УДК 004.93 *DOI 10.55421/3034-4689_2026_29_5_102*

П. А. Черкасов, Р. М. Хусаинов, Н. Г. Талипов

НЕЙРОСЕТЕВАЯ МОДЕЛЬ РАСПОЗНАВАНИЯ ДИПФЕЙКОВЫХ ИЗОБРАЖЕНИЙ

Ключевые слова: дипфейк, искусственный интеллект, нейросетевая модель, сгенерированное изображение, распознавание дипфейков, нейронная сеть.

В статье рассмотрена задача распознавания дипфейковых изображений с использованием нейросетевой модели. Проводится обзор современных исследований дипфейков на изображениях, выявляются преимущества и недостатки существующих методов распознавания. Проводится анализ существующих наборов данных (датасетов) реальных изображений и дипфейков, обосновывается выбор наиболее пригодного для обучения, валидации и тестирования нейросетевой модели. В результате анализа наборов данных в качестве исходного набора выбран датасет ArtiFact (Real and Fake Image Dataset), содержащий более 2 миллионов изображений, 19 генераторов изображений и 11 источников реальных изображений. Учитывая ограниченные вычислительные ресурсы среды, решено не использовать найденный датасет полностью, а отобрать необходимые изображения для создания собственных датасетов: final (обучающая, валидационная, тестовая выборки), controlA, controlB. Анализируются архитектуры нейронных сетей для использования в рамках задачи распознавания дипфейковых изображений. Для распознавания дипфейковых изображений использована модель Xception. Размер входного изображения для модели установлен по умолчанию – 299×299. В качестве инструментов обучения нейросетевой модели в работе задействованы облачная среда Google Kolab, локальная среда выполнения кода (Jupyter Notebook), использован язык программирования Python. Проводится обучение нейросетевой модели, подбор оптимального процесса обучения и гиперпараметров модели. Выполнено сравнение двух подходов к распознаванию дипфейковых изображений: модифицированной нейросетевой модели Xception с трансформерным заголовком и ансамбля из четырех моделей (Xception, Efficientnet-B4, ConvNeXt, Swin Transformer.). По результатам анализа выявлено, что модель CNN+tr.head демонстрирует высокие результаты метрик на тестовой выборке (F1 = 0,9105), однако чувствительна к новым типам генераторов, что выражается в падении метрики F1 на контрольной выборке до 0,7414. Ансамблевый подход, напротив, обеспечивает более высокую робастность при распознавании изображений, сгенерированных ранее неизвестными моделями (метрика F1 на контрольной выборке - 0,8082).

P. A. Cherkasov, R. M. Khusainov, N. G. Talipov

A NEURAL NETWORK MODEL FOR DEEPFAKE IMAGE RECOGNITION

Keywords: deepfake, artificial intelligence, neural network model, generated image, deepfake recognition, neural network.

This article examines the problem of deepfake image recognition using a neural network model. It provides an overview of modern research on deepfakes in images, identifying the advantages and disadvantages of existing recognition methods. Existing datasets of real images and deepfakes are analyzed, and the choice of the most suitable one for training, validating, and testing the neural network model is substantiated. As a result of the dataset analysis, the ArtiFact (Real and Fake Image Dataset) dataset was selected as the initial set. It contains over 2 million images, 19 image generators, and 11 sources of real images. Given the limited computing resources of the environment, it was decided not to use the entire found dataset, but to select the necessary images to create our own datasets: final (training, validation, test sets), controlA, and controlB. Neural network architectures for use in the deepfake image recognition problem are analyzed. The Xception model is used to recognize deepfake images. The input image size for the model is set by default - 299×299. The Google Kolab cloud environment, a local code execution environment (Jupyter Notebook), and the Python programming language were used to train the neural network model. A neural network model is trained, and the optimal training process and hyperparameters are selected. Two approaches to deepfake image recognition are compared: a modified Xception neural network model with a transformer head and an ensemble of four models (Xception, Efficientnet-B4, ConvNeXt, and Swin Transformer). The analysis revealed that the CNN+tr.head model demonstrates high metric results on the test set (F1 = 0.9105), but is sensitive to new types of generators, resulting in a drop in the F1 metric on the validation set to 0.7414. The ensemble approach, in contrast, provides higher robustness when recognizing images generated by previously unknown models (F1 metric on the validation set is 0.8082).

Введение

Впервые данный термин дипфейк («deepfake») появился в 2017 году, когда один из пользователей платформы Reddit под ником «deepfake» выложил видео порнографического содержания с замененными лицами актеров [1]. Пользователь осуществил замену с использованием моделей искусственного интеллекта (ИИ). Такой контент называется дипфейком.

Дипфейк – это синтетически созданный медиаконтент (изображения, видео или аудио), который с высокой степенью реалистичности имитирует реальные записи [2]. В более широком смысле дипфейк - любая разновидность контента, сгенерированная с использованием алгоритмов ИИ и не отличимого от реального. Из рассмотренных понятий можно сделать вывод, что для создания дипфейков обычно используют методы, а также алгоритмы ИИ. Дипфейк объединяет понятия «глубокое обучение» и «подделка», подчеркивая использование алгоритмов ИИ для создания реалистичных фальшивок. Дипфейки можно разделить на разные категории, такие как подделка голоса, создание графических изображений или видео [3].

В последнее время направление активно развивается, проводятся исследования по распознаванию дипфейковых изображений. В работе [4] исследование

направлено на выявление общих для всех моделей артефактов генерации. Исследователи заметили, что детекторы либо хорошо обучаются на знакомых артефактах генерации, пропуская более универсальные артефакты генерации, либо специально плохо обучаются на знакомых артефактах для лучшего распознавания незнакомых артефактов генерации. В ходе исследования предложена архитектура с двумя параллельными ветвями, где первая ветвь обрабатывает исходное изображение, а вторая – искаженную версию этого изображения, созданную с использованием перемешивания патчей (patch shuffling). В результате модель вынуждена обращать внимание на артефакты, найденные в обоих версиях, что повышает точность распознавания.

В работе [5] проведен анализ существующих датасетов для обучения нейронных сетей. Существующие датасеты (FaceForensics++, Celeb-DF) созданы в лабораторных условиях с использованием открытых исследовательских проектов. Исследователи использовали популярные инструменты создания дипфейков для имитации действий пользователей или злоумышленников в целях подготовки точного датасета – readface. При проверке на существующих детекторах разрыв точности между использующимися датасетами и readface достигает 40-50 %.

Однако недостаточно исследовано распознавание дипфейковых изображений. Во-первых, дипфейки можно распознать невооруженным глазом из-за сгенерированного изображения. К таким дипфейкам можно отнести картинки с неестественными физическими явлениями, ошибками генерации изображения (например, ошибочное количество пальцев) или фантастического содержания изображения. Во-вторых, выполненные дипфейки на более высоком уровне, не каждый специалист сможет с точностью распознать [6-8].

В связи с распространением дипфейков и сложностью их распознавания предложено использовать нейронные сети для решения данной задачи. Нейронные сети распознают дипфейки по артефактам генерации изображения. К таким артефактам можно отнести шумовые паттерны при генерации изображения GAN-архитектурами или неестественные текстуры кожи на сгенерированных изображениях [9-11].

Цель и задачи исследования

Цель исследования: разработать программную реализацию нейросетевой модели для бинарной классификации изображений (реальное / дипфейк), обеспечивающей высокую точность распознавания.

Достижение поставленной цели потребовало решения следующих задач:

1) проведение сравнительного анализа архитектур нейронных сетей и оценка их эффективности применительно к задаче распознавания дипфейковых изображений;

2) анализ существующих наборов данных (датасетов) синтетических и реальных изображений, обосновать выбор наиболее пригодного для обучения, валидации и тестирования нейросетевой модели;

3) реализация нейросетевой модели распознавания дипфейковых изображений и проведение анализа ее работы на тестовых и контрольных выборках.

Обзор готовых наборов данных (датасетов) для исследования

Для построения качественной нейросетевой модели необходимо выбрать и подготовить исходные данные для ее обучения и дальнейшего тестирования [12]. Для данной работы в качестве материалов для выборок рассмотрены следующие готовые наборы данных (датасеты):

1. GenImage – датасет AI-изображений (дипфейковые + реальные), содержащий более 1000000 пар изображений (дипфейковые + реальные), сгенерированные разными генераторами, включая SD, Midjourney, GAN и др. [13]. Тип датасета: общемасштабное изображение (классы похожи на ImageNet) Размещение датасета: Google Drive.

2. Fake2M – датасет (real vs fake), содержащий синтетические и реальные изображения в тренировочных / валидационных папках [14]. Размер датасета: сотни тысяч изображений Тип датасета: различные категории (не только лицо).

3. ArtiFact: Real and Fake Image Dataset – датасет, содержащий ~ 2496738 изображений (реальные + синтетические) из различных источников, созданных с помощью 25 методов генерации (GAN + Diffusion) [15]. Тип датасета: разнообразные объекты / сцены. Платформа размещения датасета: Kaggle.

4. CIFAKE: real vs AI-generated – датасет, содержащий 60000 реальных + 60000 синтетических изображений (Stable Diffusion для fake) - удобен для базового обучения и демонстраций [16]. Тип датасета: общий (без лицевой привязки).

5. Synthbuster – датасет синтетических изображений (Diffusion models), содержащий изображения, сгенерированные разными diffusion-моделями (например DALL·E 2/3, Midjourney, Stable Diffusion и др.) [17]. Размер датасета: зависит от релиза (часто несколько десятков тысяч изображений) Тип датасета: diffusion генерации.

6. SynthScars (частично доступен, аннотированные артефакты) – датасет с метками пиксельных артефактов полезен для обучения локализации аномалий [18]. Размер датасета: ~ 12236 изображений. Тип датасета: аннотации артефактов.

Рассмотрен перечень параметров датасетов. Среди них доступность для использования, размер, полнота и структура датасета. Исходя из рассмотренных параметров выбран датасет ArtiFact [15].

Для дальнейшей подготовки выборок проанализирован датасет ArtiFact. Датасет доступен на ресурсе Kaggle [15]. Датасет содержит более 2 миллионов изображений, 19 генераторов изображений и 11 источников реальных изображений. Отобраны 3 папки из датасета ArtiFact: final, controlA, controlB. Размер датасетов – 30000, 4500, 2000. Основной датасет разделен на подвыборки (train/val/test) в размере по 10500/2250/2250 изображений (70 / 15 / 15 %) для подпапок «real» и «fake» [19].

Датасет controlA содержит новые fake изображения, но уже использованные (из тестовой выборки) изображения класса «real». Датасет controlB содержит новые, неиспользованные изображения класса «fake» и «real».

Итоговое разделение данных по генераторам представлено в таблице 1.

Таблица 1 – Итоговое распределение данных в датасете

Table 1 – Final distribution of data in the dataset

Выборка / набор данных	Источник данных для класса «real»	Источник данных для класса «fake»
Обучающая выборка	imagenet, coco, lsun	stylegan1, stylegan2, pro_gan, big_gan, star_gan, ddpm, latent_diffusion, glide, lama, generative_inpainting, gau_gan, taming_transformer
Валидационная выборка	afhq, landscape	cycle_gan, gansformer, palette
Тестовая выборка	ffhq, celebahq, sfhq, metfaces	vq_diffusion, denoising_diffusion_gan, projected_gan, diffusion_gan
Датасет «ControlA»	Изображения из тестовой выборки	stylegan3, stable_diffusion, mat
Датасет «ControlB»	Общий пул из всех доступных папок, не задействованные ранее изображения	stylegan3, stable_diffusion, mat

Таким образом, данные подготовлены для дальнейшего использования в задаче распознавания дипфейковых изображений.

Выбор архитектуры нейросетевой модели

Выбор правильной архитектуры нейронной сети является одним из важных шагов для решения задачи распознавания дипфейковых изображений. Архитектура сети определяет ее структуру, количество слоев и количество нейронов в каждом слое, а также тип соединений между слоями.

Существует множество различных архитектур нейронных сетей, которые могут использоваться для решения данной задачи [20].

Сверточные нейронные сети (CNN) – это класс искусственных нейронных сетей, специально разработанных для обработки данных с пространственной информацией, таких как изображения, звуки и другие формы сигналов. Они широко используются в компьютерном зрении, обработке естественного языка, распознавании речи и других областях [21]. Основным их отличием от других типов нейронных сетей является использование операции свертки вместо полного соединения. Операция свертки позволяет извлекать локальные особенности из входных данных, что делает сверточные нейронные сети более эффективными при работе с изображениями и другими данными, имеющими пространственную структуру. Кроме того, сверточные нейронные сети обычно используют слои пулинга, которые уменьшают размерность входных данных, уменьшая количество параметров сети и ускоряя процесс обучения. Это делает их особенно эффективными для задач, где требуется высокая точность и скорость обработки данных.

Transformer – архитектура на основе механизма внимания, отказавшаяся от рекуррентных связей в пользу вычисления взвешенных связей между всеми временными позициями. Transformer кодирует временную информацию через синусоиды и обеспечивает параллельная обработка разных представлений данных [22].

Для дальнейшей работы рассмотрены варианты нейросетевых архитектур и способы их комбинации.

Проведя обзор доступных моделей, использованы два варианта: CNN (Xception) + transformer head; ансамбль из четырёх моделей, построенных на основе архитектур Xception, Efficientnet-B4, ConvNeXt, Swin Transformer.

Xception хорошо подходит для задачи распознавания дипфейковых изображений. В модели используется Depthwise Separable Convolutions, что позволяет лучше выделять текстуры и артефакты сжатия, улавливать неестественные границы между facial features (элементами лица), обнаруживать аномалии в blending (смешивании областей) [23].

Efficientnet-B4 – более современная и оптимизированная архитектура. Имеет более высокое изначальное разрешение 380×380. Также хорошо подходит для выявления глобального контекста [24].

ConvNeXt – это эволюция чистых сверток, доведенная до уровня Vision Transformer (ViT) по метрикам, но с сохранением индуктивного смещения CNN. Архитектурные преимущества ConvNeXt заключаются в большом ядре (глубинные свертки 7×7) свертки на первом этапе в каждом блоке, использовании Layer Normalization вместо BatchNorm, использовании GELU вместо ReLU. Для данной модели выбран размер исходных данных – 384×384 [25].

Swin Transformer – архитектура Vision Transformer, которая строит пирамиду признаков, как CNN, но с помощью механизма Self-Attention. Данная модель рассматривает структурную целостность и биометрическую геометрию. Преимущества модели – внимание к глобальному контексту, обнаружение границ. Для данной модели выбран размер исходных данных – 384×384 [26].

В рамках данной задачи при выборе размеров моделей приоритет отдавался размерам с «до 100 миллионов обучаемых параметров», т.к. при выборе моделей с большим количеством признаков время, затрачиваемое на обучение, превышает 4-6 часов.

Также стоит отметить, что ансамбль имеет несколько вариантов объединения моделей, таких как мягкое голосование, взвешенное голосование, метаклассификатор. Ансамбль состоит из нескольких архитектур, что влияет на необходимость подбора самих архитектур, а также их последующего обучения.

Так как архитектур несколько, то и время на их обучение увеличивается пропорционально количеству задействованных архитектур [27].

Также данные архитектуры возможно оптимизировать под возможности вычислительной среды (GPU T4).

В рамках решаемой задачи решено провести тестирование данных архитектур и определить более подходящую для распознавания дипфейковых изображений.

Построение и обучение нейросетевой модели

Для распознавания дипфейковых изображений использована нейросетевая модель Xception, предобученная на наборе данных ImageNet. Размер входного изображения для модели установлен по умолчанию – 299×299. Используются Focal Attention механизмы. Transformer Head – 2 слоя, 8 голов внимания, 4 CLS токена. Обучение проведено в 2 стадии. Стадия 1 – Focal Attention + Transformer + Classifier, Backbone полностью заморожен, 8 эпох. Стадия 2 – доразморозка модели, Разморожены слои: conv4, exit_flow, block12. Используемый оптимизатор – AdamW. Применены различные методы аугментации данных.

Дополнительный данные при обучении модели. Learning Rate (Stage 1) - 1e-4; Learning Rate (Stage 2) - 5e-5; Weight Decay - 1e-4. Batch size – 16; Gradient accumulation steps – 2.

Проведены обучение и тестирование модели. Получены следующие значения метрик на тестовой выборке: Accuracy = 0,9056; Precision = 0,8654; Recall = 0,9604; F1-score = 0,9105; ROC-AUC = 0,9293; PR-AUC =0,893; MCC = 0,816.

Матрица ошибок модели Xception, полученная с тестовой выборки, приведена на рисунке 1.

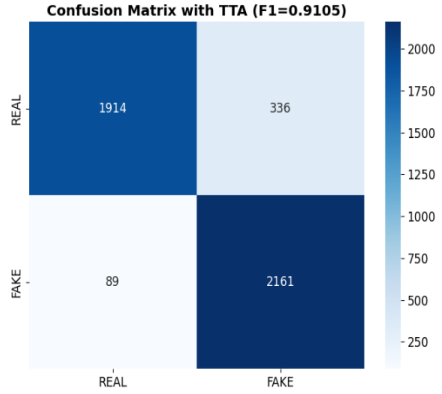

Рис. 1 – Матрица ошибок модели Xception, полученная с тестовой выборки

Fig. 1 – Error matrix for the Xception model, obtained from a test sample

Проведен анализ изображений из контрольных датасетов. Получены следующие значения метрик на контрольной выборке: Accuracy = 0,2805; Precision = 0,3122; Recall = 0,365; F1-score = 0,3366; AUC = 0,2297.

Матрица ошибок модели, полученная с контрольной выборки, приведена на рисунке 2.

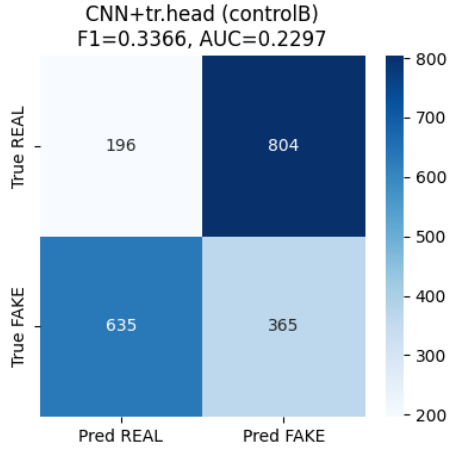

Рис. 2 – Матрица ошибок Xception, полученная с контрольной выборки

Fig. 2 – Xception error matrix derived from the test set

Низкие значения метрик Accuracy (0,2805), AUC (0,2297) и F1 (0,3366) для модели CNN+tr.head на контрольной выборке свидетельствует не просто об ошибке классификации, а о систематическом смещении предсказаний: модель склонна классифицировать новые дипфейки как реальные изображения (или наоборот, в зависимости от порога), что характерно для переобучения на специфические артефакты обучающей выборки.

Проанализированы виды моделей (Xception; Efficientnet-B4; ConvNeXt; Swin Transformer) для формирования ансамбля с мета-классификатором.

Первая модель – Xception. Взята предобученная модель из timm, обучена на imagenet. Применен CBAM для данной модели. Условия обучения – 20 эпох, с постепенной разморозкой слоев. Разрешение для входных изображений выставлено на 299×299.

Вторая модель – Efficientnet-B4. Взята предобученная модель из torchvision.models, обучена на imagenet. Условия обучения – 20 эпох, с постепенной разморозкой слоев, полное обучение модели. Разрешение для входных изображений выставлено на 380×380, стандартный размер входного изображения для данной архитектуры.

Третья модель – convnext-base. Обучение проведено с постепенной разморозкой в 4 стадии. Разрешение для входных изображений выставлено на 384×384.

Четвертая модель – Swin-base. Обучение проведено с постепенной разморозкой в 4 стадии. Разрешение для входных изображений выставлено на 384×384.

Проведено обучение данных моделей. Из данных моделей составлен ансамбль с мета-классификатором. Получены следующие значения метрик мета-классификатора на тестовой выборке: Accuracy = 0,258; Precision = 0,3222; Recall = 0,4378; F1-score = 0,3711; AUC = 0,2497.

Матрица ошибок ансамбля, полученная с тестовой выборки, приведена на рисунке 3.

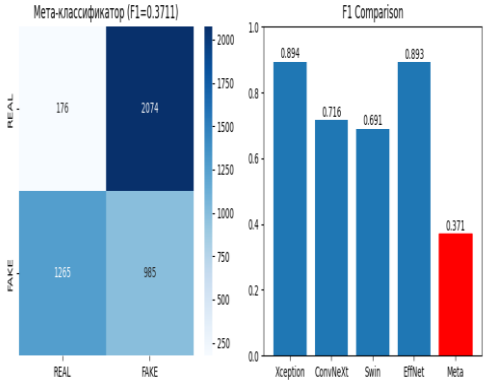

Рис. 3 – Матрица ошибок ансамбля, полученная с тестовой выборки

Fig. 3 – Ensemble error matrix obtained from the test set

Проведен анализ изображений из контрольных датасетов. Получены следующие значения метрик ансамбля на контрольной выборке: Accuracy = 0,8175; Precision = 0,8516; Recall = 0,769; F1-score = 0,8082; AUC = 0,8621.

Матрица ошибок ансамбля на контрольной выборке приведена на рисунке 4.

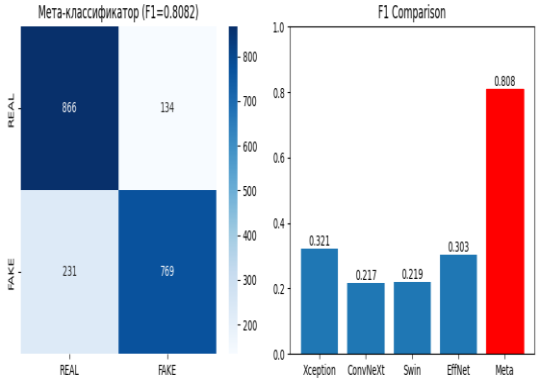

Рис. 4 – Матрица ошибок ансамбля, полученная с контрольной выборки

Fig. 4 – Ensemble error matrix obtained from the validation set

Анализ метрик модели CNN (Xception) + tr.head показывает значительное снижение значения F1 при переходе от тестовой к контрольной выборке. Это объясняется тем, что в контрольную выборку вошли изображения, созданные новыми типами генераторов. Из этого следует вывод, что модель научилась распознавать дипфейки, сгенерированные только теми алгоритмами, которые представлены в обучающей выборке. Результаты свидетельствуют о недостаточной обобщающей способности модели, несмотря на высокую точность распознавания на известных генераторах.

В ходе исследования выявлено, что мета-классификатор на основе ансамбля (Xception, EfficientNet-B4, ConvNeXt, Swin Transformer) выдает высокие метрики на контрольном наборе данных (F1 = 0,8082), однако значительное снижение качества наблюдается на тестовой выборке. Это свидетельствует о недостаточной обобщающей способности мета-классификатора. Возможные причины: различие в распределении данных между контрольным и тестовым наборами, а также возможное переобучение мета-классификатора модели. Для повышения надежности ансамбля рекомендуется использовать мягкое голосование или взвешенное голосование.

Анализ полученных результатов с существующими методами распознавания дипфейковых изображений

Проведено сравнение полученных метрик с результатами современных исследований, представленных в работах [4, 5]. В работе [4] авторы предлагают архитектуру с перемешиванием патчей (patch shuffling), достигая точности ~ 92 % на известных генераторах, однако отмечают снижение robustness на новых моделях диффузии. В исследовании [5] выявлено, что стандартные детекторы теряют до 40-50 % точности при переходе на кросс-датасетное тестирование (например, с FaceForensics++ на WildDeepfake).

Результаты проведенного исследования демонстрируют, что модель CNN+tr.head имеет сопоставимую с современными аналогами точность на обучающей выборке (F1 = 0.91), но подтверждает проблему обобщения (падение метрики F1 до 0.33 на новых генераторах). Предложенный ансамблевый подход, несмотря на сложность мета-классификации, обеспечивает более высокую устойчивость (F1 = 0.8082 на контрольной выборке ControlB).

Заключение

В ходе выполнения работы решены следующие задачи:

1. Проведен анализ современных архитектур нейронных сетей (CNN, Transformer) и их результативности для задачи детекции дипфейков. Выявлено, что сверточные и трансформерные архитектуры наиболее применимы для бинарной классификации изображений.

2. Выполнен анализ открытых наборов данных (GenImage, Fake2M, ArtiFact, CIFAKE, Synthbuster, SynthScars). На основе критериев доступности, размера и структуры выбран датасет ArtiFact. Подготовлены и структурированы обучающая (70 %), валидационная (15 %) и тестовая (15 %) выборки, а также выделены контрольные группы (controlA, controlB) для оценки устойчивости модели к сдвигу распределения данных.

3. Программно реализованы два подхода к распознаванию дипфейковых изображений: модель на основе Xception с интегрированным механизмом внимания (Transformer Head); ансамблевая модель, объединяющая сверточные (Xception, EfficientNet-B4, ConvNeXt) и трансформерную (Swin Transformer) архитектуры с мета-классификатором.

Анализ результатов показал, что модель CNN+tr.head демонстрирует высокие значения метрик на тестовой выборке (F1 = 0,9105), однако оказывается чувствительной к новым типам генераторов, что выражается в падении метрики F1 на контрольной выборке до 0,3366. Ансамблевый подход, напротив, обеспечивает более высокую робастность при распознавании изображений, сгенерированных ранее неизвестными моделями (метрика F1 на контрольном

наборе составил 0,8082). Эти результаты позволяют сделать вывод, что использование мета-классификатора в ансамбле является сложной задачей, так как его легко переобучить. Поэтому методы «мягкого» или «взвешенного» голосования остаются более предпочтительными благодаря своей простоте и устойчивости.

Таким образом, цель работы, заключающаяся в программной реализации нейросетевой модели для бинарной классификации изображений (реальное / дипфейк), достигнута.

Литература

1. Дипфейк. URL: https://ru.wikipedia.org/wiki/ Дипфейк (дата обращения 02.03.2026).
2. Что такое дипфейки, для чего их используют и чем они опасны. URL: https://practicum.yandex.ru/blog/chto-takoe-deepfake-i-kak-zashchititsya/#chto-takoye-dipfeyk (дата обращения 24.04.2026).
3. Неренц Д.В., *Филология: научные исследования*, 9, 96-111 (2025).
4. Yong Y., Zhihao Q., Ye Z., Russakovsky O., Yu W., *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 23850–23859 (2025).
5. Junyu S., Minghui L., Junguo Z., Zhifei Yu, Yipeng L., Shengshan H., Ziqi Z., Yechao Z., Wei W., Yinzhe X., Leo Yu Z, *39th Conference on Neural Information Processing Systems (NeurIPS)*, 1-13 (2025).
6. Пикуль А.С., *Безопасность информационных технологий*, **31**, 4, 116-127 (2024).
7. Воронов С.А., *Информация и безопасность*, **28**, 1, 103-110 (2025).
8. Андриянов А.М., *Научно-технический вестник Поволжья*, 1, 169-172 (2025).
9. Гришаев Д.А., *Вестник Пензенского государственного университета*, 1 (49), 28-31 (2025).
10. Гарифуллин Н.Б., *Молодежная школа-семинар по проблемам управления в технических системах имени А.А. Вавилова*. **1**, 13-15 (2024).
11. Мантуленко А.И., *Математические методы в технологиях и технике*, 5, 97-101 (2025).
12. Петрин Д.А., *Известия Института инженерной физики*, 1 (59), 56-60 (2021).
13. Набор данных genimage. URL: https://drive.google.com/drive/folders/1jGt10bwTbhEZuGXLyvrCuxOI0cBqQ1FS (дата обращения 09.03.2026).
14. Datasets: InfImagine / FakeImageDataset. URL: https://huggingface.co/datasets/InfImagine/FakeImageDataset (дата обращения 09.03.2026).
15. ArtiFact: Real and Fake Image Dataset. URL: https://elibrary.ru/item.asp?id=49175611 (дата обращения 09.03.2026).
16. Datasets: yanbax / CIFAKE_autotrain_compatible. URL: https://huggingface.co/datasets/yanbax/CIFAKE_autotrain_compatible (дата обращения 09.03.2026).
17. Synthbuster: Towards Detection of Diffusion Model Generated Images. URL: https://zenodo.org/records/10066460 (дата обращения 09.03.2026).
18. GitHub: opendatalab / LEGION. URL: https://github.com/opendatalab/LEGION (дата обращения 09.03.2026).
19. Гагарина А.И., *Современная педагогика и научные исследования в образовательной организации высшего образования: материалы Всероссийской научно-методической конференции*, 694-704 (2022).
20. Акобия В.З., *Современные тенденции развития и перспективы внедрения инновационных технологий в машиностроении, образовании и экономике*, **7**, 1, 136-138 (2025).
21. Сверточная нейронная сеть, часть 1: структура, топология, функции активации и обучающее множество // habr.com [сайт]. – URL: https://habr.com/ru/articles/348000/ (дата обращения: 24.04.2026).
22. Иванов П.П., *Цифровой регион: опыт, компетенции, проекты: сборник статей VII Международной научно-практической конференции, посвященной 95-летнему Юбилею Брянского государственного инженерно-технологического университета*, 314-318 (2025).
23. Xception: компактная глубокая нейронная сеть. URL: https://habr.com/ru/articles/347564/ (дата обращения: 24.04.2026).
24. Использование моделей EfficientNet для классификации изображений. URL: https://habr.com/ru/companies/sberbank/articles/828842/ (дата обращения: 24.04.2026).
25. Обзор – ConvNet для 2020. URL: https://habr.com/ru/companies/otus/articles/654279/ (дата обращения: 24.04.2026).
26. Обзор архитектуры Swin Transformer. URL: https://habr.com/ru/articles/599057/ (дата обращения: 24.04.2026).
27. Ансамблевые методы машинного обучения. URL: https://habr.com/ru/articles/571296/ (дата обращения: 24.04.2026).

References

1. Deepfake. URL: https://ru.wikipedia.org/wiki/ Deepfake (accessed March 2, 2026).
2. What Are Deepfakes, What Are They Used For, and Why Are They Dangerous? URL: https://practicum.yandex.ru/blog/chto-takoe-deepfake-i-kak-zashchititsya/#chto-takoye-dipfeyk (accessed April 24, 2026).
3. Nerenz D.V., *Philology: Scientific Research*, 9, 96–111 (2025).
4. Yong Y., Zhihao Q., Ye Z., Russakovsky O., Yu W., *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 23850–23859 (2025).
5. Junyu S., Minghui L., Junguo Z., Zhifei Yu, Yipeng L., Shengshan H., Ziqi Z., Yechao Z., Wei W., Yinzhe X., Leo Yu Z, *39th Conference on Neural Information Processing Systems (NeurIPS)*, 1–13 (2025).
6. Pikul A.S., *Information Technology Security*, **31**, 4, 116–127 (2024).
7. Voronov S.A., *Information and Security*, **28**, 1, 103–110 (2025).
8. Andriyanov A.M., *Scientific and Technical Bulletin of the Volga Region*, 1, 169–172 (2025).
9. Grishaev D.A., *Bulletin of Penza State University*, 1 (49), 28–31 (2025).
10. Garifullin N.B., *A.A. Vavilov Youth School-Seminar on Control Problems in Technical Systems*. **1**, 13–15 (2024).
11. Mantulenko A.I., *Mathematical Methods in Technology and Engineering*, 5, 97–101 (2025).
12. Petrin D.A., *Proceedings of the Institute of Engineering Physics*, 1 (59), 56–60 (2021).
13. genimage dataset. URL: https://drive.google.com/drive/folders/1jGt10bwTbhEZuGXLyvrCuxOI0cBqQ1FS (accessed 03/09/2026).
14. Datasets: InfImagine / FakeImageDataset. URL: https://huggingface.co/datasets/InfImagine/FakeImageDataset (accessed March 9, 2026).
15. ArtiFact: Real and Fake Image Dataset. URL: https://elibrary.ru/item.asp?id=49175611 (accessed March 9, 2026).
16. Datasets: yanbax / CIFAKE_autotrain_compatible. URL: https://huggingface.co/datasets/yanbax/CIFAKE_autotrain_compatible (accessed March 9, 2026).

17. Synthbuster: Towards Detection of Diffusion Model Generated Images. URL: https://zenodo.org/records/10066460 (accessed 03/09/2026).
18. GitHub: opendatalab / LEGION. URL: https://github.com/opendatalab/LEGION (accessed March 9, 2026).
19. Gagarina A.I., *Modern Pedagogy and Scientific Research in Higher Education Institutions: Proceedings of the All-Russian Scientific and Methodological Conference*, 694–704 (2022).
20. Akobiya V.Z., *Current Trends in Development and Prospects for the Implementation of Innovative Technologies in Mechanical Engineering, Education, and the Economy*, **7**, 1, 136–138 (2025).
21. Convolutional Neural Network, Part 1: Structure, Topology, Activation Functions, and Training Set // habr.com [website]. – URL: https://habr.com/ru/articles/348000/ (accessed: 04/24/2026).
22. Ivanov P.P., *Digital Region: Experience, Competencies, Projects: Collection of Articles from the VII International Scientific and Practical Conference Dedicated to the 95th Anniversary of Bryansk State University of Engineering and Technology*, 314–318 (2025).
23. Xception: A Compact Deep Neural Network. URL: https://habr.com/ru/articles/347564/ (accessed: April 24, 2026).
24. Using EfficientNet models for image classification. URL: https://habr.com/ru/companies/sberbank/articles/828842/ (accessed: April 24, 2026).
25. Overview – ConvNet for 2020. URL: https://habr.com/ru/companies/otus/articles/654279/ (accessed: 04/24/2026).
26. Overview of the Swin Transformer architecture. URL: https://habr.com/ru/articles/599057/ (accessed: 04/24/2026).
27. Ensemble methods in machine learning. URL: https://habr.com/ru/articles/571296/ (accessed: 04/24/2026).

© П. А Черкасов – магистрант кафедры Систем информационной безопасности (СИБ), Казанский национальный исследовательский технический университет имени А.Н. Туполева (КНИТУ им. А.;Н. Туполева), Казань, Россия, 76239bb@mail.ru; **Р. М. Хусаинов** – ассистент кафедры СИБ, КНИТУ им. А.;Н. Туполева, rumil_husainov98@mail.ru; **Н. Г. Талипов** – к-т. техн. наук, доцент кафедры СИБ, КНИТУ им. А.;Н. Туполева, nafis.talipov@mail.ru.

© P. A. Cherkasov – Master-student of the Information Security Systems (ISS) deepartment, Kazan National Research Technical University named after A.N. Tupolev (KNRTU named after A.N. Tupolev), Kazan, Russia, 76239bb@mail.ru; **R. M. Khusainov** – Assistant of the ISS department, KNRTU named after A.N. Tupolev, rumil_husainov98@mail.ru; **N. G. Talipov** – PhD (Technical Sci.), Associate Professor of the ISS department, KNRTU named after A.N. Tupolev, nafis.talipov@mail.ru.

Дата поступления рукописи в редакцию – 06.05.26.
Дата принятия рукописи в печать – 20.05.26.