Научная статья
DOI 10.66424/2071-8217-2026-2-9
УДК 004.056

ЗАЩИТА ОТ СОСТЯЗАТЕЛЬНЫХ АТАК НА БАЗЕ ДИНАМИЧЕСКИ ПЕРЕСТРАИВАЕМОГО АНСАМБЛЯ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ

Г. Д. Гавва, М. О. Калинин*

Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург, Россия

✉ *max@ibks.spbstu.ru

ДЛЯ ЦИТИРОВАНИЯ

Гавва Г. Д., Калинин М. О. Защита от состязательных атак на базе динамически перестраиваемого ансамбля моделей машинного обучения // Проблемы информационной безопасности. Компьютерные системы. 2026. № 2. С. 113–120. DOI: 10.66424/2071-8217-2026-2-9

ПОСТУПИЛА 27.04.2026
ПРИНЯТА 07.05.2026
ОПУБЛИКОВАНА 15.06.2026

Издатель: Санкт-Петербургский политехнический университет Петра Великого

АННОТАЦИЯ

Рассмотрена проблема защиты моделей машинного обучения от состязательных атак. Представлен метод защиты, основанный на динамически перестраиваемом ансамбле классификаторов с механизмом отказа, который объединяет: случайную комбинацию гетерогенных подмоделей, онлайн-анализ дисперсии прогнозов, имитацию правдоподобного ответа при атаке и механизм моделей-ловушек. Анализ согласованности выходов внутри ансамбля и отказ от выдачи наиболее вероятного прогноза снижает результативность действий нарушителя при анализе им обратной связи, получаемой от целевой модели, и генерации состязательных образцов. Экспериментальная оценка, проведенная на наборе данных UNSW-NB15, показала, что разработанный метод сохраняет высокую исходную точность защищаемой модели при воздействии состязательных атак (85–95 %) при минимальном ее снижении на 1–3 п.п. Метод позволяет устранить до 98 % атак, что значительно превосходит показатели таких широко распространенных аналогов.

КЛЮЧЕВЫЕ СЛОВА

Защита машинного обучения, ансамбль моделей, классификация, механизм отказа, состязательные атаки

Original article
DOI 10.66424/2071-8217-2026-2-9

PROTECTION AGAINST ADVERSARIAL ATTACKS BASED ON A DYNAMICALLY RECONFIGURABLE ENSEMBLE OF MACHINE LEARNING MODELS

G. D. Gavva, M. O. Kalinin*

Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia

✉ *max@ibks.spbstu.ru

FOR CITATION

Gavva G. D., Kalinin M. O. Protection against adversarial attacks based on a dynamically reconfigurable ensemble of machine learning models. *Problems of information security. Computer systems*. 2026. No. 2, pp. 113–120. DOI: 10.66424/2071-8217-2026-2-9 (In Russian)

RECEIVED 27.04.2026
ACCEPTED 07.05.2026
PUBLICATION 15.06.2026

ABSTRACT

The paper reviews the problem of protecting machine learning models from adversarial attacks. A protection method is presented based on a dynamically reconfigurable ensemble of classifiers with a failure mechanism that combines a random combination of heterogeneous sub-models, online analysis of forecast variance, simulation of a plausible attack response, and a decoy model mechanism. Analysis of the consistency of outputs in the ensemble and failure to issue the most probable output reduces the effectiveness of an attacker when analyzing feedback received from the target model and generating adversarial samples. An experimental evaluation conducted on the UNSW-NB15 dataset showed that the developed method maintains high initial accuracy of the protected model under adversarial attacks (85–95 %) with a minimal decrease of 1–3 percentage points. The method can eliminate up to 98 % of attacks, significantly exceeding the performance of similar widely used methods.

KEYWORDS

Protection of machine learning, ensemble of models, classification, mechanism for rejecting, adversarial attacks

1. ВВЕДЕНИЕ

Машинное обучение стало неотъемлемой частью практически всех сфер человеческой деятельности. Технологии искусственного интеллекта активно применяются в здравоохранении, транспорте, финансах, кибербезопасности и во многих других областях, где они демонстрируют результаты, сопоставимые или превосходящие человеческие.

Как показывают результаты исследований, модели глубокого обучения могут значительно улучшить оказание медицинской помощи в части выявления и мониторинга прогнозов различных заболеваний [1]. В транспортной отрасли цифровые интеллектуальные технологии лежат в основе развития автономных транспортных средств и «умных» транспортных систем, включая адаптивное управление движением и анализ дорожных событий в реальном времени [2]. В финансовом секторе модели машинного обучения применяются для инвестиционного консультирования, управления рисками и обнаружения мошенничества, а также для обслуживания клиентов с помощью многоязычных чат-ботов [3]. Повсеместная интеграция механизмов машинного обучения в критические информационные инфраструктуры усилила привлекательность новых цифровых сервисов для злоумышленников и спровоцировала рост специфических атак на эти технологии. Одними из ключевых являются атаки извлечения данных, моделей машинного обучения, а также атаки уклонения и отравления моделей [4–8]. Разнообразие новых классов атак на машинное обучение и их широкий спектр воздействия определили разработку и внедрение новых подходов и средств защиты, которые смогли бы противодействовать новой угрозе безопасности (табл. 1).

Несмотря на разнообразие существующих решений они имеют общие характерные недостатки: вынужденный компромисс между точностью, устойчивостью и вычислительной эффективностью и недостаточную адаптивность. Это обусловливает необходимость создания легких, гибких и действенных защитных механизмов, которые способны обнаруживать и нейтрализовать как известные, так и ранее не встречавшиеся атаки без существенного снижения производительности защищаемой модели [18].

Для преодоления ограничений известных подходов и повышения устойчивости к адаптивным атакам, универсальности и результирующей точности защищаемой модели машинного обучения авторами предложен метод защиты моделей машинного обучения, основанный на использовании динамически перестраиваемого ансамбля классификаторов с механизмом отказа.

Таблица 1 | Сопоставление атак на модели машинного обучения и методов защиты

Table 1 | Matching of attacks on machine learning models and protection methods

Разновидности атак	Методы защиты
Атака уклонения	1. Защитное дистиллирование [9, 10] усложняет вычисление градиентов. 2. Сжатие признаков [11], обфускация входа [12, 13], защита на основе рандомизации [9, 14], упрощение выходных данных [15] затрудняют подбор нарушителем входных данных. 3. Ограничение скорости запросов [15] замедляет нарушителя
Атака отравления	1. Дифференциальная приватность [9, 14] снижает влияние выбросов. 2. Безопасные вычисления для нескольких сторон [9, 16] при совместном обучении понижают риск целенаправленных атак на конкретные данные
Атака извлечения данных	1. Дифференциальная приватность затрудняет понимание извлеченных данных. 2. Гомоморфное шифрование [9] обеспечивает сокрытие данных даже в случае их извлечения
Атака извлечения модели	1. Ограничение скорости запросов делает процесс извлечения непрактично долгим для нарушителя. 2. Упрощение выходных данных снижает точность восстановления, значительно увеличивает необходимое число запросов нарушителем. 3. Внедрение паспортных слоев [12, 17] делает извлечение бесполезным для нарушителя, поскольку без ключа скопированная модель выдает бессмысленные результаты

2. МЕТОДЫ

Разработанный метод объединяет защитные механизмы на архитектурном уровне и реализует активную стратегию противодействия (см. рисунок).

Целевая модель машинного обучения представлена в виде пула из N подмоделей-экземпляров, каждый из которых обучен на модифицированных данных с использованием методов аугментации, добавления шума, отбора подмножеств признаков либо обладает отличной от других архитектурой. Все подмодели-экземпляры решают одну и ту же задачу. Вариативность ансамбля усложняет проведение атаки нарушителем, поскольку наличие некорректных данных среди обучающих не обеспечивает искажение каждой подмодели из ансамбля.

Для каждого легитимного запроса ансамбль случайным образом формирует подмножество из K моделей ($K<N$), агрегирует их прогнозы посредством усреднения и выдает итоговый прогноз общей модели. Случайный характер комбинации подмоделей в ансамбле и гетерогенность подмоделей в пуле существенно затрудняют для нарушителя создание универсального состязательного примера, эффективного в произвольный момент времени.

Механизм отказа функционирует на уровне обработки данных. Параллельно с основным вычислительным процессом каждый входной вектор X направляется в модуль согласованности, который пропускает его через все N подмоделей пула и вычисляет дисперсию полученных предсказаний. Для легитимных данных все подмодели, обученные на едином наборе, демонстрируют высокую согласованность. В данном случае дисперсия будет низкой, и запрос будет классифицирован как доверенный, и пользователю возвращается прогноз, полученный от быстрого ансамбля из K моделей.

При подаче состязательного примера малые, незаметные для человека искажения по-разному дестабилизируют различные подмодели анасамбля, что приводит к росту дисперсии прогнозов от подмоделей. При превышении заданного порога срабатывает механизм отказа. Система не выдает явного сообщения об ошибке, а имитирует нестандартное, но правдоподобное поведение (например, в задачах

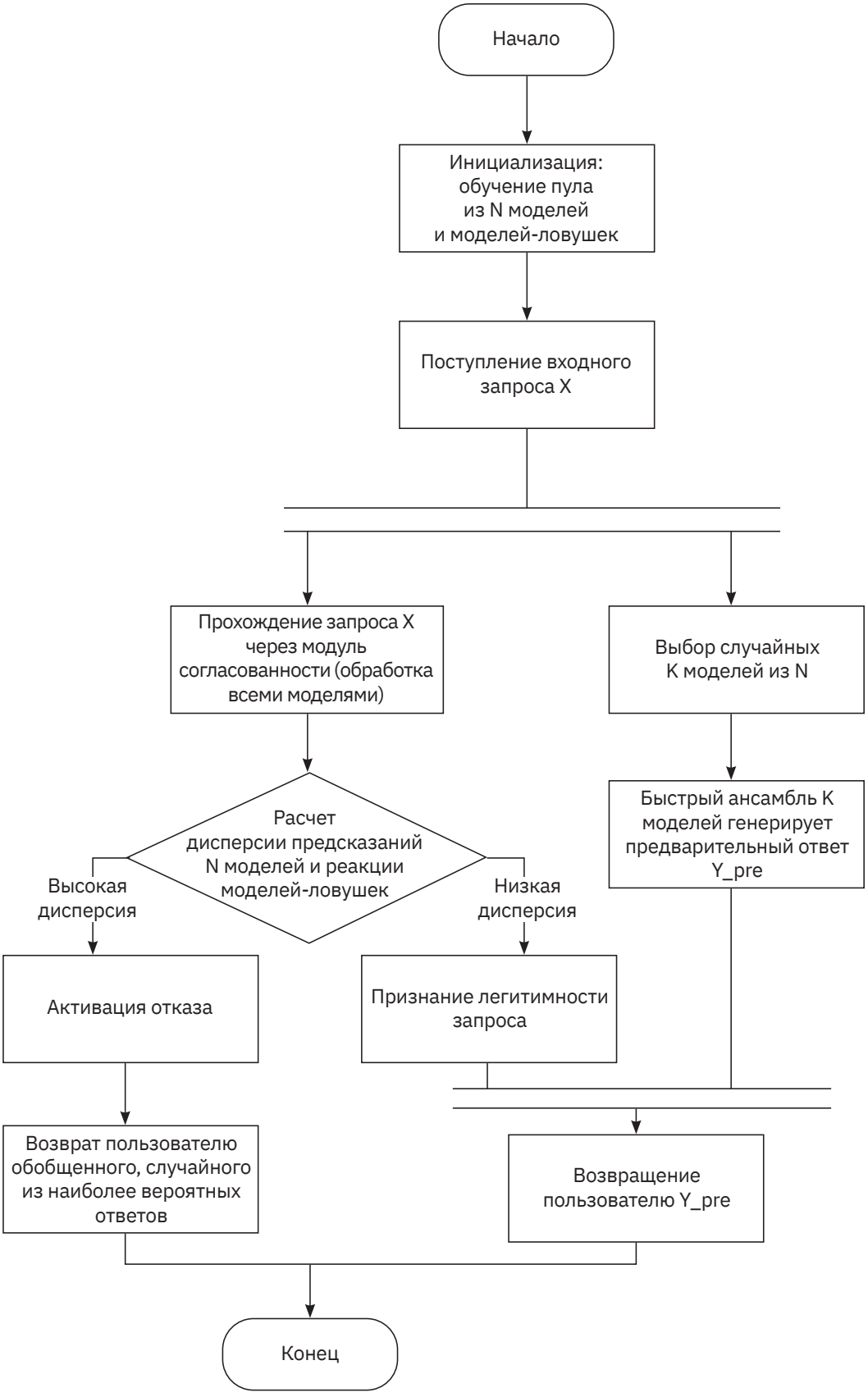

Схема разработанного метода

Scheme of the developed method

классификации ансамбль возвращает случайный, но семантически близкий ответ из числа наиболее вероятных классов). Тем самым нарушитель лишается устойчивого сигнала обратной связи (например, градиента), необходимого ему для итеративной корректировки состязательны образцов, что значительно усложняет эксплуатацию атак извлечения данных и извлечения модели.

Дополнительно в данном методе используется уровень подмоделей-ловушек. В ансамбль включаются одна или несколько специально обученных «хрупких» подмоделей-ловушек, обладающих повышенной чувствительностью к незначительным искажениям данных. Их аномальная реакция, а именно резкое отклонение выдаваемых прогнозов, служит триггером для активации механизма отказа даже в тех случаях, когда основные подмодели еще не фиксируют значимой дисперсии. Обучение подмоделей-ловушек производится либо на обратной задаче предсказания заведомо неверного, но детерминированного класса для части данных, либо на состязательных выборках данных. В результате при предъявлении нового состязательного образца подмодель-ловушка с высокой вероятностью активирует указанный атакующий класс либо выдает хаотичный прогноз, существенно расходящийся с предсказаниями основных моделей.

3. РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Для экспериментальной оценки разработанного метода построен ансамбль из восьми моделей, включающий рекуррентные нейросети, многослойный персептрон, трансформеры и модели-ловушки по две модели каждого типа. Подмножество быстрого вывода составляет три подмодели ($K=3$). Тестирование проведено на наборе данных UNSW-NB15, содержащем обширный набор примеров атак (фаззеры, бэкдоры, DoS-атаки, эксплойты, общие атаки, разведывательные атаки, шеллкоды и черви). Результаты анализа эффективности разработанного метода представлены в табл. 2. Предложенное решение

Таблица 2 | Анализ применения методов на наборе данных UNSW-NB15

Table 2 | Analysis of the application of methods in the UNSW-NB15 dataset

Метод	Исходная точность защищаемой модели при воздействии атаки, %	Снижение точности защищаемой модели при внедрении метода защиты при воздействии атаки, п. п.	Прирост вычислительной стоимости (процент времени), %	Доля устраняемых атак, %
Разработанный метод	85–95	1–3	15–25	95–98
Гомоморфное шифрование/безопасные вычисления для нескольких сторон	95 (для конфиденциальности данных)	0	≥1000	86–92
Дифференциальная приватность	70–90	3–10	10–25	88–90
Защитное дистиллярование	30–70	2–8	<5	30–35
Обфускация входа/рандомизация	20–60	1–5	1–10	1–10
Сжатие признаков	10–40	2–8	1–5	1–10
Упрощение выходов/ограничение скорости запросов	0–30	0–4	<1	27–35
Дополнительные слои	60–80	0,5–2,0	<2	90–92

обеспечивает баланс между точностью, устойчивостью и эффективностью, кроме того, усложняет задачу компрометации модели при попытке создания и эксплуатации нарушителем состязательных примеров.

4. ЗАКЛЮЧЕНИЕ

Полученные результаты свидетельствуют о том, что предложенный метод обеспечивает высокую точность защищаемой модели под воздействием атаки (85–95 %) при минимальном ее снижении на 1–3 п. п. и позволяет устранить до 98 % атак, что значительно превосходит показатели таких популярных аналогов, как защитное дистиллирование и дифференциальная приватность. При этом вычислительные затраты возрастают лишь на 15–25 %, тогда как, например, гомоморфное шифрование требует более чем 1000 %-ного увеличения времени обработки, что подтверждает достижение методом наилучшего баланса между точностью, устойчивостью и эффективностью среди рассмотренных решений.

Разработанный метод реализует активную стратегию, затрудняющую для нарушителей процесс генерации и подбора состязательных примеров по сравнению с известными аналогами за счет достигнутого эффекта скрытности. В качестве перспективного направления дальнейших исследований рассматривается адаптация разработанного метода для использования в легковесных вычислительных системах.

КОНФЛИКТ ИНТЕРЕСОВ / CONFLICT OF INTERESTS

Авторы заявляют об отсутствии конфликта интересов / The authors declare no conflict of interests.

СПИСОК ИСТОЧНИКОВ

1. **Okeibunor J. C., Jaca A., Iwu-Jaja C. J. et al.** The use of artificial intelligence for delivery of essential health services across WHO regions: a scoping review // Frontiers in Public Health. 2023. Vol. 11. P. 1102185. DOI: 10.3389/fpubh.2023.1102185.
2. **Buenaventura M., Shenk A., Nergui A. et al.** Artificial Intelligence Adoption and Sectoral Transformation: Implications for Health Care, Financial Services, Climate and Energy, and Transportation. 2025. № RR-A3888-1. DOI: 10.7249/rra3888-1.
3. **Беспалов Д. А., Богатырева М. В.** Роль искусственного интеллекта в финансовом секторе // Вестник Алтайской академии экономики и права. 2023. № 7–1. С. 10.
4. **Abomakhelb A., Jalil K. A., Buja A. G. et al.** A Comprehensive Review of Adversarial Attacks and Defense Strategies in Deep Neural Networks // Technologies. 2025. Vol. 13. № 5. P. 202. DOI: 10.3390/technologies13050202.
5. **Жуковский Е. В., Огнев Р. А.** Анализ возможности реализации состязательных атак на средства проактивной защиты, использующие машинное обучение // Методы и технические средства обеспечения безопасности информации. 2021. № 30. С. 28–29. DOI: 10.31799/2949-0693-2023-161-71.
6. **Беззатеев С. В., Афанасьева А. В., Супрун А. Ф.** Атаки на обучающие выборки в системах машинного обучения и защита от них // Инновационное приборостроение. 2023. Т. 2. № 1. С. 61–71.
7. **Jha P. K.** Adversarial Machine Learning: Attacks, Defenses, and Open Challenges // arXiv preprint arXiv:2502.05637. 2025.
8. **Намиот Д. Е.** Введение в атаки отравлением на модели машинного обучения // International Journal of Open Information Technologies. 2023. Т. 11. № 3. С. 58–68.
9. **El-Husseini F., Noura H. N., Vernier F.** Security and privacy-preserving for machine learning models: attacks, countermeasures, and future directions // Annals of Telecommunications. 2025. P. 1–22. DOI: 10.1109/CSNet64211.2024.10851722.

10. **Kuzlu M., Catak F. O., Cali U. et al.** Adversarial security mitigations of mmWave beamforming prediction models using defensive distillation and adversarial retraining // International Journal of Information Security. 2023. Vol. 22. № 2. P. 319–332.
11. **Xu W., Evans D., Qi Y.** Feature squeezing: Detecting adversarial examples in deep neural networks // arXiv preprint arXiv:1704.01155. 2017.
12. **Lederer I., Mayer R., Rauber A.** Identifying appropriate intellectual property protection mechanisms for machine learning models: A systematization of watermarking, fingerprinting, model access, and attacks // IEEE Transactions on Neural Networks and Learning Systems. 2023. Vol. 35. № 10. P. 13082–13100.
13. **Chen M., Wu M.** Protect your deep neural networks from piracy // 2018 IEEE International Workshop on Information Forensics and Security (WIFS). IEEE, 2018. P. 1–7.
14. **Lecuyer M., Atlidakis V., Geambasu R. et al.** Certified robustness to adversarial examples with differential privacy // 2019 IEEE symposium on security and privacy (SP). IEEE, 2019. P. 656–672.
15. **Isakov M., Bu L., Cheng H., Kinsy M. A.** Preventing neural network model exfiltration in machine learning hardware accelerators // 2018 Asian Hardware Oriented Security and Trust Symposium (AsianHOST). IEEE, 2018. P. 62–67.
16. **Tiwari S. S., Dhasmana G., Al-Jawahry H. M. et al.** Federated Learning Strategies for Privacy-Preserving Machine Learning Models in Cloud Computing Environments // 2024 International Conference on Communication, Computer Sciences and Engineering (IC3SE). IEEE, 2024. P. 1457–1462. DOI: 10.1109/IC3SE62002.2024.10593458.
17. **Fan L., Ng K. W., Chan C. S.** Rethinking deep neural network ownership verification: Embedding passports to defeat ambiguity attacks // Advances in neural information processing systems. 2019. Vol. 32. DOI: 10.48550/arXiv.1909.07830.
18. **Югай П. Э., Москвин Д. А.** Способы выявления состязательных атак на алгоритмы машинного обучения в системах обнаружения вторжений // Методы и технические средства обеспечения безопасности информации. 2023. № 32. С. 21–22.

REFERENCES

1. **Okeibunor J. C., Jaca A., Iwu-Jaja C. J. et al.** The use of artificial intelligence for delivery of essential health services across WHO regions: a scoping review. *Frontiers in Public Health*. 2023. Vol. 11, pp. 1102185. DOI: 10.3389/fpubh.2023.1102185.
2. **Buenaventura M., Shenk A., Nergui A. et al.** Artificial Intelligence Adoption and Sectoral Transformation: Implications for Health Care, Financial Services, Climate and Energy, and Transportation. 2025. No. RR-A3888-1. DOI: 10.7249/rra3888-1.
3. **Bespalov D. A., Bogatyreva M. V.** The role of artificial intelligence in the financial sector. *Journal of the Altai Academy of Economics and Law*. 2023. No. 7–1, pp. 10. (In Russian)
4. **Abomakhelb A., Jalil K. A., Buja A. G. et al.** A Comprehensive Review of Adversarial Attacks and Defense Strategies in Deep Neural Networks. *Technologies*. 2025. Vol. 13. No. 5, pp. 202. DOI: 10.3390/technologies13050202.
5. **Zhukovsky E. V., Ognev R. A.** Analysis of the possibility of implementing adversarial attacks on proactive defense tools using machine learning. *Metody i tehnicheskie sredstva obespecheniya bezopasnosti informacii*. 2021. No. 30. C. 28–29. DOI: 10.31799/2949-0693-2023-161-71. (In Russian)
6. **Bezzateev S. V., Afanasyeva A. V., Suprun A. F.** Attacks on data sets in machine learning systems and protection against them. *Innovacionnoe priborostroenie = Innovative Instrumentation*. 2023. Vol. 2. No. 1, pp. 61–71. DOI: 10.31799/2949-0693-2023-161-71. (In Russian)
7. **Jha P. K.** Adversarial Machine Learning: Attacks, Defenses, and Open Challenges. *arXiv preprint arXiv:2502.05637*. 2025.
8. **Namiot D. E.** Introduction to poisoning attacks on machine learning models. *International Journal of Open Information Technologies*. 2023. Vol. 11. No. 3, pp. 58–68. (In Russian)
9. **El-Husseini F., Noura H. N., Vernier F.** Security and privacy-preserving for machine learning models: attacks, countermeasures, and future directions. *Annals of Telecommunications*. 2025, pp. 1–22. DOI: 10.1109/CSNet64211.2024.10851722.

10. **Kuzlu M., Catak F. O., Cali U. et al.** Adversarial security mitigations of mmWave beamforming prediction models using defensive distillation and adversarial retraining. *International Journal of Information Security*. 2023. Vol. 22. No. 2, pp. 319–332.
11. **Xu W., Evans D., Qi Y.** Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv preprint arXiv:1704.01155*. 2017.
12. **Lederer I., Mayer R., Rauber A.** Identifying appropriate intellectual property protection mechanisms for machine learning models: A systematization of watermarking, fingerprinting, model access, and attacks. *IEEE Transactions on Neural Networks and Learning Systems*. 2023. Vol. 35. No. 10, pp. 13082–13100.
13. **Chen M., Wu M.** Protect your deep neural networks from piracy. 2018 IEEE International Workshop on Information Forensics and Security (WIFS). IEEE, 2018, pp. 1–7.
14. **Lecuyer M., Atlidakis V., Geambasu R. et al.** Certified robustness to adversarial examples with differential privacy. 2019 IEEE symposium on security and privacy (SP). IEEE, 2019, pp. 656–672.
15. **Isakov M., Bu L., Cheng H., Kinsy M. A.** Preventing neural network model exfiltration in machine learning hardware accelerators. 2018 Asian Hardware Oriented Security and Trust Symposium (AsianHOST). IEEE, 2018, pp. 62–67.
16. **Tiwari S. S., Dhasmana G., Al-Jawahry H. M. et al.** Federated Learning Strategies for Privacy-Preserving Machine Learning Models in Cloud Computing Environments. 2024 International Conference on Communication, Computer Sciences and Engineering (IC3SE). IEEE, 2024, pp. 1457–1462. DOI: 10.1109/IC-3SE62002.2024.10593458.
17. **Fan L., Ng K. W., Chan C. S.** Rethinking deep neural network ownership verification: Embedding passports to defeat ambiguity attacks. *Advances in neural information processing systems*. 2019. Vol. 32. DOI: 10.48550/arXiv.1909.07830.
18. **Yugai P. E., Moskvin D. A.** Methods of detecting adversarial attacks on machine learning algorithms in intrusion detection systems. *Metody i tehnicheskie sredstva obespecheniya bezopasnosti informacii*. 2023. No. 32, pp. 21–22. (In Russian)

СВЕДЕНИЯ ОБ АВТОРАХ / INFORMATION ABOUT AUTHORS

ГАВВА Георгий Дмитриевич – магистр, ассистент, Санкт-Петербургский политехнический университет Петра Великого, Россия, 195251, Санкт-Петербург, ул. Политехническая, д. 29
E-mail: gavva_gd@spbstu.ru

GAVVA Georgij D. – Master's Student, assistant, Peter the Great St. Petersburg Polytechnic University, Russia, 195251, St. Petersburg, Polytechnicheskaya str., 29

КАЛИНИН Максим Олегович – д-р техн. наук, профессор, Санкт-Петербургский политехнический университет Петра Великого, Россия, 195251, Санкт-Петербург, ул. Политехническая, д. 29
E-mail: max@ibks.spbstu.ru
ORCID: 0000-0002-9732-0099

KALININ Maxim O. – Doctor of Engineering Sciences, Professor, Peter the Great St. Petersburg Polytechnic University, Russia, 195251, St. Petersburg, Polytechnicheskaya str., 29