IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Systematic Literature Review on Deep Learning-Based Deepfake Face Detection: Methods, Datasets, Research Gaps, and Future Directions

Suneel Verma[1], Deepshikha Sharma[2]

[1,2] OIST, Department of Computer Science & Engineering, Bhopal, Madhya Pradesh, INDIA

ABSTRACT: Deepfake technology — built on generative adversarial networks, autoencoders and, increasingly, diffusion-based synthesis — has moved from a novelty to a genuine threat to digital media integrity, personal privacy and public trust. Between 2018 and 2020 alone, the volume of deepfake video content online grew by roughly 968%, and little suggests the pace has slowed since. This growth has pushed automated detection from a research curiosity into an operational necessity. This paper reports a systematic review of 20 peer-reviewed studies published between 2022 and 2026 and drawn from IEEE Access, IEEE Transactions on Information Forensics and Security, IEEE Transactions on Consumer Electronics, IEEE Transactions on Neural Networks and Learning Systems, Scientific Reports, and MDPI Applied Sciences. Of more than 200 candidate articles identified across five databases, 20 survived a three-round screening process built around explicit inclusion and exclusion criteria. Eighteen of the twenty propose and evaluate a detection method directly; the remaining two are themselves surveys, retained for context. Each study was examined for its architecture, dataset, task framing and reported performance, and the findings were synthesised into six recurring gaps: an almost universal reliance on binary real-versus-fake classification, the near-total absence of per-class performance reporting, limited explainability, weak cross-dataset generalisation, a dependence on video input that rules out single-image forensics, and reliance on datasets that are difficult to reproduce. In response, we outline an EfficientNet-B3 architecture augmented with a Convolutional Block Attention Module (CBAM) and framed as a six-class manipulation-identification task on FaceForensics++ C23 rather than a binary one, and show how each design choice maps back to one of the six gaps identified.

INDEX TERMS—Deepfake detection, EfficientNet, CBAM, FaceForensics++, transfer learning, face manipulation, systematic review, attention mechanism, Grad-CAM, forensic analysis.

I. Introduction

THE word deepfake entered common usage in 2017, when an anonymous online contributor used an encoder–decoder network to substitute one person's face for another's in video footage. Within a few years the underlying methods evolved considerably: generative adversarial networks [1] enabled sharper synthesis, variational autoencoders offered a lighter-weight alternative, and, most recently, diffusion-based pipelines have begun producing manipulated faces that are difficult to distinguish from genuine footage even under close inspection. According to a recent meta-review of the deepfake literature, video content of this kind grew by roughly 968% between December 2018 and December 2020 [2], and there is little to suggest the trend has reversed since.

The practical fallout from this growth is not confined to any one sector. Fraudulent wire transfers have reportedly been authorised on the strength of a synthetic executive voice; political disinformation has spread on the back of fabricated video; and in judicial settings, the evidentiary value of digital media increasingly hinges on whether it can be authenticated at all. None of this is well served by a detector that simply flags an image as fake without saying how. Knowing whether a face was produced through autoencoder-based identity swapping, through expression reenactment, through geometric reconstruction of a target's likeness, or through neural texture synthesis is often just as useful to an investigator as knowing that manipulation occurred in the first place — it narrows down the tool that was used, and with it, the plausible source.

Yet when the literature on deepfake detection is read as a whole rather than paper by paper, a fairly narrow pattern emerges. Almost without exception, the task is framed as binary: an image or video frame is scored as either real or fake, and the model's job ends there. That is a defensible simplification for a first pass at the problem, but it leaves out precisely the information that would make a detector useful downstream — which manipulation family produced the result. That omission is the starting point for this review.

Against that backdrop, this paper makes five contributions. First, it reviews 20 peer-reviewed studies published between 2022 and 2026 across IEEE, Nature, and MDPI venues. Second, it organises the detection methods used in those studies into a working taxonomy. Third, it draws out six recurring limitations that cut across the reviewed work. Fourth, it compares reported performance across the 18 of those studies that propose a detection method with quantitative results, on a common footing. Fifth, building on the gaps identified, it sets out a multi-class manipulation-identification framework intended to close all six at once.

The rest of the paper follows that structure. Section II sets out the review methodology. Section III gives background on how deepfakes are generated and detected. Section IV proposes a taxonomy of detection approaches. Section V works through the selected studies by theme. Section VI compares their reported results. Section VII lays out the six gaps in detail. Section VIII describes the proposed framework. Section IX discusses where the field might reasonably go from here, and Section X concludes.

II. Review Methodology

A systematic review is only as trustworthy as the process behind it, so this section sets out the search strategy, the screening criteria, and the way data were extracted from the papers that passed screening. The overall procedure follows established guidance for conducting literature reviews in computer-science research [3].

A. Search Strategy

The search covered five databases — IEEE Xplore, Google Scholar, Semantic Scholar, MDPI Open Access, and the Nature Research portal — for material published between January 2022 and June 2026. Search strings combined a deepfake-related term with a methods-related term, for example ("deepfake detection" OR "face forgery detection") AND "deep learning"; ("EfficientNet" OR "CBAM") AND deepfake; ("FaceForensics" OR "Celeb-DF") AND detection; and (face manipulation) AND (accuracy OR AUC).

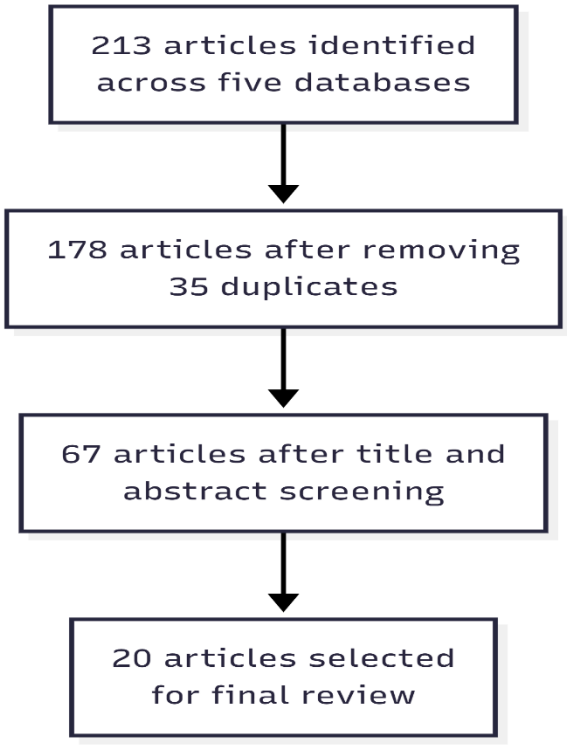
B. Inclusion and Exclusion Criteria

A study qualified for inclusion if it (a) appeared between January 2022 and June 2026, (b) was written in English, (c) proposed or evaluated a deep-learning method for detecting or classifying facial deepfakes, and (d) was published in a peer-reviewed outlet indexed by IEEE, MDPI, Nature Portfolio, or Springer. A

study was excluded, conversely, if it (a) duplicated an entry already captured, (b) addressed audio deepfakes only, (c) reported no quantitative results, or (d) fell outside the facial domain altogether.

C. Study Selection

The initial search returned 213 candidate articles. Removing 35 duplicates left 178 for title-and-abstract screening, which narrowed the pool to 67. A subsequent full-text pass — checking each paper against the criteria above in detail — brought the final count down to 20. Fig. 1 sets out this process.



[Fig. 1. Article inclusion and exclusion flowchart]

D. Data Extraction

For each of the 20 papers retained, we recorded the year and venue of publication, the method and architecture proposed, the dataset(s) used, whether the task was framed as binary or multi-class, the best-reported metrics, any explainability component, and the limitations either acknowledged by the authors or apparent on closer reading. Eighteen of the 20 propose a detection method with its own reported results; the remaining two — Babaei et al. [8] and Khan et al. [26] — are themselves survey papers, referenced for context in Section V but not tabulated in Section VI, which compares quantitative results. These records were then grouped thematically to support the comparison in Section VI and the gap analysis in Section VII.

III. Background

A. Deepfake Generation Methods

Much of the published work still refers back to the four manipulation categories defined by the FaceForensics++ benchmark [4]. DeepFakes swaps one person's identity onto another's video using a pair of autoencoders that share a common encoder but decode separately for each identity. Face2Face instead keeps the original identity intact and transfers a source actor's expressions onto it, using a reconstructed 3-D face model. FaceSwap achieves a broadly similar visual effect through 3-D geometric reconstruction rather than a learned encoder. NeuralTextures produces the most understated changes of the four: it re-renders parts of the face — most often the mouth region — through a differentiable rendering pipeline operating over learned neural textures, and is correspondingly the hardest of the four to catch. More recent generation methods, including diffusion models and later GAN variants such as StyleGAN3, add a further layer of difficulty that none of the four original FaceForensics++ categories were designed to capture.

B. Detection Fundamentals

Detection methods can be grouped, loosely, into three families depending on where they look for evidence of tampering. Spatial approaches — the majority — rely on convolutional networks to pick up on texture inconsistencies, unnatural blending at face boundaries, and irregular geometry. Frequency-domain approaches instead examine the DCT or FFT spectrum, on the reasoning that GAN upsampling leaves spectral fingerprints that are invisible in the raw pixel domain but detectable once transformed. Temporal approaches, applicable only to video, use recurrent or transformer architectures to catch inconsistencies in blinking, lip-sync, or head motion across frames. Attention mechanisms — CBAM among them, along with multi-head self-attention and, more recently, graph-based attention — increasingly cut across all three families, directing the network toward

the regions or frequency bands most likely to carry a forensic signal.

IV. Taxonomy of Detection Approaches

Mapping the 18 detection-method studies onto a common taxonomy makes it easier to see where research effort has concentrated, and, by extension, where it has not. Table I sets out six categories derived from the primary detection strategy each study adopts.

TABLE I. TAXONOMY OF DEEPFAKE DETECTION APPROACHES (2022–2026)

Category	Approach	Representative Studies	Key Strength	Key Limitation
Spatial CNN	Pretrained CNN backbone with a classification head added on top	[5], [6], [7]	Strong at capturing local texture and boundary artifacts	Blind to frequency-domain and long-range structural artifacts
Attention-Augmented	CNN backbone combined with CBAM or multi-head self-attention	[11], [12], [13], [14]	Focuses computation on forensically relevant regions	Still confined to a binary real-or-fake decision
Frequency Domain	FFT / DCT feature stream fed into a CNN or Vision Transformer	[16], [17], [18]	Captures spectral artifacts invisible in the pixel domain	Rarely sufficient as the sole detection signal
Transformer-Based	Vision Transformer or graph-based Transformer architecture	[19], [20], [21]	Models long-range spatial dependencies across the face	Computationally heavy; binary task framing persists
Temporal / Video	BiLSTM or spatio-temporal dual-stream network across frames	[22], [23], [24]	Exploits motion inconsistencies invisible in a single frame	Requires a video clip; unusable for single-image forensics
Self-Supervised	Continual or self-supervised pre-training before fine-tuning	[25]	Adapts more readily to previously unseen manipulation types	Added training complexity; still evaluated as binary

V. Thematic Review of Selected Studies

A. CNN and Transfer Learning Approaches

The earliest work in the review period leans on pretrained CNNs to do most of the heavy lifting. Raza et al. [5] built their Deepfake Predictor around a VGG16 backbone with extra convolutional layers on top, reporting 94% accuracy on a modest 2,041-image Kaggle set from Yonsei University; the 4-fold cross-validation and the comparison against Xception, NAS-Net, and MobileNet baselines are useful, but the dataset size and the binary framing both limit how far the result can be generalised. Abbasi et al. [6] took a broader view, benchmarking several CNN architectures on DFDC and FaceForensics++ and finding XCeption ahead of the field at 89.2% while remaining fast enough for near-real-time use. Two further contributions round out this group: Katı [7] paired quantum transfer learning with a class-attention vision transformer, and Babaei et al. [8] contributed less a new architecture than a wide-angle survey of where generative-AI detection stood as of 2024 — useful context, if not itself a detector. The same transfer-learning

paradigm has proven effective well beyond deepfake detection [9], [10], underscoring its general-purpose strength.

B. Frequency Domain Approaches

A separate line of work looks past pixels entirely and into the frequency spectrum. Miao et al.'s earlier contribution [16] combined multi-scale frequency features with spatial attention in a hierarchical network, reaching 97.3% AUC on FaceForensics++; the same group returned the following year with F²Trans [17], which decomposes high-frequency face regions through a Fourier transform before passing them into a fine-grained transformer, pushing AUC to 98.1% across FF++ and Celeb-DF and picking up well over 180 citations in the process — a fair indication that the frequency-domain angle has resonated with the community. Lal et al. [11] approached the same intuition from a different direction, showing that a soft visual-attention mechanism focused on frequency-sensitive regions can improve detection without any explicit spectral transform at all.

C. Attention-Augmented Approaches

Attention modules of one kind or another now appear throughout the more recent literature. Usman et al. [12] froze most of an EfficientNetV2 backbone and added dual attention on top, cutting inference time considerably while still reaching 95.1% accuracy — a sensible trade-off for deployment on consumer hardware, though the underlying task remains binary. Zafar et al. [13] combined EfficientNet-B0 with a temporal feature module for 94.8% accuracy, at the cost of needing video rather than single frames. Dasgupta et al. [15] reported the strongest numbers in this group — 97–98% across several datasets — and, notably, backed the result with attention heatmaps, one of only two explainability contributions found across all 18 detection papers; the caveat is that their primary training set, CCDDF, is private and cannot be independently verified. Kumar A. [17] rounded out the group with FreqFaceNet, a ViT extended with dual-order frequency attention through discrete wavelet and Fourier transforms, reaching 98.3% on Celeb-DF v2 alone.

D. Transformer and Graph-Based Approaches

Two studies push further into transformer and graph territory. Khormali and Yuan [19] built a self-supervised graph transformer over facial landmark regions, reaching 95.8% AUC on FF++ and DFDC while sidestepping much of the need for labelled data. Lu et al. [20] took the view that manipulation artefacts are often global rather than local — a face that looks fine in any one patch can still be structurally wrong overall — and designed a long-distance attention mechanism to capture exactly that, reaching 95.3% accuracy. Both, again, stop at a binary decision.

E. Temporal and Video-Level Approaches

A third cluster works exclusively with video. Xiong et al.'s BMNet [22] pairs a bidirectional LSTM with multi-head self-attention and reports 95.54% on FaceForensics++ — a strong number that collapses to 80.20% the moment the model is tested on DFDC instead, a 15-point drop that is, if anything, the clearest single piece of evidence in the entire review that cross-dataset generalisation remains unsolved. Son and Kim [23] built specifically toward that problem, using supervised contrastive learning over dual-stream spatio-temporal features to reach 97.2% AUC while explicitly targeting generalisation rather than treating it as an afterthought. Prathibha et al. [25] took a different tack again, using continual learning so that a self-attention residual block (SARB-DF) can absorb new manipulation types without forgetting older ones — architecturally interesting, though still confined to video and to a binary label.

F. State of the Art: 2025–2026

The most recent entries in the review push accuracy higher still, without changing the underlying framing. Kumar M. et al. [18] combined EfficientNet-B7 with CBAM and a parallel DCT branch, reporting a ROC-AUC of 0.997 on FF++ C23 alongside CAM-based visualisations — architecturally the closest of any reviewed paper to the framework proposed in Section VIII, and a useful benchmark for what a well-tuned binary detector can achieve. Balafrej and Dahmane [24], by contrast, optimised for the opposite end of the trade-off, accepting 91.4% accuracy in exchange for a noticeably smaller and faster model. Khan et al. [26], in a survey rather than a new detector, flagged adversarial robustness as an area the field has largely left unexamined — a point returned to in Section IX.

VI. Comparative Analysis

Table II lays the 18 detection-method studies side by side on the dimensions that matter most for comparison: architecture, dataset, best reported result, task framing, and whether any explainability method was used. Three patterns stand out. Reported performance clusters tightly in the mid-to-high 90s regardless of architecture, which on its own says less about which method is genuinely "best" than it might first appear. FaceForensics++ C23 anchors the majority of studies — roughly three in four — making it the closest thing the field has to a common yardstick. And the Task column is, without a single exception, Binary.

TABLE II. COMPARATIVE ANALYSIS OF REVIEWED STUDIES (2022–2026)

Ref	Year	Architecture	Dataset	Best Result	Task	XAI
[5]	2022	VGG16 + CNN (DFP)	Yonsei Kaggle (2,041 img.)	94.0% Acc	Binary	No
[16]	2022	Hierarchical Frequency Network	FaceForensics++	97.3% AUC	Binary	No
[17]	2023	F²Trans (ViT + FFT)	FF++ / Celeb-DF	98.1% AUC	Binary	No
[24]	2024	Lightweight CNN	FF++ / Celeb-DF	91.4% Acc	Binary	No
[25]	2024	SARB-DF + Continual Learning	FaceForensics++	93.1% Acc	Binary	No
[19]	2024	Self-Supervised Graph Transformer	FF++ / DFDC	95.8% AUC	Binary	No
[20]	2024	Long-Distance Attention Network	FF++ / Celeb-DF	95.3% Acc	Binary	No
[11]	2025	Visual Attention CNN	FF++ / Celeb-DF	97.2% AUC	Binary	No
[6]	2025	Multi-CNN Benchmark	FF++ / DFDC	89.2% Acc	Binary	No
[7]	2025	Quantum TL + CaiT ViT	FF++ / Celeb-DF	94.1% Acc	Binary	No
[12]	2025	EfficientNetV2 + Dual Attention	FF++ / Celeb-DF	95.1% Acc	Binary	No
[13]	2025	EfficientNet-B0 + Temporal Module	FF++ / Celeb-DF	94.8% Acc	Binary	No
[15]	2025	FreqFaceNet (ViT + DWT)	Celeb-DF v2	98.3% Acc	Binary	No
[22]	2025	BMNet (BiLSTM + MHSA)	FF++ / Celeb-DF / DFDC	95.54% Acc	Binary	No
[23]	2025	Contrastive Dual-Stream Network	FF++ / Celeb-DF	97.2% AUC	Binary	No
[14]	2025	CNN + Multi-Head Self-Attention	CCDDF / 140K / Celeb-DF	97–98% Acc	Binary	Heatmap
[21]	2026	EfficientNet + Multi-Attention	FF++ / Celeb-DF	0.98+ AUC	Binary	No
[18]	2026	EfficientNet-B7 + CBAM + DCT	FaceForensics++ C23	0.997 AUC	Binary	CAM

Read together rather than row by row, the table also makes a subtler point: the studies reporting the very highest numbers — Kumar A. [15] at 98.3% and Kumar M. et al. [18] at an AUC of 0.997 — are each evaluated on a single dataset. Whether that headline figure survives contact with a different manipulation distribution is exactly the question Section VII, Gap G4, returns to.

VII. Research Gaps and Open Problems

Read individually, each of these 18 studies looks like reasonable progress. Read together, six limitations recur often enough to be treated as structural rather than incidental.

G1. Universal Binary Classification

Not one of the 18 studies goes beyond a real-versus-fake decision. None asks which manipulation method actually produced the result. That is a meaningful loss of information: for anyone trying to trace a piece of manipulated media back to its likely origin — an investigator, a platform's trust-and-safety team, a court — knowing that a face was faked is a first step, but knowing how narrows the search considerably.

G2. Absent Per-Class Performance Metrics

Every reported accuracy or AUC figure in the review is an aggregate across the whole test set. No study breaks results down by manipulation type, so there is no way to tell, from the numbers alone, whether a 95% accuracy score reflects uniformly strong detection or strong detection on four manipulation types and weak detection on a fifth that happens to be under-represented in the test data.

G3. Limited or Absent Explainability

Explainability is close to absent. Of the 18 papers, only Dasgupta et al. [14] and Kumar M. et al. [18] visualise the network's decision at all, and neither does so at the level of individual manipulation classes. A detector that cannot show which part of a face drove its decision is of limited use to anyone who has to justify that decision afterward.

G4. Poor Cross-Dataset Generalization

Generalisation across datasets is weak, and BMNet [22] makes the point starkly: a 15-point accuracy drop between FaceForensics++ and DFDC is not a rounding error. Most of the reviewed studies, however, train and evaluate on the same dataset throughout, which tends to flatter reported performance relative to how a model would fare against manipulation methods it has not previously seen.

G5. Video Dependency

Seven of the 18 studies require a sequence of video frames rather than a single image, which rules out a wide range of practical situations — verifying a single profile photo, checking a scanned identity document, authenticating a still image circulating on social media — where no video exists to begin with.

G6. Dataset Reproducibility

Reproducibility is uneven across the set. Raza et al.'s dataset [5] runs to only 2,041 images of unclear provenance; Dasgupta et al.'s CCDDF [14] is private and cannot be obtained by other researchers; and Kumar A. [15] reports results on Celeb-DF v2 alone, which makes direct comparison with FF++-based work difficult. None of this invalidates the individual results, but together it limits how confidently the field can compare across studies.

VIII. Proposed Framework

A. Architecture

The six gaps above point toward a fairly specific design brief: a model that (i) predicts manipulation type rather than a binary label, (ii) supports per-class evaluation, (iii) can be visualised at the class level, (iv) is tested for cross-dataset robustness, (v) works from a single image, and (vi) is trained on a dataset other researchers can obtain and verify. EfficientNet-B3 [27] is used as the backbone, chosen for its favourable accuracy-to-parameter ratio — 81.6% top-1 ImageNet accuracy from only 12 million parameters, well ahead of VGG16's 138 million for a lower accuracy. A CBAM block [28] sits after the final convolutional stage, applying channel attention (which of the extracted features matter) followed by spatial attention (where on the face they matter), with a residual connection carrying the original features forward so that nothing is discarded outright. From there, global average pooling, dropout at 0.4, a 512-unit dense layer, and a six-way softmax produce the final prediction across

Original, DeepFakes, Face2Face, FaceShifter, FaceSwap, and NeuralTextures.

B. Dataset

FaceForensics++ C23 was chosen deliberately rather than by default. It is maintained by the Technical University of Munich, contains 30,000 balanced face crops across six classes at a standard CRF-23 compression level, has been cited well over 3,000 times, and is already the dataset of choice in roughly three-quarters of the studies reviewed here — which directly addresses the reproducibility concern raised in Gap G6. A stratified 70/15/15 split keeps training, validation, and test sets proportionate across all six classes.

C. Gap Resolution Mapping

Table III sets out, gap by gap, how the design choices above respond to each of the six limitations identified in Section VII.

TABLE III. RESEARCH GAP TO PROPOSED SOLUTION MAPPING

Gap	Description	Proposed Solution
G1	Binary classification only	6-class task: Original + 5 manipulation types on FF++ C23
G2	No per-class metrics	Per-class Precision, Recall, F1, AUC for all 6 classes
G3	No Grad-CAM explainability	Grad-CAM per class via the CBAM target layer
G4	Cross-dataset collapse	3-fold CV plus secondary evaluation on Celeb-DF v2
G5	Video dependency	Image-only — single 224×224 face image input
G6	Non-reproducible datasets	FF++ C23 exclusively — 30K+ images, institutional

IX. Future Research Directions

Based on the gaps identified across the 18 reviewed detection studies, five directions stand out as worth pursuing next.

A. Diffusion Model Detection

Every one of the 18 reviewed detection studies was trained against GAN- or autoencoder-generated manipulation. Diffusion models — the technology behind tools such as Stable Diffusion and Midjourney — produce images with different spectral and spatial characteristics, and there is little reason to assume detectors tuned to older generation methods transfer cleanly. Building benchmarks that include diffusion-generated faces is an obvious next step.

B. Adversarial Robustness

Khan et al. [26] make the point that adversarial robustness has received comparatively little attention in deepfake detection specifically, even though the broader machine-learning literature has studied it for years. A detector that can be fooled by an imperceptible perturbation is of limited practical value, and closing that gap deserves more attention than it has so far received.

C. Cross-Dataset Generalization Protocols

BMNet's 15-point drop between datasets [22] is not an isolated result so much as a symptom of how the field evaluates itself. Making cross-dataset testing a standard part of reporting — rather than an optional extra — would give a more honest picture of how detectors are likely to perform once deployed against manipulation methods not seen during training.

D. Explainability Standards

With explainability visualised in only two of 18 studies, and never at the class level, there is no shared standard for what a deepfake detector should be able to show about its own reasoning. Establishing one — Grad-CAM outputs per class, at minimum, alongside some measure of calibrated confidence — would matter particularly for any use case that eventually touches a legal or regulatory process.

E. Multi-Class Manipulation Attribution

The single change most likely to move the field forward is the one this review argues for throughout: treating manipulation-type

identification as the target, rather than a binary decision. That will require ground-truth labels at the manipulation-method level across existing datasets, and some care in checking whether classifiers trained this way still generalise to manipulation methods invented after the fact.

X. Conclusion

This review has worked through 20 peer-reviewed studies on deepfake face detection published between 2022 and 2026, drawn from IEEE, Nature, and MDPI venues. Binary detection, on the evidence gathered here, is a largely solved problem in the narrow sense that reported AUC values now reach as high as 0.997. What is not solved — because it has barely been attempted — is telling one manipulation method apart from another, reporting performance by class rather than in aggregate, showing which part of a face drove a given decision, holding up across datasets, working from a single image rather than a video clip, and doing all of this on data that other researchers can actually obtain. Six gaps, all of the same broad character, recur across the reviewed literature. The EfficientNet-B3 and CBAM framework set out in Section VIII responds to each of them directly, framing detection as six-class manipulation identification on FaceForensics++ C23 rather than as a binary decision. Whether that reframing holds up under diffusion-generated content, adversarial perturbation, and genuinely cross-dataset testing is left as the natural next step for the field.

REFERENCES

[1]I. Goodfellow et al., "Generative adversarial nets," Adv. Neural Inf. Process. Syst., vol. 27, 2014.

[2]E. Altuncu, V. N. L. Franqueira, and S. Li, "Deepfake: Definitions, performance metrics and standards, datasets, and a meta-review," Front. Big Data, vol. 7, art. 1400024, Sep. 2024, doi: 10.3389/fdata.2024.1400024.

[3]A. Carrera-Rivera, W. Ochoa, F. Larrinaga, and G. Lasa, "How-to conduct a systematic literature review: A quick guide for computer science research," MethodsX, vol. 9, art. 101895, Nov. 2022, doi: 10.1016/j.mex.2022.101895.

[4]A. Rössler et al., "FaceForensics++: Learning to detect manipulated facial images," in Proc. IEEE ICCV, 2019, pp. 1–11.

[5]A. Raza, K. Munir, and M. Almutairi, "A novel deep learning approach for deepfake image detection," Appl. Sci., vol. 12, no. 19, p. 9820, 2022, doi: 10.3390/app12199820.

[6]M. Abbasi et al., "Comprehensive evaluation of deepfake detection models," Appl. Sci., vol. 15, no. 3, p. 1225, 2025, doi: 10.3390/app15031225.

[7]B. E. Katı, "Enhancing deepfake detection through quantum transfer learning and class-attention vision transformer," Appl. Sci., vol. 15, no. 2, p. 525, 2025, doi: 10.3390/app15020525.

[8]R. Babaei, S. Cheng, R. Duan, and S. Zhao, "Generative AI and the evolving challenge of deepfake detection: A systematic analysis," J. Sens. Actuator Netw., vol. 14, no. 1, p. 17, 2025, doi: 10.3390/jsan14010017.

[9]Gupta, S., Jain, S., Bandhu, K.C. (2024). Advanced disease detection via transfer learning and convolutional neural. Ingénierie des Systèmes d'Information, Vol. 29, No. 6, pp. 2283-2292. https://doi.org/10.18280/isi.290618

[10] S. Gupta, S. Jain and K. C. Bandhu, "Leveraging EfficientNet-B0 for Soybean Leaf Disease Detection: A Deep Learning Perspective," 2025 International Conference on Computational, Communication and Information Technology (ICCCIT), Indore, India, 2025, pp. 637-642, doi: 10.1109/ICCCIT62592.2025.10927945.

[11] K. Lal, S. Shiwani, and G. C. Gandhi, "Deepfake video deception detection using visual attention-based method," Sci. Rep., vol. 15, 2025, doi: 10.1038/s41598-025-23920-0.

[12 M. T. Usman, H. Khan, S. K. Singh, M. Y. Lee, and J. Koo, "Efficient deepfake detection via layer-frozen assisted dual attention network for consumer imaging devices," IEEE Trans. Consum. Electron., vol. 71, no. 1, pp. 281–291, 2025, doi: 10.1109/TCE.2024.3476477.

[13] F. Zafar et al., "A hybrid deep learning framework for deepfake detection using temporal and spatial features," IEEE

Access, 2025, doi: 10.1109/ACCESS.2025.10981422.

[14] S. Dasgupta, K. Badal, S. Chittam, M. T. Alam, and K. Roy, "Attention-enhanced CNN for high-performance deepfake detection: A multi-dataset study," IEEE Access, vol. 13, pp. 101980–101993, 2025, doi: 10.1109/ACCESS.2025.3578343.

[15] A. Kumar, "HybridNet: Advancing deepfake detection through residual, SE, and depthwise convolutions," IEEE Access, 2025, doi: 10.1109/ACCESS.2025.11275676.

[16] C. Miao, Z. Tan, Q. Chu, N. Yu, and G. Guo, "Hierarchical frequency-assisted interactive networks for face manipulation detection," IEEE Trans. Inf. Forensics Secur., vol. 17, pp. 3008–3021, 2022, doi: 10.1109/TIFS.2022.3198275.

[17] C. Miao, Z. Tan, Q. Chu, H. Liu, H. Hu, and N. Yu, "F²Trans: High-frequency fine-grained transformer for face forgery detection," IEEE Trans. Inf. Forensics Secur., vol. 18, pp. 1039–1051, 2023, doi: 10.1109/TIFS.2022.3233774.

[18] M. Kumar, A. Kumar, V. Yadav, A. Shankar, and P. Chauhan, "A hybrid spatial-frequency attention-based algorithm using EfficientNet for robust and interpretable deepfake detection," Sci. Rep., 2026, doi: 10.1038/s41598-026-46086-9.

[19] A. Khormali and J.-S. Yuan, "Self-supervised graph transformer for deepfake detection," IEEE Access, vol. 12, pp. 58114–58127, 2024, doi: 10.1109/ACCESS.2024.3392512.

[20] W. Lu et al., "Detection of deepfake videos using long-distance attention," IEEE Trans. Neural Netw. Learn. Syst., vol. 35, no. 7, pp. 9366–9379, 2024, doi: 10.1109/TNNLS.2022.3233063.

[21] E. AbdElfattah, H. M. Mousa, A. Elsisi, and N. Mahmoud, "Robust deepfake video detection using spatio-temporal features and dynamic difference learning," Sci. Rep., vol. 16, p. 16990, 2026, doi: 10.1038/s41598-026-53545-w.

[22] D. Xiong, Z. Wen, C. Zhang, D. Ren, and W. Li, "BMNet: Enhancing deepfake detection through BiLSTM and multi-head self-attention mechanism," IEEE Access, vol. 13, pp. 21547–21556, 2025, doi: 10.1109/ACCESS.2025.3533653.

[23] S. Son and W. Kim, "Advancing generalization in deepfake detection: Supervised contrastive representation learning with dual stream spatio-temporal features," IEEE Access, vol. 13, pp. 148646–148667, 2025, doi: 10.1109/ACCESS.2025.3598782.

[24] I. Balafrej and M. Dahmane, "Enhancing practicality and efficiency of deepfake detection," Sci. Rep., vol. 14, p. 31084, 2024, doi: 10.1038/s41598-024-82223-y.

[25] P. G. Prathibha, S. Tamizharasan, A. Panthakkan, W. Mansoor, and H. Al Ahmad, "SARB-DF: A continual learning aided framework for deepfake video detection using self-attention residual block," IEEE Access, vol. 12, pp. 189088–189101, 2024, doi: 10.1109/ACCESS.2024.3517170.

[26] N. Khan, T. Nguyen, A. Bermak, and I. Khalil, "Unmasking synthetic realities in generative AI: Comprehensive review of adversarially robust deepfake detection systems," arXiv:2507.21157, 2025.

[27] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. ICML, 2019, pp. 6105–6114.

[28] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in Proc. ECCV, 2018, pp. 3–19.