ТЕХНИКАЛЫҚ ҒЫЛЫМДАР
ТЕХНИЧЕСКИЕ НАУКИ
TECHNICAL SCIENCE

CENTRAL ASIAN SCIENTIFIC JOURNAL
ISSN 3135-3061
2026. №4 (32), Том 3. С. 54–57
DOI: https://doi.org/10.5281/zenodo.24557240
Сайт журнала: https://cajournal.kz

УДК 004.932

СТАТИСТИЧЕСКИ ОБОСНОВАННАЯ ОЦЕНКА КРОСС-ДОМЕННОЙ РОБАСТНОСТИ МОДЕЛЕЙ ОБНАРУЖЕНИЯ DEEPFAKE-КОНТЕНТА

М.А. Косолапов
Международный университет Астана, Астана, Казахстан

Косолапов Максим Александрович — магистрант 2 курса, Высшая школа информационных технологий и инженерии, Международный университет Астана, Астана, Казахстан.

© М.А. Косолапов, 2026

Аннотация: Модели обнаружения deepfake-контента показывают высокую точность на данных, близких к обучающим, и резко теряют ее на новых наборах данных. При этом выводы о сравнительной робастности моделей часто делаются по точечным покадровым метрикам без учета дисбаланса классов и зависимости кадров одного видео. В статье предложен протокол статистической оценки, включающий сравнение с тривиальным классификатором, доверительные интервалы на основе кластерного bootstrap по видео, парное сравнение моделей и анализ калибровки. Протокол применен к трем парадигмам (CNN, Vision Transformer, спектральный анализ), обученным на FaceForensics++ и протестированным на Celeb-DF. Показано, что учет кластерной структуры расширяет доверительные интервалы в 3–5 раз и меняет выводы о значимости различий между моделями.

Ключевые слова: deepfake, обнаружение подделок, робастность, кросс-доменная оценка, кластерный bootstrap, доверительный интервал, тест DeLong, калибровка

Развитие генеративных моделей сделало синтез поддельных изображений и видео лиц (deepfake) доступным широкому кругу пользователей, что создает риски дезинформации и мошенничества. Для обнаружения подделок предложено множество детекторов, которые на эталонных наборах данных, таких как FaceForensics++ (FF++) [1], достигают площади под ROC-кривой (AUC) выше 0,95. Однако при тестировании на данных из другого источника, например на наборе Celeb-DF [2], качество большинства детекторов снижается почти до случайного уровня [3, 4]. Поэтому ключевым свойством детектора считается кросс-доменная робастность — способность сохранять качество на данных, не участвовавших в обучении.

Сравнение детекторов по робастности обычно проводится по точечным значениям метрик, рассчитанным по отдельным кадрам. Такой подход имеет два недостатка. Во-первых, наборы данных несбалансированы: поддельных видео в них в 2–6 раз больше, чем настоящих, поэтому Accuracy и F1-score могут быть высокими даже у бесполезного классификатора [5]. Во-вторых, кадры одного видео сильно коррелированы (общее лицо, освещение и метод подделки), а стандартные статистические процедуры предполагают независимость наблюдений. Цель данной работы — предложить и апробировать протокол статистически обоснованной оценки кросс-доменной робастности, учитывающий обе особенности.

Эксперимент проводился для трех парадигм обнаружения: сверточной сети EfficientNet-B4 (CNN) [6], трансформера ViT-B/16 [7] и спектрального подхода, в котором

логарифмический амплитудный спектр двумерного преобразования Фурье обрабатывается сетью ResNet-18 [8]. Чтобы сравнение отражало качество признаков, а не емкость классификатора, все экстракторы использовались с замороженными весами ImageNet, а поверх извлеченных векторов обучался одинаковый многослойный перцептрон. Обучение проводилось на подвыборке FF++ (сжатие c23). Тестирование выполнялось на Celeb-DF (v2) без дообучения: 518 видео (178 настоящих и 340 поддельных), по 30 кадров с лицом, выделенным детектором MTCNN, всего 15 540 кадров. На этапе аудита данных было обнаружено пересечение обучающей и валидационной частей FF++, поэтому внутридоменные оценки из анализа исключены, и все выводы сделаны по полностью независимому набору Celeb-DF.

Протокол оценки включает четыре элемента. Первый — сравнение с тривиальным классификатором, относящим любой кадр к классу «fake». При доле поддельных кадров π такой классификатор получает:

$$F1_{triv} = \frac{2\pi}{1+\pi}, BAcc_{triv} = 0{,}5, MCC_{triv} = 0 \quad (1)$$

где BAcc — сбалансированная точность (среднее чувствительности и специфичности), MCC — коэффициент корреляции Мэтьюса, устойчивый к дисбалансу классов [5]. Второй элемент — доверительные интервалы на основе кластерного bootstrap [9, 10]. На каждой из B = 1000 итераций с возвращением выбираются целые видео (отдельно среди настоящих и поддельных), метрика θ пересчитывается по всем их кадрам, и интервал строится по перцентилям бутстреп-распределения:

$$CI_{1-\alpha} = \left[\theta^{*}_{(\alpha/2)}; \theta^{*}_{(1-\alpha/2)}\right] \quad (2)$$

Влияние кластеризации оценивается дизайн-эффектом — отношением дисперсий кластерной и наивной (покадровой) бутстреп-оценок, через которое определяется эффективный размер выборки:

$$DEFF = \frac{Var_{кл}(\theta^{*})}{Var_{наив}(\theta^{*})}, N_{eff} = \frac{N}{DEFF} \quad (3)$$

Третий элемент — парное сравнение моделей. Тест DeLong [11] для коррелированных ROC-кривых применяется на уровне кадров и на уровне видео (после усреднения вероятностей), а основным критерием служит парный кластерный bootstrap разности AUC. Четвертый элемент — анализ калибровки по ожидаемой ошибке калибровки (ECE) [12] и проверка переноса порога решения, выбранного на обучающем домене.

Результаты оценки на Celeb-DF представлены в Таблице 1.

Таблица 1 – Метрики моделей на наборе Celeb-DF (уровень кадров, порог 0,5)

Модель	AUC	95% ДИ (по кадрам)	95% ДИ (по видео)	DEFF	F1	BAcc	MCC
Тривиальный	0,500	–	–	–	0,793	0,500	0,000
CNN	0,537	[0,527; 0,547]	[0,492; 0,582]	22,9	0,777	0,516	0,059
ViT	0,603	[0,593; 0,612]	[0,558; 0,649]	24,5	0,795	0,511	0,106
Spectral	0,497	[0,487; 0,507]	[0,463; 0,529]	12,0	0,789	0,504	0,028

Значения F1-score всех моделей (0,777–0,795) совпадают с F1 тривиального классификатора (0,793), а сбалансированная точность не превышает 0,516. Следовательно, высокий F1 в кросс-доменном тесте отражает лишь долю поддельных кадров в наборе. Дизайн-эффект составил от 12 до 25: 15 540 кадров по объему информации эквивалентны примерно 630–1300 независимым наблюдениям, что сопоставимо с числом видео. В результате кластерные интервалы в 3–5 раз шире наивных (Рисунок 1, а). Для CNN наивный интервал [0,527; 0,547] указывает на значимое превышение случайного уровня, однако кластерный интервал [0,492; 0,582] содержит 0,5. Значимо выше случайного уровня оказывается только ViT: AUC = 0,603, 95% ДИ [0,558; 0,649].

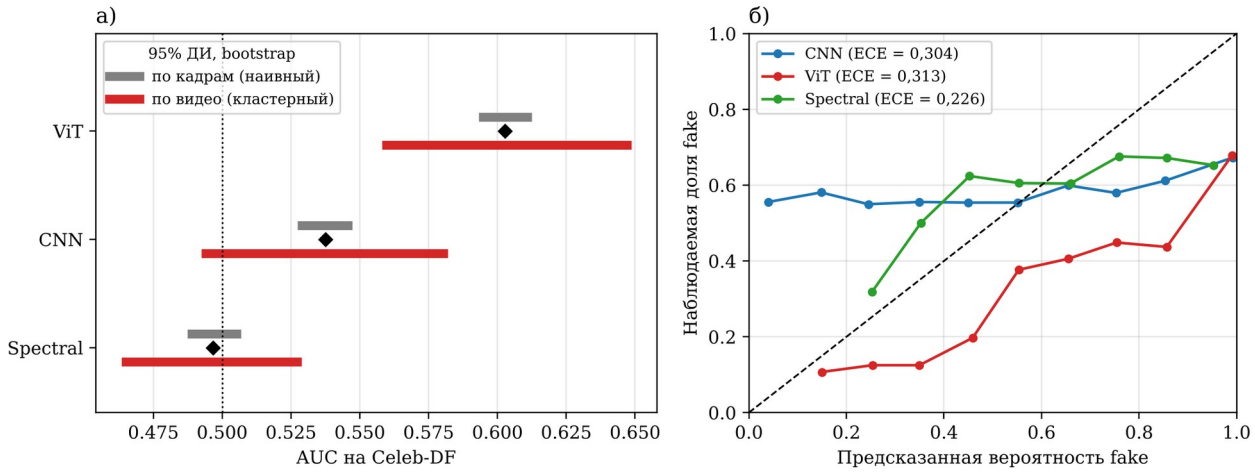


Рисунок 1 - Оценка моделей на Celeb-DF: а) 95% доверительные интервалы AUC, построенные по кадрам и по видео; б) диаграмма надежности (пунктир — идеальная калибровка)

Результаты парного сравнения моделей приведены в Таблице 2.

Таблица 2 – Попарное сравнение моделей по AUC на Celeb-DF

Пара моделей	ΔAUC	p (DeLong, кадры)	95% ДИ ΔAUC (кластерный bootstrap)	p (bootstrap)	p (DeLong, видео)
ViT – CNN	0,065	< 0,001	[0,021; 0,112]	≤ 0,001	0,0091
ViT – Spectral	0,106	< 0,001	[0,054; 0,164]	≤ 0,001	0,0015
CNN – Spectral	0,041	< 0,001	[−0,018; 0,100]	0,152	0,265

Покадровый тест DeLong признает значимыми все три разности ($p < 0,001$). При учете кластерной структуры значимыми остаются только преимущества ViT над CNN ($\Delta AUC = 0,065$, 95% ДИ [0,021; 0,112]) и над спектральной моделью ($\Delta AUC = 0,106$, 95% ДИ [0,054; 0,164]). Разница между CNN и спектральной моделью статистически не подтверждается ($p = 0,15$). Тест DeLong на уровне видео дает согласованный результат ($p = 0,009$, $p = 0,0015$ и $p = 0,26$ соответственно).

Анализ калибровки показывает, что на новом домене все модели систематически завышают вероятность подделки (Рисунок 1, б): ECE составляет 0,23–0,31. При пороге 0,5 к поддельным отнесено 90% настоящих кадров Celeb-DF для CNN и около 98% для ViT и спектральной модели. Перенос порога, соответствующего точке равных ошибок (EER) на FF++, снижает долю ложных срабатываний лишь до 69–91%. Даже порог, оптимально подобранный на самом Celeb-DF, обеспечивает сбалансированную точность не выше 0,57. Следовательно, деградация связана с самими признаками, а не только с положением порога. Дополнительный стресс-тест на подвыборке Celeb-DF с типовыми искажениями (сжатие JPEG, снижение разрешения, размытие, шум) по протоколу [13] показал, что по точечным оценкам ViT сохраняет преимущество при всех искажениях, хотя на подвыборке из 50 видео доверительные интервалы остаются широкими. Спектральная модель при размытии и снижении разрешения дает систематически инвертированные решения (AUC = 0,32–0,40), то есть ее признаки чувствительны к качеству изображения, а не к следам манипуляции.

Таким образом, для корректной оценки робастности детекторов deepfake-контента рекомендуется: сопоставлять результаты с тривиальным классификатором и использовать метрики, устойчивые к дисбалансу (BAcc, MCC, AUC); строить доверительные интервалы bootstrap по видео, а не по кадрам; сравнивать модели парным кластерным bootstrap или тестом DeLong на уровне видео; проверять отсутствие пересечений между обучающей и тестовой выборками; оценивать калибровку и перенос порога решения. Применение протокола показало, что из трех исследованных парадигм только трансформер ViT сохраняет статистически значимую обобщающую способность на новом домене, хотя и на низком

уровне. Дальнейшая работа будет направлена на разработку гибридной модели, сочетающей глобальные признаки трансформера с частотными признаками, устойчивыми к сжатию, и на ее оценку по предложенному протоколу.

Список литературы:

1. Rössler A., Cozzolino D., Verdoliva L., Riess C., Thies J., Nießner M. FaceForensics++: Learning to Detect Manipulated Facial Images // Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). – 2019. – P. 1–11.
2. 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.
3. Yan Z., Zhang Y., Yuan X., Lyu S., Wu B. DeepfakeBench: A Comprehensive Benchmark of Deepfake Detection // Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track. – 2023. – Vol. 36.
4. Dolhansky B., Bitton J., Pflaum B., Lu J., Howes R., Wang M., Ferrer C. C. The DeepFake Detection Challenge (DFDC) Dataset [Электронный ресурс] // arXiv:2006.07397. – 2020.
5. Chicco D., Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation // BMC Genomics. – 2020. – Vol. 21. – Art. 6.
6. Tan M., Le Q. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks // Proceedings of the 36th International Conference on Machine Learning (ICML). – 2019. – P. 6105–6114.
7. Dosovitskiy A., Beyer L., Kolesnikov A. et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale // International Conference on Learning Representations (ICLR). – 2021.
8. Frank J., Eisenhofer T., Schönherr L., Fischer A., Kolossa D., Holz T. Leveraging Frequency Analysis for Deep Fake Image Recognition // Proceedings of the 37th International Conference on Machine Learning (ICML). – 2020. – P. 3247–3258.
9. Efron B., Tibshirani R. J. An Introduction to the Bootstrap. – New York: Chapman & Hall/CRC, 1993.
10. Field C. A., Welsh A. H. Bootstrapping clustered data // Journal of the Royal Statistical Society: Series B (Statistical Methodology). – 2007. – Vol. 69, No. 3. – P. 369–390.
11. DeLong E. R., DeLong D. M., Clarke-Pearson D. L. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach // Biometrics. – 1988. – Vol. 44, No. 3. – P. 837–845.
12. Guo C., Pleiss G., Sun Y., Weinberger K. Q. On Calibration of Modern Neural Networks // Proceedings of the 34th International Conference on Machine Learning (ICML). – 2017. – P. 1321–1330.
13. Jiang L., Li R., Wu W., Qian C., Loy C. C. DeeperForensics-1.0: A Large-Scale Dataset for Real-World Face Forgery Detection // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). – 2020. – P. 2889–2898.

© Central Asian Scientific Journal, 2026