УДК 004.056 **DOI: 10.18413/2518-1092-2022-8-2-0-7**

Герасимов В.М. [1]
Маслова М.А. [2,3]
Халилаева Э.И. [2]

ЗАЩИТА ОТ СОСТЯЗАТЕЛЬНЫХ АТАК НА АУДИО И ИЗОБРАЖЕНИЯ В МОДЕЛЯХ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА С ПРИМЕНЕНИЕМ МЕТОДА SGEC

[1)] Федеральное государственное автономное образовательное учреждение высшего образования «Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики», Кронверкский пр., д. 49, г. Санкт-Петербург, 197101, Россия
[2)] Севастопольский государственный университет, ул. Университетская, д. 33, г. Севастополь, 299053, Россия
[3)] Ростовский государственный экономический университет (РИНХ), ул. Большая Садовая, д. 69, 7. Ростов-на-Дону, 344002, Россия

e-mail: my.virus.kaspersky@gmail.com, mashechka-81@mail.ru, emine.halilaeva@yandex.ru

Аннотация
В современном мире использование искусственного интеллекта (ИИ) все чаще сталкивается с риском состязательных атак на аудио и изображения. Данная статья исследует эту проблему и представляет метод SGEC как средство минимизации этих рисков. Рассматриваются различные виды атак на аудио и изображения, такие как искажение разметки, атаки "белого ящика" и "черного ящика", утечки через обученные модели и атаки на уровне железа. Основной акцент делается на методе SGEC, который предлагает шифрование данных и обеспечение их целостности в моделях ИИ. Статья также рассматривает другие способы защиты аудио и изображений, включая двойную проверку и ансамбли методов, ограничение доступа и анонимизацию данных, а также использование доказуемо устойчивых моделей ИИ.
Ключевые слова: состязательные атаки; защита голосовых отпечатков; защита биометрических данных; стеганография; шифрование данных; риски состязательных атак
Для цитирования: Герасимов В.М., Маслова М.А., Халилаева Э.И. Защита от состязательных атак на аудио и изображения в моделях искусственного интеллекта с применением метода SGEC // Научный результат. Информационные технологии. – Т.8, №2, 2023. – С. 53-60. DOI: 10.18413/2518-1092-2022-8-2-0-7

Gerasimov V.M.[1]
Maslova M.A.[2,3]
Khalilayeva E.I. [2]

PROTECTION AGAINST ADVERSARIAL ATTACKS ON AUDIO AND IMAGES IN ARTIFICIAL INTELLIGENCE MODELS USING THE SGEC METHOD

[1)] Saint Petersburg National Research University of Information Technologies, Mechanics and Optics, 49 Kronverkskiy prospekt, St. Petersburg, 197101, Russia
[2)] Sevastopol State University, 33 Universitetskaya St., Sevastopol, 299053, Russia
[3)] Rostov State Economic University (RINH), 69 Bolshaya Sadovaya St., Rostov-on-Don, 344002, Russia

e-mail: my.virus.kaspersky@gmail.com, mashechka-81@mail.ru, emine.halilaeva@yandex.ru

Abstract
In the modern world, the use of artificial intelligence (AI) is increasingly facing the risk of adversarial attacks on audio and images. This article explores this issue and presents the SGEC method as a means to minimize these risks. Various types of attacks on audio and images are discussed, including label manipulation, white-box and black-box attacks, leakage through trained models, and hardware-level attacks. The main focus is on the SGEC method, which offers data encryption and ensures their integrity in AI models. The article also examines other approaches to protect audio and images, such as dual verification and ensemble methods, access restriction and data anonymization, as well as the use of provably robust AI models.
Keywords: adversarial attacks; voiceprint protection; biometric data protection; steganography; data encryption; risks of adversarial attacks

For citation: Gerasimov V.M., Maslova M.A., Khalilayeva E.I. Protection against adversarial attacks on audio and images in artificial intelligence models using the SGEC method // Research result. Information technologies. – T.8, №2, 2023. – P. 53-60. DOI: 10.18413/2518-1092-2022-8-2-0-7

ВВЕДЕНИЕ

Развитие моделей искусственного интеллекта (ИИ), основанных на аудио и изображениях, открывает новые перспективы во множестве областей, включая медицину, автономные системы и мультимедийные приложения. Однако, с возросшим применением таких моделей, возникают и новые риски, связанные с состязательными атаками на аудио и изображения.

Состязательные атаки на аудио [1] и изображения [2] представляют собой специально разработанные методы введения искажений или шумов во входные данные моделей с целью обмана и искажения их результатов. Такие атаки могут быть использованы злоумышленниками для изменения выводов моделей ИИ и даже подрыва их безопасности.

В данной статье рассмотрим возможные риски состязательных атак на аудио и изображения в моделях искусственного интеллекта, обсудим способы манипуляции с данными, которые могут привести к неправильным результатам и искажениям, изучим последствия подобных атак и их потенциальные угрозы в различных сферах применения моделей на основе аудио и изображений.

В контексте защиты от этих рисков представлен метод SGEC [3], который предлагает шифрование данных, предназначенный для защиты системы ИИ от состязательных атак. Рассмотрим принципы работы и преимущества этого метода, а также его важность в обеспечении безопасности и надежности моделей ИИ на основе аудио и изображений.

Разбор возможных рисков состязательных атак на аудио и изображения и представление метода SGEC позволит нам получить более полное представление о безопасности и защите систем ИИ, а также определить наилучшие практики и меры предосторожности для предотвращения таких атак.

ОСНОВНАЯ ЧАСТЬ

Использование незащищённых данных в системах искусственного интеллекта (ИИ) сопряжено с рядом серьезных рисков. Вот некоторые из них:

1. Утечка конфиденциальной информации [4]. Если данные, содержащие личную или чувствительную информацию, попадут в руки злоумышленников, это может привести к утечке конфиденциальных данных. Например, в случае использования биометрических данных, таких как голос или лицо пользователей, утечка таких данных может привести к незаконному доступу к личным аккаунтам или идентификационным системам.

2. Нарушение приватности [5]. Незащищённые данные могут раскрывать личную информацию о пользователях, их привычках, предпочтениях и поведении. Это может нарушить их приватность и привести к нежелательной рекламе, мошенничеству или другим формам злоупотребления личной информацией.

3. Манипуляция и подделка данных [6]. Злоумышленники могут внести изменения в незащищённые данные, предоставленные системе ИИ, с целью искажения результатов или обмана системы. Например, подделка данных об обучающей выборке может привести к неправильным выводам или искаженным предсказаниям модели ИИ.

4. Атаки на систему ИИ [7]. Незащищённые данные могут использоваться для проведения атак на систему ИИ. Например, злоумышленники могут внедрить вредоносный код или манипулировать данными, чтобы нарушить работу или контроль над системой ИИ.

5. Утрата доверия пользователей [8]. Если данные пользователей не защищены должным образом, это может привести к потере доверия со стороны пользователей. Пользователи могут отказаться от использования системы ИИ или отказаться от предоставления своих данных из-за опасений по поводу их безопасности и конфиденциальности.

6. Юридические последствия [9]. Использование незащищённых данных может нарушать законодательство о защите данных, такое как общий регламент по защите данных

(GDPR) в Европейском союзе. Нарушение таких правил может привести к юридическим последствиям, включая штрафы и судебные разбирательства.

Учитывая эти риски, защита данных и обеспечение безопасности системы ИИ становятся критически важными аспектами, чтобы защитить пользователей, предотвратить утечки информации и обеспечить доверие к технологии ИИ [12].

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ И ИХ ОБСУЖДЕНИЕ

Применение рисков, связанных с состязательными атаками на аудио и изображения, позволяет системе нормально функционировать путем обеспечения ее безопасности и защиты от потенциальных угроз.

Путем изучения и понимания возможных рисков, связанных с состязательными атаками, система ИИ может принять соответствующие меры для защиты своих моделей и данных. Это позволяет обнаруживать и предотвращать попытки искажения или подмены аудио и изображений, которые могут повлиять на точность и достоверность результатов моделей ИИ.

Таблица

Риски состязательных атак и их описание

Table

Risks of adversarial attacks and their description

Риски информационной безопасности	**Описание**
1. Искажение разметки	Злоумышленники могут изменять метаданные и разметку данных, вводя искажения, чтобы ввести в заблуждение систему ИИ.
2. Искажение обучающей выборки	Злоумышленники могут изменять обучающую выборку данных, вводя искажения или злонамеренные примеры, для искажения работы системы ИИ.
3. Атаки "белого ящика" и "черного ящика"	Злоумышленники могут использовать атаки "белого ящика" и "черного ящика" для взлома системы ИИ путем доступа к ее внутренним параметрам или ввода вредоносных данных.
4. Дискретные правила обнаружения на основе ML	Злоумышленники могут использовать методы обхода и обмана системы ИИ, чтобы избежать обнаружения искусственным интеллектом, работающим на основе дискретных правил.
5. Атаки на предобученные и аутсорсинговые ML-модели	Злоумышленники могут нацелиться на предобученные модели и аутсорсинговые модели, пытаясь получить доступ к их конфиденциальным данным или вводя вредоносные данные.
6. Утечки через обученные модели	Злоумышленники могут использовать утечки информации через обученные модели, чтобы получить доступ к конфиденциальным данным, используемым в системе ИИ.
7. Атаки на уровне железа	Злоумышленники могут проводить физические атаки на инфраструктуру системы ИИ, включая атаки на железо и перехват электромагнитных излучений.

Таблица представляет основные риски информационной безопасности, связанные с защитой от состязательных атак на аудио и изображения в моделях искусственного интеллекта. Они могут нанести вред работе системы ИИ и утечке конфиденциальных данных. Применение метода SGEC позволяет снизить вероятность возникновения данных рисков и обеспечить безопасность системы ИИ.

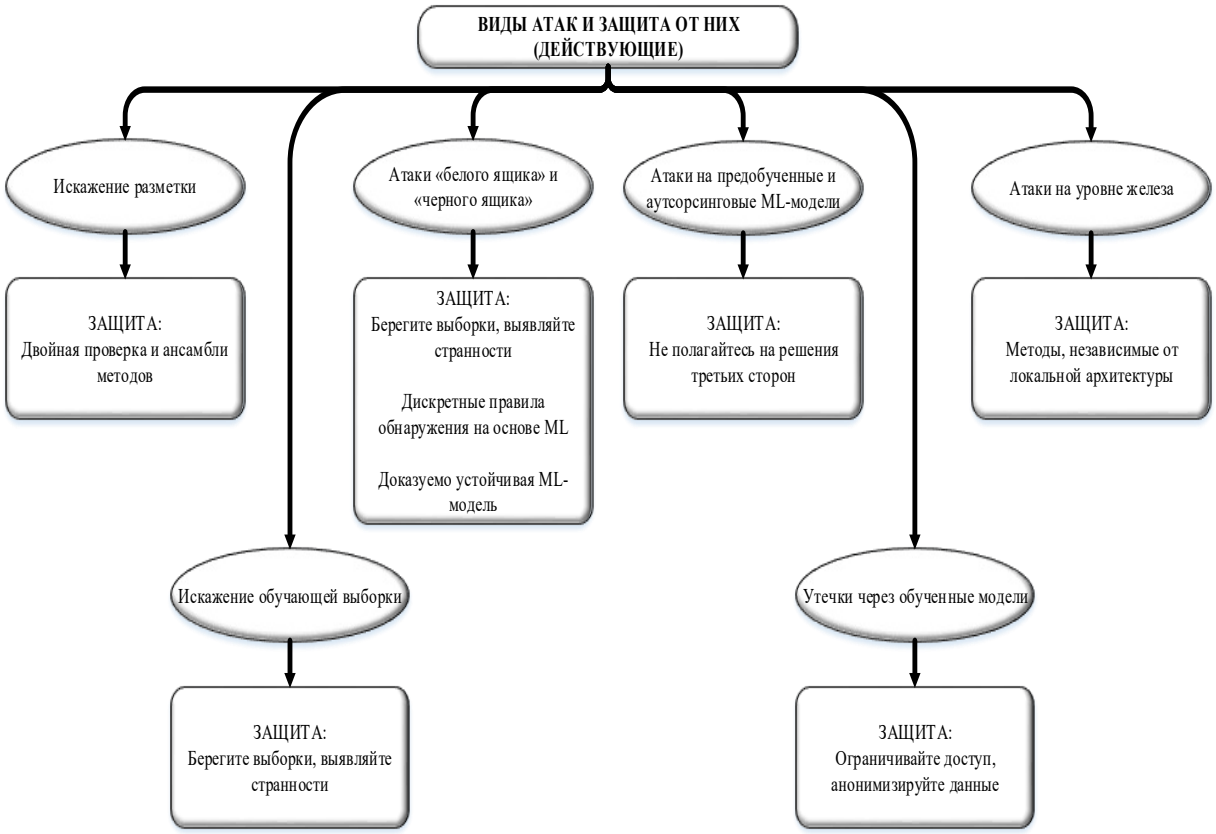

Рис. 1. Виды атак на систему ИИ и защита от них
Fig. 1. Types of attacks on AI systems and their defense

Использование метода SGEC, который предлагает шифрование данных [10], позволяет системе ИИ обеспечить безопасность и надежность при обработке аудио и изображений. Этот метод помогает предотвратить несанкционированный доступ к данным, а также защищает их от внешних вмешательств и состязательных атак.

Благодаря реализации этих мер защиты система ИИ может сохранять свою функциональность и надежность. Она может продолжать точно обрабатывать и анализировать аудио и изображения, основываясь на надежных данных и получая достоверные результаты. Это позволяет системе ИИ успешно выполнять свои задачи и быть полезной в различных сферах применения, включая медицину, автономные системы и мультимедийные приложения.

Таким образом, применение данных рисков и метода SGEC способствует нормальному функционированию системы ИИ, обеспечивая ее защиту от возможных атак, сохранение безопасности данных и достоверность результатов. Это важные аспекты для дальнейшего развития и успешного применения моделей ИИ на основе аудио и изображений.

Использование метода SGEC. Защита от состязательных атак на аудио и изображения является критически важным аспектом в области искусственного интеллекта (ИИ). Состязательные атаки – это попытки искажения, изменения или подмены данных в целях обмана системы ИИ и получения нежелательных результатов. Эти атаки могут иметь серьезные последствия, включая неверные диагнозы в медицинской области, ошибочные решения в автономных системах или фальсификацию мультимедийного контента.

Если акцентировать внимание на аудио и изображениях, то состязательные атаки могут применяться для изменения звуковых сигналов, таких как речь [11], или частичное (полное) искажение изображения. Например, злоумышленники создают аудио-сигналы, которые похожи на белый шум, позволяющий обмануть систему распознавания речи и дестабилизировать работу системы. Также атаки на изображения могут включать добавление или удаление определенных

элементов внутри определенного кадра. Подобные манипуляции над медиа-контентом может привести к ошибочным результатам анализа или распознавания объектов.

В итоге, подобные действия злоумышленников, которые используют состязательные атаки на системы ИИ не вызывает доверия и надёжность у пользователей — система может давать некорректные (неверные) результаты. Поэтому необходимо принимать меры для защиты моделей ИИ от таких атак.

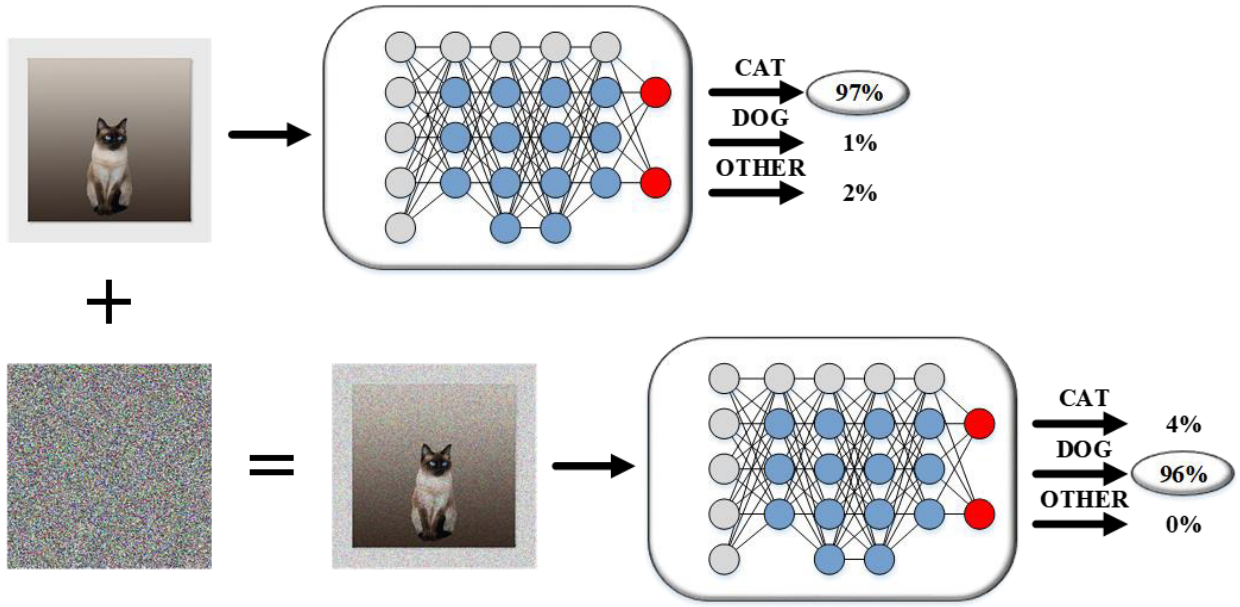

Рис. 2. Возможная атака на ИИ, путём зашумления данных

Fig. 2. Potential attack on AI through data perturbation

Метод SGEC представляет собой один из подходов, предлагающих решение для защиты данных аудио и изображений в системах ИИ. Этот метод основывается на шифровании данных путем генерации криптографических ключей и использования их для защиты и аутентификации данных. SGEC позволяет предотвратить несанкционированный доступ к данным и обеспечить их конфиденциальность и целостность.

Применение метода SGEC позволяет системе ИИ нормально функционировать, так как это обеспечивает ее защиту от возможных состязательных атак и гарантирует правильность и достоверность результатов. Защищенные данные позволяют системе ИИ принимать обоснованные решения на основе надежных и целостных аудио и изображений.

Однако следует отметить, что защита от состязательных атак – это непрерывный процесс, поскольку злоумышленники постоянно разрабатывают новые методы и алгоритмы для обмана систем ИИ. Поэтому необходимо постоянно совершенствовать методы защиты и проводить регулярные аудиты системы для выявления потенциальных уязвимостей и улучшения механизмов защиты.

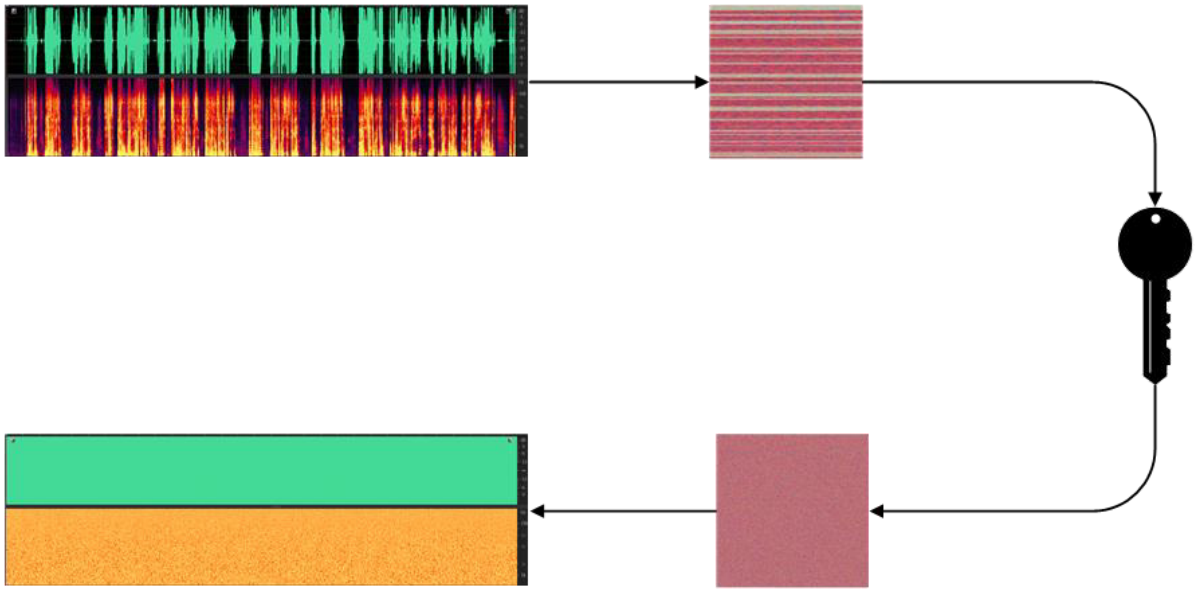

Рис. 3. Пример защиты голосовых данных

Fig. 3. Example of voice data protection

В целом, защита от состязательных атак на аудио и изображения в моделях ИИ является важной задачей для обеспечения надежности и безопасности систем ИИ. Применение метода SGEC и других подобных подходов способствует нормальному функционированию системы, обеспечивая ее защиту от атак, сохранение конфиденциальности и правильность обработки аудио и изображений.

ЗАКЛЮЧЕНИЕ

В заключении можно подчеркнуть важность защиты от состязательных атак на аудио и изображения в моделях искусственного интеллекта. Эти атаки могут серьезно подорвать надежность и безопасность систем ИИ, особенно тех, которые работают с конфиденциальными данными пользователей.

Применение метода SGEC, предложенного в данной статье, представляет эффективный подход к защите от таких атак. Шифрование данных и обеспечение их целостности помогают предотвратить изменение искажения данных, а также предотвращают несанкционированный доступ к информации.

Однако следует отметить, что разработка методов защиты информационной безопасности является непрерывным процессом. С появлением новых атак и методов обхода необходимо продолжать исследования и развивать новые подходы для обеспечения безопасности систем ИИ.

В целом, применение метода SGEC и других мер безопасности поможет снизить риски состязательных атак на аудио и изображения в моделях искусственного интеллекта. Это способствует созданию более надежных и безопасных систем ИИ, которые могут быть успешно применены в различных областях, сохраняя конфиденциальность и целостность данных.

Благодарность. Работа выполнена в рамках Соглашения от 30.06.2022 г. № 40469-21/2022-к.

Список литературы

1. Esmaeilpour M., Cardinal P., Koerich A. L. A robust approach for securing audio classification against adversarial attacks //IEEE transactions on information forensics and security. – 2019. – Т. 15. – С. 2147-2159.

2. Xu H. et al. Adversarial attacks and defenses in images, graphs and text: A review //International Journal of Automation and Computing. – 2020. – Т. 17. – С. 151-178.

3. Свидетельство о государственной регистрации программы для ЭВМ № 2022663168 Российская Федерация. SGEC-система "BIOM" для шифрования и сокрытия голосовых данных пользователей на сервере: № 2022662279: заявл. 27.06.2022: опубл. 12.07.2022 / В. М. Герасимов, М. А. Маслова; заявитель Федеральное государственное автономное образовательное учреждение высшего образования «Севастопольский государственный университет». – EDN FJQWGB.

4. Clark D., Hunt S., Malacaria P. Quantitative analysis of the leakage of confidential data // Electronic Notes in Theoretical Computer Science. – 2002. – Т. 59. – №. 3. – С. 238-251.

5. Martin K. The penalty for privacy violations: How privacy violations impact trust online // Journal of Business Research. – 2018. – Т. 82. – С. 103-116.

6. Yang J. et al. Msta-net: forgery detection by generating manipulation trace based on multi-scale self-texture attention //IEEE transactions on circuits and systems for video technology. – 2021. – Т. 32. – №. 7. – С. 4854-4866.

7. Li G. et al. DeSVig: Decentralized swift vigilance against adversarial attacks in industrial artificial intelligence systems //IEEE Transactions on Industrial Informatics. – 2019. – Т. 16. – №. 5. – С. 3267-3277.

8. Meeßen S. M. et al. Trust is essential: positive effects of information systems on users' memory require trust in the system //Ergonomics. – 2020. – Т. 63. – №. 7. – С. 909-926.

9. Lupton M. Some ethical and legal consequences of the application of artificial intelligence in the field of medicine //Trends Med. – 2018. – Т. 18. – №. 4. – С. 100147.

10. Герасимов, В. М. Комплексная система защиты биометрического голосового отпечатка от воздействия кибермошенников / В. М. Герасимов // XI Конгресс молодых учёных: Сборник научных трудов, Санкт-Петербург, 04–08 апреля 2022 года. – Санкт-Петербург: федеральное государственное автономное образовательное учреждение высшего образования "Национальный исследовательский университет ИТМО", 2022. – С. 72-76. – EDN VTVBBS.

11. Герасимов, В. М. Возможные угрозы и атаки на систему голосовой идентификации пользователя / В. М. Герасимов, М. А. Маслова // Научный результат. Информационные технологии. – 2022. – Т. 7, № 1. – С. 32-37. – DOI 10.18413/2518-1092-2022-7-1-0-4. – EDN JBCXMF.

12. Разработка программного модуля системы распознавания лиц с использованием метода Виолы – Джонса / М. И. Ожиганова, С. М. Арванова, А. А. Абитов, И. А. Уначев // Цифровая трансформация науки и образования: Сборник научных трудов II Международной научно-практической конференции, НАЛЬЧИК, 01–04 октября 2021 года. – НАЛЬЧИК, 2021. – С. 271-277. – EDN NRFFLF.

References

1. Esmaeilpour M., Cardinal P., Koerich A.L. A robust approach for securing audio classification against adversarial attacks //IEEE transactions on information forensics and security. – 2019. – T. 15. – P. 2147-2159.

2. Xu H. et al. Adversarial attacks and defenses in images, graphs and text: A review // International Journal of Automation and Computing. – 2020. – T. 17. – P. 151-178.

3. Certificate of state registration of the computer program No. 2022663168 Russian Federation. SGEC-system "BIOM" for encrypting and hiding the voice data of users on the server: No. 2022662279: App. 06/27/2022: publ. July 12, 2022 / V.M. Gerasimov, M.A. Maslova; applicant Federal State Autonomous Educational Institution of Higher Education "Sevastopol State University". – EDN FJQWGB.

4. Clark D., Hunt S., Malacaria P. Quantitative analysis of the leakage of confidential data // Electronic Notes in Theoretical Computer Science. – 2002. – T. 59. – №. 3. – P. 238-251.

5. Martin K. The penalty for privacy violations: How privacy violations impact trust online // Journal of Business Research. – 2018. – T. 82. – P. 103-116.

6. Yang J. et al. Msta-net: forgery detection by generating manipulation trace based on multi-scale self-texture attention // IEEE transactions on circuits and systems for video technology. – 2021. – T. 32. – №. 7. – P. 4854-4866.

7. Li G. et al. DeSVig: Decentralized swift vigilance against adversarial attacks in industrial artificial intelligence systems //IEEE Transactions on Industrial Informatics. – 2019. – T. 16. – №. 5. – P. 3267-3277.

8. Meeßen S. M. et al. Trust is essential: positive effects of information systems on users' memory require trust in the system //Ergonomics. – 2020. – T. 63. – №. 7. – P. 909-926.

9. Lupton M. Some ethical and legal consequences of the application of artificial intelligence in the field of medicine //Trends Med. – 2018. – T. 18. – №. 4. – P. 100147.

10. Gerasimov, V. M. Comprehensive system for protecting a biometric voice print from the effects of cyber fraudsters / V.M. Gerasimov // XI Congress of Young Scientists: Collection of scientific papers, St. Petersburg,

April 04–08, 2022. - St. Petersburg: Federal State Autonomous Educational Institution of Higher Education "National Research University ITMO", 2022. – P. 72-76. – EDN VTVBBS.

11. Gerasimov, V.M. Possible threats and attacks on the user's voice identification system / V.M. Gerasimov, M.A. Maslova // Scientific result. Information Technology. – 2022. – V. 7, No. 1. – P. 32-37. – DOI 10.18413/2518-1092-2022-7-1-0-4. – EDN JBCXMF.

12. Ozhiganova M. I., Arvanova S. M., Abitov A. A., Unachev I. A. Development of a software module for a face recognition system using the Viola-Jones method // Digital transformation of science and education: Collection of scientific papers II International Scientific and Practical Conference, NALCHIK, October 01–04, 2021. - NALCHIK, 2021. – P. 271-277. – EDN NRFFLF.

Герасимов Виктор Михайлович, инженер, студент первого курса магистратуры направления «Безопасность систем искусственного интеллекта» факультета Безопасности Информационных Технологий (БИТ)
Маслова Мария Александровна, старший преподаватель кафедры Информационная безопасность Института информационных технологий, аспирант, младший научный сотрудник Ростовского государственного экономического университета (РИНХ)
Халилаева Эмине Илимдаровна, студент первого курса магистратуры кафедры «Информационная безопасность» Института информационных технологий

Gerasimov Viktor Mikhailovich, engineer, a first-year master's student in the field of "Security of Artificial Intelligence Systems" faculty of the Security Information Technology (SIT)
Maslova Maria Alexandrovna, Senior Lecturer of the Department Information security Institute of Information Technologies, postgraduate student, junior researcher Rostov State Economic University (RINH)
Khalilayeva Emine Ilimdarovna, a first-year master's student in the field of the Department «Information security», Institute of Information Technology