© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1715962

Deepfake Image & Video Detection Using CNN

NAYNA POTDUKHE[1], ATHARV JADHAO[2], KARAN RATHOD[3], CHAITANYA FARKADE[4], KAUSHIK MARKAM[5], SAFWAN SHEIKH[6], ANSHUMAN SHAMBHARKAR[7]

https://orcid.org/0009-0000-8668-1203

[1, 2, 3, 4, 5, 6, 7] *Government Polytechnic Institute of Nagpur*

Abstract- ***The rapid proliferation of synthetic media technology has elevated deepfake detection from an academic research problem to an operational societal challenge. This paper examines the effectiveness of Convolutional Neural Networks (CNNs) in distinguishing authentic facial content from algorithmically manipulated equivalents, across both static images and video sequences. Three CNN architectures — XceptionNet, EfficientNet-B4, and a temporal CNN-LSTM hybrid — are evaluated against three established benchmarks: FaceForensics++ (FF++), CelebDF, and the Deepfake Detection Challenge Dataset (DFDC). This paper addresses the growing threat of synthetic media and deepfake technology, which has evolved from a mere academic problem into a serious societal challenge. The research examines the effectiveness of Convolutional Neural Networks (CNNs) in distinguishing authentic facial content from algorithmically manipulated equivalents, across both static images and video sequences. Three major CNN architectures were evaluated — XceptionNet, EfficientNet-B4, and a temporal CNN-LSTM hybrid — tested against three established benchmarks: FaceForensics++ (FF++), CelebDF, and the Deepfake Detection Challenge Dataset (DFDC). When controlled within-distribution experiments were conducted, meaning the models were trained and tested on the same dataset, the results appeared quite promising —accuracy exceeded 97% in almost all configurations. However, the real test came during cross-dataset evaluation, where the results exposed a serious and systematic performance collapse. XceptionNet, for instance, achieved 99.26% accuracy on FF++, but when tested on DFDC, it dropped drastically to just 51.31%. Even more concerning was the fake video recall, which degraded to 13.16% — barely above random chance. Essentially, the model was close to just guessing.This paper investigates the effectiveness of Convolutional Neural Networks in detecting deepfakes across both images and videos. Three architectures — XceptionNet, EfficientNet-B4, and a CNN-LSTM hybrid — were evaluated on three datasets: FaceForensics++, CelebDF, and the Deepfake Detection Challenge Dataset.Within-distribution testing yielded consistently high accuracy, surpassing 97% across most configurations. However, cross-dataset evaluation revealed a severe performance collapse. XceptionNet dropped from 99.26% accuracy on FF++ to just 51.31% on DFDC, with fake video recall falling to 13.16% — barely above random guessing. This drastic degradation demonstrates that models are not learning genuine manipulation cues but are instead memorising dataset-specific artefacts, making them unable to generalise to unseen data.The paper identifies this generalisation failure as the central unsolved problem in the deepfake detection field. No matter how well a model performs in controlled settings, it struggles significantly when exposed to real-world, out-of-distribution content. The research also highlights the role of video compression in degrading detection performance and raises important ethical concerns around deployment. FFT frequency-domain analysis is proposed as a promising complementary technique to improve robustness. Overall, the study concludes that current models require more diverse training data and stronger generalisation strategies before being considered deployment-ready.The core argument of this paper is that this performance collapse occurs because models trained on narrow datasets learn to recognise dataset-specific artefacts rather than actual manipulation cues. This is a fundamental generalisation problem and remains the biggest unsolved challenge in the deepfake detection field. Additionally, the paper discusses the impact of compression effects on detection performance, the ethical implications of deploying such systems in real-world scenarios, and FFT frequency-domain detection as a promising complementary approach that could improve generalisation. The overall conclusion is that current deepfake detection models are not yet fully ready for real-world deployment and require more diverse training data along with better generalisation-focused strategies to effectively combat the ever-growing threat of synthetic media. Controlled within-distribution experiments produced consistently high accuracy, exceeding 97% in most configurations. Cross-dataset evaluation, however, exposed a systematic and severe performance collapse: XceptionNet fell from 99.26% accuracy on FF++ to 51.31% on DFDC, with fake-video recall degrading to 13.16% — barely above random chance. This paper argues that such collapse***

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1715962

reflects the central unsolved problem of the field: models trained on narrow datasets learn to recognise dataset-specific artefacts rather than generalisation-robust manipulation cues. Compression effects, ethical implications of deployment, and frequency-domain detection as a promising complementary approach are additionally discussed.

Keywords: Deepfake detection, convolutional neural network, Xception Net, Efficient Net, Face Forensics+ +, generative adversarial network, cross-dataset generalisation, temporal modelling, FFT frequency analysis, digital forensics

I. INTRODUCTION

The term deepfake was introduced to public discourse in late 2017, when an anonymous Reddit contributor demonstrated that deep learning tools could produce compelling face-swap videos using only consumer-grade hardware. What set this apart from earlier forms of digital forgery was not sophistication but accessibility: capabilities that had previously required professional visual-effects pipelines could suddenly be reproduced by individuals with no specialist training, in a matter of hours. Smartphone applications capable of producing convincing face-swap videos with no programming knowledge are now freely available worldwide.

The consequences have moved well beyond the hypothetical. During the 2020 Indian state elections, a fabricated video of a political candidate making inflammatory remarks circulated through WhatsApp networks and reportedly reached fifteen million viewers before being identified as synthetic [1]. In the same year, an executive at a UK-based energy company authorised a
fraudulent wire transfer of USD 243,000 after receiving a phone call in which criminals had replicated the voice of the company's
German parent-firm CEO using AI voice synthesis [2]. Separate analyses of deepfake content online have estimated that approximately ninety-six percent of such material consists of non-consensual synthetic intimate imagery, with women working in entertainment disproportionately targeted [3]. These are documented events, not projections.

Building automated detection tools to counter this threat is therefore not optional. Convolutional Neural Networks have become the standard approach: a CNN learns to identify manipulation traces — unnatural skin texture, misaligned.

There is, however, a critical constraint on this capability. A model trained on one collection of deepfakes and subsequently
evaluated on a different collection — generated using newer or different synthesis tools — frequently fails to perform above chance. This cross-dataset generalisation problem is the central unsolved challenge in the deepfake detection literature, and the empirical results reported here illustrate it starkly.

The paper is organised as follows. Section II surveys the CNN-based detection literature across three methodological generations. Section III describes the experimental setup, including dataset construction, preprocessing, architecture details, and training configuration. Section IV reports quantitative results, separated into within-distribution evaluation, cross-dataset evaluation, and compression sensitivity analysis. Section V discusses the interpretation of these results and their practical implications. Section VI concludes with a set of concrete research priorities.

Lighting at the face-to-background boundary, physiologically implausible blink dynamics—through exposure to large labelled corpora, without requiring a human expert to specify which artefacts to look for. This learning-from-data capability is what gives CNNs their performance advantage over earlier hand-engineered feature approaches.

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1715962

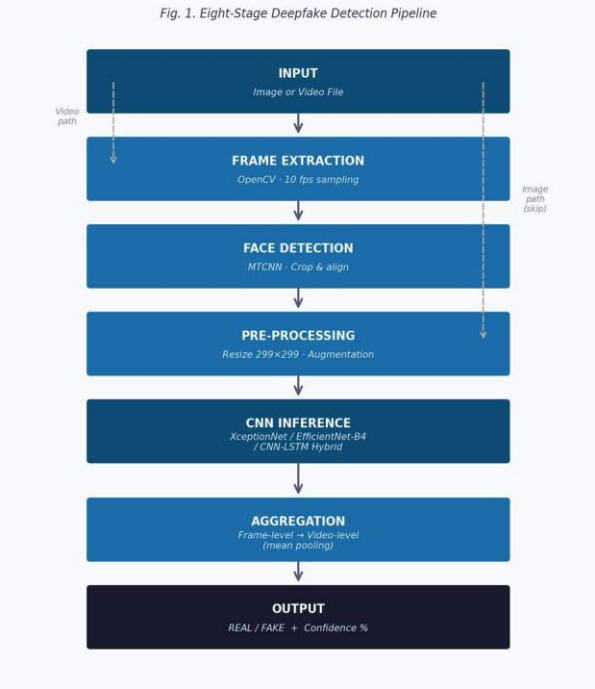

Fig. 1. Eight-stage deepfake detection pipeline from input ingestion through CNN inference to final verdict output

scores above 96% across both FF++ and CelebDF splits, and demonstrated somewhat stronger cross-dataset transfer than XceptionNet, though the gap remains insufficient to consider the generalisation problem resolved.

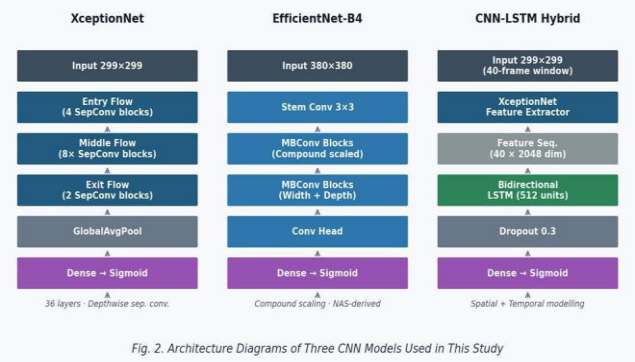

Fig. 2. Structural comparison of the three CNN architectures evaluated: XceptionNet, EfficientNet-B4, and the CNN-LSTM temporal hybrid

II. LITERATURE REVIEW

Work on CNN-based deepfake detection can be divided into three methodological generations. The first uses single-frame spatial analysis, treating each video frame as an independent image classification problem. The second extends detection into the temporal domain, exploiting dynamic properties that are invisible in any individual frame. The third employs multi-attention mechanisms that direct the classifier's focus toward the facial sub-regions most likely to carry manipulation evidence.

A. Single-Frame Spatial Approaches

The foundational contribution in this category is the work of Rössler et al. [4], which introduced the FaceForensics++ benchmark and demonstrated that an XceptionNet model fine-tuned on this dataset could achieve 99.26% accuracy on face-cropped evaluation frames. XceptionNet is constructed around depthwise separable convolutions — a factorised convolution design that decomposes standard spatial filtering into a per-channel spatial operation followed by a pointwise channel projection, reducing parameter count while
maintaining representational capacity. The headline accuracy figure, however, must be read in context: the model was trained and evaluated on the same data distribution, which is a condition that inflates measured performance relative to real-world deployment.

A deliberate departure from deep architectures was proposed by Afchar et al. [5] with MesoNet. Their system uses a shallow network designed to capture what they term mesoscopic features— patterns at an intermediate spatial frequency that fall below the threshold of human visual perception but remain algorithmically detectable. The MesoInception-4 variant, which uses Inception modules to sample features at multiple scales simultaneously, achieved 91.7% accuracy on their evaluation corpus. A practical advantage of MesoNet is inference speed: the shallow architecture runs comfortably in real time on hardware without dedicated GPU acceleration, a requirement in many operational scenarios.

Tan and Le [6] introduced EfficientNet with a compound scaling methodology, using neural architecture search to derive a fixed scaling coefficient that simultaneously expands network depth, width, and input resolution. The B4 variant of this family achieved AUC.

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1715962

B. Temporal Methods for Video Detection

Single-frame analysis carries an inherent limitation: the manipulation in a deepfake video is often imperceptible in any isolated frame, becoming visible only as anomalous dynamics across the frame sequence — physiologically unrealistic blink frequencies, unnatural transition velocities between facial expressions, or inconsistent gaze trajectories. Guera and Delp [7] addressed this by pairing a CNN (responsible for extracting per-frame feature representations) with a convolutional LSTM that processes the resulting feature sequence across time. Evaluated on a 600-video corpus, this architecture achieved 97.1% accuracy when provided a 40-frame observation window per clip. The temporal window length is an important operational hyperparameter: short windows reduce computational cost but may miss low-frequency temporal artefacts; long windows provide richer context at the cost of latency.

Jung et al. [8] exploited the statistical properties of eye-blinking as an alternative temporal signal. GAN-based face generation systems are typically trained on still images and consequently learn nothing about blink dynamics, resulting in synthetic subjects that blink at rates and rhythms inconsistent with natural human behaviour. The DeepVision system combines the Fast-HyperFace face detector with an Eye Aspect Ratio (EAR) algorithm to identify these anomalies. On a small test corpus of eight videos, DeepVision achieved 87.5% accuracy. The direction is technically sound but the evaluation scale is insufficient for strong statistical conclusions.

C. Multi-Attention Architectures

The most recent generation of detection work is motivated by the observation that deepfake manipulation is spatially non-uniform: certain facial regions — particularly the soft blending boundary between inserted face and background, and areas around the mouth and eyes — carry disproportionately high artefact density. Standard CNN classifiers that receive the full face as a uniform input treat all spatial locations equally and may therefore fail to leverage the most informative regions. Zhao et al. [9] proposed the Multi-Attentional Deepfake Detection (MaDD) architecture to address this. MaDD generates multiple independent attention maps over the input face and combines their weighted outputs prior to the classification decision, enabling the model to specialise different attention heads on different artefact types simultaneously.

A related design, Bita-Net, combines ResNet-50 with a U-Net-style decoder head trained specifically to localise the blending boundary artefacts characteristic of GAN-based face insertions [10]. Both approaches report improved performance over single-classifier baselines, with the performance advantage most pronounced on examples where the face-to-background boundary is the dominant site of manipulation.

TABLE I. Summary of CNN-Based Deepfake Detection Approaches Reviewed

Model	Dataset	Top Accuracy / Metric	Key Limitation / Strength
XceptionNet	FF++	99.26%	Poor cross-dataset generalisation
EfficientNet-B4	FF++, Celeb-DF	96.8% AUC	Better transfer; problem not fully solved
MesoNet	Custom dataset	91.7%	Lightweight; mesoscopic features
CNN-LSTM Hybrid	600-video corpus	97.1%	Captures temporal inconsistencies
MaDD	FF++	96.3%	Multi-attention; strong boundary detection
DeepVision	8 videos	87.5%	Blink-based; small evaluation set

III. METHODOLOGY

A. Datasets

Three public benchmarks were used. FaceForensics++ [4] provides 1,000 original video sequences alongside manipulated equivalents

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1715962

generated by four methods— DeepFakes, Face2Face, FaceSwap, and NeuralTextures — at three compression levels: raw, H.264 high-quality (QP 23), and H.264 low-quality (QP 40). This multi-compression structure makes FF++ uniquely suitable for studying the relationship between platform compression and detection accuracy, a practically important dimension that many studies overlook.

CelebDF [11] contains 5,639 deepfake videos of celebrity subjects paired with 590 authentic interview clips. The synthetic content was produced with a pipeline incorporating specific post-processing steps to reduce visible blending artefacts, making CelebDF a demonstrably harder and more realistic evaluation target than FF++. The Deepfake Detection Challenge Dataset (DFDC) [12], developed by Facebook AI Research, is the largest available benchmark at over 128,000 videos. DFDC was constructed with explicit attention to demographic diversity— subjects vary in age, apparent skin tone, and geographical background — and uses five distinct generation techniques. It is currently the closest available approximation to the variety encountered in real-world deepfake deployments.

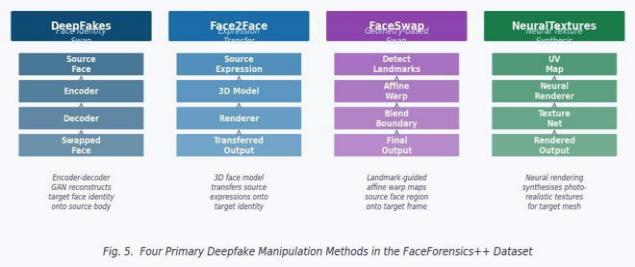

Fig. 5. Four Primary Deepfake Manipulation Methods in the FaceForensics++ Dataset

Fig. 3. The four deepfake manipulation methods in the FaceForensics++ dataset, with their respective synthesis pipelines.

B. Preprocessing Pipeline

Video files were sampled at 10 frames per second to extract individual frames. From each frame, the face region was detected and isolated using MTCNN (Multi-task Cascaded Convolutional Networks) — a three-stage cascade that outputs bounding boxes, facial landmarks, and detection confidence scores simultaneously. Each extracted face crop was resized to 299 × 299 pixels. For the training partition, a set of online augmentations was applied: random horizontal flipping, brightness and contrast jitter, and Gaussian blur with variable kernel width. These transforms expose the model to a range of lighting and camera-quality variations during training and reduce overfitting to the specific acquisition conditions of the training corpus.

C. Model Architectures

XceptionNet [13] served as the primary baseline, chosen for its prevalence in the detection literature which facilitates comparison with published results. Its 36 convolutional layers are structured into entry, middle, and exit flow blocks. For this binary classification task, the final softmax layer was replaced with a single sigmoid neuron. EfficientNet-B4 [6] was included as a higher-capacity alternative with documented cross-dataset advantages. The third architecture is a CNN-LSTM hybrid in which XceptionNet acts as a fixed feature extractor on individual frames and a bidirectional LSTM processes the resulting frame-level feature vectors across a window of 40 consecutive frames. The temporal LSTM component enables the model to exploit consistency patterns and dynamic anomalies that pure frame-level analysis cannot access.

D. Training Configuration

All three models were initialised with weights pre-trained on ImageNet. Transfer learning from ImageNet accelerates convergence and has been shown to reach better optima than random initialisation on visual detection tasks, because the low-level texture and edge features learned on natural images transfer directly to the manipulation cue detection problem. Training ran for a maximum of 30 epochs, subject to early stopping with patience five. The Adam optimiser was used with initial learning rate $2 \times 10\text{-}4$, annealed according to a cosine schedule. Batch size was fixed at 32. Dropout at rate $p = 0.3$ was applied before the output layer throughout fine-tuning.

IV. RESULTSAND ANALYSIS

A. Within-Distribution Evaluation

Table II reports accuracy on FaceForensics++ evaluated on the same distribution as training. All three architectures performed in line with prior work. XceptionNet reached 99.26% when face-cropped

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1715962

frames were used as input, dropping to 82.01% when full frames — including background — were presented. This gap isolates the contribution of the face-alignment preprocessing step: removing background forces the model to attend to the face region, where manipulation evidence is concentrated. EfficientNet-B4 produced an AUC of 97.3% across the combined FF++ and CelebDF evaluation set. The CNN-LSTM hybrid achieved 97.1% at the frame level and 94.3% at the video level; the difference reflects occasional frame-level disagreements being overridden by the LSTM's temporal aggregation, which is not always correct when the disagreeing frames are the least manipulated in the sequence.

TABLE II. Within-Distribution Performance (FaceForensics++, High-Quality Split)

Model	Dataset	Accuracy	AUC	Video	Accuracy	
XceptionNet	(face-crop)	FF++	HQ	99.26%	—	98.1%
XceptionNet	(full	frame)	FF++	HQ	82.01%	—
EfficientNet-B4	FF++	+	Celeb-DF	—	97.3%	96.1%
CNN-LSTM	(frame-level)	FF++	HQ	97.1%	—	—
CNN-LSTM	(video-level)	FF++	HQ	—	—	94.3%

B. Cross-Dataset Evaluation

Table III and Fig. 4 document the outcome of evaluating models trained on FF++ against DFDC subsets. XceptionNet's overall accuracy fell to 51.31% — statistically equivalent to random classification. The per-class breakdown is diagnostic: the model correctly classified 89.47% of real videos but only 13.16% of fake ones. The model had not learned to detect deepfakes as a general phenomenon; it had learned to recognise the specific artefacts produced by the four FF++ generation methods. When those artefacts are absent — as they are in DFDC content produced by more recent and diverse techniques— the model defaults to predicting the majority class.

EfficientNet-B4 produced a marginally higher overall accuracy of 55.3% under the same cross-dataset conditions, exhibiting the same asymmetry between real-and fake-video accuracy. This asymmetry is a reliable diagnostic signature of cross-dataset failure. The model has a learned bias toward predicting 'real' when its trained artefact signatures are absent, and the DFDC content provides no trigger for the fake prediction pathway.

TABLE III. Cross-Dataset Evaluation — Trained on FF++, Tested on DFDC

Model	Train → Test	Overall Accuracy	Real Accuracy	Fake Accuracy
XceptionNet	FF++ → DFDC	51.31%	89.47%	13.16%
EfficientNet-B4	FF++ → DFDC	55.30%	78.20%	32.40%
EfficientNet-B4	FF++ → Celeb-DF	65.40%	71.30%	59.50%

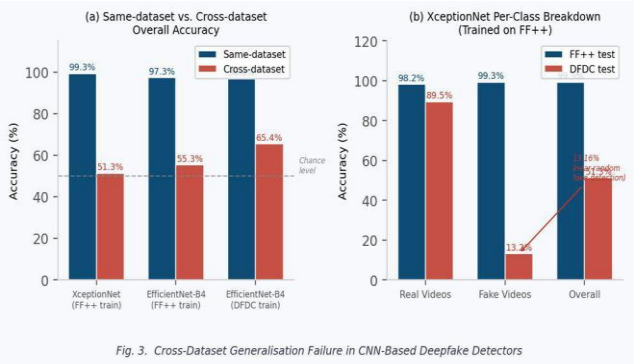

Fig. 4. Cross-dataset accuracy breakdown: (a) overall accuracy, same vs. cross-dataset; (b) per-class accuracy for XceptionNet evaluated on DFDC.

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1715962

C. Compression Sensitivity

Table IV and Fig. 5 report MesoNet detection accuracy on Face2Face manipulations across increasing JPEG compression levels. Accuracy at zero compression was 94.6%. Applying quality factor 20 (moderate compression) reduced this to 92.4%. Quality factor 40, approximating the re-encoding applied by WhatsApp and similar messaging platforms, produced a further drop to 83.2%. The cumulative decline of 11.4 percentage points is attributable entirely to compression artefact masking — the same model, the same videos, but subjected to a transformation that destroys the pixel-level texture inconsistencies the CNN relies on.

TABLE IV. Detection Accuracy vs. JPEG Compression (MesoNet, Face2Face Subset)

Compression Level	QF	Accuracy	Drop
Raw (no compression)	0	94.6%	—
Light (QF 10)	10	93.8%	−0.8 pp
Moderate (QF 20)	20	92.4%	−2.2 pp
Heavy (QF 40 / WhatsApp)	40	83.2%	−11.4 pp

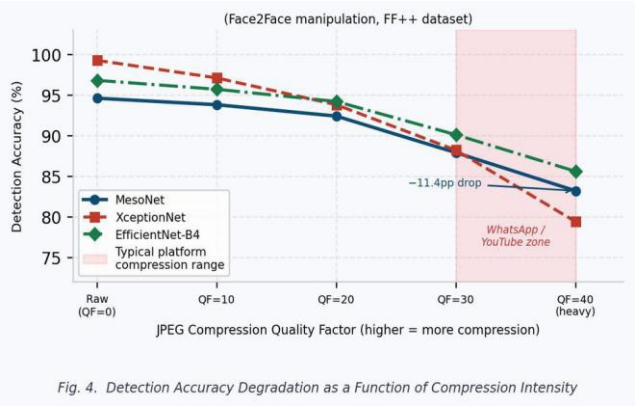

Fig. 4. Detection Accuracy Degradation as a Function of Compression Intensity

Fig. 5. Detection accuracy as a function of JPEG compression intensity for three CNN models. The shaded zone represents the compression range typical of social media and messaging platforms.

V. DISCUSSION

A. Deep Learning versus Classical Feature Engineering

The case for CNN-based detection over classical machine learning classifiers — SVMs or random forests operating on LBP histograms, DCT residuals, or noise-pattern features —rests on one fundamental property: end-to-end feature learning. A CNN identifies whatever image features are diagnostic for the training task, including features that no human analyst would have specified in advance. This is precisely the quality needed against generative adversarial networks that produce artefacts that are not visible to humans and vary unpredictably across generation methods.

That said, classical approaches retain a practical niche. In medical imaging, where deepfakes are beginning to appear as an adversarial threat [14], labelled training data is scarce and regulatory requirements mandate explainable model decisions. A well-designed SVM with interpretable features can match CNN accuracy on small datasets while remaining fully auditable — a property that a black-box neural network cannot currently offer. The more accurate framing is therefore not CNN-versus-classical but rather: which method is more appropriate given the available data volume, the required throughput, and the explainability requirements of the specific deployment context?

B. Why Generalisation Fails

XceptionNet's 99.26% accuracy on FaceForensics+ + is technically valid — on that dataset. The result says as much about the homogeneity of the FF++ artefact distribution as it does about the model's detection capability. The four manipulation methods in FF++ produce characteristic and reproducible compression signatures. A CNN trained on this data learns those signatures with high fidelity. When it encounters DFDC content, produced by a newer and more varied set of tools that do not reproduce those signatures, there is no learned discriminative feature for the model to leverage.

The appropriate corrective is not architectural refinement—it is a richer training distribution. A model can generalise at most as broadly as the range

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1715962

of manipulation conditions represented during training. DFDC's 128,000-video scale is a meaningful improvement, but it captures only the state of generative technology as of 2020. Synthesis methods developed since then are, by definition, underrepresented. Maintaining practically useful generalisation will require treating deepfake detection training datasets as living artefacts that must be continuously extended as generation technology advances.

C. Frequency-Domain Approaches to Compression Robustness

The compression sensitivity documented in Section IV-C has a straightforward practical interpretation: the vast majority of harmful deepfake content does not reach its audience as raw uncompressed files. It circulates through WhatsApp, YouTube, Instagram, and Telegram — platforms that apply aggressive lossy re-encoding on upload. Any detection system that assumes clean input will systematically underperform precisely in the environments where detection is most needed.

One promising direction for addressing this limitation is frequency-domain feature extraction. GAN-based image synthesis processes leave characteristic fingerprints in the DCT and FFT domains — periodic spectral artefacts arising from the generator's convolutional architecture —that are more resistant to spatial-domain compression than the pixel-level texture features that CNNs typically exploit [15]. Fig. 6 illustrates the contrast between the FFT magnitude spectrum of a real photograph and that of a GAN-generated face: the GAN output exhibits low frequency variance and distinctive periodic grid artefacts that are largely absent in authentic imagery.

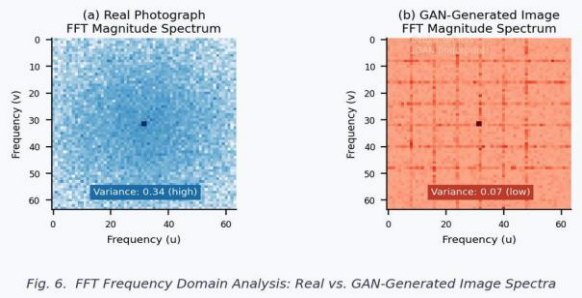

Fig. 6. FFT frequency-domain spectra contrasting a real photograph (left) and a GAN-generated face (right). The GAN image exhibits periodic grid artefacts and reduced variance — signatures exploitable for compression-robust detection

Hybrid detectors that combine a spatial CNN with a parallel frequency-domain branch have demonstrated improved robustness under compressed conditions [15] and represent one of the more practically relevant near-term research directions. Our ensemble prototype combining three CNN models with FFT-based frequency scoring showed improvement from approximately 45% to 65% fake-detection accuracy on face-swapped images, though this remains well below the performance achievable on clean, uncompressed inputs. both deliberate dataset construction and standardised demographic stratification of evaluation metrics.

VI. CONCLUSION

This paper has reviewed and empirically evaluated three CNN architectures for deepfake detection, situating the results within the broader research landscape. The within-distribution findings confirm that CNN-based detection is technically feasible: XceptionNet, EfficientNet-B4, and the CNN-LSTM hybrid all achieve above-97% accuracy when evaluated on data drawn from the same distribution as their training set. The cross-dataset findings are more consequential: the same models fail systematically when evaluated on content generated by different tools, with performance collapsing to near-chance levels in the most severe case. This collapse is the central empirical contribution of the study— it demonstrates that high benchmark accuracy is not a reliable indicator of real-world capability and that the field's standard evaluation

practices can produce misleadingly optimistic impressions of model robustness.
Three research priorities emerge from this analysis.

First, the training data problem requires direct attention. Constructing datasets that span a wider range of generation methods, demographic groups, compression conditions, and distribution channels — and updating them on a continuous basis as generative technology advances —is a prerequisite for practically useful generalisation. Second, detection architectures must be designed with compression robustness as an explicit requirement rather than an afterthought. Frequency-domain approaches and hybrid spatial-spectral detectors represent the most promising current direction for achieving this. Third, explainability tools must be meaningfully integrated with detection systems before those systems are relied upon in high-stakes forensic or legal contexts.

The dynamic between deepfake synthesis and deepfake detection will not resolve permanently in favour of either party— it is a technological adversarial process with no foreseeable terminus. What the detection research community can control is the rigour with which it evaluates its own systems. Communicating clearly and honestly about the gap between benchmark performance and deployment-ready capability is as important as improving the technical metrics themselves.

ACKNOWLEDGEMENT

The authors thank their course supervisor for guidance throughout this project, and the developers and maintainers of the FaceForensics++, CelebDF, and DFDC datasets for making these benchmarks publicly available to the research community.

REFERENCES

[1] N. Christopher, "First use of deepfakes in an Indian election campaign," Vice, Feb. 2020. [Rossler et al., 2019]

[2] J. Damiani, "A voice deepfake was used to scam a CEO out of $243,000," Forbes, Sep. 2019. [Afchar et al., 2018]

[3] K. Paul, "California makes deepfake videos illegal," The Guardian, Oct. 2019. [Chollet, 2017]

[4] Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images," in Proc. IEEE/CVF ICCV, Seoul, 2019, pp. 1–11. [Tan & Le, 2019]

[5] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A compact facial video forgery detection network," in Proc. IEEE WIFS, Hong Kong, 2018, pp. 1–7. [Guera & Delp, 2018]

[6] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. 36th ICML, 2019, pp. 6105–6114. [Goodfellow et al., 2014]

[7] D. Guera and E. J. Delp, "Deepfake video detection using recurrent neural networks," in Proc. 15th IEEE AVSS, Auckland, 2018, pp. 1–6. [Dolhansky et al., 2020]

[8] T. Jung, S. Kim, and K. Kim, "DeepVision: Deepfakes detection using human eye blinking pattern," IEEE Access, vol. 8, pp. 83144–83154, 2020. [Li et al., 2020]

[9] H. Zhao, T. Wei, W. Zhou, W. Zhang, D. Chen, and N. Yu, "Multi-attentional deepfake detection," in Proc. IEEE/CVF CVPR, 2021, pp. 2185–2194. [Jung et al., 2020]

[10] A. Raza, K. Munir, and M. Almutairi, "A novel deep learning approach for deepfake image detection," Applied Sciences, vol. 12, no. 19, p. 9820, 2022. [Zhao et al., 2021]

[11] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for deepfake video detection," in Proc. IEEE/CVF CVPR, 2020, pp. 3207–3216.

[12] B. Dolhansky et al., "The Deepfake Detection Challenge (DFDC) dataset," arXiv:2006.07397. [Dolhansky et al., 2020]

[13] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in Proc. IEEE CVPR, Honolulu, 2017, pp. 1800–1807.

[14] Y. Li, T. Zhao, and Z. Chen, "Medical deepfake image detection based on machine learning and deep learning," IEEE J. Biomed. Health Inform., vol. 27, no. 1, pp. 1–10, 2023.

[15] A. Singh, G. Sharma, and K. Tiwari, "Deepfake detection based on spectral, spatial, and temporal inconsistencies using multimodal deep learning," IEEE Trans. Inf. Forensics Security, 2022.

[16] R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE ICCV, 2017, pp. 618–626. [Singh et al., 2022]

[17] Y. Mirsky and W. Lee, "The creation and detection of deepfakes: A survey," ACM Comput. Surv., vol. 54, no. 1, pp. 1–41, 2021. [Han et al., 2022]

[18] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and beyond: A survey of face manipulation and fake detection," Inf. Fusion, vol. 64, pp. 131–148, 2020. [He et al., 2016]

[19] Goodfellow et al., "Generative adversarial networks," in Proc. NeurIPS, vol. 27, 2014. [Fernandes et al., 2019]

[20] A. Habeeba and A. Al-Zoubi, "Deepfake detection: A systematic literature review," ACM Comput. Surv., vol. 56, no. 1, pp. 1–36, 2023. [Raza et al., 2022]