

Анализ влияния дисбаланса классов на качество трёхклассовой классификации изображений

А.Н. Жданова, С.И. Саблин, А.В. Трухова

Аннотация— В последние годы развитие генеративных моделей искусственного интеллекта привело к резкому росту количества синтетических изображений и дипфейков. Современные инструменты позволяют создавать реалистичные изображения людей, что усложняет задачу проверки подлинности визуального контента и снижает доверие людей к цифровым медиа. В статье рассматривается проблема классификации аватаров пользователей по типу происхождения (реальные, рисованные, сгенерированные) в условиях выраженного дисбаланса классов. Исследование направлено на выявление влияния неравномерного распределения данных на точность классификации изображений по типу происхождения.

В исследовании сравниваются архитектуры ResNet-50, ResNet-18, MobileNetV3, EfficientNet-B0 и ConvNeXt-Tiny, обученные на наборах изображений из открытых источников. Анализ показал, что при высокой доле доминирующего класса ($\approx 80\%$) модели демонстрируют значительное снижение макро-F1 при переходе к независимому тестированию, что свидетельствует о переобучении и потере обобщающей способности. ConvNeXt-Tiny показала наибольшую устойчивость к классовому и доменному смещению, а MobileNetV3 – оптимальное соотношение точности и вычислительной эффективности. Полученные результаты подчёркивают важность учёта дисбаланса классов при построении систем направленных на классификацию изображений.

Ключевые слова— дипфейки, классификация изображений, дисбаланс классов, нейронные сети, синтетические изображения, трёхклассовая классификация

I. ВВЕДЕНИЕ

В последние годы наблюдается стремительный рост количества синтетических изображений и дипфейков – контента, созданного с помощью генеративных моделей искусственного интеллекта. Раньше такие технологии использовались преимущественно в развлекательных целях, но сегодня они все чаще становятся инструментом мошенничества и дезинформации.

По данным компании BI.Zone AntiFraud, за период с 1 января по 1 июня текущего года число мошенничеств с

использованием дипфейков выросло в 2,3 раза по сравнению с аналогичным периодом прошлого года. Социальная инженерия все еще остается более распространенным инструментом кибермошенников, однако доля дипфейков в подобных атаках растет и уже достигла 3–5% [1]. На глобальном уровне статистика ещё тревожнее: в 2023 году сектор финансовых технологий столкнулся с ростом подобных инцидентов на 700%, а количество случаев мошенничества с использованием дипфейков во всём мире увеличилось в десять раз. Прогнозируется, что к 2027 году потери от мошенничества с применением генеративного ИИ в США могут достичь 40 миллиардов долларов [2].

Одновременно с ростом числа злоупотреблений стремительно увеличивается доступность инструментов для генерации дипфейков и качество созданных изображений. Сегодня в сети существует большое количество программ и онлайн-сервисов для генерации изображений. Современные генеративные модели, такие как StyleGAN и Stable Diffusion, делают подделки всё более реалистичными. Наиболее известный пример – проект This Person Does Not Exist, создающий реалистичные изображения людей [3], [4].

В таких условиях становится особенно важно развивать системы распознавания подлинности изображений, способные отличать реальные фотографии от синтетических [5], [6].

А с учётом того, что государственные органы и компании обязаны вести официальные страницы в социальных сетях, например «ВКонтакте» и «Одноклассниках», необходимость автоматической проверки подлинности изображений профиля становится важным элементом цифровой безопасности.

Современные методы кибербезопасности всё чаще используют технологии искусственного интеллекта – для анализа данных, прогнозирования угроз и повышения эффективности защиты информационных систем [7]. При этом сами модели машинного обучения становятся объектом атак, что дополнительно повышает требования к их устойчивости [8]. Принципы машинного обучения и методы распознавания, заложенные в этих системах, могут быть применены и к задаче детекции синтетических изображений.

Несмотря на большое количество существующих решений, задача классификации изображений по типу их происхождения остаётся недостаточно изученной. В отличие от работ, ориентированных на бинарную классификацию «реальное / синтетическое», в данной статье рассматривается трёхклассовая постановка задачи в условиях сильного дисбаланса классов, что позволяет исследовать влияние этого дисбаланса на качество классификации аватаров.

Статья получена 30 апреля 2026.

А.Н. Жданова – к.т.н., доцент кафедры программных систем Самарского национального исследовательского университета имени академика С.П. Королева, Самара, Российская Федерация (e-mail: danilenko@ssau.ru).

С.И. Саблин – студент магистратуры Самарского национального исследовательского университета имени академика С.П. Королева, Самара, Российская Федерация (e-mail: serjjiniuss@gmail.com).

А.В. Трухова – студент магистратуры Самарского национального исследовательского университета имени академика С.П. Королева, Самара, Российская Федерация (e-mail: trukhova-s@yandex.ru).

II. ЦЕЛЬ ИССЛЕДОВАНИЯ

Цель исследования – определение влияния дисбаланса классов на качество классификации аватаров по типу происхождения для задач цифровой идентификации.

Задачи исследования:

1. формирование обучающей выборки на основе датасетов с платформы Kaggle;
2. обучение моделей, основанных на выбранных архитектурах;
3. проведение экспериментов по оценке влияния дисбаланса классов на качество классификации;
4. формирование выводов по результатам эксперимента.

III. СУЩЕСТВУЮЩИЕ ПОДХОДЫ К ДЕТЕКЦИИ СГЕНЕРИРОВАННЫХ ИЗОБРАЖЕНИЙ

Существующие подходы к детекции и классификации изображений можно разделить на несколько групп.

Статистические подходы стали одним из первых направлений в задачах анализа подлинности изображений. Они основаны на исследовании различных компонентов изображения, таких как например распределений пиксельных значений, что позволяет выявлять несоответствия, возникающие при редактировании или генерации изображений.

К данной группе относятся, например, анализ уровня ошибок (Error Level Analysis, ELA). Данный метод выявляет области изображения с различной степенью сжатия, что позволяет обнаруживать следы редактирования или генерации. Однако метод ELA имеет существенные ограничения: он применим исключительно к форматам сжатия с потерями, в первую очередь к JPEG, и, кроме того, практическая эффективность метода снижается из-за его уязвимости к повторному сжатию [9].

Другим примером статистических методов является закон Бенфорда (Benford's Law) (также известный как закон первой цифры или закон значащей цифры). Данный закон часто применяется в криминалистике для обнаружения фальсификации изображений. Он может быть использован и для детекции сгенерированных изображений, однако работает не для всех генераторов. Так, например, изображения, сгенерированные StyleGAN2, часто не соответствуют закону Бенфорда, а изображения, сгенерированные StyleGAN3, имеют практически тот же процент соответствия, что и реальные изображения [10].

Методы глубокого обучения используют многослойные нейронные сети, которые самостоятельно извлекают информативные признаки из изображений. В качестве примеров таких методов можно привести:

Сверточные нейронные сети (CNN) – они обучаются на больших наборах данных и способны выявлять различные паттерны, характерные для тех или иных объектов. Так, например, CNN используются для детектирования и распознавания лиц в системах видеонаблюдения [11]. Эта особенность может быть использована и для выявления паттернов, характерных для реальных и сгенерированных изображений.

GAN-детекторы и fingerprinting – методы из данной группы используют особенности (отпечатки), оставляемые различными генераторами что, позволяет

определить не только синтетическое происхождение изображения, но и генеративную модель, с помощью которой оно было создано. Такие методы показывают высокую точность на известных генераторах, но плохо работают на новых [12], [13].

С распространением диффузионных моделей появились работы, посвященные детекции создаваемых ими изображений, а также построению универсальных детекторов, сохраняющих работоспособность на новых типах генеративных моделей [14], [15], [16]. Для обучения и сопоставления таких детекторов опубликованы крупномасштабные наборы данных, содержащие реальные и сгенерированные изображения [17], [18].

Другие методы: анализ метаданных (может быть легко обойден удалением или подделкой данных), поиск оригиналов изображений (хорошо работает на частично-измененных изображениях, но бесполезен для полностью сгенерированных), анализ водяных знаков (добавление меток при генерации не слишком распространено, что делает данный подход малоэффективным на данный момент. Также метка может быть удалена или искажена).

Отдельное направление исследований связано с проблемой дисбаланса классов при обучении глубоких нейронных сетей. Для его компенсации применяются методы передискретизации выборки и взвешивания функции потерь, включая focal loss, а также специализированные стратегии обучения на длиннохвостых распределениях [19], [20], [21], [22]. При ограниченном объеме размеченных данных используются подходы few-shot learning, а устойчивость моделей к смене источника данных исследуется в работах по доменной генерализации [23], [24]. Вместе с тем влияние дисбаланса классов на многоклассовую классификацию изображений по типу происхождения изучено недостаточно, что определяет предмет настоящего исследования.

IV. МАТЕМАТИЧЕСКАЯ ПОСТАНОВКА ЗАДАЧИ

Классификация изображений (Image Classification) – это фундаментальная задача в области компьютерного зрения (Computer Vision), заключающаяся в анализе и отнесении изображения к одной или нескольким заранее определенным категориям (классам) путем присвоения соответствующих меток.

Пусть имеется множество изображений $X = \{x_1, x_2, \dots, x_n\}$ каждое из которых принадлежит одному из классов $Y = \{y_1, y_2, y_3\}$, где:

- y_1 – реальные фотографии;
- y_2 – нарисованные изображения;
- y_3 – сгенерированные изображения.

Необходимо построить модель $f(x) = y$, которая по входному изображению x предсказывает метку класса y .

V. ИСПОЛЬЗУЕМЫЕ АРХИТЕКТУРЫ

В рамках данного исследования использовались архитектуры глубокого обучения, представляющие различные поколения свёрточных нейронных сетей [25], [26], [27], [28]. Цель экспериментов заключалась не в максимизации точности отдельных архитектур, а в оценке их устойчивости к дисбалансу классов в

обучающей выборке. Информация об использованных архитектурах представлена в таблице 1.

Таблица 1. Использованные архитектуры и их особенности

Название архитектуры	Ключевая особенность	Практическое преимущество
ResNet-18	Остаточные связи (<i>skip connections</i>)	Быстрая сходимость при умеренной глубине
ResNet-50	Глубокие остаточные связи (<i>Bottleneck Blocks</i>)	Лучшая способность извлекать сложные признаки
MobileNetV3	<i>HardSwish</i> , <i>Squeeze-and-Excitation</i> , автооптимизация	Высокая точность при малом числе параметров
EfficientNet-B0	Сбалансированное масштабирование слоёв (<i>MBConv</i>)	Минимальные вычисления при высокой точности
ConvNeXt-Tiny	Гибрид CNN и ViT (патчи, GELU, LayerNorm)	Улучшенное качество при компактной структуре

VI. ОБУЧАЮЩАЯ ВЫБОРКА

Для обучения моделей собраны изображения лиц из датасетов с платформы Kaggle [29]. Для каждого класса изображений было выбрано несколько датасетов, чтобы обеспечить большее разнообразие выборки и повысить обобщающую способность модели за счёт включения изображений, отличающихся по стилю, качеству и условиям съёмки. Описание состава обучающей выборки представлено в таблице 2. Все изображения представлены в формате JPG.

Таблица 2. Описание состава обучающей выборки

Класс	Название	Размер изображений
Реальные фотографии	Real vs Fake Faces	600×600
	Deepfake and Real Images	256×256
	CelebFaces Dataset	64×64
Сгенерированные изображения	Detect AI-Generated Faces: High-Quality Dataset	768×768
	Fake-Vs-Real-Faces (Hard)	300×300
	Person Face Dataset	1024×1024
Нарисованные изображения	Anime Face Dataset	63×63
	LEGO Minifigure Faces	1024×1024
	CUHK Face Sketch Database (CUFS)	64×80

Примеры изображений для каждого класса представлены на рисунке 1.



Рис. 1: Примеры изображений для каждого класса

Анализ распределения классов показал, что основную часть датасета составляют рисованные изображения (около 82 %), а классы реальных и сгенерированных изображений распределены примерно равномерно.

Диаграмма распределения классов представлена на рисунке 2.

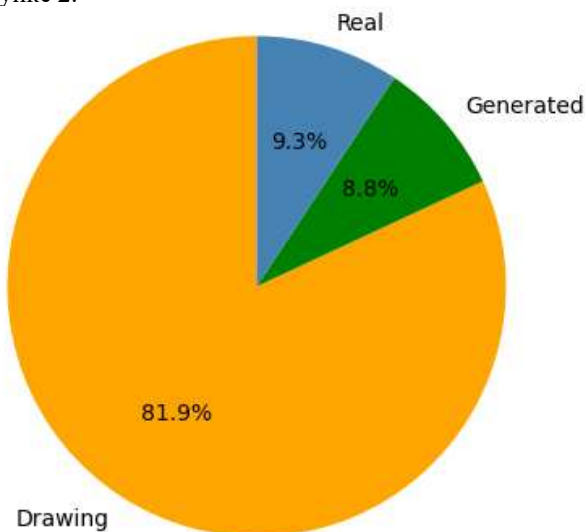


Рис. 2: Диаграмма распределения классов

VII. ОБУЧЕНИЕ МОДЕЛЕЙ

Обучение выполнялось на объединённом наборе данных (разделение 80/10/10), предварительно нормализованных до разрешения 224×224 px. Для компенсации дисбаланса классов применялись взвешенные коэффициенты в функции потерь, оптимизатор AdamW и планировщик ReduceLROnPlateau [30].

Детекция лиц и обрезка фона не выполнялись, что частично обусловило использование фоновых признаков при классификации.

VIII. ЭКСПЕРИМЕНТ

В рамках исследования проведено сравнение нейросетевых архитектур ResNet-50, MobileNetV3, EfficientNet-B0 и ConvNeXt-Tiny, включая их Few-Shot-варианты.

Цель эксперимента – исследовать влияние дисбаланса классов на качество классификации аватаров по типу происхождения (real, drawing, generated) для задач цифровой идентификации.

На рисунке 3 представлено сравнение F1-метрик для различных архитектур.

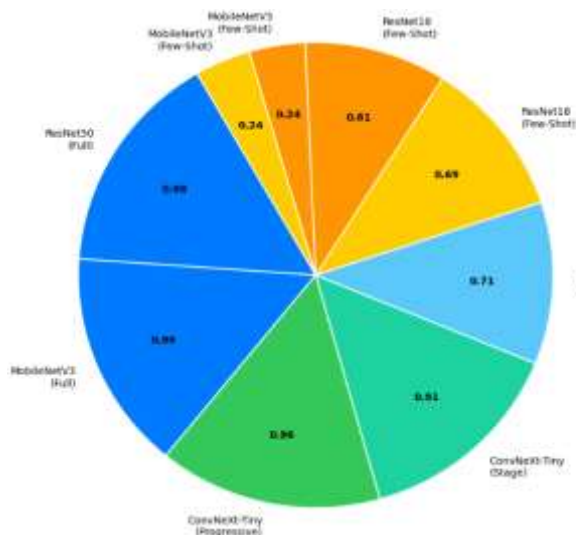


Рис. 3: Сравнение F1-метрики для различных архитектур

Наибольшие значения F1-метрики получены у ResNet-50, ConvNeXt-Tiny Stage 2 и MobileNetV3 Full, что подтверждает эффективность предобученных свёрточных архитектур при ограниченном количестве данных.

Few-Shot-модели закономерно показывают снижение F1, отражающее зависимость качества классификации от объёма выборки и стратегии заморозки весов.

В таблице 3 представлены результаты обучения на основном датасете.

Таблица 3 – Результаты обучения на основном датасете

Модель	Тип обучения	F1	Accuracy
ResNet-50	Полное обучение	0,98	0,99
MobileNetV3	Полное обучение	0,96	0,96
ConvNeXt-Tiny (Stage 2)	Прогрессивное размораживание	0,96	0,98

Для оценки обобщающей способности использовался независимый тестовый набор из 1340 изображений, исключённых из обучающего процесса. Метрики рассчитывались по макро-усреднению (Macro F1, Balanced Accuracy) с учётом стратифицированного разбиения по источникам данных, что позволило контролировать возможное class shift и domain imbalance.

В таблице 4 представлены результаты независимого тестирования.

Таблица 4. Результаты независимого тестирования (Macro F1 и сбалансированная точность)

Модель	ResNet-50	MobileNetV3	ConvNeXt-Tiny (Stage 2)
Macro F1	Полное обучение	Полное обучение	250 изображений на класс
Balanced Accuracy	0,230	0,235	0,104
F1 (AI)	0,000	0,000	0,032
F1 (Drawing)	0,032	0,014	0,014
F1 (Real)	0,732	0,647	0,373

Падение F1 при переходе на независимую выборку демонстрирует высокую чувствительность моделей к дисбалансу классов и источнику данных, подтверждая ограниченную способность свёрточных архитектур к обобщению при доменных смещениях.

При доминировании класса drawing ($\approx 82\%$) модели сохраняют высокую точность на основном наборе, но теряют устойчивость на сбалансированных данных, где макро-F1 снижается почти в четыре раза.

Этот результат отражает закономерную реакцию CNN-архитектур на классовое и доменное смещение и подтверждает ключевую цель исследования – выявление влияния дисбаланса данных на поведение моделей.

На рисунке 4 представлено сравнение F1-метрик по классам для различных архитектур.

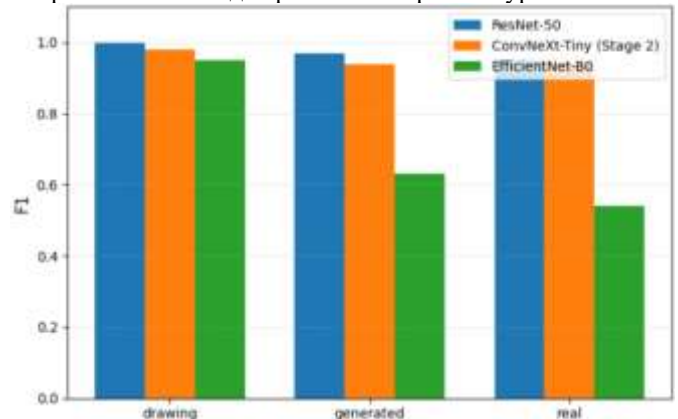


Рис. 4: Сравнение F1-метрик по классам для различных архитектур

Поклассовый анализ выявил выраженное смещение моделей в сторону доминирующего класса. При доле порядка 80% изображений категории drawing все архитектуры демонстрируют стабильно высокие значения метрик для этой группы, тогда как показатели для real и generated заметно ниже.

Особенно сильное падение наблюдается у компактных и few-shot-моделей, что отражает склонность свёрточных сетей переобучаться на большинство и игнорировать редкие категории.

ConvNeXt-Tiny благодаря прогрессивному размораживанию весов достигает на основном наборе качества, сопоставимого с полностью обученными моделями, при существенно меньшем объёме обучающих данных.

ResNet-50, несмотря на высокую точность, быстрее переобучается из-за жёсткой иерархии остаточных блоков и ограниченной адаптивности фильтров.

MobileNetV3 при few-shot-обучении сохраняет относительную устойчивость благодаря оптимизированным активациям и глубинным свёрткам, обеспечивая лучшее соотношение скорости и точности при малом объёме данных.

На рисунке 5 отражены карты внимания ResNet-50 для изображений класса generated [31].



Рис. 5: Карты внимания ResNet-50 для изображений класса generated

Карты внимания ResNet-50 для изображений класса generated, ошибочно классифицированных как drawing. Модель фокусируется на текстуре и цветовых переходах, а не на лицевых чертах, что приводит к переоценке «рисованных» признаков.

На рисунке 6 отражены карты внимания ConvNeXt-Tiny для изображений класса real.



Рис. 6: Карты внимания ConvNeXt-Tiny для изображений класса real

Карты ConvNeXt-Tiny для изображений класса real, ошибочно отнесённых к generated. Активации смещены в сторону фона и бликов, что указывает на влияние нерелевантных визуальных факторов.

Визуальный анализ подтверждает, что компактные сети чаще путают фон с лицом, тогда как глубокие архитектуры избирательно реагируют на текстурные признаки, усиливая bias к доминирующему классу.

На рисунке 7 отражено t-SNE-встраивание признаков для модели ConvNeXt-Tiny Stage 1 [32].

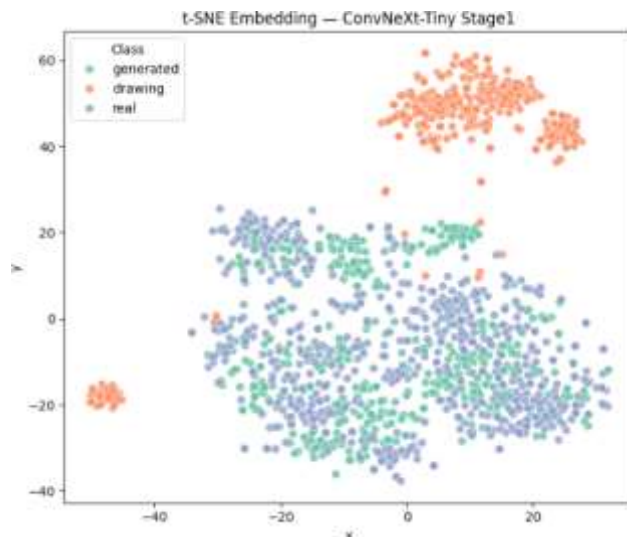


Рис. 7: t-SNE-встраивание признаков для модели ConvNeXt-Tiny Stage 1

Диаграмма демонстрирует плотное разделение рисованных аватаров и частичное пересечение областей real и generated, отражающее визуальное сходство синтетических изображений с реальными лицами.

Ошибки чаще наблюдались у изображений с нетипичным освещением, низкой детализацией и у аватаров аниме-стиля, что подтверждает зависимость моделей от стиливых характеристик данных.

В таблице 5 отражены данные об изменении Accuracy и Macro-F1 при росте доли доминирующего класса drawing.

Таблица 5. Изменение Accuracy и Macro-F1 при росте доли доминирующего класса drawing

Сценарий	Модель	Accuracy (micro)	Macro-F1
50-25-25	EfficientNet-B0	0,123	0,145
70-15-15	ResNet18 FewShot 12 ep	0,068	0,094
80-10-10	MobileNetV3 FewShot 4 ep	0,129	0,083
80-10-10	ConvNeXt-Tiny Stage 2	0,023	0,018

Дополнительная оценка при изменении долей классов в обучающей выборке показала закономерное снижение точности и Macro-F1 с ростом доли drawing.

При превышении 70 % доли доминирующего класса Macro-F1 падает ниже 0,3 для всех архитектур.

ConvNeXt-Tiny демонстрирует минимальное снижение (~25 %), что указывает на её большую устойчивость к классовому и доменному смещению.

IX. РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТА

Проведённые эксперименты подтвердили высокую чувствительность CNN-архитектур к классовому и доменному дисбалансу. Рост доли доминирующего класса ведёт к переобучению и потере способности к распознаванию редких категорий.

Сравнение архитектур ResNet-50, MobileNetV3, EfficientNet-B0 и ConvNeXt-Tiny показало, что при $F1 \approx 0,96-0,98$ на несбалансированных данных макро-F1 снижается до 0,13–0,26 на независимом тесте, отражая переобучение к доминирующему классу drawing.

ConvNeXt-Tiny достигает сопоставимого качества при меньшем объёме обучающих данных, ResNet-50 быстрее переобучается, а MobileNetV3 сохраняет лучший баланс точности и вычислительной эффективности.

Х. ЗАКЛЮЧЕНИЕ

В работе реализован полный пайплайн классификации аватаров по типу происхождения (real, drawing, generated) с учётом классового и доменного дисбаланса.

Проведённое исследование подтверждает, что дисбаланс классов оказывает ключевое влияние на устойчивость нейронных сетей при классификации изображений. При доминировании одного класса (свыше 70%) макро-F1 снижается ниже 0,3 для всех архитектур, однако ConvNeXt демонстрирует наименьшее падение ($\approx 25\%$), что указывает на его устойчивость к классовому смещению.

ConvNeXt-Tiny обеспечила сопоставимое качество при меньшем объёме обучающих данных, MobileNetV3 – лучший баланс скорости и качества.

Результаты подтверждают необходимость учёта class shift и domain imbalance при обучении и перспективность гибридных моделей CNN-ViT для повышения устойчивости классификации.

Эти результаты важны для задач цифровой идентификации, где количество подлинных изображений ограничено, а объём синтетического контента постоянно растёт.

БИБЛИОГРАФИЯ

- [1] Ведомости. "Число случаев мошенничества с использованием дипфейков выросло в 2,3 раза" [Online]. Available: <https://www.vedomosti.ru/technology/articles/2025/06/19/1118123-chislo-sluchaev-moshennichstva-s-ispolzovaniem-dipfeikov-viroslo> Accessed: Oct. 30, 2025.
- [2] L. H. Singh, P. Charanarur, and N. K. Chaudhary, "Advancements in detecting Deepfakes: AI algorithms and future prospects – a review", *Discover Internet of Things*, vol. 5, p. 53, 2025. doi: 10.1007/s43926-025-00154-0.
- [3] T. Karras, S. Laine, and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks", in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4396–4405. doi: 10.1109/CVPR.2019.00453.
- [4] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models", in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10674–10685. doi: 10.1109/CVPR52688.2022.01042.
- [5] M. S. Rana, M. N. Nobil, B. Murali, and A. H. Sung, "Deepfake Detection: A Systematic Literature Review", *IEEE Access*, vol. 10, pp. 25494–25513, 2022. doi: 10.1109/ACCESS.2022.3154404.
- [6] A. Heidari, N. Jafari Navimipour, H. Dag, and M. Unal, "Deepfake detection using deep learning methods: A systematic and comprehensive review", *WIREs Data Mining and Knowledge Discovery*, vol. 14, no. 2, art. e1520, 2024. doi: 10.1002/widm.1520.
- [7] Д.Е. Намиот, Е.А. Илюшин, И.В. Чижов, "Искусственный интеллект и кибербезопасность", *International Journal of Open Information Technologies*, vol. 10, no. 9, pp. 135–147, 2022. [Online]. Available: <https://www.injoit.org/index.php/j1/article/view/1402>
- [8] Д.Е. Намиот, "Схемы атак на модели машинного обучения", *International Journal of Open Information Technologies*, vol. 11, no. 5, pp. 68–86, 2023. [Online]. Available: <https://www.injoit.org/index.php/j1/article/view/1566>
- [9] В.Е. Пономарев, "Криминалистическое исследование цифровых изображений, изменённых ИИ: применимость новых и классических методов анализа", *Юридическая наука*, no. 9, pp. 430–434, 2025. doi: 10.24412/2220-5500-2025-9-430-434.
- [10] С.П. Никитенкова, "Применение закона Бенфорда в обнаружении deepfake-изображений", *Вестник Томского государственного университета. Управление, вычислительная техника и информатика*, no. 64, pp. 128–137, 2023. doi: 10.17223/19988605/64/13.
- [11] А.Б. Мудрич, К.В. Ежова, "Детектирование и распознавание лиц в системах видеонаблюдения", *International Journal of Open Information Technologies*, vol. 12, no. 3, pp. 82–87, 2024. [Online]. Available: <https://www.injoit.org/index.php/j1/article/view/1798>
- [12] N. Yu, L. Davis, and M. Fritz, "Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints", in *Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, 2019, pp. 7555–7565. doi: 10.1109/ICCV.2019.00765.
- [13] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "CNN-Generated Images Are Surprisingly Easy to Spot... for Now", in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 8692–8701. doi: 10.1109/CVPR42600.2020.00872.
- [14] R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, and L. Verdoliva, "On the Detection of Synthetic Images Generated by Diffusion Models", in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5. doi: 10.1109/ICASSP49357.2023.10095167.
- [15] Z. Wang et al., "DIRE for Diffusion-Generated Image Detection", in *Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, 2023, pp. 22388–22398. doi: 10.1109/ICCV51070.2023.02051.
- [16] U. Ojha, Y. Li, and Y. J. Lee, "Towards Universal Fake Image Detectors that Generalize Across Generative Models", in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 24480–24489. doi: 10.1109/CVPR52729.2023.02345.
- [17] M. Zhu et al., "GenImage: A Million-Scale Benchmark for Detecting AI-Generated Image", in *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, Datasets and Benchmarks Track, 2023. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/hash/f4d4a021f9051a6c18183b059117e8b5-Abstract-Datasets_and_Benchmarks.html
- [18] J. J. Bird and A. Lotfi, "CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images", *IEEE Access*, vol. 12, pp. 15642–15650, 2024. doi: 10.1109/ACCESS.2024.3356122.
- [19] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance", *Journal of Big Data*, vol. 6, art. 27, 2019. doi: 10.1186/s40537-019-0192-5.
- [20] S. Rezvani and X. Wang, "A broad review on class imbalance learning techniques", *Applied Soft Computing*, vol. 143, art. 110415, 2023. doi: 10.1016/j.asoc.2023.110415.
- [21] Y. Zhang, B. Kang, B. Hooi, S. Yan, and J. Feng, "Deep Long-Tailed Learning: A Survey", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10795–10816, 2023. doi: 10.1109/TPAMI.2023.3268118.
- [22] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, 2020. doi: 10.1109/TPAMI.2018.2858826.
- [23] Y. Song, T. Wang, P. Cai, S. K. Mondal, and J. P. Sahoo, "A Comprehensive Survey of Few-shot Learning: Evolution, Applications, Challenges, and Opportunities", *ACM Computing Surveys*, vol. 55, no. 13, art. 271, 2023. doi: 10.1145/3582688.
- [24] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain Generalization: A Survey", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4396–4415, 2023. doi: 10.1109/TPAMI.2022.3195549.
- [25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition", in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.
- [26] A. Howard et al., "Searching for MobileNetV3", in *Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, 2019, pp. 1314–1324. doi: 10.1109/ICCV.2019.00140.
- [27] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks", in *Proc. 36th Int. Conf. on Machine Learning (ICML)*, PMLR, vol. 97, 2019, pp. 6105–6114. [Online]. Available: <https://proceedings.mlr.press/v97/tan19a.html>
- [28] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s", in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 11966–11976. doi: 10.1109/CVPR52688.2022.01167.

- [29] Kaggle. [Online]. Available: <https://www.kaggle.com/> (accessed: Aug. 12, 2026).
- [30] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization", in Proc. 7th Int. Conf. on Learning Representations (ICLR), 2019. [Online]. Available: <https://openreview.net/forum?id=Bkg6RiCqY7>
- [31] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization", International Journal of Computer Vision, vol. 128, no. 2, pp. 336–359, 2020. doi: 10.1007/s11263-019-01228-7.
- [32] L. van der Maaten and G. Hinton, "Visualizing Data using t-SNE", Journal of Machine Learning Research, vol. 9, pp. 2579–2605, 2008. [Online]. Available: <https://jmlr.org/papers/v9/vandermaaten08a.html>

Analysis of the Impact of Class Imbalance on the Quality of Three-Class Image Classification

A.N. Zhdanova, S.I. Sablin, A.V. Trukhova

Abstract—Recent advances in generative artificial intelligence have led to a rapid increase in synthetic imagery and deepfakes, posing new challenges to the reliability of visual information. Modern AI tools can produce highly realistic images of non-existent people, making it increasingly difficult to verify media content and maintain public trust in visual data. This study explores the task of classifying user avatars by their origin type (real, drawing, generated) under conditions of significant class imbalance. The research focuses on analyzing how uneven data distribution affects the accuracy of image classification.

In this work, several deep learning architectures (including ResNet-50, MobileNetV3, EfficientNet-B0, and ConvNeXt-Tiny) were trained and evaluated using datasets from open sources. Experimental results demonstrate that when the dominant “drawing” class accounts for about 80% of the dataset, all models show a substantial decrease in macro-F1 scores during independent testing, indicating overfitting and reduced generalization. Among the evaluated models, ConvNeXt-Tiny exhibited the highest stability to class and domain shifts, while MobileNetV3 provided the most efficient balance between classification accuracy and computational cost. The findings highlight the importance of considering class imbalance when training and evaluating image classification models.

Keywords—deepfakes, image classification, class imbalance, neural networks, synthetic images, three-class classification.

REFERENCES

- [1] Vedomosti. “Chislo sluchaev moshennichestva s ispol'zovaniem dipfeikov vyroslo v 2,3 raza” [Online]. Available: <https://www.vedomosti.ru/technology/articles/2025/06/19/1118123-chislo-sluchaev-moshennichestva-s-ispolzovaniem-dipfeikov-viroslo> Accessed: Oct. 30, 2025.
- [2] L. H. Singh, P. Charanarur, and N. K. Chaudhary, “Advancements in detecting Deepfakes: AI algorithms and future prospects – a review”, *Discover Internet of Things*, vol. 5, p. 53, 2025. doi: 10.1007/s43926-025-00154-0.
- [3] T. Karras, S. Laine, and T. Aila, “A Style-Based Generator Architecture for Generative Adversarial Networks”, in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4396–4405. doi: 10.1109/CVPR.2019.00453.
- [4] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-Resolution Image Synthesis with Latent Diffusion Models”, in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10674–10685. doi: 10.1109/CVPR52688.2022.01042.
- [5] M. S. Rana, M. N. Nobil, B. Murali, and A. H. Sung, “Deepfake Detection: A Systematic Literature Review”, *IEEE Access*, vol. 10, pp. 25494–25513, 2022. doi: 10.1109/ACCESS.2022.3154404.
- [6] A. Heidari, N. Jafari Navimipour, H. Dag, and M. Unal, “Deepfake detection using deep learning methods: A systematic and comprehensive review”, *WIREs Data Mining and Knowledge Discovery*, vol. 14, no. 2, art. e1520, 2024. doi: 10.1002/widm.1520.
- [7] D. Namiot, E. Ilyushin, and I. Chizov, “Artificial Intelligence and Cybersecurity”, *International Journal of Open Information Technologies*, vol. 10, no. 9, pp. 135–147, 2022. [Online]. Available: <https://www.injoit.org/index.php/j1/article/view/1402>
- [8] D. Namiot, “Schemes of Attacks on Machine Learning Models”, *International Journal of Open Information Technologies*, vol. 11, no. 5, pp. 68–86, 2023. [Online]. Available: <https://www.injoit.org/index.php/j1/article/view/1566>
- [9] V. E. Ponomarev, “Forensic Examination of Digital Images Altered by AI: Applicability of New and Classical Analysis Methods”, *Legal Science*, no. 9, pp. 430–434, 2025. doi: 10.24412/2220-5500-2025-9-430-434.
- [10] S. P. Nikitenkova, “Application of Benford’s Law in Deepfake Image Detection”, *Tomsk State University Journal of Control and Computer Science*, no. 64, pp. 128–137, 2023. doi: 10.17223/19988605/64/13.
- [11] A. B. Mudrich and K. V. Ezhova, “Face Detection and Recognition in Video Surveillance Systems”, *International Journal of Open Information Technologies*, vol. 12, no. 3, pp. 82–87, 2024. [Online]. Available: <https://www.injoit.org/index.php/j1/article/view/1798>
- [12] N. Yu, L. Davis, and M. Fritz, “Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints”, in *Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, 2019, pp. 7555–7565. doi: 10.1109/ICCV.2019.00765.
- [13] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, “CNN-Generated Images Are Surprisingly Easy to Spot... for Now”, in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 8692–8701. doi: 10.1109/CVPR42600.2020.00872.
- [14] R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, and L. Verdoliva, “On the Detection of Synthetic Images Generated by Diffusion Models”, in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5. doi: 10.1109/ICASSP49357.2023.10095167.
- [15] Z. Wang et al., “DIRE for Diffusion-Generated Image Detection”, in *Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, 2023, pp. 22388–22398. doi: 10.1109/ICCV51070.2023.02051.
- [16] U. Ojha, Y. Li, and Y. J. Lee, “Towards Universal Fake Image Detectors that Generalize Across Generative Models”, in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 24480–24489. doi: 10.1109/CVPR52729.2023.02345.
- [17] M. Zhu et al., “GenImage: A Million-Scale Benchmark for Detecting AI-Generated Image”, in *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, Datasets and Benchmarks Track, 2023. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/hash/f4d4a021f9051a6c18183b059117e8b5-Abstract-Datasets_and_Benchmarks.html
- [18] J. J. Bird and A. Lotfi, “CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images”, *IEEE Access*, vol. 12, pp. 15642–15650, 2024. doi: 10.1109/ACCESS.2024.3356122.
- [19] J. M. Johnson and T. M. Khoshgoftaar, “Survey on deep learning with class imbalance”, *Journal of Big Data*, vol. 6, art. 27, 2019. doi: 10.1186/s40537-019-0192-5.
- [20] S. Rezvani and X. Wang, “A broad review on class imbalance learning techniques”, *Applied Soft Computing*, vol. 143, art. 110415, 2023. doi: 10.1016/j.asoc.2023.110415.
- [21] Y. Zhang, B. Kang, B. Hooi, S. Yan, and J. Feng, “Deep Long-Tailed Learning: A Survey”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10795–10816, 2023. doi: 10.1109/TPAMI.2023.3268118.
- [22] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal Loss for Dense Object Detection”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, 2020. doi: 10.1109/TPAMI.2018.2858826.
- [23] Y. Song, T. Wang, P. Cai, S. K. Mondal, and J. P. Sahoo, “A Comprehensive Survey of Few-shot Learning: Evolution, Applications, Challenges, and Opportunities”, *ACM Computing Surveys*, vol. 55, no. 13s, art. 271, 2023. doi: 10.1145/3582688.

- [24] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain Generalization: A Survey", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4396–4415, 2023. doi: 10.1109/TPAMI.2022.3195549.
- [25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition", in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.
- [26] A. Howard et al., "Searching for MobileNetV3", in *Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, 2019, pp. 1314–1324. doi: 10.1109/ICCV.2019.00140.
- [27] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks", in *Proc. 36th Int. Conf. on Machine Learning (ICML)*, PMLR, vol. 97, 2019, pp. 6105–6114. [Online]. Available: <https://proceedings.mlr.press/v97/tan19a.html>
- [28] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s", in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 11966–11976. doi: 10.1109/CVPR52688.2022.01167.
- [29] Kaggle. [Online]. Available: <https://www.kaggle.com/> (accessed: Aug. 12, 2026).
- [30] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization", in *Proc. 7th Int. Conf. on Learning Representations (ICLR)*, 2019. [Online]. Available: <https://openreview.net/forum?id=Bkg6RiCqY7>
- [31] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization", *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, 2020. doi: 10.1007/s11263-019-01228-7.
- [32] L. van der Maaten and G. Hinton, "Visualizing Data using t-SNE", *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008. [Online]. Available: <https://jmlr.org/papers/v9/vandermaaten08a.html>