УДК 004.852 Вестник УрФО № 2(60) / 2026, с. 76–87

Частикова В.А., Марченко В.А.

DOI: 10.14529/secur260208

СОСТЯЗАТЕЛЬНАЯ РОБАСТНОСТЬ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ: СИСТЕМАТИЧЕСКИЙ ОБЗОР УНИВЕРСАЛЬНЫХ АТАК, МЕТОДОВ ЗАЩИТЫ И ТРЕБОВАНИЙ К ИХ АУДИТУ

В работе обобщены результаты исследований состязательной (adversarial) робастности моделей глубокого обучения. Обзор подготовлен по протоколу PRISMA 2020: из 540 первоначально найденных публикаций формальным критериям включения удовлетворили 32 источника. Основное внимание уделено универсальным состязательным возмущениям (UAP) и способам защиты от них. Авторы прослеживают происхождение явления, приводят таксономию атак и разбирают основные алгоритмы построения UAP, после чего сопоставляют между собой современные защиты – состязательное обучение, TRADES, случайное сглаживание – и эталонные бенчмарки AutoAttack, RobustBench и ImageNet-C. Отдельный раздел соотносит международные результаты с действующей в России нормативной базой: приказами ФСТЭК России № 17, № 21 и № 239, ФЗ-187, ГОСТ Р ИСО/МЭК 27001-2021 и ГОСТ Р 59276-2020. Как показывает анализ, универсальность возмущений вместе с их переносимостью между архитектурами делает атаки практически значимыми, тогда как формальные гарантии среди известных защит дают только методы сертифицированной робастности. В заключение выделены вопросы, которые в национальной методической базе требуют первоочередной проработки: физически реализуемые UAP, защита крупных мультимодальных моделей и стандартизация процедур аудита значимых объектов критической информационной инфраструктуры по устойчивости к adversarial-воздействиям.

Ключевые слова: *глубокое обучение, состязательные атаки, универсальные состязательные возмущения, состязательное обучение, случайное сглаживание, переносимость атак, AutoAttack, RobustBench, информационная безопасность, ФСТЭК России.*

ADVERSARIAL ROBUSTNESS OF MACHINE LEARNING MODELS: A SYSTEMATIC REVIEW OF UNIVERSAL ATTACKS, DEFENSE METHODS, AND AUDIT REQUIREMENTS

The paper presents a systematic review of adversarial robustness of deep learning models, conducted per the PRISMA 2020 protocol. Of 540 initially identified publications, 32 sources passed the formal inclusion and exclusion criteria. The focus is on Universal Adversarial Perturbations (UAP) and defense approaches. The origins and taxonomy of attacks are systematized; key UAP generation algorithms are reviewed; modern defense methods (adversarial training, TRADES, randomized smoothing) and benchmarks (AutoAttack, RobustBench, ImageNet-C) are analyzed. A separate section maps international results onto the Russian regulatory framework – FSTEC orders No. 17, No. 21, No. 239, Federal Law 187-FZ, GOST R ISO/IEC 27001-2021 and GOST R 59276-2020. Universality and cross-architecture transferability make these attacks practically significant; among existing defenses, only certified methods provide formal guarantees. Priority directions for the national methodological framework are formulated: physically realizable UAPs, defense of large multimodal models, and standardization of adversarial-robustness audit procedures for significant objects of critical information infrastructure.

Keywords: *deep learning, adversarial attacks, universal adversarial perturbations, adversarial training, randomized smoothing, transferability, AutoAttack, RobustBench, information security, FSTEC of Russia.*

Введение

Сегодня глубокое обучение применяется в компьютерном зрении, обработке естественного языка и системах поддержки принятия решений, в том числе в таких ответственных областях, как медицинская диагностика, автономный транспорт, биометрия и средства информационной безопасности [1]. За последнее десятилетие, однако, у нейросетевых моделей обнаружился новый класс уязвимостей – состязательные примеры (adversarial examples). Это специально подобранные входные данные, которые отличаются от исходных небольшими, незаметными для человека возмущениями, но при этом способны полностью изменить решение классификатора [2, 3]. Число работ по этой теме идет на сотни, поэтому обобщение накопленных результатов само по себе превратилось в отдельную научную задачу [4, 5].

Среди таких атак выделяются универсальные состязательные возмущения (UAP) – единый шумовой паттерн, не привязанный к конкретному входу: будучи наложенным на произвольное изображение из распределения данных, он с высокой вероятностью вызывает у модели ошибку [6]. От классических состязательных атак UAP отличаются масштабируемостью. Возмущение достаточно вычислить один раз, после чего его можно применять многократно и даже физически – например, напечатав на наклейке или билборде либо нанеся на одежду [7]. Из-за этого UAP представляют не только академический интерес, но и реальную угрозу для биометрического контроля доступа, систем видеоанали-

тики, автономного транспорта и медицинской диагностики [8].

Отечественная нормативная база учитывает эту угрозу лишь частично. Приказы ФСТЭК России № 17, № 21 и № 239 [23, 24, 25], Федеральный закон № 187-ФЗ [26], ГОСТ Р ИСО/МЭК 27001-2021 [27] и ГОСТ Р 59276-2020 [28] задают общие требования к доверию и защите, однако в них нет ни прикладных методик оценки устойчивости моделей машинного обучения к UAP, ни пороговых значений метрик, ни обязательного перечня атакующих алгоритмов для аттестационного тестирования. Тем самым между международной исследовательской практикой и российской системой сертификации значимых объектов критической информационной инфраструктуры (КИИ) образуется методологический разрыв.

Цель настоящего обзора – привести в систему ключевые публикации 2013–2025 гг. по двум связанным направлениям: с одной стороны, природа и алгоритмы генерации универсальных состязательных возмущений, с другой – подходы к защите моделей, как эмпирические, так и сертифицированные. Попутно мы соотносим эти результаты с требованиями ФСТЭК России и национальных стандартов и указываем направления, которые в нормативно-методической базе следует проработать в первую очередь.

1. Методология обзора

Работа выполнена по рекомендациям PRISMA 2020 [29]. Литературу мы искали с января по апрель 2026 г. в базах Google Scholar, IEEE Xplore, ACM Digital Library, OpenReview, arXiv и Scopus. Запросы строились по комбинациям ключевых терминов: «adversarial examples», «universal adversarial perturbation», «adversarial robustness», «certified robustness», «adversarial training», «adversarial attack defense». Поиск охватывал период 2013–2025 гг.; нижняя граница задана работой Szegedy и др. [2]. В качестве метаисточников использованы RobustBench [13] и обзоры [4, 5]. Этапы фильтрации сведены в табл. 1.

Таблица 1

Этапы отбора публикаций (PRISMA 2020 flow)

Этап отбора	Количество записей
Identification (всего найдено)	540
Удалено дубликатов	89
Screening по заголовку и аннотации	451
Исключено по тематике / нерелевантности	241
Eligibility (оценка полного текста)	210
Исключено по критериям невключения	178
Included (вошло в обзор)	32

Критерии включения. К рассмотрению допускались работы: опубликованные в реферируемом издании или на конференциях ранга A* (NeurIPS, ICML, ICLR, CVPR, ICCV, ECCV, IEEE S&P, USENIX Security, ACM CCS, NDSS); содержащие экспериментальную либо теоретическую новизну в области состязательной робастности; имеющие открытую реализацию или включенные в RobustBench; индексируемые с DOI или arXiv-идентификатором.

Критерии исключения. Из дальнейшей обработки исключались:

1) препринты без последующей рецензируемой публикации со сроком более 24 месяцев;

2) работы по adversarial-атакам в неинформационно-визуальных доменах вне общеметодологических вопросов;

3) работы без воспроизводимых результатов (нет кода, гиперпараметров или данных);

4) дубликаты публикаций (расширенные журнальные версии учитывались как один источник);

5) тезисы и материалы воркшопов без полного рецензирования;

6) источники старше 2013 г., кроме классических работ, заложивших методологический базис;

7) работы, не прошедшие минимальную оценку качества по адаптированной шкале.

Качество каждой включенной работы оценивалось по четырем критериям: воспроизводимость (открытый код и веса), статистическая обоснованность (доверительные ин-

тервалы или тесты значимости), независимость данных (использование стандартных бенчмарков) и – при заявлении новых защитных механизмов – проверка под адаптивной атакой [14]. На публикации 2021–2025 гг. пришлось 34 % включенных источников.

2. Результаты

2.1. Истоки проблемы и таксономия атак

Впервые состязательные примеры описали Szegedy и др. в 2013 г., изучая «странные» свойства глубоких нейронных сетей [2]. Они обнаружили, что почти для каждого правильно классифицируемого изображения можно построить визуально неотличимый от него вариант, который сеть отнесет к заранее выбранному ошибочному классу. Двумя годами позже Goodfellow и др. объяснили этот эффект линейностью сетей в пространствах высокой размерности и предложили быстрый способ генерации таких примеров – Fast Gradient Sign Method (FGSM) [3].

Атаки принято классифицировать сразу по нескольким независимым осям: по уровню доступа к модели (white-box, black-box, серобоксный), по цели (нецелевые и целевые), по этапу жизненного цикла (отравление обучающих данных либо атаки на этапе вывода, evasion), по используемой метрике возмущения (L_∞, L_2, L_0), а также по специфичности – индивидуальные или универсальные [4, 5, 9]. В систематических обзорах такое деление применяется как стандартное [4].

Наряду с эмпирическими методами велась и теоретическая оценка робастности. Carlini и Wagner установили, что многие защиты 2016–2017 гг. держались на «маскировании градиента» и легко пробивались сильной итеративной атакой [7]. Athalye и др. распространили этот вывод шире: из девяти защит, представленных на ICLR 2018, семь удалось обойти, скомпенсировав обнаруженную маскировку [14]. Так сформировался методологический стандарт: каждую новую защиту следует проверять адаптивной атакой, построенной именно против нее.

2.2. Универсальные состязательные возмущения

Понятие универсального состязательного возмущения предложили Moosavi-Dezfooli и др. [6]. Для произвольной обученной модели f и распределения данных μ они построили единый вектор v с малой L_∞-нормой, при наложении которого более 80 % изображений из μ классифицируются неверно. Показательно, что возмущение, рассчитанное на VGG-19, успешно атакует и GoogLeNet, и ResNet-152: это говорит об общих геометрических уязвимостях, присущих обученным сетям [6, 9].

Геометрически существование UAP объясняют так: в высокоразмерном пространстве признаков границы принятия решения проходят систематически близко к естественному распределению данных – вдоль некоторого «направления наименьшей устойчивости» [6]. Именно эта общая для разных входов уязвимость и позволяет одному вектору обманывать модель на широком классе изображений. Переносимость UAP, в свою очередь, зависит от архитектурного сходства моделей: внутри семейства сверточных сетей возмущения переносятся хорошо, тогда как между CNN и трансформерами перенос заметно слабее [9, 30].

На практике опасность UAP усиливается тем, что их можно реализовать физически. Eykholt и др. с помощью наклеек успешно атаковали классификаторы дорожных знаков: в полевых испытаниях знак «STOP» принимался за «Speed Limit 45» в 84 % случаев – при разных ракурсах съемки и освещении [10]. Для систем автономного транспорта и видеоаналитики на значимых объектах КИИ – это прямая угроза [25].

2.3. Эволюция алгоритмов генерации UAP

В исходном алгоритме Moosavi-Dezfooli универсальное возмущение набирается итеративно – как накопление минимальных индивидуальных возмущений (DeepFool) с проекцией в шар L_∞-нормы радиуса ε [6]. Алгоритм несложен, однако требует обучающих изображений и сходится лишь за сотни итераций. Дальнейшие работы развивались в двух направлениях – ускорение сходимости и снижение требований к данным.

Первое направление – генеративные подходы. В методе NAG Mopuri и др. обучают условную генеративную сеть, которая сама порождает UAP, оптимизируя функцию обмана модели-жертвы [11]. По сравнению с итеративным методом Moosavi-Dezfooli при той же норме ε это дает более разнообразные возмущения и до 5 % более высокий fooling rate. Та же идея заложена в GAP, где к состязательной постановке задачи добавлены возможности GAN [16].

Второе направление – data-free-атаки, которым доступ к обучающему набору вообще не нужен. Вместо обучающих данных они

опираются на априорные сведения о статистике активаций внутренних слоев сети [11]. На практике это особенно опасно: злоумышленнику не требуется представительная выборка из домена жертвы, а fooling rate при ε = 10/255 удерживается на уровне 60–70 % – почти как у методов, использующих данные.

2.4. Подходы к защите

Защиты обычно делят на три группы: состязательное обучение (AT), то есть обучение на смеси чистых и состязательных примеров; методы предобработки и сглаживания входа; наконец, сертифицированные методы. Наиболее сильной эмпирической защитой остается состязательное обучение в формулировке Madry и др. [12]: модель решает min-max задачу, минимизируя потери по параметрам и одновременно максимизируя их по PGD-возмущениям в шаре радиуса ε. Защиты сопоставлены в табл. 2.

Таблица 2

Сравнение основных подходов к защите от состязательных атак

Метод	Тип гарантий	Снижение FR (UAP)	Потеря Top-1, п.п.	Стоимость обучения/инференса	Источник
AT (Madry, PGD-AT)	эмпирические	−45..−55 п.п.	10–14	обучение в 5–10× медленнее	[12]
TRADES	эмпирические	−40..−50 п.п.	8–11	обучение в 5–8× медленнее	[18]
Free AT (Shafahi)	эмпирические	−35..−45 п.п.	12–15	обучение в ~1.5× медленнее	[19]
Fast AT (Wong)	эмпирические	−30..−40 п.п.	10–13	обучение в ~2× медленнее	[17]
Gowal AT (доп. данные)	эмпирические	−50..−58 п.п.	5–8	обучение в ~6× медленнее	[31]
Distillation (Papernot)	эмпирические, обходимо	малое	≤2	обучение в 2× медленнее	[15]
Randomized Smoothing	сертифицированные	−50..−55 п.п.	8–12	инференс в 50–100× медленнее	[20]

TRADES [18] выносит баланс между чистой и робастной точностью в явный гиперпараметр β и на ImageNet и CIFAR-10 устойчиво обходит классическое AT по метрике AutoAttack. Free AT [19] и Fast AT [17] решают другую задачу – удешевляют обучение ценой небольшого снижения робастности. Наконец, Gowal и др. [31] и Rebuffi и др. [32] показали, насколько сильно AT выигрывает от дополнительных или синтезированных данных: на CIFAR-10 при ε = 8/255 робастная точность поднимается с 53 % у базовой модели Madry до 66 %.

Сертифицированные методы устроены принципиально иначе. Случайное сглаживание Cohen, Rosenfeld и Kolter [20] дает формальную гарантию: для входа x при заданном σ вычисляется радиус r, внутри которого ответ классификатора заведомо не меняется. Гарантия носит вероятностный характер, но при достаточном числе шумовых проходов превращается в практически применимую сертификацию. Цена этого – рост времени инференса в 50–100 раз, из-за чего метод плохо подходит для систем реального времени.

Отдельного внимания заслуживает проблема мнимой защиты. Дистилляцию, предложенную Papernot и др. [15], вскоре обошли Carlini и Wagner своей C&W-атакой [7]. Точно так же простые преобразования входа – квантование, фильтрация, JPEG-сжатие – пробиваются адаптивными атаками [14]. Отсюда и общее методологическое требование: защиту следует проверять не только под FGSM или PGD, но и под адаптивным противником.

2.5. Бенчмарки и стандартизированные оценки

Невоспроизводимость результатов и их сильная зависимость от выбранного атакующего алгоритма долго оставались системной проблемой области. Чтобы снять эту зависимость, Croce и Hein предложили AutoAttack [21] – ансамбль из четырех атак (APGD-CE, APGD-DLR, FAB и Square) без свободных гиперпараметров; сегодня он де-факто служит эта-

лоном при оценке робастности. На AutoAttack опирается и лидерборд RobustBench [13], где собрано свыше 100 моделей с проверенными значениями стандартной и робастной точности.

Устойчивость к естественным искажениям измеряют отдельным бенчмарком – ImageNet-C / CIFAR-10-C [22], в котором предусмотрено 19 типов искажений на пяти уровнях интенсивности. Состязательную робастность он напрямую не охватывает, зато позволяет проверить, насколько связаны обычная и адверсарная устойчивость. Связь оказалась слабой, что лишний раз подтверждает: защита от UAP – самостоятельная задача. По данным Bai и др. [30] и Sehwag и др. [33], при одинаковых условиях обучения трансформеры (ViT, Swin) робастнее сверточных сетей, однако под адаптивными атаками это преимущество в значительной степени нивелируется.

3. Обсуждение

3.1. Сводный анализ утверждений

Из всей рассмотренной литературы можно вычленить несколько устойчивых утверждений, за которыми стоит разная по силе доказательная база. В табл. 3 они сведены в формате «утверждение – оценка силы доказательств – источники».

Табл. 3 показывает, что к 2025 году по ряду фундаментальных вопросов в сообществе сложился консенсус: существование UAP, их переносимость внутри семейства CNN, превосходство AT над тривиальными защитами, польза дополнительных данных. Доказательства средней силы остаются там, где сравнивают трансформеры и CNN, оценивают защиты для новых архитектур (Swin, ConvNeXt) или обсуждают применимость сертифицированных методов в продакшене. Заметный методологический сдвиг произошел в 2018–2021 гг.: после работы Athalye и др. [14] и появления AutoAttack [21] оценка под адаптивным противником и публи-

Таблица 3

Ключевые утверждения о состязательной робастности и сила доказательной базы

Утверждение	Сила	Обоснование	Источники
Глубокие сети уязвимы к состязательным примерам	Strong (10/10)	Воспроизведено сотнями работ, теоретическое объяснение через локальную линейность	[2, 3]
Существуют UAP с FR ≥ 80 % при ε = 10/255	Strong (9/10)	Подтверждено для 6+ архитектур на ImageNet	[6, 9, 11]
UAP переносимы между CNN-архитектурами	Strong (8/10)	Подтверждено для VGG, ResNet, DenseNet, GoogLeNet	[6, 9]
UAP слабо переносятся между CNN и трансформерами	Moderate (7/10)	Несколько независимых наблюдений на ViT и Swin	[9, 30]
Состязательное обучение Madry — сильнейший эмпирический базис	Strong (9/10)	Лидер RobustBench 2018–2021, независимо воспроизведено	[12, 13, 21]
Доп. / синтетические данные существенно повышают робастность AT	Strong (8/10)	CIFAR-10: рост робастной точности с 53 % до 66 %	[31, 32]
TRADES обеспечивает лучший компромисс точности/робастности	Moderate (7/10)	Подтверждено на CIFAR-10/100, частично на ImageNet	[18]
Случайное сглаживание — единственный масштабируемый сертифицированный метод	Strong (8/10)	Вероятностные гарантии, доказанные теоремами	[20]
Простые защиты (дистилляция, фильтрация) обходятся адаптивной атакой	Strong (9/10)	Систематическая проверка 9 защит ICLR 2018	[7, 14]
Физически реализуемые атаки эффективны в реальных условиях	Moderate (7/10)	Несколько демонстраций на знаках/одежде	[10]
Состязательная робастность требует увеличенного объема данных	Moderate (6/10)	Теоретический анализ для линейных моделей	[14]

кация моделей в RobustBench стали нормой. При этом с 2020 г. лидеры лидерборда прибавляют на CIFAR-10 лишь доли процента робастной точности в год, и каждый такой шаг обходится во все большие вычисления [13, 31].

3.2. Применимость в контексте российской нормативной базы

В России защита значимых информационных систем регулируется актами ФСТЭК и национальными стандартами. Приказ № 17 [23] относится к государственным ИС, приказ № 21 [24] – к персональным данным, а приказ № 239 [25] во исполнение ФЗ-187 [26] – к значимым объектам КИИ. Общие требования к системе менеджмента ИБ задает ГОСТ Р ИСО/МЭК 27001-2021 [27], тогда как доверие к системам ИИ прямо регулирует единственный национальный стандарт – ГОСТ Р 59276-2020 [28]. На международном уровне эти вопросы затрагивают ИСО/МЭК 23894:2023 [34] и NIST AI RMF 1.0 [35].

Сопоставление этих документов с проблематикой состязательной робастности приведено в табл. 4.

Таблица 4

Покрытие угрозы состязательных атак нормативными документами

Документ	Объект регулирования	Покрытие adversarial-угроз
ГОСТ Р ИСО/МЭК 27001-2021 [27]	СМИБ организации	Отсутствует прямое регулирование
ГОСТ Р 59276-2020 [28]	Доверие к системам ИИ	Упомянуто, без методик оценки
Приказ ФСТЭК № 17 [23]	Государственные ИС	Косвенно через тестирование защищенности
Приказ ФСТЭК № 21 [24]	ИС персональных данных	Косвенно через анализ уязвимостей
Приказ ФСТЭК № 239 [25]	Значимые объекты КИИ	Косвенно через периодический аудит
ФЗ-187 [26]	Безопасность КИИ	Общие требования, без специфики ИИ
БДУ ФСТЭК [36]	Банк данных угроз	Категория угроз для ИИ формируется
ИСО/МЭК 23894:2023 [34]	Управление рисками ИИ	Общие положения по adversarial-рискам
NIST AI RMF 1.0 [35]	Управление рисками ИИ	Adversarial-устойчивость как часть Trust

Из табл. 4 видно главное: ни один из действующих российских документов не содержит формализованной методики оценки устойчивости моделей машинного обучения к UAP. Нигде не закреплены ни пороговые значения fooling rate, ни обязательный перечень атакующих алгоритмов для аттестации, ни требование публиковать воспроизводимые результаты аудита. В БДУ ФСТЭК [36] категория угроз для систем ИИ пока только формируется, и специфика adversarial-атак – UAP, отравление данных, инверсия модели – представлена лишь в общих чертах. О том же пробеле пишут Намиот и Ильюшин [37].

На практике это оборачивается разрывом между международными исследовательскими стандартами – AutoAttack, RobustBench, PRISMA – и отечественной системой аттестации. Разработчикам защищенных систем компьютерного зрения (биометрия, видеоаналитика на объектах КИИ, медицинская диагностика) приходится либо самостоятельно переносить к себе зарубежные методики, либо формально выполнять общие требования, не проверяя систему на специфические угрозы. Чтобы устранить этот разрыв, нужен отдельный методический документ ФСТЭК, ориентированный именно на adversarial-угрозы.

3.3. Матрица исследовательских пробелов

Табл. 5 сводит открытые вопросы и типы защитных методов в единую матрицу пробе-

Таблица 5

Матрица исследовательских пробелов

Открытый вопрос / Метод	AT/PGD	TRADES	Random. Smoothing	Detection
Робастность для трансформеров (ViT, Swin)	3	1	GAP	GAP
Защита от физически реализуемых UAP	2	GAP	GAP	GAP
Применимость к мультимодальным моделям (CLIP)	GAP	GAP	GAP	GAP
Стандартизация в нормативных документах ФСТЭК	GAP	GAP	GAP	GAP
Гармонизация с ГОСТ Р 59276-2020	GAP	GAP	GAP	GAP
Сертифицированный аудит больших LLM	GAP	GAP	1	GAP

лов: пометка GAP означает, что систематических исследований по соответствующей ячейке нет, а число показывает, сколько публикаций по ней удалось найти.

Заключение

К настоящему времени состязательная робастность оформилась в самостоятельную область со своей таксономией атак, набором эталонных алгоритмов и воспроизводимыми бенчмарками. Центральное место в ней занимают универсальные состязательные возмущения: сочетая свойства классических атак с масштабируемостью и переносимостью между моделями, они становятся одной из наиболее реалистичных угроз для развернутых систем компьютерного зрения, в том числе на значимых объектах КИИ.

Защиты выстраиваются в довольно ясную иерархию. На одном полюсе – эмпирические методы во главе с состязательным обучением Madry и его ускоренными вариантами (Free AT, Fast AT, Gowal AT) и компромиссным TRADES; на другом – сертифицированные методы, прежде всего случайное сглаживание. Ни один из классов не свободен от ограничений: эмпирические защиты не гарантируют устойчивости и со временем обходятся адаптивной атакой, а сертифицированные дороги в инференсе и плохо масштабируются на сложные задачи.

Сопоставление международной практики с российской нормативной базой обнажает методологический пробел: ни ГОСТ Р ИСО/МЭК 27001-2021, ни ГОСТ Р 59276-2020, ни приказы ФСТЭК № 17, № 21 и № 239 не предъявляют формализованных требований к оценке устойчивости моделей машинного обучения к UAP. Открытыми мы считаем четыре направления: (1) оценка робастности новых архитектур – трансформеров, диффузионных и мультимодальных моделей; (2) защита от физически реализуемых UAP, важная для биометрии и автономного транспорта; (3) подготовка методических документов ФСТЭК по аттестационному тестированию ИИ-компонентов значимых объектов КИИ; (4) сертифицированная робастность больших языковых моделей.

Литература

1. LeCun Y., Bengio Y., Hinton G. Deep learning // Nature. – 2015. – Vol. 521, no. 7553. – P. 436–444. DOI: 10.1038/nature14539.

2. Szegedy C., Zaremba W., Sutskever I., Bruna J., Erhan D., Goodfellow I., Fergus R. Intriguing properties of neural networks // International Conference on Learning Representations (ICLR). – 2014. – arXiv:1312.6199 [Электронный ресурс]. – URL: https://arxiv.org/abs/1312.6199 (дата обращения: 15.05.2026).

3. Goodfellow I.J., Shlens J., Szegedy C. Explaining and harnessing adversarial examples // International Conference on Learning Representations (ICLR). – 2015. – arXiv:1412.6572 [Электронный ресурс]. – URL: https://arxiv.org/abs/1412.6572 (дата обращения: 15.05.2026).

4. Biggio B., Roli F. Wild patterns: Ten years after the rise of adversarial machine learning // Pattern Recognition. – 2018. – Vol. 84. – P. 317–331. – DOI: 10.1016/j.patcog.2018.07.023.

5. Akhtar N., Mian A. Threat of adversarial attacks on deep learning in computer vision: A survey // IEEE Access. – 2018. – Vol. 6. – P. 14410–14430. – DOI: 10.1109/ACCESS.2018.2807385.

6. Moosavi-Dezfooli S.-M., Fawzi A., Fawzi O., Frossard P. Universal adversarial perturbations //

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). – 2017. – P. 1765–1773. – DOI: 10.1109/CVPR.2017.17.

7. Carlini N., Wagner D. Towards evaluating the robustness of neural networks // IEEE Symposium on Security and Privacy (S&P). – 2017. – P. 39–57. – DOI: 10.1109/SP.2017.49.

8. Finlayson S.G., Bowers J.D., Ito J., Zittrain J.L., Beam A.L., Kohane I.S. Adversarial attacks on medical machine learning // Science. – 2019. – Vol. 363, no. 6433. – P. 1287–1289. – DOI: 10.1126/science.aaw4399.

9. Tramèr F., Papernot N., Goodfellow I., Boneh D., McDaniel P. The space of transferable adversarial examples // arXiv preprint. – 2017. – arXiv:1704.03453 [Электронный ресурс]. – URL: https://arxiv.org/abs/1704.03453 (дата обращения: 15.05.2026).

10. Eykholt K., Evtimov I., Fernandes E., Li B., Rahmati A., Xiao C., Prakash A., Kohno T., Song D. Robust physical-world attacks on deep learning visual classification // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). – 2018. – P. 1625–1634. – DOI: 10.1109/CVPR.2018.00175.

11. Mopuri K.R., Ojha U., Garg U., Babu R.V. NAG: Network for adversary generation // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). – 2018. – P. 742–751. – DOI: 10.1109/CVPR.2018.00084.

12. Madry A., Makelov A., Schmidt L., Tsipras D., Vladu A. Towards deep learning models resistant to adversarial attacks // International Conference on Learning Representations (ICLR). – 2018. – arXiv:1706.06083 [Электронный ресурс]. – URL: https://arxiv.org/abs/1706.06083 (дата обращения: 15.05.2026).

13. Croce F., Andriushchenko M., Sehwag V., Debenedetti E., Flammarion N., Chiang M., Mittal P., Hein M. RobustBench: A standardized adversarial robustness benchmark // Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track. – 2021. – arXiv:2010.09670 [Электронный ресурс]. – URL: https://arxiv.org/abs/2010.09670 (дата обращения: 15.05.2026).

14. Athalye A., Carlini N., Wagner D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples // Proceedings of the 35th International Conference on Machine Learning (ICML). – 2018. – P. 274–283. – arXiv:1802.00420 [Электронный ресурс]. – URL: https://arxiv.org/abs/1802.00420 (дата обращения: 15.05.2026).

15. Papernot N., McDaniel P., Wu X., Jha S., Swami A. Distillation as a defense to adversarial perturbations against deep neural networks // IEEE Symposium on Security and Privacy (S&P). – 2016. – P. 582–597. – DOI: 10.1109/SP.2016.41.

16. Goodfellow I., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A., Bengio Y. Generative adversarial nets // Advances in Neural Information Processing Systems (NeurIPS). – 2014. – Vol. 27. – arXiv:1406.2661 [Электронный ресурс]. – URL: https://arxiv.org/abs/1406.2661 (дата обращения: 15.05.2026).

17. Wong E., Rice L., Kolter J.Z. Fast is better than free: Revisiting adversarial training // International Conference on Learning Representations (ICLR). – 2020. – arXiv:2001.03994 [Электронный ресурс]. – URL: https://arxiv.org/abs/2001.03994 (дата обращения: 15.05.2026).

18. Zhang H., Yu Y., Jiao J., Xing E.P., El Ghaoui L., Jordan M.I. Theoretically principled trade-off between robustness and accuracy // Proceedings of the 36th International Conference on Machine Learning (ICML). – 2019. – P. 7472–7482. – arXiv:1901.08573 [Электронный ресурс]. – URL: https://arxiv.org/abs/1901.08573 (дата обращения: 15.05.2026).

19. Shafahi A., Najibi M., Ghiasi M.A., Xu Z., Dickerson J., Studer C., Davis L.S., Taylor G., Goldstein T. Adversarial training for free! // Advances in Neural Information Processing Systems (NeurIPS). – 2019. – Vol. 32. – arXiv:1904.12843 [Электронный ресурс]. – URL: https://arxiv.org/abs/1904.12843 (дата обращения: 15.05.2026).

20. Cohen J., Rosenfeld E., Kolter J.Z. Certified adversarial robustness via randomized smoothing // Proceedings of the 36th International Conference on Machine Learning (ICML). – 2019. – P. 1310–1320. – arXiv:1902.02918 [Электронный ресурс]. – URL: https://arxiv.org/abs/1902.02918 (дата обращения: 15.05.2026).

21. Croce F., Hein M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks // Proceedings of the 37th International Conference on Machine Learning (ICML). – 2020. – P. 2206–2216. – arXiv:2003.01690 [Электронный ресурс]. – URL: https://arxiv.org/abs/2003.01690 (дата обращения: 15.05.2026).

22. Hendrycks D., Dietterich T. Benchmarking neural network robustness to common corruptions and perturbations // International Conference on Learning Representations (ICLR). – 2019. – arXiv:1903.12261 [Электронный ресурс]. – URL: https://arxiv.org/abs/1903.12261 (дата обращения: 15.05.2026).

23. Приказ ФСТЭК России от 11.02.2013 № 17 «Об утверждении Требований о защите информации, не составляющей государственную тайну, содержащейся в государственных информационных системах» (с изм.) [Электронный ресурс] // ФСТЭК России. – URL: https://fstec.ru/dokumenty/vse-dokumenty/prikazy/prikaz-fstek-rossii-ot-11-fevralya-2013-g-n-17 (дата обращения: 15.05.2026).

24. Приказ ФСТЭК России от 18.02.2013 № 21 «Об утверждении Состава и содержания организационных и технических мер по обеспечению безопасности персональных данных при их обработке в информационных системах персональных данных» (с изм.) [Электронный ресурс] // ФСТЭК России. – URL: https://fstec.ru/dokumenty/vse-dokumenty/prikazy/prikaz-fstek-rossii-ot-18-fevralya-2013-g-n-21 (дата обращения: 15.05.2026).

25. Приказ ФСТЭК России от 25.12.2017 № 239 «Об утверждении Требований по обеспечению безопасности значимых объектов критической информационной инфраструктуры Российской Федерации» (с изм.) [Электронный ресурс] // ФСТЭК России. – URL: https://fstec.ru/dokumenty/vse-dokumenty/prikazy/prikaz-fstek-rossii-ot-25-dekabrya-2017-g-n-239 (дата обращения: 15.05.2026).

26. Федеральный закон от 26.07.2017 № 187-ФЗ «О безопасности критической информационной инфраструктуры Российской Федерации» // Собрание законодательства Российской Федерации. – 2017. – № 31 (ч. I). – Ст. 4736.

27. ГОСТ Р ИСО/МЭК 27001-2021. Информационная безопасность, кибербезопасность и защита конфиденциальности. Системы менеджмента информационной безопасности. Требования. – М.: Стандартинформ, 2022. – 28 с.

28. ГОСТ Р 59276-2020. Системы искусственного интеллекта. Способы обеспечения доверия. Общие положения. – М.: Стандартинформ, 2021. – 16 с.

29. Page M.J., McKenzie J.E., Bossuyt P.M. et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews // BMJ. – 2021. – Vol. 372. – Article n71. – DOI: 10.1136/bmj.n71.

30. Bai Y., Mei J., Yuille A.L., Xie C. Are transformers more robust than CNNs? // Advances in Neural Information Processing Systems (NeurIPS). – 2021. – Vol. 34. – arXiv:2111.05464 [Электронный ресурс]. – URL: https://arxiv.org/abs/2111.05464 (дата обращения: 15.05.2026).

31. Gowal S., Rebuffi S.-A., Wiles O., Stimberg F., Calian D.A., Mann T. Improving robustness using generated data // Advances in Neural Information Processing Systems (NeurIPS). – 2021. – Vol. 34. – arXiv:2110.09468 [Электронный ресурс]. – URL: https://arxiv.org/abs/2110.09468 (дата обращения: 15.05.2026).

32. Rebuffi S.-A., Gowal S., Calian D.A., Stimberg F., Wiles O., Mann T. Fixing data augmentation to improve adversarial robustness // arXiv preprint. – 2021. – arXiv:2103.01946 [Электронный ресурс]. – URL: https://arxiv.org/abs/2103.01946 (дата обращения: 15.05.2026).

33. Sehwag V., Mahloujifar S., Handina T., Dai S., Xiang C., Chiang M., Mittal P. Robust learning meets generative models: Can proxy distributions improve adversarial robustness? // International Conference on Learning Representations (ICLR). – 2022. – arXiv:2104.09425 [Электронный ресурс]. – URL: https://arxiv.org/abs/2104.09425 (дата обращения: 15.05.2026).

34. ISO/IEC 23894:2023. Information technology – Artificial intelligence – Guidance on risk management. – Geneva: International Organization for Standardization, 2023. — 30 p.

35. NIST AI Risk Management Framework (AI RMF 1.0). – Gaithersburg: National Institute of Standards and Technology, 2023. – 48 p. – DOI: 10.6028/NIST.AI.100-1.

36. Банк данных угроз безопасности информации [Электронный ресурс] // ФСТЭК России. – URL: https://bdu.fstec.ru/ (дата обращения: 15.05.2026).

37 Намиот Д.Е., Ильюшин Е.А. Атаки на системы машинного обучения // International Journal of Open Information Technologies. – 2022. – Т. 10, № 3. – С. 57–66.

References

1. LeCun Y., Bengio Y., Hinton G. Deep learning. Nature, 2015, vol. 521, no. 7553, pp. 436–444. DOI: 10.1038/nature14539.

2. Szegedy C., Zaremba W., Sutskever I., Bruna J., Erhan D., Goodfellow I., Fergus R. Intriguing properties of neural networks. Proc. ICLR, 2014, arXiv:1312.6199. Available at: https://arxiv.org/abs/1312.6199 (accessed 15.05.2026).

3. Goodfellow I.J., Shlens J., Szegedy C. Explaining and harnessing adversarial examples. Proc. ICLR, 2015, arXiv:1412.6572. Available at: https://arxiv.org/abs/1412.6572 (accessed 15.05.2026).

4. Biggio B., Roli F. Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition, 2018, vol. 84, pp. 317–331. DOI: 10.1016/j.patcog.2018.07.023.

5. Akhtar N., Mian A. Threat of adversarial attacks on deep learning in computer vision: A survey. IEEE Access, 2018, vol. 6, pp. 14410–14430. DOI: 10.1109/ACCESS.2018.2807385.

6. Moosavi-Dezfooli S.-M., Fawzi A., Fawzi O., Frossard P. Universal adversarial perturbations. Proc. IEEE CVPR, 2017, pp. 1765–1773. DOI: 10.1109/CVPR.2017.17.

7. Carlini N., Wagner D. Towards evaluating the robustness of neural networks. Proc. IEEE S&P, 2017, pp. 39–57. DOI: 10.1109/SP.2017.49.

8. Finlayson S.G., Bowers J.D., Ito J., Zittrain J.L., Beam A.L., Kohane I.S. Adversarial attacks on medical machine learning. Science, 2019, vol. 363, no. 6433, pp. 1287–1289. DOI: 10.1126/science.aaw4399.

9. Tramèr F., Papernot N., Goodfellow I., Boneh D., McDaniel P. The space of transferable adversarial examples. arXiv preprint, 2017, arXiv:1704.03453. Available at: https://arxiv.org/abs/1704.03453 (accessed 15.05.2026).

10. Eykholt K., Evtimov I., Fernandes E., Li B., Rahmati A., Xiao C., Prakash A., Kohno T., Song D. Robust physical-world attacks on deep learning visual classification. Proc. IEEE CVPR, 2018, pp. 1625–1634. DOI: 10.1109/CVPR.2018.00175.

11. Mopuri K.R., Ojha U., Garg U., Babu R.V. NAG: Network for adversary generation. Proc. IEEE CVPR, 2018, pp. 742–751. DOI: 10.1109/CVPR.2018.00084.

12. Madry A., Makelov A., Schmidt L., Tsipras D., Vladu A. Towards deep learning models resistant to adversarial attacks. Proc. ICLR, 2018, arXiv:1706.06083. Available at: https://arxiv.org/abs/1706.06083 (accessed 15.05.2026).

13. Croce F., Andriushchenko M., Sehwag V., Debenedetti E., Flammarion N., Chiang M., Mittal P., Hein M. RobustBench: A standardized adversarial robustness benchmark. Proc. NeurIPS Datasets and Benchmarks Track, 2021, arXiv:2010.09670. Available at: https://arxiv.org/abs/2010.09670 (accessed 15.05.2026).

14. Athalye A., Carlini N., Wagner D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. Proc. ICML, 2018, pp. 274–283, arXiv:1802.00420. Available at: https://arxiv.org/abs/1802.00420 (accessed 15.05.2026).

15. Papernot N., McDaniel P., Wu X., Jha S., Swami A. Distillation as a defense to adversarial perturbations against deep neural networks. Proc. IEEE S&P, 2016, pp. 582–597. DOI: 10.1109/SP.2016.41.

16. Goodfellow I., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A., Bengio Y. Generative adversarial nets. Proc. NeurIPS, 2014, vol. 27, arXiv:1406.2661. Available at: https://arxiv.org/abs/1406.2661 (accessed 15.05.2026).

17. Wong E., Rice L., Kolter J.Z. Fast is better than free: Revisiting adversarial training. Proc. ICLR, 2020, arXiv:2001.03994. Available at: https://arxiv.org/abs/2001.03994 (accessed 15.05.2026).

18. Zhang H., Yu Y., Jiao J., Xing E.P., El Ghaoui L., Jordan M.I. Theoretically principled trade-off between robustness and accuracy. Proc. ICML, 2019, pp. 7472–7482, arXiv:1901.08573. Available at: https://arxiv.org/abs/1901.08573 (accessed 15.05.2026).

19. Shafahi A., Najibi M., Ghiasi M.A., Xu Z., Dickerson J., Studer C., Davis L.S., Taylor G., Goldstein T. Adversarial training for free! Proc. NeurIPS, 2019, vol. 32, arXiv:1904.12843. Available at: https://arxiv.org/abs/1904.12843 (accessed 15.05.2026).

20. Cohen J., Rosenfeld E., Kolter J.Z. Certified adversarial robustness via randomized smoothing. Proc. ICML, 2019, pp. 1310–1320, arXiv:1902.02918. Available at: https://arxiv.org/abs/1902.02918 (accessed 15.05.2026).

21. Croce F., Hein M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. Proc. ICML, 2020, pp. 2206–2216, arXiv:2003.01690. Available at: https://arxiv.org/abs/2003.01690 (accessed 15.05.2026).

22. Hendrycks D., Dietterich T. Benchmarking neural network robustness to common corruptions and perturbations. Proc. ICLR, 2019, arXiv:1903.12261. Available at: https://arxiv.org/abs/1903.12261 (accessed 15.05.2026).

23. Prikaz FSTEK Rossii ot 11.02.2013 No. 17 «Ob utverzhdenii Trebovaniy o zashchite informatsii, ne sostavlyayushchey gosudarstvennuyu taynu, soderzhashcheysya v gosudarstvennykh informatsionnykh sistemakh» (s izm.) [Order of the FSTEC of Russia No. 17 of 11.02.2013 "On approval of the Requirements for the protection of information not constituting a state secret contained in state information systems" (as amended)]. FSTEC of Russia. Available at: https://fstec.ru/ (accessed 15.05.2026). (In Russian).

24. Prikaz FSTEK Rossii ot 18.02.2013 No. 21 «Ob utverzhdenii Sostava i soderzhaniya organizatsionnykh i tekhnicheskikh mer po obespecheniyu bezopasnosti personal'nykh dannykh pri ikh obrabotke v informatsionnykh sistemakh personal'nykh dannykh» (s izm.) [Order of the FSTEC of Russia No. 21 of 18.02.2013 "On approval of the Composition and content of organizational and technical measures to ensure the security of personal data during its processing in personal data information systems" (as amended)]. FSTEC of Russia. Available at: https://fstec.ru/ (accessed 15.05.2026). (In Russian).

25. Prikaz FSTEK Rossii ot 25.12.2017 No. 239 «Ob utverzhdenii Trebovaniy po obespecheniyu bezopasnosti znachimykh ob"yektov kriticheskoy informatsionnoy infrastruktury Rossiyskoy Federatsii» (s izm.) [Order of the FSTEC of Russia No. 239 of 25.12.2017 "On approval of the Requirements for ensuring the security of significant objects of the critical information infrastructure of the Russian Federation" (as amended)]. FSTEC of Russia. Available at: https://fstec.ru/ (accessed 15.05.2026). (In Russian).

26. Federal'nyy zakon ot 26.07.2017 No. 187-FZ «O bezopasnosti kriticheskoy informatsionnoy

infrastruktury Rossiyskoy Federatsii» [Federal Law No. 187-FZ of 26.07.2017 "On the security of the critical information infrastructure of the Russian Federation"]. Sobranie zakonodatel'stva Rossiyskoy Federatsii, 2017, no. 31 (pt. I), art. 4736. (In Russian).

27. GOST R ISO/MEK 27001-2021. Informatsionnaya bezopasnost', kiberbezopasnost' i zashchita konfidentsial'nosti. Sistemy menedzhmenta informatsionnoy bezopasnosti. Trebovaniya [GOST R ISO/IEC 27001-2021. Information security, cybersecurity and privacy protection. Information security management systems. Requirements]. Moscow, Standartinform Publ., 2022. 28 p. (In Russian).

28. GOST R 59276-2020. Sistemy iskusstvennogo intellekta. Sposoby obespecheniya doveriya. Obshchie polozheniya [GOST R 59276-2020. Artificial intelligence systems. Methods for ensuring trust. General provisions]. Moscow, Standartinform Publ., 2021. 16 p. (In Russian).

29. Page M.J., McKenzie J.E., Bossuyt P.M. et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ, 2021, vol. 372, n71. DOI: 10.1136/bmj.n71.

30. Bai Y., Mei J., Yuille A.L., Xie C. Are transformers more robust than CNNs? Proc. NeurIPS, 2021, vol. 34, arXiv:2111.05464. Available at: https://arxiv.org/abs/2111.05464 (accessed 15.05.2026).

31. Gowal S., Rebuffi S.-A., Wiles O., Stimberg F., Calian D.A., Mann T. Improving robustness using generated data. Proc. NeurIPS, 2021, vol. 34, arXiv:2110.09468. Available at: https://arxiv.org/abs/2110.09468 (accessed 15.05.2026).

32. Rebuffi S.-A., Gowal S., Calian D.A., Stimberg F., Wiles O., Mann T. Fixing data augmentation to improve adversarial robustness. arXiv preprint, 2021, arXiv:2103.01946. Available at: https://arxiv.org/abs/2103.01946 (accessed 15.05.2026).

33. Sehwag V., Mahloujifar S., Handina T., Dai S., Xiang C., Chiang M., Mittal P. Robust learning meets generative models: Can proxy distributions improve adversarial robustness? Proc. ICLR, 2022, arXiv:2104.09425. Available at: https://arxiv.org/abs/2104.09425 (accessed 15.05.2026).

34. ISO/IEC 23894:2023. Information technology – Artificial intelligence – Guidance on risk management. Geneva, ISO, 2023. 30 p.

35. NIST AI Risk Management Framework (AI RMF 1.0). Gaithersburg, NIST, 2023. 48 p. DOI: 10.6028/NIST.AI.100-1.

36. Bank dannykh ugroz bezopasnosti informatsii [Information security threat data bank]. FSTEC of Russia. Available at: https://bdu.fstec.ru/ (accessed 15.05.2026). (In Russian).

37. Namiot D.E., Ilyushin E.A. Ataki na sistemy mashinnogo obucheniya [Attacks on machine learning systems]. International Journal of Open Information Technologies, 2022, vol. 10, no. 3, pp. 57–66. (In Russian).

ЧАСТИКОВА Вера Аркадьевна, кандидат технических наук, доцент, доцент кафедры кибербезопасности и защиты информации федерального государственного бюджетного образовательного учреждения высшего образования «Кубанский государственный технологический университет» (ФГБОУ ВО «КубГТУ»). 350072, г. Краснодар, ул. Московская, д. 2. E-mail: chastikova_va@mail.ru

МАРЧЕНКО Вячеслав Андреевич, магистрант кафедры кибербезопасности и защиты информации федерального государственного бюджетного образовательного учреждения высшего образования «Кубанский государственный технологический университет» (ФГБОУ ВО «КубГТУ»). 350072, г. Краснодар, ул. Московская, д. 2. E-mail: marche11a.slava@mail.ru.

CHASTIKOVA Vera Arkadievna, Candidate of Technical Sciences, Associate Professor of the Department of Cybersecurity and Information Protection at the Federal State Budgetary Educational Institution of Higher Edu-cation "Kuban State Technological University". 350072, Krasnodar, Moskovskaya str., 2. E-mail: chastikova_va@mail.ru

MARCHENKO Vyacheslav Andreevich, master's degree student of the Department of Cybersecurity and Information Protection at the Federal State Budgetary Educational Institution of Higher Edu-cation "Kuban State Technological University". 350072, Krasnodar, Moskovskaya str., 2. E-mail: marche11a.slava@mail.ru.