УДК 004.932.2

Теплов Дмитрий Иванович

бакалавр, Уфимский университет науки и технологий, г. Уфа

РАСПОЗНАВАНИЕ СГЕНЕРИРОВАННЫХ ИЗОБРАЖЕНИЙ ЛИЦ

Аннотация. В исследовании рассматривается задача автоматического определения сгенерированных изображений человеческих лиц с использованием свёрточной нейронной сети. Приведён анализ существующих методов обнаружения синтезированных изображений, описана разработанная архитектура нейросети, представлены результаты экспериментального тестирования модели – полученная точность классификации составила 83,3 процента. Работа предназначена для студентов, изучающих методы и средства защиты информации, а также для специалистов в областях информационной безопасности, машинного обучения и искусственных нейронных сетей.

Annotation. This study examines the problem of automatic detection of generated (synthesized) human face images using a convolutional neural network. The paper provides an analysis of existing methods for detecting synthesized images, describes the developed neural network architecture, and presents the results of experimental testing of the model – the achieved classification accuracy was 83.3 percent. The work is intended for students studying information security methods and tools, as well as for specialists in the fields of information security, machine learning, and artificial neural networks.

Ключевые слова: дипфейк, сверточная нейронная сеть, генератор, информационная безопасность, классификация изображений, мошенничество, искусственный интеллект, машинное обучение, искусственные нейронные сети, защита информации.

Keywords: deepfake, convolutional neural network, generator, information security, image classification, fraud, artificial intelligence, machine learning, artificial neural networks, information protection.

Введение: современные генеративные нейросетевые модели позволяют создавать фотореалистичные изображения людей, видеофрагменты с ними, на которых показаны действия, которые они никогда не совершали, или поддельные записи голоса. Такие материалы активно используются злоумышленниками для создания фальшивых аккаунтов, проведения фишинговых атак, манипуляции общественным сознанием или кражи денежных средств. Пользователям с развитием технологий становится все труднее отличать реальное и сгенерированное, поэтому возникает необходимость внедрения автоматических систем анализа, не требующих участия человека.

Формулировка проблемы исследования: Термин «дипфейк» (deepfake) состоит из двух понятий: «глубинное обучение» (deep learning) и «подделка» (fake). Дипфейк – это метод синтеза контента, основанный на машинном обучении и искусственном интеллекте. Нейросетевая модель изменяет исходные изображения, путем наложения случайных фрагментов, либо синтезирует новые [изображения] из уже имеющихся. Таким образом нейросеть может подменять лица, мимику, жесты и голос в видео или звуковой дорожке.

Данная технология может быть использована и для манипуляции общественным сознанием. Так, дипфейки с участием политических деятелей, знаменитых актёров и иных известных персон могут создавать как шутники, так и мошенники. Несанкционированные цифровые копии могут считаться посягательством на использование чьей-то личности или чужого бренда [1**Ошибка! Источник ссылки не найден.**].

Известны случаи, когда мошенники с помощью дипфейка маскировались под родственников или друзей. Пострадавшие переводили деньги злоумышленникам, потому что считали, что их попросили знакомые люди, а на самом же деле голос или изображение были сгенерированы нейросетью [2][3].

Формулировка актуальности исследования: проблема мошенничества с использованием дипфейков нуждается в комплексном решении. Сегодня, когда у абсолютного большинства населения имеется возможность общаться с помощью электронной почты, мессенджеров, аудио и видеозвонков, возникают сложности с определением реальности собеседника, так как это может быть как реальный человек, так и злоумышленник.

Взрослые люди особенно подвержены опасности, так как не обладают актуальными знаниями в областях искусственного интеллекта и машинного обучения и даже не предполагают о том, что взаимодействующий человек может быть не тем, кем является. Этот мошенник, притворяющийся знакомым, может требовать немедленного перевода денежных средств, аргументируя тем, что попал в проблему, нуждающуюся в срочном решении.

В России по данным МВД в 2024 году было совершено 765400 преступлений с использованием цифровых технологий, их число выросло на 13,1% с 2023 года, увеличилась и доля таких преступлений, при этом в 3-5% от числа таких случаев преступники отправляли дипфейки, хотя еще совсем недавно этот показатель не превышал десятых долей процента [4].

Анализ существующих подходов: далее рассмотрим существующие подходы распознавания сгенерированных изображений:

1) Традиционные методы: обнаружение артефактов, возникающих в процессе генерации или редактирования изображений, таких как несоответствия на уровне пикселей, некорректное освещение, неестественные тени или аномалии частотной области.

Методы анализируют статистические аномалии в изображении, например, несоответствия в коэффициентах дискретного косинусного преобразования (DCT), артефакты сжатия «JPEG», или следы операций ресемплинга и клонирования областей. Эти методы не используют машинное обучение, а основаны на ручных правилах и пороговых значениях.

2) Свёрточные нейронные сети: автоматическое извлечение пространственных признаков из изображений для выявления микро-артефактов, незаметных человеческому глазу.

Сверточные нейронные сети (например, предобученные «XceptionNet», «ResNeXt50», «MobileNet», «EfficientNet») обучаются на больших наборах данных, содержащих реальные и синтезированные изображения. Модель учится различать тонкие паттерны, как артефакты смешивания, нереалистичные текстуры кожи, искажения границ лица и другие.

3) Гибридные методы с анализом частотной области: комбинирование пространственных признаков (RGB) с частотными характеристиками изображения, поскольку генеративные модели оставляют различимые следы в высокочастотных компонентах.

Из изображения извлекаются частотные признаки (например, дискретное косинусное преобразование «DCT», дискретное преобразование Фурье «DFT», анализ ошибок сжатия «ELA»). Эти признаки объединяются с исходными RGB-пикселями и подаются на вход «CNN» или «Vision Transformer». Минимальное слияние между пространственной и частотной областями показывает наилучшую обобщающую способность против различных методов генерации.

4) Ансамблевые методы: комбинирование нескольких моделей или классификаторов для повышения точности и робастности детекции за счёт усреднения их результатов.

Несколько различных моделей обучаются независимо. Затем их предсказания объединяются с помощью одного из методов: голосование, усреднение вероятностей или стекинг, когда предсказания базовых моделей подаются на вход мета-классификатору.

5) «Vision Transformers» с вейвлет-кодированием: использование архитектуры «Transformer» (аналогичной той, что работает в «NLP») для захвата глобальных зависимостей в изображении, в комбинации с частотным разложением через вейвлет-преобразование.

Изображение разбивается на патчи, как в классическом «ViT». Вместо стандартных позиционных кодирований применяется двухуровневое дискретное вейвлет-преобразование (DWT). Это позволяет модели одновременно учитывать высокочастотные детали и глобальную структуру. Механизм перекрёстного внимания объединяет оба представления.

Ниже приведено сравнение рассмотренных подходов (таблица 1):

Таблица 1 – Сравнение подходов

Номер	**Название**	**Точность (%)**	**Ссылка**
1	Традиционные методы	60 – 75	[5]
2	Сверточные нейронные сети	60 – 95	[6]
3	Гибридные методы с анализом частотной области	90 – 95	[7]
4	Ансамблевые методы	95	[8]
5	«Vision Transformers» с вейвлет-кодированием	98 – 99	[9]

Цель и задачи исследования: целью работы является разработка системы автоматического определения сгенерированных (синтезированных) изображений человеческих лиц, позволяющая снять с человека обязанность по визуальной верификации подлинности собеседника. В качестве основного инструмента такой системы предлагается использование свёрточной нейронной сети (CNN), обученной на размеченном наборе данных [**Ошибка! Источник ссылки не найден.**], содержащем реальные и дипфейк-изображения.

Для достижения поставленной цели необходимо решить следующие задачи:

1) Провести анализ предметной области:

1.1) сформулировать проблему исследования;

1.2) сформулировать актуальность исследования;

1.3) выполнить анализ существующих подходов.

2) Построить нейросетевую модель:

2.1) подобрать данные для обучения модели;

2.2) выбрать библиотеки для построения модели;

2.3) подготовить данные для обучения;

2.4) построить архитектуру модели;

2.5) обучить модель;

2.6) протестировать модель.

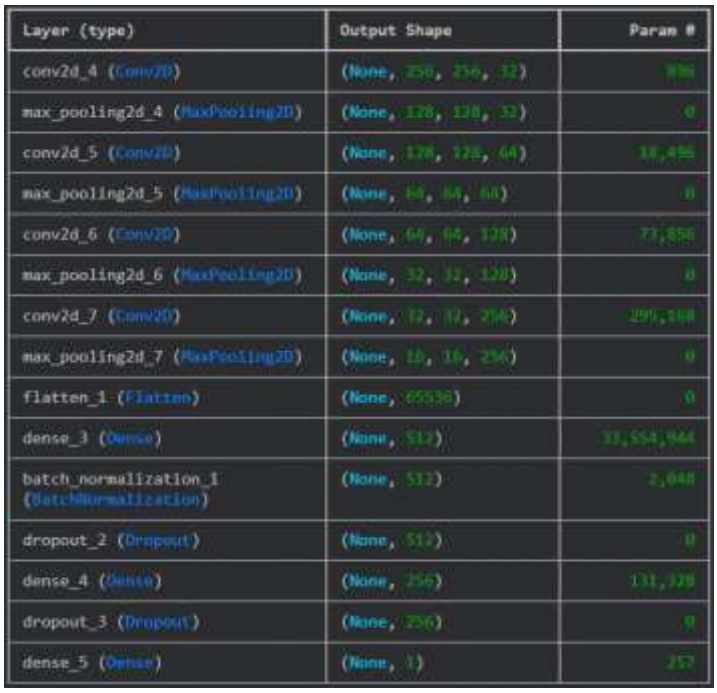
3) Осуществить анализ результатов:

3.1) оценить качество модели;

3.2) определить последствия внедрения предложенного решения для задач информационной безопасности.

Содержательное изложение подхода:

1) Структурная схема решения (рисунок 1):



Layer (type)	Output Shape	Param #
conv2d_4 (Conv2D)	(None, 256, 256, 32)	896
max_pooling2d_4 (MaxPooling2D)	(None, 128, 128, 32)	0
conv2d_5 (Conv2D)	(None, 128, 128, 64)	18,496
max_pooling2d_5 (MaxPooling2D)	(None, 64, 64, 64)	0
conv2d_6 (Conv2D)	(None, 64, 64, 128)	73,856
max_pooling2d_6 (MaxPooling2D)	(None, 32, 32, 128)	0
conv2d_7 (Conv2D)	(None, 32, 32, 256)	295,168
max_pooling2d_7 (MaxPooling2D)	(None, 16, 16, 256)	0
flatten_1 (Flatten)	(None, 65536)	0
dense_3 (Dense)	(None, 512)	33,554,944
batch_normalization_1 (BatchNormalization)	(None, 512)	2,048
dropout_2 (Dropout)	(None, 512)	0
dense_4 (Dense)	(None, 256)	131,328
dropout_3 (Dropout)	(None, 256)	0
dense_5 (Dense)	(None, 1)	257

Рисунок 1 – Структурная схема решения

2) Функциональная схема решения:

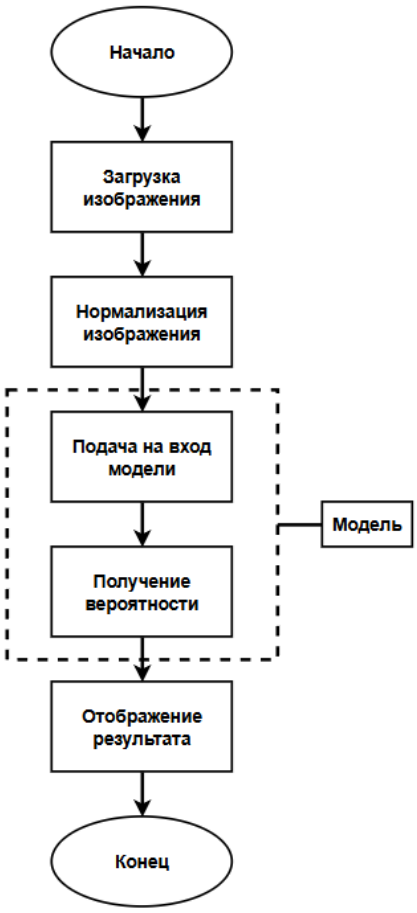


Рисунок 2 – Функциональная схема решения

3) Описание модели:

3.1) входной слой – для подачи изображения в модель;

3.2) 4 сверточных слоя с операцией «MaxPooling» – для выделения ключевых признаков;

3.3) слой «flatten» – который переведет вектор изображения в одномерный;

3.4) 2 полносвязных слоя – для обработки признаков одномерного вектора;

3.5) выходной слой с операцией «sigmoid» – для формирования ответа.

4) Описание алгоритмов:

4.1) подготовка данных – с использованием «ImageDataGenerator» осуществляется загрузка изображений из директорий (Train, Validation, Test) с автоматической нормализацией и

пакетной обработкой, что позволяет оптимально расходовать оперативную память;

4.2) обучение – реализуется методом обратного распространения ошибки. На каждой эпохе модель последовательно обрабатывает пакеты изображений, вычисляет значение функции потерь, затем с помощью оптимизатора корректирует веса модели для минимизации ошибки. Процесс повторяется в течение 10 эпох;

4.3) тестирование – после завершения обучения модель фиксирует веса и выполняет классификацию изображений из тестовой выборки без дальнейшего изменения параметров. Результаты тестирования позволяют оценить обобщающую способность модели на ранее не встречавшихся данных.

5) Описание методов:

5.1) «Adam» – адаптивный оптимизатор, который вычисляет индивидуальную скорость обучения для каждого параметра модели на основе оценок первого и второго моментов градиента;

5.2) «binary_crossentropy» – бинарная кросс-энтропия – функция потерь, используемая для задач бинарной классификации; измеряет различие между предсказанным распределением вероятностей и истинной меткой класса;

5.3) «accuracy» – метрика точности, показывающая долю правильно классифицированных примеров от общего числа примеров;

5.4) «precision» – метрика точности, показывающая, какая доля объектов, определённых моделью как «сгенерированные», действительно является «сгенерированными»;

5.5) «recall» – метрика полноты, показывающая, какая доля «сгенерированных» изображений была правильно обнаружена моделью.

6) Описание методик:

6.1) разбиение данных на тренировочную, валидационную и тестовую выборки;

6.2) нормализация данных – все изображения приводятся к единому масштабу путём деления значений пикселей (исходный диапазон [0, 255]) на 255, что преобразует их в диапазон [0, 1]. Такая нормализация ускоряет сходимость градиентного спуска и повышает устойчивость обучения, так как все входные признаки находятся в одном числовом диапазоне

Вычислительный эксперимент: полученные в ходе тестирования метрики (accuracy = 0,833, precision = 0,859, recall = 0,794) свидетельствуют о том, что модель успешно справляется с задачей бинарной классификации. Высокое значение «precision» означает низкий уровень ложных срабатываний, а значение «recall» показывает, что большинство сгенерированных изображений успешно обнаруживаются.

Выводы: в ходе выполнения работы была достигнута поставленная цель – разработана система автоматического определения сгенерированных (синтезированных) изображений человеческих лиц на основе свёрточной нейронной сети. Решены все сформулированные задачи:

1) Проведён анализ предметной области: сформулирована проблема исследования, обоснована актуальность, выполнен анализ существующих подходов к обнаружению дипфейк-изображений.

2) Построена нейросетевая модель: подобран датасет [**Ошибка! Источник ссылки не найден.Ошибка! Источник ссылки не найден.**], выбраны библиотеки, разработана архитектура «CNN», модель обучена и протестирована.

3) Осуществлён анализ результатов: получены метрики качества – модель правильно классифицирует более 83% изображений, при этом точность обнаружения сгенерированных лиц составляет 86%, а полнота – 79%.

Практическая значимость разработанной системы заключается в возможности её интеграции в существующие системы информационной безопасности таких приложений, как мессенджеры, социальные сети и банковские приложения для автоматической проверки подлинности изображений пользователей, что позволяет снизить риск мошенничества с

использованием дипфейк-технологий и снять с пользователя нагрузку по подтверждению личности собеседника.

Направления дальнейшего улучшения модели включают:

- увеличение количества эпох;
- добавление аугментации данных;
- использование предобученных архитектур;
- комбинирование пространственных признаков с частотным анализом.

Список литературы

1. Дипфейк: как распознать и как защититься [Электронный ресурс] // СберКлевер : медиаплатформа ПАО Сбербанк. 2024. URL: https://www.sberbank.ru/ru/person/kibrary/articles/deepfake-kak-raspoznat-i-kak-zashchititsya (дата обращения: 07.08.2026).
2. Grandparents top AI swindlers' list // Ecns.cn: official website of China News Service. 2026. 9 Apr. **URL:** https://www.ecns.cn/china/2026-04-09/detail-ihfcmemi3022930.shtml (дата обращения: 07.08.2026).
3. Zakreski D. 'That was my granddaughter's voice': Senior suspects scammers used AI in Regina, Saskatoon // CBC News. 2025. 12 Dec. **URL:** https://www.cbc.ca/news/canada/saskatoon/grandparent-scam-ai-9.7010483 (дата обращения: 07.08.2026).
4. Загвоздкина К. Эксперты сообщили о резком росте числа мошенничеств с использованием дипфейков [Электронный ресурс] / К. Загвоздкина // Forbes.ru. 2025. 19 июня. URL: https://www.forbes.ru/tekhnologii/539953-eksperty-soobsili-o-rezkom-roste-chisla-moshennichestv-s-ispol-zovaniem-dipfejkov (дата обращения: 07.08.2026).
5. Countering synthetic realities: Advances, challenges, and perspectives in deepfake detection / M. I. Ahmad, M. Jariwal, S. S. Choudhury [Электронный ресурс] // Computer Science Review. 2026. Vol. 60. Art. 100789. **URL:** https://www.sciencedirect.com/science/article/pii/S2772941926000396 (дата обращения: 07.08.2026).

6. Deepfake detection via spatial–temporal deep networks: leveraging CNNs and LSTMs for enhanced accuracy / Y. Jugade, Z. Syed, C. Dcosta [Электронный ресурс] // Discover Applied Sciences. 2026. Vol. 8. Art. 179. **URL:** https://link.springer.com/article/10.1007/s42452-025-08014-w (дата обращения: 07.08.2026).

7. Enhanced Deep Learning DeepFake Detection Integrating Handcrafted Features / A. Hinke-Navarro, M. Nieto-Hidalgo, J. M. Espin, J. E. Tapia // Cognitive Security Technologies (CST) Department of Fraunhofer AISEC. 2025. **URL:** https://deepfake-total.com/related_work/2507.20608 (дата обращения: 07.08.2026).

8. Ensemble techniques for deepfake detection: comparing stacking, weighted voting and averaging approaches [Электронный ресурс] // Nigeria Computer Society: Digital Library. 2026. 23 Jan. **URL:** https://library.ncs.org.ng/download/ensemble-techniques-for-deepfake-detection-comparing-stacking-weighted-voting-and-averaging-approaches/ (дата обращения: 07.08.2026)

9. Cascaded Cross-Attention Vision Transformers with Wavelet-Based Encodings for DeepFake Detection / A. Polamarasetti, S. Ahmad, M. Zaman [Электронный ресурс] // International Journal of Computer Vision. 2026. Vol. 134, no. 5. Art. 221. **URL:** https://link.springer.com/article/10.1007/s11263-026-02806-2 (дата обращения: 07.08.2026).

10. Karki M. Deepfake and real images dataset [Electronic resource] / M. Karki // Kaggle. 2024. **URL:** https://www.kaggle.com/datasets/manjilkarki/deepfake-and-real-images (дата обращения: 07.08.2026)

Resources

1. Deepfake: how to recognize and how to protect yourself [Electronic resource] // SberKlever: media platform of Sberbank PJSC. 2024. URL: https://www.sberbank.ru/ru/person/kibrary/articles/deepfake-kak-raspoznat-i-kak-zashchititsya (accessed: 07.08.2026).

2. Grandparents top AI swindlers' list // Ecns.cn: official website of China News Service. 2026. 9 Apr. URL: https://www.ecns.cn/china/2026-04-09/detail-ihfcmemi3022930.shtml (accessed: 07.08.2026).

3. Zakreski D. 'That was my granddaughter's voice': Senior suspects scammers used AI in Regina, Saskatoon // CBC News. 2025. 12 Dec. URL: https://www.cbc.ca/news/canada/saskatoon/grandparent-scam-ai-9.7010483 (accessed: 07.08.2026).

4. Zagvozdkina K. Experts report a sharp increase in the number of fraud cases using deepfakes [Electronic resource] / K. Zagvozdkina // Forbes.ru. 2025. June 19. URL: https://www.forbes.ru/tekhnologii/539953-eksperty-soobsili-o-rezkom-roste-cisla-mosennicestv-s-ispol-zovaniem-dipfejkov (accessed: 07.08.2026).

5. Countering synthetic realities: Advances, challenges, and perspectives in deepfake detection / M. I. Ahmad, M. Jariwal, S. S. Choudhury [Electronic resource] // Computer Science Review. 2026. Vol. 60. Art. 100789. URL: https://www.sciencedirect.com/science/article/pii/S2772941926000396 (accessed: 07.08.2026).

6. Deepfake detection via spatial–temporal deep networks: leveraging CNNs and LSTMs for enhanced accuracy / Y. Jugade, Z. Syed, C. Dcosta [Electronic resource] // Discover Applied Sciences. 2026. Vol. 8. Art. 179. URL: https://link.springer.com/article/10.1007/s42452-025-08014-w (accessed: 07.08.2026).

7. Enhanced Deep Learning DeepFake Detection Integrating Handcrafted Features / A. Hinke-Navarro, M. Nieto-Hidalgo, J. M. Espin, J. E. Tapia // Cognitive Security Technologies (CST) Department of Fraunhofer AISEC. 2025. URL: https://deepfake-total.com/related_work/2507.20608 (accessed: 07.08.2026).

8. Ensemble techniques for deepfake detection: comparing stacking, weighted voting and averaging approaches [Electronic resource] // Nigeria Computer Society: Digital Library. 2026. 23 Jan. URL: https://library.ncs.org.ng/download/ensemble-techniques-for-deepfake-detection-comparing-stacking-weighted-voting-and-averaging-approaches/ (accessed: 07.08.2026).

9. Cascaded Cross-Attention Vision Transformers with Wavelet-Based Encodings for DeepFake Detection / A. Polamarasetti, S. Ahmad, M. Zaman [Electronic resource] // International Journal of Computer Vision. 2026. Vol. 134, no. 5. Art. 221. URL: https://link.springer.com/article/10.1007/s11263-026-02806-2 (accessed: 07.08.2026).

10. Karki M. Deepfake and real images dataset [Electronic resource] / M. Karki // Kaggle. 2024. URL: https://www.kaggle.com/datasets/manjilkarki/deepfake-and-real-images (accessed: 07.08.2026).