ТЕХНИКАЛЫҚ ҒЫЛЫМДАР
ТЕХНИЧЕСКИЕ НАУКИ
TECHNICAL SCIENCE

CENTRAL ASIAN SCIENTIFIC JOURNAL
ISSN 3135-3061
2026. №4 (32), Том 2. С. 139–145
DOI: https://doi.org/10.5281/zenodo.23949135
Сайт журнала: https://cajournal.kz

УДК 004.056:004.932

МУЛЬТИМОДАЛЬНЫЙ ПОДХОД К ВЫЯВЛЕНИЮ DEEPFAKE-ВИДЕО: АРХИТЕКТУРА СИСТЕМЫ DEEPCHECK BOT И МЕТОДИКА ЕЕ ЭКСПЕРИМЕНТАЛЬНОЙ ОЦЕНКИ

Ж.К. Ербол, Н.Д. Жолдасбек
Satbayev University, Алматы, Казахстан

Ербол Жулдызай Камбарқызы — студентка 4 курса, кафедра «Кибербезопасность, обработка, и хранение информации», Satbayev University, Алматы, Казахстан.

Жолдасбек Назым Досжанқызы — студентка 4 курса, кафедра «Кибербезопасность, обработка, и хранение информации», Satbayev University, Алматы, Казахстан.

© Ж.К. Ербол, Н.Д. Жолдасбек, 2026

Аннотация: В статье рассматривается задача автоматического выявления видеозаписей, в которых лицо или речь человека изменены с помощью генеративных нейронных сетей (deepfake). Показано, что распространенные покадровые детекторы теряют надежность при обработке сжатого видео из мессенджеров и социальных сетей и при встрече с ранее не известными способами генерации. Предложена архитектура мультимодальной системы DeepCheck Bot, которая объединяет три независимых канала анализа: поиск пространственных артефактов кадра сверточной нейронной сетью, оценку межкадровой динамики мимики и проверку согласованности движения губ с речевым сигналом. Описано правило слияния решений, корректно учитывающее отсутствие или низкое качество звуковой дорожки. Разработана методика экспериментальной оценки на открытых наборах данных FaceForensics++, Celeb-DF, DFDC и FakeAVCeleb, включающая междатасетное тестирование и полный факторный эксперимент 2^3 для анализа влияния степени сжатия, числа анализируемых кадров и веса аудиоканала на качество обнаружения. Работа носит проектный характер: программный прототип и численные результаты испытаний являются предметом следующего этапа исследования.

Ключевые слова: deepfake, обнаружение подделок, мультимодальный анализ, аудиовизуальная синхронизация, сверточные нейронные сети, информационная безопасность, планирование эксперимента.

Введение

Генеративные модели глубокого обучения за несколько лет прошли путь от лабораторных экспериментов до общедоступных приложений. Начиная с генеративно-состязательных сетей [1] и заканчивая современными диффузионными моделями, такие методы позволяют подменить лицо на видеозаписи, изменить мимику или синтезировать речь конкретного человека. Получаемый контент принято называть deepfake – от сочетания английских слов deep learning (глубокое обучение) и fake (подделка).

Для информационной безопасности deepfake представляет угрозу сразу в нескольких направлениях. Поддельные видеообращения используются для распространения дезинформации и подрыва доверия к цифровым источникам [2]. Синтетическое изображение и голос руководителя применяются в схемах социальной инженерии, когда сотрудника убеждают перевести деньги или раскрыть конфиденциальные данные. Наконец, подмена

лица создает риски для систем удаленной биометрической идентификации, которые все шире используются в банковских и государственных сервисах, в том числе в Республике Казахстан.

Визуальное качество подделок уже достигло уровня, при котором человек не способен надежно отличить их от оригинала без специальных средств. Поэтому возникает потребность в автоматических системах проверки, которые были бы одновременно достаточно точными, устойчивыми к реальным условиям распространения видео и доступными обычному пользователю. Существующие решения, как показывают обзоры [3; 4], обычно удовлетворяют только части этих требований.

Цель настоящей работы – обосновать архитектуру мультимодальной системы выявления deepfake-видео DeepCheck Bot и разработать методику ее экспериментальной оценки. Для достижения цели решаются следующие задачи: анализ существующих подходов и выявление их ограничений; формулировка проверяемой гипотезы; проектирование модулей системы и правила объединения их решений; выбор наборов данных, метрик и плана эксперимента.

1. Анализ существующих подходов к обнаружению deepfake

Методы выявления синтетического видео можно условно разделить на четыре группы в зависимости от того, какую информацию они используют.

Покадровые методы на основе сверточных нейронных сетей. Сверточная сеть (Convolutional Neural Network, CNN) обучается отличать настоящие изображения лица от измененных по локальным признакам: неестественной текстуре кожи, размытию на границе вставленной области, ошибкам освещения. Публикация набора данных FaceForensics++ [5] показала, что сеть XceptionNet [6] достигает высокой точности на видео с умеренным сжатием. Существуют и облегченные архитектуры, например MesoNet [7], и капсульные сети [8]. Общий недостаток этой группы – каждый кадр рассматривается изолированно, а мелкие артефакты, на которые опирается модель, в значительной степени разрушаются при сильном сжатии.

Методы анализа временной динамики. Поддельное видео часто выглядит правдоподобно на отдельном кадре, но выдает себя в движении: мерцанием границы лица, неестественным морганием, рывками мимики. Сочетание сверточной сети с рекуррентной (CNN + RNN/LSTM) позволяет учитывать последовательность кадров [9]. Модели на основе механизма внимания (Transformer) [10] дают возможность анализировать более длинный временной контекст, однако требуют больших вычислительных ресурсов.

Мультимодальные методы. При подмене лица или синтезе речи нарушается естественная связь между звуком и артикуляцией. Модель SyncNet [11] обучается оценивать, насколько движения губ соответствуют звуковому сигналу, а работа [12] показала, что признаки, связанные с движением рта, обобщаются на новые методы генерации лучше, чем признаки текстуры. Для обучения таких моделей созданы аудиовизуальные наборы данных, например FakeAVCeleb [13].

Методы цифровой криминалистики. Эта группа анализирует шум сенсора камеры, спектральные характеристики и следы повторного сжатия [3]. Такие признаки полезны как дополнительный источник информации, но сами по себе чувствительны к обработке видео платформами.

Сопоставление групп методов приведено в таблице 1.

Таблица 1 – Сравнение основных подходов к обнаружению deepfake

Подход	Источник признаков	Достоинства	Ограничения
CNN (Xception, MesoNet) [5–7]	Отдельный кадр	Высокая точность на знакомых данных, простота обучения	Чувствительность к сжатию, слабое обобщение на новые генераторы
CNN + RNN/LSTM [9]	Последовательность кадров	Учитывает моргание и динамику мимики	Ограниченное временное окно, больше вычислений

Подход	Источник признаков	Достоинства	Ограничения
Transformer [10]	Длинная последовательность	Широкий временной контекст	Высокая ресурсоемкость, большой объем обучающих данных
Аудиовизуальные методы [11–13]	Губы и речь	Устойчивость к сжатию, лучшее обобщение	Требуют речи в кадре и приемлемого качества звука
Цифровая криминалистика [3]	Шум, спектр, следы сжатия	Не зависит от содержания кадра	Признаки стираются при перекодировании

Анализ литературы позволяет выделить четыре нерешенные проблемы. Во-первых, детекторы, обученные на одном наборе данных, заметно теряют точность на подделках, созданных другими методами; в обзорах [3; 4] это явление называют проблемой обобщения. Во-вторых, большинство решений использует один канал информации – только изображение или только звук. В-третьих, видео в мессенджерах и социальных сетях проходит повторное сжатие, которое уничтожает часть признаков. В-четвертых, пользователь обычно получает лишь итоговую вероятность без объяснения, что вызывает недоверие к результату.

Сравнение доступных онлайн-сервисов, выполненное авторами на предыдущем этапе работы, показало схожую картину: бесплатные сервисы ошибаются на качественных подделках, а решения с проверкой звука и изображения, как правило, коммерческие и закрытые. Это подтверждает практическую потребность в открытом и простом инструменте, объединяющем несколько каналов анализа.

2. Постановка задачи и гипотеза исследования

Объектом исследования являются цифровые видеозаписи, содержащие изображение лица человека и, при наличии, его речь. Предметом исследования являются методы извлечения и объединения пространственных, временных и аудиовизуальных признаков синтетической модификации видео.

Научная проблема состоит в противоречии между быстрым развитием методов генерации реалистичного видеоконтента и недостаточной устойчивостью существующих детекторов к новым видам подделок и к сжатию видео при распространении.

Рабочая гипотеза: объединение трех независимых групп признаков – пространственных, временных и аудиовизуальных – позволит повысить качество обнаружения по сравнению с покадровым детектором, причем выигрыш будет наибольшим на сильно сжатом видео, где пространственные артефакты деградируют, а нарушение синхронности губ и речи сохраняется. Гипотеза является предположением и подлежит экспериментальной проверке по методике, изложенной в разделе 4.

3. Архитектура системы DeepCheck Bot

DeepCheck Bot проектируется как сервис, в который пользователь отправляет видеофайл или ссылку через Telegram-бот либо веб-форму и получает заключение о вероятности подделки. Обобщенная архитектура системы показана на рисунке 1.

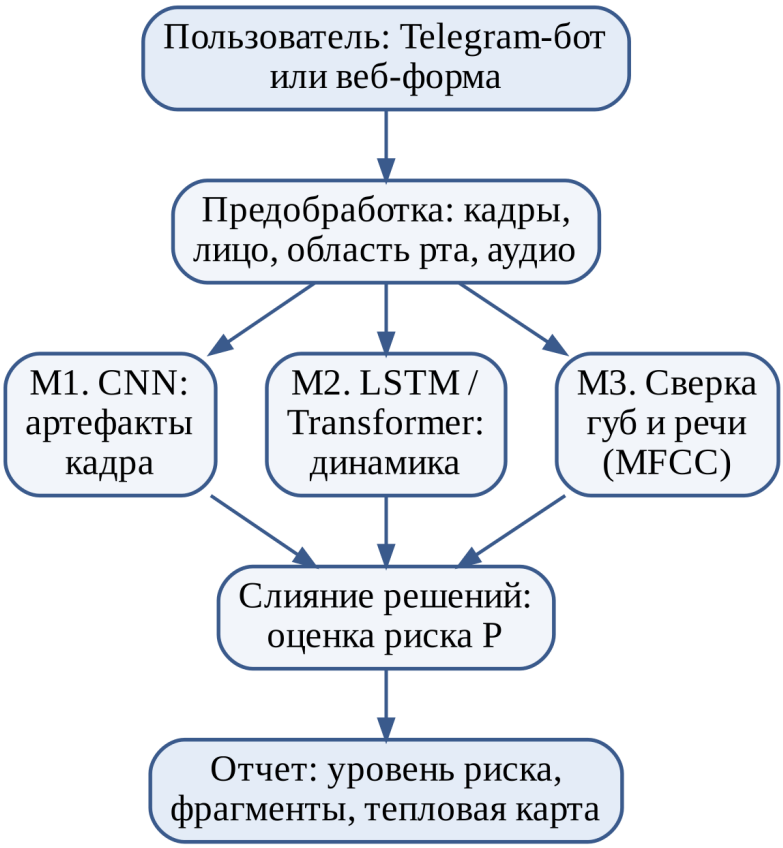


Рисунок 1 – Обобщенная архитектура системы DeepCheck Bot (разработано авторами)

Модуль предобработки декодирует видео и выбирает N кадров с равным шагом. На каждом кадре выполняется обнаружение и выравнивание лица по ключевым точкам, например каскадной сетью MTCNN [14], и выделяется область рта. Звуковая дорожка извлекается отдельно и преобразуется в последовательность мел-частотных кепстральных коэффициентов (MFCC). На этом же этапе проверяется наличие речи и оценивается качество звука.

Модуль M1 (пространственный анализ) использует сверточную сеть, предобученную на изображениях и дообученную на кадрах лица, например EfficientNet [15]. Для каждого кадра он вычисляет вероятность подделки, а итоговая оценка p_s получается усреднением по кадрам. Для объяснения решения строится тепловая карта значимых областей методом Grad-CAM [16].

Модуль M2 (временной анализ) получает последовательность векторов признаков кадров от M1 и обрабатывает ее двунаправленной LSTM-сетью или облегченным Transformer-кодировщиком. Выход модуля – вероятность p_t, отражающая аномалии динамики: неестественную частоту моргания, мерцание границы лица, рывки мимики.

Модуль M3 (аудиовизуальная сверка) сравнивает последовательность признаков области рта v_t с последовательностью звуковых признаков a_t, приведенных к общему пространству. По аналогии с [11] вычисляется среднее расстояние между ними при временном сдвиге Δ:

$$D(\Delta) = (1/T) \cdot \Sigma_t \, \| a_t - v_{t+\Delta} \|, \qquad (1)$$

где T – число временных шагов, Δ – сдвиг в кадрах. Для естественной речи минимум D(Δ) выражен и находится около нулевого сдвига, а при подмене лица или голоса кривая D(Δ) становится плоской. Мерой уверенности синхронизации служит разность $C = \mathrm{med}_\Delta D(\Delta) - \min_\Delta D(\Delta)$; малые значения C преобразуются в высокую вероятность подделки p_a.

Модуль слияния решений формирует итоговую оценку риска как взвешенную сумму оценок модулей:

$$P = (w_s \cdot p_s + w_t \cdot p_t + m \cdot w_a \cdot p_a) / (w_s + w_t + m \cdot w_a), \qquad (2)$$

где w_s, w_t, w_a – веса модулей, подбираемые на валидационной выборке; m – индикатор пригодности звука (m = 1, если в видео есть речь достаточного качества, иначе m = 0). Нормировка в знаменателе важна: отсутствие звука не должно трактоваться ни как признак подделки, ни как признак подлинности – система просто опирается на оставшиеся каналы и сообщает пользователю об ограничении.

Итоговая оценка P сопоставляется с двумя порогами $\tau_1 < \tau_2$ и переводится в один из трех уровней: «низкий риск», «требуется дополнительная проверка», «высокий риск». Пользователь получает не только уровень риска, но и вклад каждого модуля, временные отрезки с наибольшими аномалиями и тепловую карту. Такое заключение носит вероятностный характер и не может рассматриваться как самостоятельное доказательство подделки. Последовательность обработки сведена в таблицу 2.

Таблица 2 – Этапы обработки видео в системе DeepCheck Bot

Этап	Входные данные	Результат
1. Прием запроса	Видеофайл или ссылка	Файл и метаданные
2. Предобработка	Видео- и аудиопоток	Кадры лица, область рта, MFCC, индикатор m
3. Модуль M1	Кадры лица	Оценка p_s, тепловая карта
4. Модуль M2	Признаки кадров	Оценка p_t
5. Модуль M3	Область рта и речь	Оценка p_a
6. Слияние	p_s, p_t, p_a, m	Итоговая оценка P и уровень риска
7. Отчет	P и данные модулей	Заключение для пользователя

4. Методика экспериментальной оценки

Данные. Для обучения и проверки планируется использовать открытые наборы FaceForensics++ [5], Celeb-DF [17], DFDC [18] и FakeAVCeleb [13]. Последний необходим для модуля M3, так как содержит видео с поддельным изображением, поддельным звуком и их сочетанием. Разбиение на обучающую, валидационную и тестовую выборки выполняется по личностям: один и тот же человек не должен встречаться в разных частях выборки, иначе оценка качества окажется завышенной.

Метрики. Качество классификации оценивается долей правильных ответов (Accuracy), точностью (Precision), полнотой (Recall), F1-мерой и площадью под ROC-кривой (AUC):

$$\text{Precision} = TP / (TP + FP), \quad \text{Recall} = TP / (TP + FN), \tag{3}$$

$$F1 = 2 \cdot \text{Precision} \cdot \text{Recall} / (\text{Precision} + \text{Recall}), \tag{4}$$

где TP – верно обнаруженные подделки, FP – настоящие видео, ошибочно признанные поддельными, FN – пропущенные подделки. Для сервиса проверки важны оба вида ошибок: ложная тревога подрывает доверие к настоящему видео, а пропуск подделки оставляет пользователя без защиты. Дополнительно фиксируются время обработки одной минуты видео и потребление памяти.

Сравнение вариантов. Эксперимент проводится по схеме последовательного наращивания: базовый покадровый детектор (M1), затем M1 + M2 и полная система M1 + M2 + M3. Все варианты проверяются на одной и той же тестовой выборке. Для оценки обобщающей способности выполняется междатасетный тест: обучение на FaceForensics++ и проверка на Celeb-DF и DFDC без дообучения.

Факторный эксперимент. Чтобы проверить ключевую часть гипотезы – что аудиоканал особенно полезен на сжатом видео, – планируется полный факторный эксперимент 2^3 [19]. Откликом служит F1-мера, факторами – степень сжатия видео, число анализируемых кадров и вес аудиоканала. Степень сжатия задается параметром CRF кодека H.264: чем он выше, тем сильнее сжатие. Уровни факторов приведены в таблице 3.

Таблица 3 – Факторы и уровни варьирования

Фактор	Нижний уровень (−1)	Верхний уровень (+1)
x_1 – степень сжатия, CRF	23	35
x_2 – число анализируемых кадров N	16	48
x_3 – вес аудиоканала w_a	0,2	0,6

Значение CRF = 23 соответствует видео почти без видимых потерь, CRF = 35 – сильному сжатию, характерному для пересылки через мессенджеры. План эксперимента включает 8 опытов – все сочетания уровней трех факторов, каждый опыт повторяется трижды на разных подвыборках тестовых данных. По результатам строится уравнение регрессии с учетом взаимодействий:

$$Y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_{12}x_1x_2 + b_{13}x_1x_3 + b_{23}x_2x_3 + b_{123}x_1x_2x_3, \quad (5)$$

где коэффициенты вычисляются как $b_i = (1/8)\cdot\Sigma\ x_{iu}\cdot\bar{y}_u$, а $\bar{y}_u$ – средний отклик в u-м опыте. Однородность дисперсий проверяется по критерию Кохрена, значимость коэффициентов – по критерию Стьюдента, адекватность модели – по критерию Фишера. Ортогональность плана гарантирует, что оценки коэффициентов не зависят друг от друга.

Особый интерес представляет коэффициент взаимодействия b_{13}. Если гипотеза верна, он окажется значимым и положительным: увеличение веса аудиоканала будет давать тем больший прирост F1, чем сильнее сжато видео. Значимый отрицательный b_1 при этом подтвердит деградацию качества при сжатии, а знак и величина b_2 покажут, оправдано ли увеличение числа анализируемых кадров с учетом роста времени обработки.

5. Обсуждение и ограничения

Предлагаемый подход опирается на идею, что разные виды признаков разрушаются при сжатии и меняются при появлении новых генераторов в разной степени. Мелкие текстурные артефакты наиболее уязвимы, тогда как нарушение синхронности речи и артикуляции является свойством содержания видео, а не его качества. Поэтому объединение каналов потенциально повышает устойчивость системы, а раздельные оценки модулей делают заключение более понятным для пользователя.

Вместе с тем у подхода есть ограничения. Модуль М3 работает только при наличии речи говорящего в кадре; для видео без звука или с закадровым голосом система опирается на два визуальных модуля. Последовательная обработка тремя моделями увеличивает время ответа, что требует оптимизации для формата экспресс-проверки в Telegram-боте. Генераторы, специально обученные обходить детекторы, со временем могут научиться воспроизводить и синхронность губ, поэтому модель нуждается в регулярном дообучении на новых подделках. Наконец, тепловая карта показывает, на какие области опиралась модель, но не доказывает причинно-следственную связь между ними и фактом подделки.

Важно подчеркнуть, что настоящая работа представляет проектное обоснование системы. Утверждать о превосходстве DeepCheck Bot над существующими решениями можно будет только после реализации прототипа и проведения описанных экспериментов.

Заключение

В работе проанализированы основные подходы к выявлению deepfake-видео и показано, что их ключевыми ограничениями являются слабая обобщающая способность, зависимость от качества видео и использование единственного канала информации. Предложена архитектура мультимодальной системы DeepCheck Bot, объединяющей пространственный, временной и аудиовизуальный анализ, и правило слияния решений, корректно учитывающее отсутствие пригодного звука. Разработана методика экспериментальной оценки, включающая междатасетное тестирование и полный факторный эксперимент 2^3, позволяющий количественно оценить влияние сжатия, числа кадров и веса аудиоканала на качество обнаружения.

Дальнейшая работа включает программную реализацию прототипа, обучение модулей на открытых наборах данных, проведение запланированных экспериментов и проверку сформулированной гипотезы. Практическая ценность результатов состоит в создании

доступного инструмента проверки видео для пользователей, журналистов и специалистов по информационной безопасности.

Список литературы:

1. Goodfellow I., Pouget-Abadie J., Mirza M. [et al.]. Generative Adversarial Nets // Advances in Neural Information Processing Systems. – 2014. – Vol. 27. – P. 2672–2680.

2. Kietzmann J., Lee L. W., McCarthy I. P., Kietzmann T. C. Deepfakes: Trick or Treat? // Business Horizons. – 2020. – Vol. 63, № 2. – P. 135–146. – DOI: 10.1016/j.bushor.2019.11.006.

3. Verdoliva L. Media Forensics and DeepFakes: An Overview // IEEE Journal of Selected Topics in Signal Processing. – 2020. – Vol. 14, № 5. – P. 910–932. – DOI: 10.1109/JSTSP.2020.3002101.

4. Tolosana R., Vera-Rodriguez R., Fierrez J. [et al.]. DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection // Information Fusion. – 2020. – Vol. 64. – P. 131–148. – DOI: 10.1016/j.inffus.2020.06.014.

5. Rössler A., Cozzolino D., Verdoliva L. [et al.]. FaceForensics++: Learning to Detect Manipulated Facial Images // Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). – 2019. – P. 1–11. – DOI: 10.1109/ICCV.2019.00009.

6. Chollet F. Xception: Deep Learning with Depthwise Separable Convolutions // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). – 2017. – P. 1251–1258.

7. Afchar D., Nozick V., Yamagishi J., Echizen I. MesoNet: a Compact Facial Video Forgery Detection Network // Proceedings of the IEEE International Workshop on Information Forensics and Security (WIFS). – 2018. – P. 1–7.

8. Nguyen H. H., Yamagishi J., Echizen I. Capsule-Forensics: Using Capsule Networks to Detect Forged Images and Videos // Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – 2019. – P. 2307–2311.

9. Sabir E., Cheng J., Jaiswal A. [et al.]. Recurrent Convolutional Strategies for Face Manipulation Detection in Videos // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). – 2019. – P. 80–87.

10. Vaswani A., Shazeer N., Parmar N. [et al.]. Attention Is All You Need // Advances in Neural Information Processing Systems. – 2017. – Vol. 30. – P. 5998–6008.

11. Chung J. S., Zisserman A. Out of Time: Automated Lip Sync in the Wild // Computer Vision – ACCV 2016 Workshops. – Cham: Springer, 2017. – (Lecture Notes in Computer Science; Vol. 10117). – P. 251–263.

12. Haliassos A., Vougioukas K., Petridis S., Pantic M. Lips Don't Lie: A Generalisable and Robust Approach to Face Forgery Detection // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). – 2021. – P. 5039–5049.

13. Khalid H., Tariq S., Kim M., Woo S. S. FakeAVCeleb: A Novel Audio-Video Multimodal Deepfake Dataset [Электронный ресурс] // arXiv. – 2021. – URL: https://arxiv.org/abs/2108.05080 (дата обращения: 04.11.2026).

14. Zhang K., Zhang Z., Li Z., Qiao Y. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks // IEEE Signal Processing Letters. – 2016. – Vol. 23, № 10. – P. 1499–1503.

15. Tan M., Le Q. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks // Proceedings of the 36th International Conference on Machine Learning (ICML). – 2019. – Vol. 97. – P. 6105–6114.

16. Selvaraju R. R., Cogswell M., Das A. [et al.]. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization // Proceedings of the IEEE International Conference on Computer Vision (ICCV). – 2017. – P. 618–626.

17. Li Y., Yang X., Sun P., Qi H., Lyu S. Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). – 2020. – P. 3207–3216.

18. Dolhansky B., Bitton J., Pflaum B. [et al.]. The DeepFake Detection Challenge (DFDC) Dataset [Электронный ресурс] // arXiv. – 2020. – URL: https://arxiv.org/abs/2006.07397 (дата обращения: 04.11.2026).

19. Montgomery D. C. Design and Analysis of Experiments. – 9th ed. – Hoboken: Wiley, 2017. – 640 p.