Оригинальная статья / Original article

УДК 681.51

https://doi.org/10.21869/2223-1560-2025-29-4-53-69



# Гибридный двухуровневый метод автоматического выявления подмены лица оператора на изображении

**М.Д. Халеев** [1] ✉

[1] Санкт-Петербургский Федеральный исследовательский центр Российской академии наук
ул. Корпусная, д. 18, г. Санкт-Петербург 199178, Российская Федерация

✉ e-mail: Haleev.M@iias.spb.su

## Резюме

***Цель исследования:*** *Разработка гибридного двухуровневого метода для повышения как точности, так и устойчивости выявления подмены лица оператора на изображениях, что является актуальной задачей в условиях постоянного роста и усложнения угроз со стороны дипфейк-технологий.*

***Методы.*** *Предложена архитектура, объединяющая сверточную нейронную сеть EfficientNet для извлечения глубоких паттернов и ансамбль из четырех классификаторов. Эти классификаторы целенаправленно анализируют специфические группы признаков: экспертные, текстурные, статистические и основанные на координатах лицевых ориентиров, что позволяет выявлять конкретные артефакты синтеза. Для обучения и тестирования был сформирован обширный и репрезентативный комплексный набор данных объемом 34 000 изображений, включающий как сгенерированные дипфейки, так и публичные датасеты.*

***Результаты.*** *Экспериментально подтверждена высокая эффективность предложенного метода: точность составила 0,921, а F1-мера – 0,914. Эти показатели значительно превосходят результаты любой из моделей, использованных по отдельности, что доказывает ярко выраженный и практически значимый синергетический эффект от их объединения.*

***Заключение.*** *Работа демонстрирует, что синергия глубокого обучения и классических признаковых моделей позволяет создать действительно более надежный и точный детектор. Предложенный метод повышает общую точность и увеличивает надежность системы, эффективно компенсируя индивидуальные слабости отдельных классификаторов. Это подтверждает гипотезу о том, что сочетание способности нейросети извлекать сложные, неявные паттерны и способности признаковых моделей анализировать конкретные, заранее известные специфические артефакты (например, геометрические искажения) ведет к созданию более мощного и устойчивого детектора.*

***Ключевые слова:*** *подмена лиц; компьютерное зрение; искусственный интеллект; глубокое обучение; дипфейк; информационная безопасность.*

***Конфликт интересов:*** *Автор декларирует отсутствие явных и потенциальных конфликтов интересов, связанных с публикацией настоящей статьи.*

***Финансирование.*** *Исследования выполнены в рамках бюджетной темы FFZF-2025-0003.*

**Для цитирования:** Халеев М.Д. Гибридный двухуровневый метод автоматического выявления подмены лица оператора на изображении // Известия Юго-Западного государственного университета. 2025; 29(4): 53-69. https://doi.org/ 10.21869/-2223-1560-2025-29-4-53-69.

*Поступила в редакцию 14.10.2025* *Подписана в печать 19.11.2025* *Опубликована 22.12.2025*

# Hybrid two-level method for automatic detection of face substitution in an image

**Mikhail D. Haleev [1] ✉**

[1] St. Petersburg Federal Research Center of the Russian Academy of Sciences
18, Korpusnaya str., St. Petersburg 199178, Russian Federation

✉ e-mail: Haleev.M@iias.spb.su

## Abstract

***Purpose of research.*** *The development of a hybrid, two-level method to enhance both the accuracy and robustness of detecting operator face spoofing in images, which is a pressing issue given the constant growth and sophistication of threats from deepfake technologies.*

***Methods.*** *A novel architecture is proposed, combining the EfficientNet convolutional neural network for deep pattern extraction with an ensemble of four classifiers. These classifiers specifically analyze distinct feature groups: expert-based, textural, statistical, and those based on facial landmark coordinates, enabling the detection of specific synthesis artifacts. For training and testing, an extensive and representative dataset of 34,000 images was compiled, including deepfakes generated by several modern tools as well as public datasets.*

***Results.*** *The high efficacy of the proposed method was experimentally confirmed: accuracy reached 0.921 and the F1-score was 0.914. These metrics significantly surpass the performance of any of the individual models used separately, demonstrating a pronounced and practically significant synergistic effect from their combination.*

***Conclusion.*** *This work demonstrates that the synergy between deep learning and classical feature-based models allows for the creation of a genuinely more reliable and precise detector. The proposed method improves overall accuracy and enhances system robustness by effectively compensating for the individual weaknesses of separate classifiers. This validates the hypothesis that combining a neural network's ability to extract complex, implicit patterns with feature-based models' capacity to analyze specific, predefined artifacts (such as geometric distortions) leads to a more powerful and resilient detector.*

***Keywords:*** *face swapping; computer vision; artificial intelligence; deep learning; deepfake; information security.*

***Conflict of interest.*** *The Author declare the absence of obvious and potential conflicts of interest related to the publication of this article.*

***Funding.*** *Research was supported by Russian State Research FFZF-2025-0003.*

**For citation:** Haleev M. D. Hybrid two-level method for automatic detection of face substitution in an image. *Izvestiya Yugo-Zapadnogo gosudarstvennogo universiteta = Proceedings of the Southwest State University*. 2025; 29(4): 53-69 (In Russ.). https:// doi.org/10.21869/2223-1560-2025-29-4-53-69.

*Received 14.10.2025* *Accepted 19.11.2025* *Published 22.12.2025*

***

## Введение

Создание и распространение синтетических медиафайлов, в частности изображений с подмененными лицами (дипфейков), стало одной из ключевых проблем информационной безопасности. Алгоритмы на основе глубокого обучения достигли такого уровня реализма, что визуально отличить подделку от оригинала становится практически невозможно. Это создает значительные риски, связанные с распространением дезинформации, манипуляцией общественным мнением, мошенничеством и компрометацией личных данных, что требует разработки эффективных средств автоматического противодействия.

В предыдущем исследовании [1] был предложен метод, основанный на анализе геометрических признаков, извлеченных из лицевых ориентиров. Этот метод, используя модель персептрона, позволил подтвердить эффективность признакового анализа и достичь точности 0.682. Однако оставался значительный потенциал для дальнейшего повышения качества детекции за счет использования более сложных архитектур и комбинирования различных источников информации.

Настоящая работа является развитием предыдущих исследований и предлагает гибридный ансамблевый подход к обнаружению дипфейков. Ключевая гипотеза заключается в том, что максимальной эффективности можно достичь путем объединения сильных сторон двух парадигм: способности глубоких сверточных сетей самостоятельно извлекать сложные текстурные и высокоуровневые признаки из пиксельного пространства и способности классических моделей целенаправленно анализировать конкретные, заранее определенные аномалии (геометрические, статистические и т.д.). Для проверки этой гипотезы был разработан двухуровневый метод, объединяющий нейросетевую модель EfficientNet и несколько классификаторов, обученных на различных наборах признаков.

Далее в работе представлен обзор релевантных исследований в данной области, подробно описан предложенный метод, методика формирования набора данных и разделения выборок для обучения. В заключительных разделах анализируются результаты проведенных экспериментов и формулируются итоговые выводы о преимуществах разработанного гибридного подхода.

## Материалы и методы

### Анализ современного состояния исследования

Чтобы обозначить место данной работы среди существующих решений, был проведен анализ ключевых исследований в области обнаружения дипфейков. Современные подходы можно условно разделить на несколько категорий в зависимости от используемых методов и архитектур.

Одной из основополагающих работ является исследование [2], в котором

авторы классифицируют манипуляции с лицами на четыре типа: генерация, замена, модификация и изменение выражения. Эта таксономия помогает систематизировать как сами угрозы, так и методы их обнаружения. В работе [3] предлагается унифицированная архитектура для выявления всех типов дипфейков, которая достигает точности 98% на наборе данных FaceForensics++, демонстрируя возможность создания универсальных детекторов. Исследование [4] концентрируется непосредственно на задаче обнаружения изображений с замененными лицами, что наиболее близко к тематике нашей работы, и представляет сравнительный анализ различных архитектур.

Многие современные решения используют сложные нейросетевые архитектуры. Например, в [5] для анализа видео применяются признаки, извлеченные из трех различных архитектур-трансформеров (DaViT, iFormer и GPViT), что позволяет достигать точности до 97.72%. Другим эффективным подходом является использование ансамблей. В работе [6] предложен двухуровневый ансамбль AWARE-NET, объединяющий модели Xception, Res2Net101 и EfficientNet-B7 с помощью механизма адаптивного взвешивания, что значительно повышает устойчивость детектора. Для обработки видеоданных был предложен SFormer [7] – сквозной пространственно-временной трансформер, который эффективно моделирует как пространственные, так и временные зависимости. Некоторые исследователи также применяют графовые нейронные сети (GNN), как в работе [8], где модель DFGNN используется для агрегации информации и уточнения признаков узлов графа, что повышает интерпретируемость и обобщающую способность детектора.

В последнее время появляются подходы, использующие новейшие достижения в области генеративных моделей. Метод DiffusionFake [9] применяет предварительно обученную модель Stable Diffusion для реконструкции лиц, что заставляет детектор изучать более устойчивые и обобщенные признаки подделки. Аналогично, работа DeCLIP [10] первой использует мощные представления из большой визуально-языковой модели CLIP для локализации областей подделки на изображении. Другой инновационный подход, Real Appearance Modeling (RAM) [11], обучает автоэнкодер восстанавливать исходные лица из искусственно "искаженных" версий, что заставляет модель изучать устойчивое представление подлинных лиц и лучше обобщать его на неизвестные типы подделок. В работе [12] предложен метод JRC, который совмещает неконтролируемую реконструкцию с контролируемой классификацией, используя не только явные артефакты, но и скрытые представления генеративных несоответствий.

Другие исследователи отходят от анализа артефактов и фокусируются на альтернативных признаках. В [13] предлагается метод FreqBlender, который работает в частотной области, смешивая

частотные характеристики реальных и поддельных лиц. Развивая этот подход, метод FreqDebias [14] борется со "спектральным смещением", когда детекторы чрезмерно полагаются на определенные частотные диапазоны, вводя аугментацию "Forgery Mixup" для диверсификации частотных характеристик. В работе [15] акцент смещается на выявление семантических изменений (например, несоответствие возраста), а не специфических артефактов. Метод, описанный в [16], основан на декомпозиции изображения и анализе таких характеристик, как текстура и степень "естественности". Работа [17] представляет детектор TAD, который разделяет признаки на две взаимоисключающие группы – текстурные несоответствия и артефакты – для уменьшения взаимных помех. В [18] предлагается использовать пространственные несоответствия во взгляде (GazeForensics) в качестве биометрического признака для обнаружения подделок.

Помимо традиционного обнаружения, исследуются и смежные задачи. Например, метод TSOM [19] решает задачу последовательного обнаружения дипфейков, предсказывая упорядоченную последовательность манипуляций, что позволяет восстановить историю создания подделки. Другие работы выходят за рамки анализа только визуальных данных. Так, фреймворк ART-AVDF [20] является аудиовизуальным и использует артикуляционное представление, объединяя слуховой энкодер и энкодер для губ, что повышает надежность детекции. Существуют и проактивные методы защиты, такие как NullSwap [21], который не обнаруживает подделки, а "маскирует" исходные изображения, добавляя незаметные возмущения, чтобы предотвратить саму возможность качественной замены лица.

В то время как многие из упомянутых методов ориентированы на видеоданные или используют сложные мультимодальные подходы, наша работа сосредоточена исключительно на обнаружении подмены лиц в статичных изображениях, предлагая гибридный двухуровневый метод для повышения точности и надежности детекции.

### Набор данных

Центральной задачей при подготовке исследования было формирование комплексного набора данных, способного обеспечить как широту охвата, так и специфичность для обучения устойчивой модели детекции. Существующие публичные датасеты, несмотря на их объем, зачастую ограничены определенным набором методов генерации дипфейков. Модель, обученная исключительно на таких данных, рискует переобучиться на выявление артефактов, свойственных только этим методам, и показывать низкую эффективность на новых, неизвестных типах подделок.

Чтобы преодолеть это ограничение, был применен двухэтапный подход. На первом этапе была выполнена целевая генерация 6000 дипфейк-изображений для внесения в обучающую выборку кон-

тролируемого разнообразия. Для этой цели были задействованы три различных программных инструмента, на основе каждого из которых было сгенерировано по 2000 изображений: Faceswap v3 Segmind API, расширение Roop для Stable Diffusion и репозиторий Wuhuikai на GitHub. Использование нескольких генераторов было продиктовано необходимостью обучить модель распознавать более общие, фундаментальные признаки подделки, а не специфические артефакты одного инструмента.

Процесс генерации основывался на 6000 исходных и 6000 целевых изображениях из набора данных CelebA. Во избежание пересечений и для обеспечения чистоты данных эти два набора были сделаны полностью непересекающимися. Каждый из трех генераторов обработал по 2000 уникальных пар изображений, что позволило получить сбалансированную выборку из 6000 синтетических изображений (результат работы генератора показан на рис. 1). К ним было добавлено 8000 оригинальных фотографий для формирования первичного сбалансированного набора.

На втором этапе решалась задача масштабирования данных и обеспечения их обобщающей способности. Хотя наш сгенерированный набор и вносил разнообразие методов, его объем (14000 изображений) был недостаточен для обучения глубокой нейронной сети, устойчивой к переобучению. Именно поэтому первичный набор был объединен с 20 000 изображений из авторитетного тематического датасета DF40 [22]. Добавление данных из DF40 позволило не только значительно увеличить итоговый размер выборки, но и обогатить ее изображениями из другого источника, что критически важно для повышения способности модели к генерализации.

Таким образом, итоговый гибридный набор данных объемом 34000 изображений объединил в себе преимущества двух подходов: специфичность и разнообразие методов из контролируемой генерации и масштаб с широким охватом из устоявшегося публичного датасета.

**Рис. 1.** Входные оригинальные изображения (**1** и **2**) и изображение подмены лица с помощью Roop (**3**)

**Fig. 1.** Original input images (**1** and **2**) and the face-swapped image generated using Roop (**3**)

Для корректного обучения и объективной оценки метода была применена многоуровневая стратегия разделения данных, направленная на предотвращение утечки информации между базовыми моделями и мета-моделью (как показано на рис. 2).

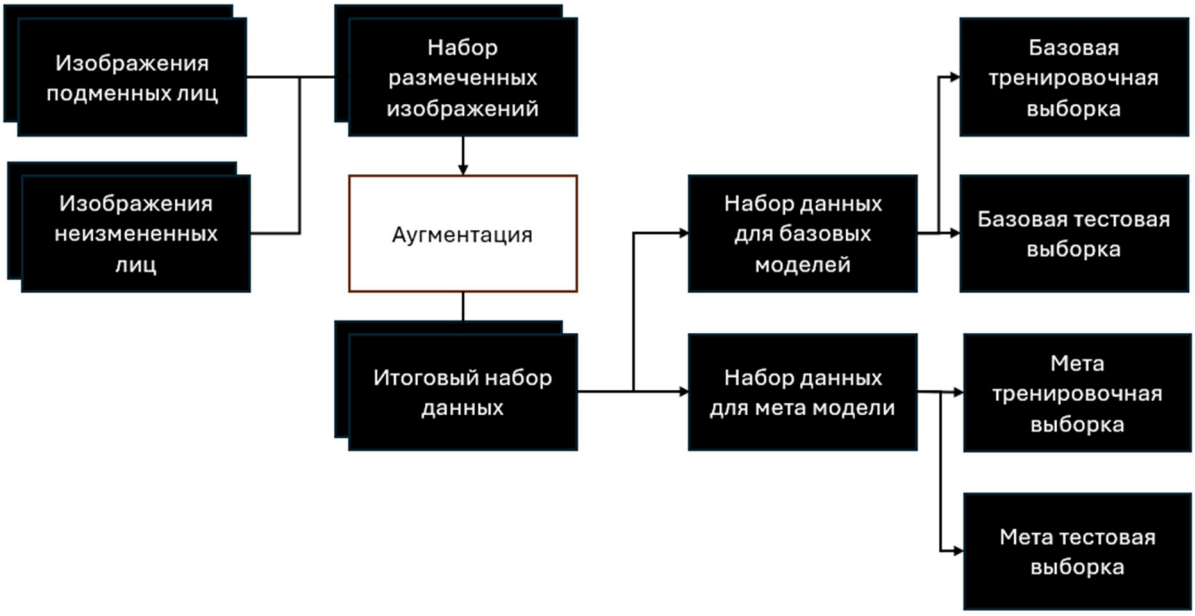


**Рис. 2.** Схема разделения набора данных для обучения и оценки

**Fig. 2.** Dataset splitting scheme for ensemble training and evaluation

Процесс был организован следующим образом:

• Формирование итогового набора таким образом, чтобы исходные изображения с подмененными и неизмененными лицами были объединены в единый размеченный набор данных. К этому набору был применен комплекс техник аугментации (повороты, отражения, изменение яркости) для увеличения его объема и разнообразия, что способствует повышению устойчивости моделей.

• Первичное разделение выполнено таким образом, что итоговый аугментированный набор данных был разделен на два основных непересекающихся подмножества:

– Набор данных для базовых моделей использовался исключительно для обучения и тестирования четырех признаковых классификаторов и глубокой нейронной сети.

– Набор данных для мета-модели является второй, полностью независимой частью, и был зарезервирован для обучения и финальной оценки мета-модели.

• Вторичное разделение выполнено так, что каждое из двух основных подмножеств было разделено на тренировочную и тестовую выборки в стандартной пропорции.

Такая строгая изоляция гарантирует, что мета-модель обучается на предсказаниях, которые базовые модели

сделали на совершенно новых для них данных. Это позволяет мета-классификатору эффективно учиться обобщать и корректировать ошибки базовых моделей, не подстраиваясь под артефакты их обучающей выборки, что ведет к созданию более надежной итоговой системы.

## Метод определения подмены лиц на изображении

Предложенный метод состоит из двух шагов. На первом шаге предполагается, что значимые для детекции артефакты проявляются в нарушении естественной геометрии лица и антропометрических соотношений, которые можно зафиксировать с помощью набора признаков. Второй шаг заключается в использовании сверточной нейронной сети (в данном случае EfficientNet), которая способна самостоятельно, без предварительной инженерии признаков, извлекать сложные паттерны и текстурные аномалии непосредственно из пиксельного пространства изображения.

Предложенный метод объединяет предсказания глубокой нейросетевой модели и четырех моделей, построенных на различных группах признаков. Для построения этих признаковых моделей рассматривались два основных алгоритма: случайный лес (Random Forest) и градиентный бустинг (CatBoost). Выбор итоговой модели для каждой из четырех групп признаков производился на основе экспериментальной оценки. В рамках метода предложено использовать тот классификатор, который продемонстрировал наибольшую прогностическую точность на соответствующем наборе данных, что позволило подобрать наиболее подходящую архитектуру для каждого типа признаков, максимизируя их индивидуальный вклад в итоговое решение.

Схема предложенного метода, включающая обучение базовых моделей и последующее обучение мета-модели на их предсказаниях, представлена на рис. 3.

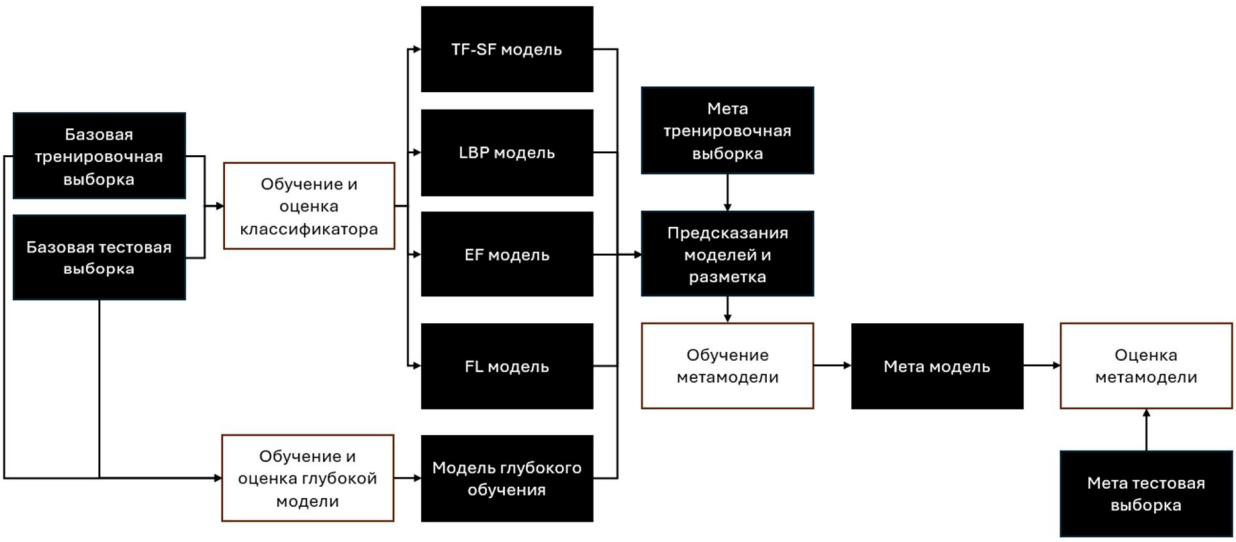


**Рис. 3.** Метод определения подмены лиц на изображении

**Fig. 3.** The proposed face spoofing detection method

В рамках метода были разработанные следующие модели:

1. Модели на основе экспертные признаках (EF), которые были рассчитаны на основе лицевых ориентиров, полученных с помощью фреймворка Google MediaPipe. Они включают в себя метрики, оценивающие форму лица, такие как его условная площадь и объем, а также набор соотношений ключевых межлицевых расстояний. В частности, анализировались отношения расстояний между глазами, носом, ртом и ушами для оценки пространственного расположения черт лица. Для классификации на основе этого набора признаков использовалась модель случайного леса.

2. Признаки на основе лицевых ориентиров (FL), которые использовались в качестве отдельного набора признаков использовались непосредственно координаты ключевых точек лица, извлеченные с помощью Google MediaPipe. Для этого набора также была обучена отдельная модель случайного леса.

3. Статистические и текстурные признаки (TF-SF) извлекались с помощью специализированных библиотек Python. Статистические признаки фиксируют распределение и изменчивость интенсивности пикселей, в то время как текстурные анализируют паттерны и градиенты изображения для выявления неестественных переходов и артефактов. Признаки были объединены в один набор, для которого была обучена модель на основе градиентного бустинга CatBoost.

4. Локальные бинарные паттерны (LBP) использовались для кодирования микротекстурной информации изображения путем сравнения значений пикселей в локальной окрестности. Этот подход обеспечивает детальное представление мелкозернистых структур, часто искажаемых при создании дипфейков. Для данного набора признаков была обучена модель CatBoost.

Итоговый сценарий работы системы, реализующей предложенный метод в режиме предсказания выглядит следующим образом (рис. 4): по одному входному изображению формируется вектор из пяти предсказаний (одно от глубокой модели EfficientNet и по одному от каждой из четырех описанных признаковых моделей). Этот вектор подается на вход мета-модели, также реализованной на основе CatBoost. Выбор CatBoost в качестве мета-классификатора обусловлен его способностью эффективно агрегировать предсказания от разнородных базовых моделей и строить сильную решающую границу. В результате мета-модель формирует окончательное, более точное и устойчивое предсказание.

## Результаты и их обсуждение

В данном разделе представлены результаты экспериментальной оценки эффективности предложенного гибридного подхода к детекции дипфейков. Исследование было разделено на два ключевых этапа. Сначала была проведена оценка производительности каждой из четырех

признаковых моделей по отдельности, чтобы определить базовое качество каждого набора признаков. Затем была проанализирована работа глубокой нейросетевой модели EfficientNet и итогового ансамбля, объединяющего все базовые классификаторы. В качестве основных метрик для всех экспериментов использовались точность (Accuracy) и F1-мера (F1-score).

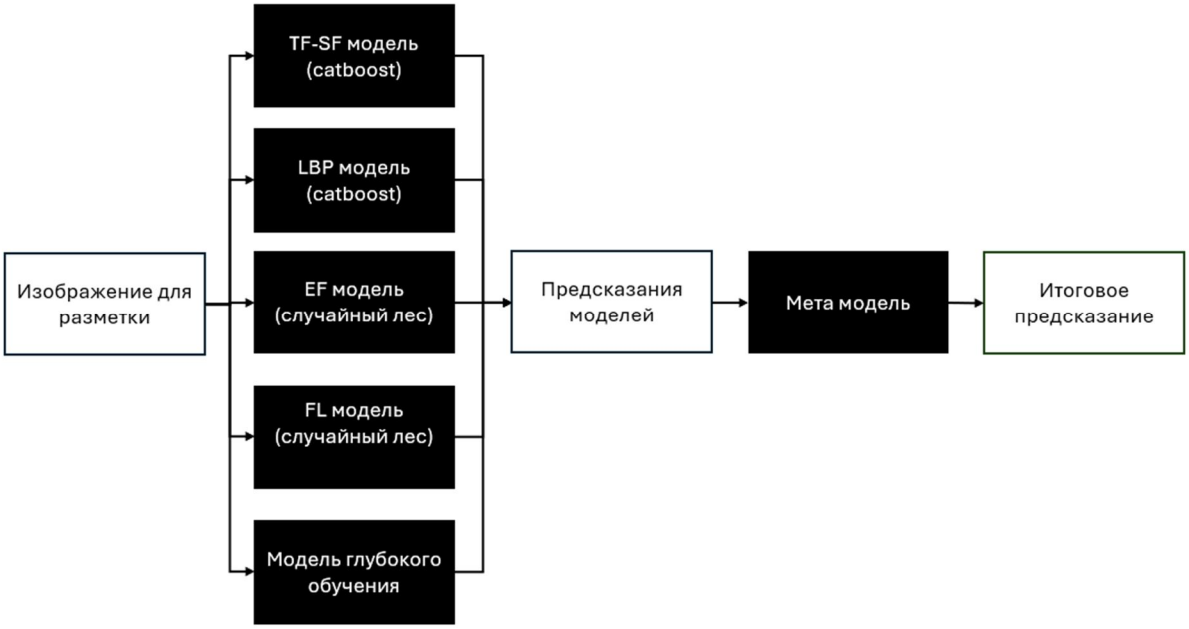


**Рис. 4.** Сценарий определения подмены лиц на изображении

**Fig. 4.** Face swap detection scenario

## Результаты тестирования признаковых моделей

В табл. 1 представлены результаты оценки эффективности четырех признаковых моделей, обученных для задачи детекции дипфейков. Для каждой модели были рассчитаны метрики Accuracy и F1-score.

Анализ полученных метрик показывает, что наилучшую производительность продемонстрировала модель, использующая в качестве признаков необработанные координаты лицевых ориентиров (facial landmarks). Данная модель достигла значения Accuracy 0,772 и F1-score 0,711, что является самым высоким показателем среди всех рассмотренных подходов. Этот результат свидетельствует о том, что прямое использование координат ключевых точек лица содержит наиболее сильный и явный сигнал для различения аутентичных и синтетических изображений.

Остальные три модели показали сопоставимые, хотя и более низкие, результаты. Модели, основанные на локальных бинарных паттернах (LBP) и объединенных текстурно-статистических признаках, продемонстрировали близкие значения Accuracy (0,727 и 0,716 соответственно), подтверждая важность анализа микротекстур и распределения пикселей для выявления аномалий.

Наименее эффективной оказалась модель, построенная на экспертных признаках (Accuracy 0,704, F1-score 0,614). Заметное снижение метрики F1-score по сравнению с Accuracy может указывать на сложности в классификации одного из классов. Можно предположить, что, хотя вычисление высокоуровневых признаков (площадей, объемов, соотношений) обеспечивает интерпретируемость, оно приводит к потере части важной информации, содержащейся в исходных координатах ориентиров.

В целом, полученные результаты, хоть и недостаточны для самостоятельного практического применения, подтверждают корректность выбора всех четырех наборов признаков. Каждый из них улавливает уникальные аспекты данных, что делает их ценными компонентами для последующего ансамблирования в рамках мета-модели.

**Таблица 1.** Метрики признаковых моделей

**Table 1.** Feature-Based Model Metrics

| Экспертные характеристики / Expert characteristics | | Лицевые ориентиры / Facial landmarks | | Текстурные и статистические признаки / Textural and statistical features | | Локальные бинарные паттерны / Local binary patterns | |
|---|---|---|---|---|---|---|---|
| Точность | F1-мера | Точность | F1-мера | Точность | F1-мера | Точность | F1-мера |
| 0,703874 | 0,61442 | 0,771722 | 0,71132 | 0,716349 | 0,661619 | 0,727293 | 0,684077 |

### Результаты тестирования предложенного метода

В табл. 2 представлены результаты экспериментов по оценке производительности глубокой модели (EfficientNet) и ансамбля, объединяющего ее с четырьмя признаковыми классификаторами. Исследовалось влияние различных методов предобработки изображений на итоговые метрики, а также эффективность самого ансамблевого подхода.

Для оптимизации вычислительных ресурсов и эффективной проверки гипотез эксперименты были разделены на два этапа. Первый, поисковый, этап включал в себя широкое тестирование всех семи методов предобработки на сокращенном наборе данных (половина от общего объема, half) с ограниченным числом эпох обучения (8). Такой подход позволил быстро выявить общие тенденции и определить наиболее перспективные методы, поскольку этой конфигурации было достаточно для выявления относительной эффективности подходов при меньших временных затратах. На втором, основном, этапе лучшие из выявленных методов (RGB и Grayscale) были подвергнуты полноценному обучению на полном наборе данных (full) в течение 25 эпох для достижения максимально возможного качества и точной оценки их потенциала.

**Таблица 2.** Метрики глубоких моделей и предложенного метода

**Table 2.** Metrics of deep models and the proposed method

| Формат обработки / Processing format | Число эпох / Number of epochs | Размер набора данных / Dataset size | Точность EfficientNet / Accuracy EfficientNet | F1-мера EfficientNet / / F1-Measure EfficientNet | Точность метода / Method accuracy | F1-мера метода / F1-method measure |
|---|---|---|---|---|---|---|
| rgb | 8 | 0,5 | 0,852046 | 0,8375 | 0,888159 | 0,878362 |
| grayscale | 8 | 0,5 | 0,85336 | 0,844186 | 0,883563 | 0,87156 |
| srm | 8 | 0,5 | 0,738236 | 0,735515 | 0,839352 | 0,818676 |
| dct | 8 | 0,5 | 0,644342 | 0,653148 | 0,81265 | 0,781298 |
| ela | 8 | 0,5 | 0,641935 | 0,672407 | 0,815933 | 0,784635 |
| svd | 8 | 0,5 | 0,603195 | 0,635578 | 0,811556 | 0,780077 |
| dft | 8 | 0,5 | 0,528781 | 0,577097 | 0,811337 | 0,778974 |
| rgb | 25 | 1 | 0,90545 | 0,897192 | 0,920552 | 0,914164 |
| grayscale | 25 | 1 | 0,904137 | 0,896601 | 0,913548 | 0,90673 |

Анализ результатов одиночной модели EfficientNet, особенно на полном наборе данных, показал, что наиболее эффективными являются базовые представления изображений. Модели, обученные на стандартных RGB-изображениях и их Grayscale-версиях, продемонстрировали наилучшие и практически идентичные результаты, достигнув Accuracy 0,905 и F1-score 0,897. Это говорит о том, что архитектура EfficientNet способна самостоятельно извлекать необходимые признаки из стандартного пиксельного пространства без необходимости в сложной предварительной обработке.

Напротив, методы предобработки, основанные на частотных преобразованиях (DCT, DFT) или анализе уровня ошибок (ELA), показали значительное снижение производительности. Это может быть связано с тем, что такие преобразования нарушают локальные пространственные зависимости в изображении, которые критически важны для сверточных нейронных сетей.

Ключевым выводом экспериментов является стабильное и повсеместное улучшение метрик при использовании метода. Во всех без исключения конфигурациях результирующая модель превзошла по качеству соответствующую ей одиночную глубокую модель.

Наивысший результат всего исследования был достигнут именно ансамблем, использующим глубокую модель на RGB-изображениях: Accuracy 0,921 и F1-score 0,914. Это представляет собой

значимое улучшение по сравнению с лучшей одиночной моделью.

Особенно ярко преимущество предложенного метода проявляется на слабых глубоких моделях. Например, для предобработки DFT, где EfficientNet показала Accuracy всего 0,528, метод смог поднять этот показатель до 0,811. Это демонстрирует, что признаковые модели вносят устойчивый и независимый вклад, компенсируя недостатки глубокой модели и делая итоговую систему более надежной.

Таким образом, эксперименты подтвердили, что гибридный подход, совмещающий глубокое обучение с классическими признаковыми моделями, позволяет достичь более высокой точности и устойчивости по сравнению с использованием исключительно глубокого ней-росетевого подхода.

## Выводы

В настоящей работе была поставлена и решена задача повышения эффективности автоматического выявления подмены лица на изображениях на статичных изображениях. В качестве развития предыдущих исследований, основанных на признаковых методах, был разработан и протестирован гибридный двухуровневый метод, объединяющий сильные стороны глубокого обучения и классических моделей машинного обучения.

Предложенная архитектура нейронной сети, в рамках разработанного метода состоит из пяти базовых моделей (сверточной нейронной сети EfficientNet и четырех классификаторов на разнородных наборах признаков: геометрических, текстурных, статистических и основанных на лицевых ориентирах) и мета-модели CatBoost, агрегирующей их предсказания. Эксперименты, проведенные с использованием строгой методологии разделения данных для предотвращения переобучения, убедительно продемонстрировали преимущество предложенного метода.

Ключевым результатом исследования является достижение итоговой точности 0.921 и F1-меры 0.914, что существенно превосходит показатели любой из моделей, использованных по отдельности. Было установлено, что предложенный метод не только повышает общую точность, но и значительно увеличивает надежность системы, эффективно компенсируя слабости отдельных классификаторов. Это подтверждает основную гипотезу работы о том, что синергия между способностью нейросети извлекать сложные паттерны из пикселей и способностью признаковых моделей анализировать специфические артефакты (такие как геометрические искажения) ведет к созданию более мощного и устойчивого детектора.

Таким образом, данное исследование вносит вклад в область информационной безопасности, предлагая валидированную и эффективную архитектуру для противодействия распространению синтетического контента.

В качестве направлений для будущей работы планируется дальнейшее

расширение набора данных за счет использования более разнообразных генераторов дипфейков, а также исследование других архитектур глубокого обучения в качестве базовой модели. Дополнительно, перспективным видится анализ различных мета-классификаторов и добавление новых групп признаков для дальнейшего повышения точности и обобщающей способности метода.

## Список литературы

1. Халеев М.Д. Интеллектуальный метод автоматического выявления подмены лица на изображении // Системы анализа и обработки данных. 2025. Т. 97, №1. С. 105-120.

2. Tolosana R., Vera-Rodriguez R., Fierrez J., Morales A., Ortega-Garcia J. Deepfakes and beyond: A Survey of face manipulation and fake detection // Information Fusion. 2020. Vol. 64. P. 131–148. https://doi: 10.1016/J.INFFUS.2020.06.014.

3. Nawaz M., Javed A., Irtaza A. A deep learning model for FaceSwap and face-reenactment deepfakes detection // Applied Soft Computing. 2024. Vol. 162. P. 111854. https://doi: 10.1016/J.ASOC.2024.111854.

4. Ding X. и др. Swapped face detection using deep learning and subjective assessment // EURASIP Journal on Information Security. 2020. Vol. 2020, № 1. https://doi: 10.1186/S13635-020-00109-8.

5. Essa E. Feature fusion Vision Transformers using MLP-Mixer for enhanced deepfake detection // Neurocomputing. 2024. Vol. 598. P. 128128. https://doi: 10.1016/ J.NEUCOM.2024.128128

6. Salman M., et al. AWARE-NET: Adaptive Weighted Averaging for Robust Ensemble Network in Deepfake Detection // Computer Vision and Pattern Recognition. 2025.

7. Kingra S., Aggarwal N., Kaur N. SFormer: An end-to-end spatio-temporal transformer architecture for deepfake detection // Forensic Science International: Digital Investigation. 2024. Vol. 51. P. 301817. https:// doi: 10.1016/J.FSIDI.2024.301817

8. Khalid F., Javed A., ul ain Q., Ilyas H., Irtaza A. DFGNN: An interpretable and generalized graph neural network for deepfakes detection // Expert Systems with Applications. 2023. Vol. 222. P. 119843. https://doi: 10.1016/J.ESWA.2023.119843

9. Sun K., et al. DiffusionFake: Enhancing Generalization in Deepfake Detection via Guided Stable Diffusion // 38th Conference on Neural Information Processing Systems (NeurIPS 2024). 2024.

10. Smeu S., Oneata E., Oneata D. DeCLIP: Decoding CLIP Representations for Deepfake Localization // Proceedings of the Winter Conference on Applications of Computer Vision (WACV). 2025. С. 149-159.

11. Tian J., et al. Real Appearance Modeling for More General Deepfake Detection // ECCV. 2025.

12. Yan B., Li C. T., Lu X. JRC: Deepfake detection via joint reconstruction and classification // Neurocomputing. 2024. Vol. 598. P. 127862. https://doi: 10.1016/J.NEUCOM.2024.127862

13. Li H., et al. FreqBlender: Enhancing DeepFake Detection by Blending Frequency Knowledge // 38th Conference on Neural Information Processing Systems (NeurIPS 2024). 2024.

14. Kashiani H., Talemi N. A., Afghah F. FreqDebias: Towards Generalizable Deepfake Detection via Consistency-Driven Frequency Debiasing // CVPR. 2025.

15. Zou M., et al. Semantic Contextualization of Face Forgery: A New Definition, Dataset, and Detection Method // Computer Vision and Pattern Recognition. 2025.

16. Chew C. J., et al. Preserving manipulated and synthetic Deepfake detection through face texture naturalness // Journal of Information Security and Applications. 2024. Vol. 83. P. 103798. https://doi: 10.1016/J.JISA.2024.103798.

17. Gao J., at al. Texture and artifact decomposition for improving generalization in deep-learning-based deepfake detection // Engineering Applications of Artificial Intelligence. 2024. Vol. 133. P. 108450. https://doi: 10.1016/J.ENGAPPAI.2024.108450

18. He Q., Peng C., Liu D., Wang N., Gao X. GazeForensics: DeepFake detection via gaze-guided spatial inconsistency learning // Neural Networks. 2024. Vol. 180. P. 106636. https://doi: 10.1016/J.NEUNET.2024.106636

19. Li Y., at al. Texture Shape and Order Matter: A New Transformer Design for Sequential DeepFake Detection // Proceedings of the Winter Conference on Applications of Computer Vision (WACV). 2025. P. 202-211.

20. Wang Y., Huang H. Audio–visual deepfake detection using articulatory representation learning // Computer Vision and Image Understanding. 2024. Vol. 248. P. 104133. https://doi: 10.1016/J.CVIU.2024.104133

21. Wang T., Cheng H., Zhang X., Wang Y. NullSwap: Proactive Identity Cloaking Against Deepfake Face Swapping // CVPR. 2025.

22. Yan Z., et al. DF40: Toward Next-Generation Deepfake Detection // arXiv preprint arXiv:2406.13495. 2024.

## References

1. Haleev M. D. Intelligent method for automatic detection of face substitution in an image. *Sistemy analiza i obrabotki dannykh = Analysis and Data Processing Systems*. 2025; 97(1): 105-120. (In Russ.).

2. Tolosana R., Vera-Rodriguez R., Fierrez J., Morales A., Ortega-Garcia J. Deepfakes and beyond: A Survey of face manipulation and fake detection. *Information Fusion.* 2020; 64: 131–148. https://doi: 10.1016/J.INFFUS.2020.06.014.

3. Nawaz M., Javed A., A. Irtaza A deep learning model for FaceSwap and face-reenactment deepfakes detection. *Appl Soft Comput.* 2024; 162: 111854. https://doi: 10.1016/J.ASOC.2024.111854.

4. Ding X., Raziei Z., Larson E. C., E Olinick. V., Krueger P., Hahsler M. Swapped face detection using deep learning and subjective assessment. *EURASIP J Inf Secur*. 2020; 2020(1). https://doi: 10.1186/S13635-020-00109-8.

5. Essa E. Feature fusion Vision Transformers using MLP-Mixer for enhanced deepfake detection. *Neurocomputing*. 2024; 598: 128128. https://doi: 10.1016/J.NEUCOM.2024.128128.

6. Muhammad Salman, Iqra Tariq, Mishal Zulfiqar, Muqadas Jalal, Sami Aujla, Sumbal Fatima, AWARE-NET: Adaptive Weighted Averaging for Robust Ensemble Network in Deepfake Detection. *Computer Vision and Pattern Recognition*. 2025.

7. Kingra S., N Aggarwal., Kaur N. SFormer: An end-to-end spatio-temporal transformer architecture for deepfake detection. *Forensic Science International: Digital Investigation*. 2024; 51: 301817. https://doi: 10.1016/J.FSIDI.2024.301817.

8. Khalid F., Javed A., ul ain Q., Ilyas H., Irtaza A. DFGNN: An interpretable and generalized graph neural network for deepfakes detection. *Expert Syst Appl*. 2023; 222: 119843. https://doi: 10.1016/J.ESWA.2023.119843.

9. Ke Sun, Shen Chen, Taiping Yao, Hong Liu, Xiaoshuai Sun, Shouhong Ding, Rongrong Ji, DiffusionFake: Enhancing Generalization in Deepfake Detection via Guided Stable Diffusion. *38th Conference on Neural Information Processing Systems (NeurIPS 2024)*. 2024.

10. Stefan Smeu, Elisabeta Oneata, Dan Oneata, DeCLIP: Decoding CLIP Representations for Deepfake Localization. *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*. 2025. P. 149-159.

11. Jiahe Tian, Cai Yu, Xi Wang, Peng Chen, Zihao Xiao, Jiao Dai, Jizhong Han, and Yesheng Chai, Real Appearance Modeling for More General Deepfake Detection. *ECCV*, 2025.

12. Yan B., Li C. T., Lu X. JRC: Deepfake detection via joint reconstruction and classification, *Neurocomputing*. 2024; 598: 127862. https://doi: 10.1016/J.NEUCOM.2024.127862.

13. Hanzhe Li, Jiaran Zhou, Yuezun Li, Baoyuan Wu, Bin Li, Junyu Dong, FreqBlender: Enhancing DeepFake Detection by Blending Frequency Knowledge. *38th Conference on Neural Information Processing Systems (NeurIPS 2024)*. 2024.

14. Hossein Kashiani, Niloufar Alipour Talemi, Fatemeh Afghah, FreqDebias: Towards Generalizable Deepfake Detection via Consistency-Driven Frequency Debiasing. CVPR, 2025.

15. Mian Zou, Baosheng Yu, Yibing Zhan, Siwei Lyu, and Kede Ma, Semantic Contextualization of Face Forgery: A New Definition, Dataset, and Detection Method. *Computer Vision and Pattern Recognition*, 2025.

16. Chew C. J., Lin Y. C., Chen Y. C., Fan Y. Y., Lee J. S. Preserving manipulated and synthetic Deepfake detection through face texture naturalness. *Journal of Information Security and Applications*. 2024; 83: 103798. https://doi: 10.1016/J.JISA.2024.103798.

17. Gao J., et al. Texture and artifact decomposition for improving generalization in deep-learning-based deepfake detection. Eng Appl Artif Intell. 2024; 133: 108450. https://doi: 10.1016/J.ENGAPPAI.2024.108450.

18. He Q., Peng C., Liu D., Wang N., Gao X. GazeForensics: DeepFake detection via gaze-guided spatial inconsistency learning. *Neural Networks*. 2024; 180: 106636. https://doi: 10.1016/J.NEUNET.2024.106636.

19. Yunfei Li, Yuezun Li, Xin Wang, Baoyuan Wu, Jiaran Zhou, Junyu Dong, Texture Shape and Order Matter: A New Transformer Design for Sequential DeepFake Detection. *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, 2025. P. 202-211.

20. Wang Y., H. Huang Audio–visual deepfake detection using articulatory representation learning. *Computer Vision and Image Understanding*. 2024; 248: 104133. https://doi: 10.1016/J.CVIU.2024.104133.

21. Tianyi Wang, Harry Cheng, Xiao Zhang, Yinglong Wang, NullSwap: Proactive Identity Cloaking Against Deepfake Face Swapping, CVPR, 2025.

22. Yan Z., et al. DF40: Toward Next-Generation Deepfake Detection. *arXiv preprint arXiv:2406.13495*, 2024.

---

## Информация об авторе / Information about the Author

**Халеев Михаил Дмитриевич,** кандидат технических наук, младший научный сотрудник, Санкт-Петербургский федеральный исследовательский центр Российской академии наук, г. Санкт-Петербург, Российская Федерация, e-mail: Haleev.M@iias.spb.su

**Mikhail D. Haleev,** Cand. of Sci. (Engineering), Junior Research Fellow, St. Petersburg Federal Research Center of the Russian Academy of Sciences, St. Petersburg, Russian Federation, e-mail: Haleev.M@iias.spb.su