial attacks, wireless ad hoc networks, machine learning, MIMO.*

References

1. Bose A. J., Aarabi P. Adversarial attacks on face detectors using neural net based constrained optimization // 2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP). – IEEE, 2018. – P. 1-6. DOI: 10.1109/MMSP.2018.8547128
2. Laidlaw C., Feizi S. Functional adversarial attacks // arXiv preprint arXiv:1906.00001. – 2019. DOI: https://doi.org/10.48550/arXiv.1906.00001
3. Guo C. et al. Simple black-box adversarial attacks // International Conference on Machine Learning. – PMLR, 2019. – P. 2484-2493. DOI: https://doi.org/10.48550/arXiv.1905.07121
4. Wallace E. et al. Universal adversarial triggers for attacking and analyzing NLP // arXiv preprint arXiv:1908.07125. – 2019. DOI: https://doi.org/10.48550/arXiv.1908.07125
5. Morris J. X. et al. Textattack: A framework for adversarial attacks in natural language processing. – 2020. DOI: https://doi.org/10.48550/arXiv.2005.05909
6. Lepekhin M., Sharoff S. Experiments with adversarial attacks on text genres // arXiv preprint arXiv:2107.02246. – 2021. DOI: https://doi.org/10.48550/arXiv.2107.02246
7. Fursov I. et al. Adversarial Attacks on Deep Models for Financial Transaction Records // arXiv preprint arXiv:2106.08361. – 2021. DOI: https://doi.org/10.1145/3447548.3467145
8. Finlayson S. G. et al. Adversarial attacks on medical machine learning // Science. – 2019. – V. 363. – №. 6433. – P. 1287-1289. DOI: 10.1126/science.aaw4399
9. Newaz A. K. M. I. et al. Adversarial attacks to machine learning-based smart healthcare systems // GLOBECOM 2020-2020 IEEE Global Communications Conference. – IEEE, 2020. – P. 1-6. DOI: 10.1109/GLOBECOM42002.2020.9322472
10. Liu Q. et al. Adversarial attack on DL-based massive MIMO CSI feedback // Journal of Communications and Networks. – 2020. – V. 22. – №. 3. – P. 230-235. DOI: 10.1109/JCN.2020.000016
11. Wang X., Zheng Z., Fei Z. ASAP: Adversarial Learning Based Secure Autoprecoder Design for MIMO Wiretap Channels // IEEE Wireless Communications Letters. – 2022. – V. 11. – №. 9. – P. 1915-1919. DOI: 10.1109/LWC.2022.3187089
12. Catak E., Catak F. O., Moldsvor A. Adversarial machine learning security problems for 6G: mmWave beam prediction use-case // 2021 IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom). – IEEE, 2021. – P. 1-6. DOI: 10.1109/BlackSeaCom52164.2021.9527756
13. Kuzlu M. et al. The Adversarial Security Mitigations of mmWave Beamforming Prediction Models using Defensive Distillation and Adversarial Retraining // arXiv preprint arXiv:2202.08185. – 2022. DOI: https://doi.org/10.1007/s10207-022-00644-0
14. Karabulut M. A., Shah A. F. M. S., Ilhan H. A Novel MIMO-OFDM Based MAC Protocol for VANETs // IEEE Transactions on Intelligent Transportation Systems. – 2022. DOI: 10.1109/TITS.2022.3180697
15. 15. Guo H. et al. Vehicular intelligence in 6G: Networking, communications, and computing // Vehicular Communications. – 2022. – V. 33. – P. 1-19. DOI: https://doi.org/10.1016/j.vehcom.2021.100399
16. 16. Cheng X., Huang Z., Chen S. Vehicular communication channel measurement, modelling, and application for beyond 5G and 6G // IET Communications. – 2020. – V. 14. – №. 19. – P. 3303-3311. DOI: https://doi.org/10.1049/iet-com.2020.0531
17. 17. Ballet V. et al. Imperceptible adversarial attacks on tabular data // arXiv preprint arXiv:1911.03274. – 2019. DOI: https://doi.org/10.48550/arXiv.1911.03274
18. 18. Alkhateeb A. DeepMIMO: A generic deep learning dataset for millimeter wave and massive MIMO applications // arXiv preprint arXiv:1902.06435. – 2019. DOI: https://doi.org/10.48550/arXiv.1902.06435