ДВОРЕЦКАЯ П. С., БАЗАРОВА И. А.
МЕТОДЫ И СРЕДСТВА ОБНАРУЖЕНИЯ СИНТЕТИЧЕСКОГО КОНТЕНТА, ГЕНЕРИРУЕМОГО НЕЙРОННЫМИ СЕТЯМИ

УДК 004.896:004.93'1, ГРНТИ 28.23.15

Методы и средства обнаружения синтетического контента, генерируемого нейронными сетями

Methods and tools for detecting synthetic content generated by neural networks

П. С. Дворецкая[1], И. А. Базарова[2]

[1]ООО «ГазИнформСервис», филиал в городе Ухта,
[2] Ухтинский государственный технический университет, г. Ухта

P. S. Dvoretskaya[1], I. A. Bazarova[2]

[1]LLC "GazInformService", branch in the city of Ukhta;
[2] Ukhta State Technical University, Ukhta

В статье представлен обзор современных инструментов и методов обнаружения синтетического медиаконтента (дипфейков), проведен анализ рынка готовых решений для детектирования, сформулированы ключевые проблемы доступности и практического применения таких средств.

This article presents an overview of modern tools and methods for detecting synthetic media content (deepfakes), analyzes the market for ready-made detection solutions, and identifies the main issues with the availability and practical application of such tools.

Ключевые слова: *синтетический медиаконтент, дипфейк, обнаружение дипфейков, информационная безопасность*

Keywords: *synthetic media content, deepfake, deepfake detection, information security.*

Введение

Эпоха цифровых технологий вступила в фазу, когда увиденное на видео или изображениях больше не может считаться безоговорочной правдой. Технологии глубокого обучения позволили создавать гиперреалистичные поддельные видео с изображением человека, называемые дипфейками, которые представляют прямую угрозу безопасности личности, стабильности финансовых систем и общественно-политическому порядку.

Спектр угроз, связанных с использованием синтетического контента, постепенно растет. Если в 2024 году, согласно отчету Sensity AI [1], атаки были связаны с финансовым сектором, обходом систем биометрии и влиянием на общественное мнение, а основными целями являлись публичные лица, то к 2025 году по отчету Resemble AI [2] вектор смещается в сторону частных лиц, которые

составили 34% всех жертв, что превышает долю финансового мошенничества (23%) и политических манипуляций (14%).

Причиной столь широкого распространения дипфейков является полная демократизация технологий их создания. Специалисты компании Sensity произвели подсчет инструментов для генерации медиаконтента, находящихся в открытом доступе – публичные репозитории, проекты с открытым исходным кодом, бесплатные и платные инструменты стоимостью до $50 в месяц. В результате исследования на начало 2024 года было выявлено порядка 13 тысяч инструментов для генерации изображений, замены лиц, создания цифровых аватаров и поддельных аудио. Уже по данным на 2025 год от MWS AI (входит в MTC Web Services, ранее MTS AI), в интернете доступно уже около 50 тыс. бесплатных инструментов для создания дипфейков [3].

Согласно исследованию «Testing Human Ability To Detect “Deepfake” Images of Human Faces» [4], люди плохо распознают дипфейки. По результатам эксперимента, в котором приняло участие 280 респондентов, средняя точность отличия человеком реальных изображений от сгенерированных составила 62%. Данный результат выше случайного угадывания (50%), но является недостаточным для надежной защиты от угроз в реальной жизни.

В связи с этим, разработка и совершенствование автоматизированных методов обнаружения синтетического контента становятся критически важной задачей в области компьютерных наук и информационной безопасности.

Целью данного исследования является проведение системного анализа современных средств обнаружения синтетического визуального контента, генерируемого нейронными сетями. Основной упор в работе делается на анализ методов детектирования видеоконтента. Данный выбор обусловлен высокой убедительностью и потенциалом воздействия, а также доминированием данного формата в кибератаках.

Классификация и технологии создания синтетического видеоконтента

В современной научной литературе синтетический медиаконтент классифицируются по нескольким ключевым признакам – модальность и тип манипуляции.

По модальности выделают визуальные, включающие из изображения и видео, и аудиальные дипфейки.

Наиболее детально разработана классификация по типу манипуляции, при которой визуальные дипфейки делятся на пять основных категорий [5]:

- замена лиц (Face-Swapping);
- синхронизация губ (Lip-Syncing);
- лицевая анимацию (Facial Reenactment);
- полный синтез лиц (Entire Face Synthesis);
- манипуляция атрибутами лица (Facial Attribute manipulation).

Эволюция методов генерации дипфейк контента прошла несколько этапов, начиная от кропотливой ручной работы заканчивая мощными алгоритмами искусственного интеллекта.

Генеративно-состязательные сети (GAN), представленные Ian J. Goodfellow и его командой в 2014 году [6], стали фундаментом для создания фотореалистичных изображений. Основная идея GAN – использование двух нейронных сетей – генератора, создающего изображение, и дискриминатора, отличающего поддельное изображение от сгенерированного, которые обучаются и одновременно конкурируют друг с другом для получения наилучшего результата. GAN продемонстрировали эффективность в задачах полного синтеза фотореалистичных лиц и манипуляции атрибутами, однако их обучение часто отличается нестабильностью и требует баланса между двумя сетями.

Широкое применение автоэнкодеров (AE) и вариационных автоэнкодеров (VAE) [7] для создания дипфейков началось несколько позже, чем GAN. Классические автоэнкодеры, основанные на энкодер-декодер архитектуре, оказались эффективны для решения задач замены лиц и переноса мимики, в то время как VAE позволяет осуществлять контролируемые манипуляции с атрибутами лица и даже синтезировать новые изображения.

Наиболее современным и актуальным подходом к созданию дипфейков является применение диффузионных моделей, использующих процесс итеративного зашумления и восстановления данных [8]. Диффузионные модели, такие как Sora и Stable Diffusion, на данный момент позволяют достигать высокого качества фотореалистичных изображений.

Эволюция методов генерации синтетического контента, выражающаяся в уменьшении визуально заметных артефактов, подтверждает необходимость разработки инструментов обнаружения дипфейков.

Описание методологии обнаружения дипфейков

Наравне с активным развитием технологии генерации разрабатываются методы обнаружения дипфейков.

Все многообразие современных методов детектирования можно систематизировать по ключевым признакам, таким как тип анализируемых артефактов и используемая архитектура детектора.

С точки зрения обнаруживаемых артефактов, подходы делятся на четыре основные группы [9]:

- Детекторы пространственных артефактов, выполняющие обнаружение аномалий в пределах одного кадра (несовершенство смешивания исходного и подставляемого изображения, неестественные тени и блики).
- Детекторы временных артефактов, выявляющие несоответствия между последовательными кадрами видео (дрожание, мерцание, нарушение плавности и реалистичности движений).
- Детекторы частотных артефактов, выполняющие поиск аномалий, невидимых человеческим глазом и оставляемых генеративными моделями в спектральной области изображения.
- Анализ специальных артефактов нацелен на специфические "отпечатки" методов генерации, такие как рассогласование аудио и видео (несоответствие движения губ и речи) или уникальные шумовые паттерны (PRNU).

С архитектурной точки зрения в основе детекторов лежат:

- Сверточные нейронные сети (CNN), такие как XceptionNet и EfficientNet, для анализа статических изображений.
- Рекуррентные сети (RNN, LSTM) и трансформеры (Vision Transformer или Transformer Encoder), предназначенные для обработки видеопоследовательностей и выявления временных несоответствий.
- Специализированные архитектуры, выходя за рамки классических CNN, предлагают иные способы обработки представления данных для выявления более сложных артефактов. Так Capsule Network анализирует пространственные взаимоотношения между частями лица, выявляя анатомические несоответствия. U-Net решает задачу определения поддельных областей через создание пиксельных масок. Графовые сети (GNN) моделируют сложные связи между регионами лица, позволяя выявить не соответствие движения одной области лица (например, рта во время улыбки) с ожидаемой реакцией других областей (щек, глаз).

Наиболее эффективными на сегодняшний день являются гибридные подходы, комбинирующие анализ пространственных и временных артефактов с использованием современных архитектур (например, Vision Transformer), что позволяет значительно повысить устойчивость детекторов к новым, неизвестным ранее типам дипфейков.

Описанные методологические подходы находят свое практическое применение также в виде готовых программных решений и сервисов, которые будут рассмотрены в следующем разделе.

Анализ рынка готовых решений для детектирования дипфейков

Как упоминалось ранее, современный рынок информационных технологий характеризуется наличием обширного и разнообразного инструментария для генерации синтетического медиаконтента, представленного как открытыми программными решениями с низким порогом входа, так и коммерциализированными сервисами.

Данная ситуация закономерно актуализирует вопрос о существовании сопоставимого по масштабам и доступности рынка решений для детектирования дипфейк контента.

Проведенный анализ рынка готовых решений для распознавания синтетического медиаконтента основывался на следующих критериях, позволяющих оценить их практическую применимость в исследовательской деятельности и повседневности:

- поддерживаемые типы контента;
- стоимость и условия использования;
- наличие и условия тестового доступа.

В финальную выборку (Таблицы 1) вошли только решения, работоспособность которых можно было подтвердить через официальные каналы, такие как актуальные ресурсы разработчиков, сайты проектов.

В анализ включались исключительно решения, предназначенные для распознавания дипфейков в видео-контенте, а также универсальные системы, поддерживающие анализ видео наряду с другими типами данных.

Таблица 1. Инструменты детектирования дипфейков и их характеристики

Инструмент	**Тип контента**	**Мин. стоимость**	**Способ доступа**	**Публичный демонстра-ционный доступ**	**Условия демонстрацион ного доступа**
Attestiv	Текст, видео, изображения	10$/месяц	Веб-сервис, API	+	5 проверок только видео в месяц, только через веб-сайт
Reality Defender	Аудио, видео, изображения	399$/месяц	API, SDK	+	50 проверок аудио и изображений в месяц, только через API
Sensity	Аудио, видео, изображения	Индивидуаль-ный расчет	Веб-сервис, API, SDK	-	Только для коммерческого использования
DuckDuck Goose.AI	Аудио, видео, изображения	Индивидуаль-ный расчет	Веб-сервис, API	-	Только для коммерческого использования
HIVE Moderation	Аудио, видео, изображения	Индивидуаль-ный расчет	Веб-сервис, API,	+	Без ограничений, только через Веб-сервис
Vision Labs DeepFake detection	Аудио, видео, изображения	Индивидуаль ный расчет	Модуль к desktop-приложени ю Luna Pass	-	Только для коммерческого использования
Ai or not	Аудио, видео, изображения	15$/месяц	Веб-сервис, API	+	10 проверок изображений и аудио через веб-сайт или API
Deepfake-o-Meter	Аудио, видео, изображения	Бесплатно	Веб-сервис, исходный код на GitHub	+	Без ограничений, через Веб-сервис и исходный код
Deepware Scanner	Видео	Бесплатно	Веб-сервис, API	+	Без ограничений, через Веб-сервис и API

На основании собранных данных инструменты по детектированию можно разделить на две категории:

- коммерческие SaaS-решения, предоставляющие услуги в формате подписки и являющиеся готовыми к использованию и интеграции в бизнес-процессы платформами;
- академические работы – сервисы, разработанные в исследовательских лабораториях, направленные на достижении прорыва в точности детектирования и публикации результатов в открытом доступе.

На рынке наблюдается выраженная коммерциализация и закрытость решений. Большинство высокоточных систем распознавания, таких как Sensity, DuckDuckGoose.AI, HIVE и Vision Labs, существуют в формате коммерческих SaaS-платформ или являются разработками, доступными исключительно для корпоративных клиентов и правоохранительных органов.

О данном аспекте свидетельствует непрозрачное ценообразование, где стоимость услуг раскрывается только после персональных консультаций (Рисунок 1), а также ограничение доступа к функционалу, включая использование форм обратной связи и регистрацию в системе, для пользователей, использующих бесплатные домены электронной почты (Рисунок 2). Данная мера, вероятно, служит защитой от спама, но на практике создаёт ограничения для широкого круга потенциальных пользователей, таких как независимые исследователи, представители малого бизнеса и частные лица.

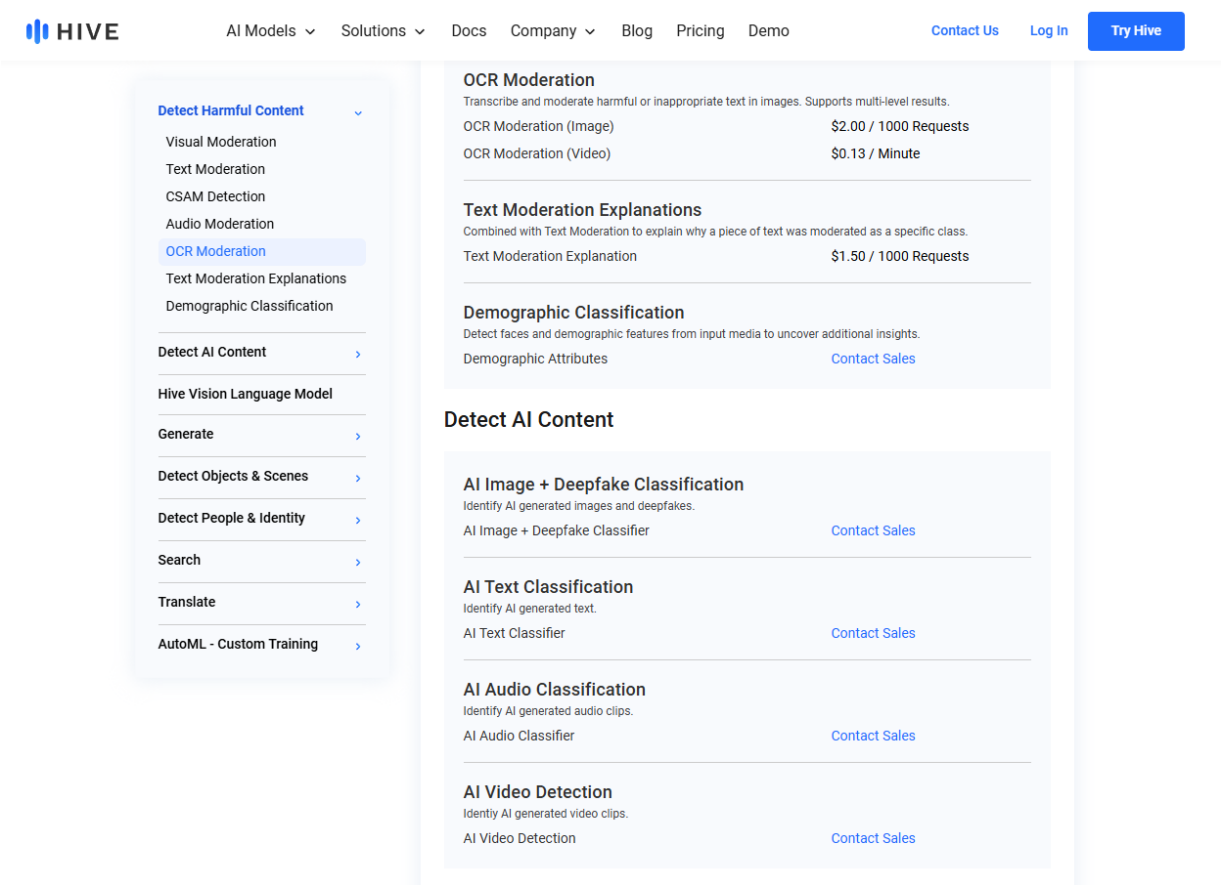

Рисунок 1. HIVE MODELS ограниченный доступ к ценообразованию функций

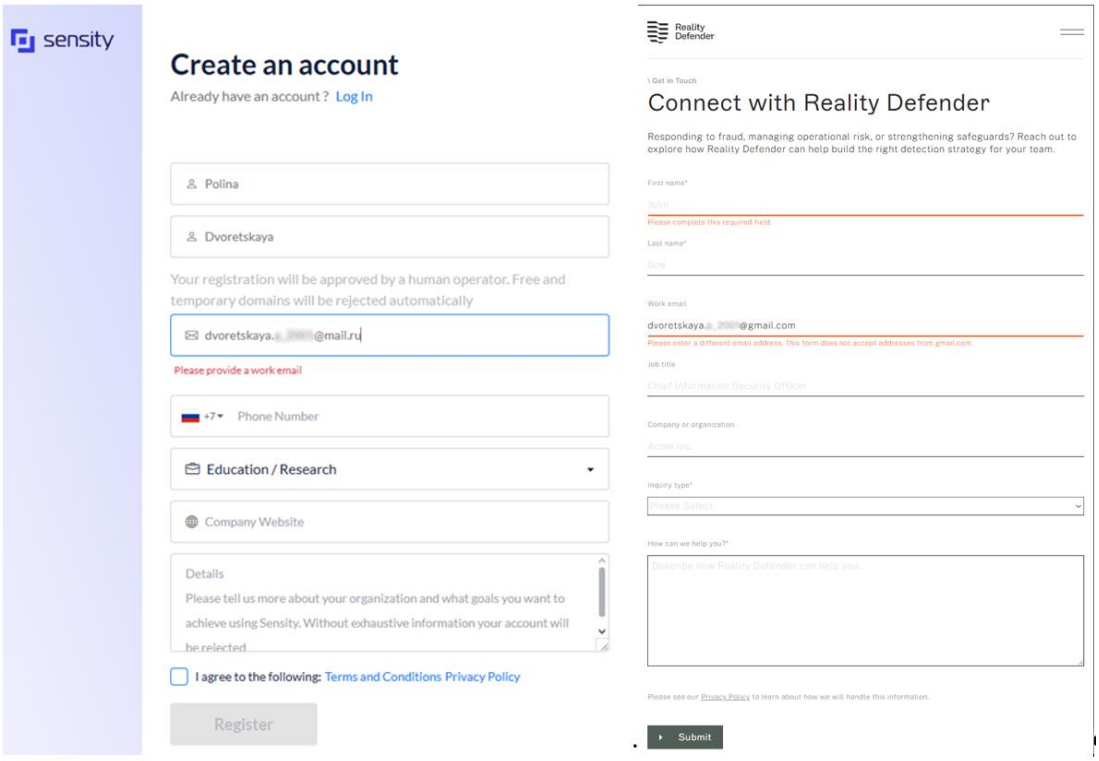

Рисунок 2. Реализация фильтра корпоративных клиентов через проверку email-адресов

Еще одной ключевой проблемой является ограниченная практическая ценность демонстрационного доступа сервисов. Бесплатные тарифные планы оказываются настолько ограничены функционально, что не способны реализовать свою ключевую задачу – предоставить пользователю полноценный опыт взаимодействия с сервисом.

К примеру, Attestiv предлагает лишь 5 проверок видео в месяц, что недостаточно даже для ознакомительного тестирования. В свою очередь ресурс Ai or not в рамках пробного периода ограничивает как количество проверок и тип проверяемого контента, так и предоставляет только бинарную оценку (дипфейк или нет), что снижает информативность результата.

Стоит отметить, что подавляющее большинство инструментов детектирования являются зарубежными решениями. Учитывая тот факт, что международные платежные системы приостановили свою работу в России, приобретение подписок и оплата услуг зарубежных продуктов становится существенным ограничением применимости рассматриваемых ресурсов.

Заключение

Проведенный анализ рынка готовых инструментов детектирования дипфейков выявил ряд ограничений для их применения как в научно-исследовательской деятельности, так и для рядовых пользователей. Особенно актуален данный вопрос стоит для российской аудитории на фоне ограничения международных платежных систем.

При этом наблюдается значительный дисбаланс между растущим количеством инструментов для создания и распознавания синтетического

контента. В данных условиях разработка открытого, некоммерческого инструментария для детектирования поддельных медиа становится одной из важных задач в области информационной безопасности.

Положительной динамикой является то, что крупные компании осознают масштаб угрозы и начинают активно инвестировать в данное направление. Ярким примером служит компания РВБ (объединенная компания Wildberries & Russ), которая уже опубликовала бета-версию своего онлайн-детектора, способного на данном этапе анализировать изображения на предмет синтетического происхождения.

Перспективным направлением является создание комплексных платформ для мультимодального анализа (видео, аудио, текст) и интеграция подобных систем распознавания в соцсети и новостные порталы для блокировки поддельного контента в реальном времени.

Список использованных источников и литературы:

1. Статистика фонда «Общественное мнение» в области Интернет-ресурсов [Электронный ресурс]: Фонд «Общественное мнение» – URL: http://bd.fom.ru/report/cat/smi/smi_int (дата обращения: 05.02.2012).

2. Состояние дипфейков 2024 [Электронный ресурс]: Sensity AI – URL: https://sensity.ai/reports (дата обращения: 13.10.2025).

3. Отчет об инцидентах с Deepfake за первый квартал 2025 г.: Картирование инцидентов с Deepfake [Электронный ресурс]: Resemble AI – URL: https://www.resemble.ai/wp-content/uploads/2025/04/ResembleAI-Q1-Deepfake-Threats.pdf (дата обращения: 13.10.2025).

4. Крылова Е. На одно лицо: в Сети стремительно выросло число программ по созданию дипфейков / Е. Крылова – Текст: электронный // Известия : интернет-портал. – URL: https://iz.ru/1893474/elizaveta-krylova/na-odno-lico-v-seti-stremitelno-vyroslo-chislo-programm-po-sozdaniyu-dipfejkov (дата обращения: 13.10.2025).

5. Брэй С. Д., Джонсон С. Д., Клейнберг Б. Проверка способности человека обнаруживать «дипфейковые» изображения человеческих лиц [Электронный ресурс] // Журнал кибербезопасности. 2023. Том. 9, вып. 1. П. тяд011. – URL: https://doi.org/10.1093/cybsec/tyad011 (дата обращения: 20.10.2025).

6. Мирский Ю., Ли В. Создание и обнаружение дипфейков: опрос [Электронный ресурс] // ACM Computing Surveys. 2021. Том. 54, № 1. С. 1–41. – URL: https://doi.org/10.1145/3425780. (дата обращения: 20.10.2025).

7. Гудфеллоу И. Дж., Пуже-Абади Дж., Мирза М., Сюй Б. и др. Генеративно-состязательные сети [Электронный ресурс] // Достижения в нейронных системах обработки информации 27 (НИПС 2014) / под ред. З. Гахрамани, М. Веллинг, К. Кортес, Н. Лоуренс, К. К. Вайнбергер. – 2014. – С. 2672–2680. – URL: https://doi.org/10.1145/3422622 (дата обращения: 20.10.2025).

8. Балди П. Автоэнкодеры, обучение без учителя и глубокие архитектуры [Электронный ресурс] // Материалы 29-й Международной конференции по машинному обучению. Мастер-классы. – Эдинбург, Шотландия, 2012 г. – URL: https://proceedings.mlr.press/v27/baldi12a.html. (дата обращения: 20.10.2025).

9. Хо Дж., Джайн А., Аббил П. Вероятностно-диффузионные модели шумоподавления [Электронный ресурс] // Достижения в области нейронных систем обработки информации. 2020. Том. 33. С. 6840–6851. – URL: https://www.semanticscholar.org/paper/Denoising-Diffusion-Probabilistic-Models-Ho-Jain/5c126ae3421f05768d8edd97ecd44b1364e2c99a (дата обращения: 25.10.2025).

10. Ле Б.М. и др. SoK: Систематизация и бенчмаркинг детекторов Deepfake в единой системе [Электронный ресурс] // 10-й Европейский симпозиум IEEE по безопасности и конфиденциальности, 2025 г. IEEE, 2025. С. 883-902. – URL: https://ieeexplore.ieee.org/document/11129367 (дата обращения: 25.10.2025).

List of references

1. Statistics of the Public Opinion Foundation in the field of Internet resources [Electronic resource]: Public Opinion Foundation. – URL: http://bd.fom.ru/report/cat/smi/smi_int (access date: 02/05/2012).

2. State of deepfakes 2024 [Electronic resource]: Sensity AI. – URL: https://sensity.ai/reports (access date: 10/13/2025).

3. Deepfake Incident Report for the first quarter of 2025: Mapping Deepfake Incidents [Electronic resource]. – URL: https://www.resemble.ai/wp-content/uploads/2025/04/ResembleAI-Q1-Deepfake-Threats.pdf (access date: 10/13/2025).

4. Krylova E. On the same page: the number of programs for creating deepfakes on the Internet has rapidly grown / E. Krylova – Text: electronic // Izvestia: Internet portal. – URL: https://iz.ru/1893474/elizaveta-krylova/na-odno-lico-v-seti-stremitelno–vyroslo-chislo-programm-po-sozdaniyu-dipfejkov (date of access: 10/13/2025).

5. Bray S. D., Johnson S. D., Kleinberg B. Testing a person's ability to detect "deepfake" images of human faces [Electronic resource] // Journal of Cyber Security. 2023. Vol. 9, no. 1. P. tjad011. – URL: https://doi.org/10.1093/cybsec/tyad011 (access date: 10/20/2025).

6. Mirsky Yu., Li V. Creation and detection of deepfakes: a survey [Electronic resource] // ACM Computing Surveys. 2021. Vol. 54, no. 1. pp. 1–41. – URL: https://doi.org/10.1145/3425780. (access date: 10/20/2025).

7. Goodfellow I. J., Pouget-Abadie J., Mirza M., Xu B. et al. Generative adversarial networks [Electronic resource] // Advances in neural information processing systems 27 (NIPS 2014) / ed. Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, K. K. Weinberger. – 2014. – P. 2672–2680. – URL: https://doi.org/10.1145/3422622 (access date: 10/20/2025).

8. Baldi P. Autoencoders, unsupervised learning and deep architectures [Electronic resource] // Proceedings of the 29th International Conference on Machine Learning (ICML). Master classes. – Edinburgh, Scotland, 2012 – URL: https://proceedings.mlr.press/v27/baldi12a.html. (access date: 10/20/2025).

9. Ho J., Jain A., Abbil P. Probabilistic diffusion models of noise reduction [Electronic resource] // Advances in the field of neural information processing systems. 2020. – URL: https://www.semanticscholar.org/paper/Denoising-Diffusion-

Probabilistic-Models-Ho-Jain/5c126ae3421f05768d8edd97ecd44b1364e2c99a (access date: 10/25/2025).

10. Le B.M. et al. SoK: Systematization and benchmarking of Deepfake detectors in a single system [Electronic resource] // 10th IEEE European Symposium on Security and Privacy, 2025. IEEE, 2025. pp. 883-902. – URL: https://ieeexplore.ieee.org/document/11129367 (access date: 10/25/2025).