

A Comparative Study of Deep Learning Techniques for Deepfake Detection

Prof. Rana Afreen Sheikh¹, Ms. Nikita Pramod Hambarde²,
Ms. Zarnain Tasneem Shahid Hussain³, Ms. Madiha Fatema Syed Irshad Ali⁴

Professor, Department of Master of Computer Applications (MCA)¹

Students, Department of Master of Computer Applications (MCA)²⁻⁴

Vidyabharati Mahavidyalaya, Amravati, Maharashtra, India

Abstract: Deepfake media generated or altered using artificial intelligence can imitate real people and events with increasing visual realism, creating challenges for digital trust, privacy, cybersecurity, journalism, and forensic investigation. This paper presents a comparative literature-based study of deep learning techniques used for deepfake detection. Representative approaches are examined across spatial, temporal, frequency-aware, transformer-based, and hybrid categories. Major public datasets, including FaceForensics++, Celeb-DF, and the DeepFake Detection Challenge dataset, are also compared. The analysis shows that CNN-based methods provide efficient spatial feature extraction, temporal models capture frame-to-frame inconsistencies, frequency-aware methods expose complementary forensic traces, and transformer-based methods offer broader contextual modelling at higher computational cost. Recent benchmark and generalization studies indicate that performance measured only within the training distribution can overstate practical reliability. The paper therefore emphasizes cross-dataset evaluation, robustness to compression, computational efficiency, explainability, and reproducibility as important criteria for practical deepfake detection.

Keywords: Deepfake detection, deep learning, CNN, temporal modelling, frequency-domain analysis, transformers, FaceForensics++, Celeb-DF, DFDC, generalization.

I. INTRODUCTION

The rapid growth of generative artificial intelligence has made it increasingly possible to synthesize or modify visual media so that manipulated faces appear authentic. Deepfake generation commonly involves identity replacement, facial reenactment, expression manipulation, and other forms of synthetic editing. Earlier work such as FaceForensics++ established the importance of benchmark-based evaluation for facial manipulation detection and demonstrated that large-scale manipulated-media datasets can support automated forensic analysis [1].

The detection problem is difficult because modern manipulations can preserve many visual characteristics of genuine media while hiding or weakening traditional forensic traces. Compression, resizing, changes in lighting, camera characteristics, and social-media processing can further alter the evidence available to a detector. MesoNet demonstrated that compact deep neural networks could detect facial video manipulation efficiently, while later work moved toward representations designed to generalize across manipulation methods [2], [3].

A central concern in the literature is generalization. A detector may perform well when training and testing data are drawn from similar distributions but degrade when exposed to unseen manipulation techniques or different datasets. Face X-Ray addressed this issue by learning a representation associated with blending boundaries rather than relying on artifacts tied to one specific manipulation family [4]. More recent work has continued to investigate frequency-domain bias, domain adaptation, and cross-domain generalization [5], [6].



This paper does not claim a new benchmark score. Instead, it synthesizes published research and compares representative approaches using common dimensions: forensic evidence exploited, architecture, computational characteristics, data requirements, robustness, generalization, and practical deployment considerations.

A. Problem Statement

The central problem is reliable identification of manipulated facial images and videos when the detector cannot assume that the manipulation method, dataset, compression level, or content distribution is the same as that used during training. A practical system should minimize false decisions while retaining useful performance on previously unseen media.

B. Aim

The aim of this study is to compare major deep learning techniques for deepfake detection and identify the factors that influence their effectiveness, robustness, generalization, computational requirements, and suitability for practical deployment.

C. Objectives

- Study the nature of deepfake generation and the forensic evidence available for detection.
- Classify representative detection methods into spatial, temporal, frequency-aware, transformer-based, and hybrid categories.
- Compare widely used public datasets and evaluation practices.
- Examine the strengths and limitations of major deep learning approaches.
- Identify generalization, compression, explainability, and computational challenges.
- Develop an evaluation-oriented framework for selecting and validating detection approaches.

II. LITERATURE REVIEW

The literature on deepfake detection has evolved from compact CNN classifiers toward increasingly diverse systems that use temporal information, frequency representations, attention mechanisms, semi-supervised learning, and domain adaptation. Early approaches established that facial manipulation leaves detectable artifacts, but subsequent research has shown that detector performance is highly dependent on data distribution and evaluation protocol.

A. CNN-Based Spatial Detection

CNN-based detectors learn spatial representations directly from face crops. MesoNet used relatively shallow networks to focus on mesoscopic facial properties and reported high detection rates on the evaluated Deepfake and Face2Face data [2]. FaceForensics++ further demonstrated the value of deep learning and domain-specific forensic knowledge under varying compression conditions [1]. CNNs remain attractive because they are comparatively simple to train and deploy, and pretrained image backbones can provide strong feature extraction.

However, a CNN can learn cues correlated with a particular dataset rather than cues intrinsic to manipulation. This becomes problematic when the generation pipeline changes. Gong and Li report that conventional CNN-based approaches can have limited generalization to unseen data and emphasize evaluating detectors across datasets and manipulation types [7].

B. Temporal and Video-Based Detection

A video contains information unavailable in an isolated frame. Temporal approaches attempt to capture inconsistencies across adjacent frames, including changes in facial appearance, motion, or other signals. CNN features can be combined with recurrent or sequence-processing components, while 3D CNNs can jointly model spatial and short-term temporal information. The advantage is a richer evidence source, but the cost is greater data and computational requirements. Frame sampling strategy can also influence results.



C. Frequency-Aware Detection

Frequency-domain analysis provides a complementary view of image evidence. Qian et al. proposed F³-Net, which uses frequency-aware decomposed components and local frequency statistics to detect face forgery [5]. Their findings indicate that frequency information can remain useful under compression, where subtle spatial artifacts may be weakened. Recent research has extended this direction by studying frequency bias itself. Kashiani et al. identified spectral bias in deepfake detectors and proposed FreqDebias to reduce over-reliance on specific frequency bands, reporting improved cross-domain generalization [6].

D. Generalization-Oriented Representations

Face X-Ray represents a notable attempt to learn manipulation-related evidence that is less tied to a particular generator. The method models blending boundaries between image sources and was evaluated on unseen manipulation techniques, where the authors reported stronger generalization than several existing detectors [4]. This direction is important because a useful detector should respond to characteristics shared by different manipulations rather than memorizing one generation pipeline.

E. Transformer and Attention-Based Methods

Transformer-based architectures provide mechanisms for modelling long-range relationships and global contextual information. Recent surveys group transformer-based detection as a distinct research direction and report growing interest in attention-based modelling [7]. Their potential advantage is richer global representation, particularly when artifacts are distributed across a face or across a sequence. The trade-off is computational complexity, memory demand, and a greater need for carefully controlled training data.

F. Benchmark and Standardization Research

DeepfakeBench highlighted a major methodological issue: results reported by different studies can be difficult to compare when preprocessing, datasets, experimental settings, and evaluation metrics differ [8]. The benchmark integrates multiple datasets and detection methods under a more unified evaluation framework. This supports a key principle of the present study: a detector should not be ranked using one accuracy value from one dataset alone.

Study	Main contribution	Evidence / approach	Main implication
FaceForensics++ [1]	Large-scale benchmark for facial manipulation	CNN and forensic cues; compression-aware benchmark	Standardized evaluation is essential
MesoNet [2]	Compact facial video detector	Mesoscopic spatial features	Efficient CNNs can be effective
Face X-Ray [4]	More general forgery representation	Blending-boundary representation	Generalization beyond known manipulation is possible
F ³ -Net [5]	Frequency-aware detection	DCT and local frequency statistics	Frequency cues complement spatial cues
FreqDebias [6]	Frequency-domain generalization	Frequency debiasing + consistency	Avoiding spectral shortcuts improves robustness
DeepfakeBench [8]	Unified benchmark	Multiple datasets/methods/protocols	Fair comparison requires standardized evaluation
Gong & Li [7]	Contemporary survey	CNN, semi-supervised, transformer, biological-signal groups	Generalization remains a major challenge

III. METHODOLOGY

This study uses a structured literature-review and comparative-analysis methodology. The purpose is not to report a new model trained by the authors, but to organize published evidence into a consistent framework. Numerical results



from different papers are not directly combined because datasets, preprocessing, train-test splits, and evaluation measures differ.

A. Literature Selection

The review prioritizes peer-reviewed conference and journal publications, established benchmark papers, dataset papers, and recent survey literature. Foundational studies were retained when they introduced widely used datasets or detection concepts. Recent work was included when it addressed generalization, frequency bias, domain adaptation, or standardized **evaluation**.

B. Inclusion Criteria

- Studies addressing automated detection of manipulated facial images or videos.
- Methods using deep learning, including CNNs, temporal models, frequency-aware models, transformers, or hybrid systems.
- Research introducing or evaluating recognized deepfake datasets or benchmarks.
- Studies examining generalization, robustness, compression, explainability, or deployment.
- Survey and benchmark studies that synthesize or standardize the detection field.

C. Comparative Dimensions

Each method is compared using six dimensions: (1) primary forensic evidence, (2) architecture, (3) computational characteristics, (4) data requirements, (5) robustness/generalization potential, and (6) deployment and explainability considerations. This multi-dimensional view avoids treating accuracy as the only measure of detector quality.

D. Evaluation Principles

The literature supports a layered evaluation protocol. First, in-dataset testing measures whether a model can learn the intended task. Second, cross-dataset testing measures robustness to distribution shift. Third, compression and quality stress tests examine whether the detector remains useful after resizing, transcoding, blur, noise, or other common transformations. Metrics should include accuracy, precision, recall, F1-score, ROC-AUC, false-positive rate, and false-negative rate. For video systems, inference time and resource consumption are also relevant.



Fig. 1. Conceptual Deepfake Detection Pipeline

IV. DEEPPFAKE DETECTION TECHNIQUES

A. Spatial CNN Methods

Spatial CNN methods analyse individual frames or face crops. They are useful as baselines because their behaviour is easier to benchmark than more complex multi-stage systems. Their main limitation is sensitivity to dataset-specific visual cues and reduced access to temporal evidence.

B. Temporal Models

Temporal models process multiple frames and can learn inconsistencies that are not visible in a single image. They are suitable for video-level detection but require careful frame sampling and increased computation. A temporal model can also inherit the weaknesses of the spatial feature extractor on which it depends.

C. Frequency-Aware Models

Frequency-aware models analyse information that may be difficult to observe directly in RGB space. They can complement CNN features, particularly under compression. However, recent work shows that frequency patterns themselves can become model biases, so frequency-based systems should also be evaluated on unseen manipulation types [5], [6].

D. Transformer / Attention Models

Transformers and attention mechanisms can model relationships across larger spatial regions or sequences. Their main attraction is flexible contextual modelling, but the increased computational and data requirements can make them less suitable for resource-constrained deployment.



E. Hybrid and Fusion Approaches

Hybrid systems combine complementary evidence sources, such as spatial and frequency features or spatial and temporal representations. The rationale is that a manipulation may leave different traces in different domains. The drawback is additional model complexity and the need for careful fusion so that one biased feature source does not dominate the final decision.

V. COMPARATIVE ANALYSIS

The comparative analysis indicates that no single architecture can be considered universally best for all deepfake detection conditions. CNNs offer a strong baseline and lower complexity; temporal models provide additional video evidence; frequency-aware methods contribute complementary forensic information; transformers provide broader contextual modelling; and hybrid systems can combine these advantages at the cost of greater complexity.

Approach	Main evidence	Strength	Limitation	Compute	Generalization	Use case
CNN	Spatial artifacts	Simple, efficient	Dataset shortcuts	Low–Med	Moderate	Baseline / edge
CNN + temporal	Frame sequence	Motion consistency	More compute	Med–High	Moderate	Video analysis
3D CNN	Spatio-temporal	Joint feature learning	Data intensive	High	Moderate	Video
Frequency-aware	Spectral cues	Compression complement	Spectral bias possible	Med	Potentially strong	Forensics
Transformer	Global/contextual	Long-range modelling	Memory/data cost	High	Promising	Research / servers
Hybrid fusion	Multiple cues	Complementary evidence	Complex pipeline	High	Potentially strong	High-assurance systems

A. Dataset Comparison

Dataset choice strongly influences reported detector performance. FaceForensics++ contains manipulated material generated with several facial manipulation methods and was designed to support standardized evaluation [1]. Celeb-DF was developed with a focus on realistic visual quality and contains 590 original videos and 5,639 corresponding DeepFake videos in version 2 [9]. The DFDC dataset was created as a large-scale challenge dataset and contains more than 100,000 videos in its full version [10].

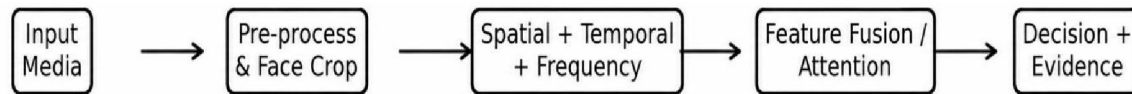
Dataset	Type	Important characteristic	Comparative purpose
FaceForensics++ [1]	Image/video	Multiple manipulation methods; compression settings	Baseline and controlled comparison
Celeb-DF v2 [9]	Video	High visual quality; 5,639 fake + 590 real source videos	Generalization / realistic media
DFDC [10]	Video	Large-scale challenge dataset; >100k videos	Large-scale video evaluation
DeepfakeBench [8]	Benchmark framework	Multiple datasets and standardized protocols	Cross-method comparison

B. Evaluation Metrics

Accuracy is useful when class distributions are balanced, but it should not be reported alone. Precision measures the proportion of predicted fake samples that are actually fake, while recall measures the proportion of fake samples detected. F1-score balances precision and recall. ROC-AUC summarizes ranking performance over classification thresholds. False-positive and false-negative rates are especially important in applications where incorrect authenticity



decisions have different consequences. For video systems, latency, memory usage, and throughput should also be reported.



Evaluation: in-dataset → cross-dataset → compression/quality stress test

Fig. 2. Proposed Evaluation-Oriented Comparative Framework

VI. RESULTS AND DISCUSSION

Because this paper is a literature-based comparative study, the results presented here are synthesized findings rather than new experimental measurements. Reported numerical values from individual papers are not merged into a single leaderboard because the studies use different datasets and protocols.

A. Comparative Findings

First, spatial CNN methods remain important baseline systems because they can learn manipulation artifacts directly from images and can often be implemented with moderate computational cost. However, the literature identifies a gap between strong in-distribution performance and robust performance on unseen data [7], [8].

Second, temporal approaches can provide evidence that is unavailable in a single frame. Their value is greatest when manipulations create inconsistencies across time. The additional information comes with increased processing cost and dependence on frame sampling. Therefore, temporal modelling should be selected when video-level evidence justifies the additional complexity.

Third, frequency-aware methods provide a complementary signal. F³-Net showed that frequency information can expose forgery patterns and remain useful across compression qualities [5]. Recent FreqDebias research further indicates that detectors may become overly dependent on particular frequency bands, demonstrating that frequency analysis should be designed for generalization rather than used as an isolated solution [6].

Fourth, transformer-based methods provide a flexible mechanism for global relationships, but their higher resource requirements can make them less attractive for lightweight applications. Their practical advantage should therefore be demonstrated under consistent cross-dataset evaluation rather than inferred from architecture size.

Finally, hybrid approaches appear promising because deepfakes can leave multiple kinds of evidence. A system that combines spatial, temporal, and frequency information can reduce dependence on a single artifact family. However, fusion increases complexity and must be evaluated carefully to ensure that the model does not simply learn dataset-specific correlations from several input branches.

B. Reported Evidence from Selected Studies

FaceForensics++ demonstrated the value of a large benchmark and reported that domain-specific knowledge could improve detection under compression [1]. MesoNet showed that compact networks could achieve high detection rates on the evaluated manipulation types [2]. Face X-Ray specifically targeted generalizable blending evidence and reported effectiveness on unseen manipulation techniques [4]. F³-Net demonstrated the usefulness of frequency-aware cues, particularly under lower-quality media [5]. More recent generalization research has moved toward domain adaptation and frequency debiasing rather than simply increasing model size [6], [11].

These findings support a broader interpretation of detector performance: the best practical system is not necessarily the one with the highest accuracy on a familiar benchmark. Robustness to distribution shift, compression, unseen generation methods, and operational constraints should be treated as first-class evaluation objectives.



C. Practical Recommendation

For a practical research prototype, a CNN-based spatial detector can be used as the baseline. A frequency-aware branch can then be evaluated as a complementary component, followed by temporal modelling if video-level information is required. Cross-dataset testing should be mandatory before drawing conclusions about generalization. Transformer or multi-branch architectures should be considered when their additional computational cost is justified by measurable robustness improvements.

VII. RESEARCH GAP

The reviewed literature reveals several connected research gaps. The first is the gap between benchmark performance and real-world generalization. Detectors can learn shortcuts associated with identities, backgrounds, compression patterns, or particular generation pipelines. DeepfakeBench was developed in part because inconsistent preprocessing and evaluation practices can make reported results difficult to compare [8].

The second gap concerns robustness after media processing. Online platforms routinely resize and recompress media, which can remove or alter forensic traces. A detector should therefore be tested under controlled quality degradation rather than only on clean benchmark inputs.

The third gap is the absence of a universally reliable detector. As generation methods evolve, artifacts that once distinguished fake content may become less useful. Recent work on domain adaptation and frequency debiasing indicates that continual improvement and adaptation are necessary [6], [11].

The fourth gap is explainability. A binary label or probability score does not necessarily indicate why a sample was classified as fake. For high-impact applications, evidence localization, confidence calibration, and human review can make automated detection more useful and safer.

The fifth gap is efficiency. Large transformer and hybrid models may improve representation capacity but can require substantial computation. Future research should report not only accuracy-related measures but also latency, memory, throughput, and energy requirements when deployment is a goal.

VIII. CHALLENGES AND LIMITATIONS

- Dataset bias: public datasets may overrepresent specific identities, manipulation pipelines, or recording conditions.
- Cross-dataset generalization: performance can decline when the test distribution differs from the training distribution.
- Compression and low quality: transcoding, blur, resizing, and noise may remove useful forensic traces.
- Adversarial adaptation: generation systems can evolve in response to known detection strategies.
- Computational cost: video and multi-branch models can require substantial processing resources.
- Explainability: detector scores do not automatically identify the forensic evidence behind a decision.
- Privacy and ethics: face datasets contain biometric information and require responsible collection, storage, and use.
- No universal detector: deepfake detection should be treated as probabilistic forensic analysis rather than an infallible authenticity oracle.

IX. FUTURE SCOPE

- Development of generalizable detectors trained to recognize manipulation-independent evidence.
- Combination of spatial, temporal, frequency, and multimodal evidence with controlled fusion.
- Self-supervised and semi-supervised learning to reduce dependence on expensive labelled fake data.
- Domain adaptation and continual learning for new manipulation families while limiting catastrophic forgetting.
- Robustness testing under compression, resizing, noise, blur, and real social-media transformations.



- Explainable detection with frame-level or region-level evidence localization and calibrated confidence.
- Lightweight architectures suitable for mobile, edge, and real-time moderation systems.
- Standardized evaluation using multiple datasets and consistent preprocessing and metrics.

X. CONCLUSION

Deepfake detection has developed from compact CNN-based classifiers into a broad research area involving spatial, temporal, frequency-aware, transformer, and hybrid approaches. The literature reviewed in this study shows that each family provides a different form of forensic evidence. CNNs remain useful and efficient baselines, temporal models add information across frames, frequency-aware methods provide complementary spectral evidence, and transformer-based approaches offer broader contextual modelling. Hybrid systems can combine these signals but introduce additional complexity.

The most important finding of the comparative study is that detector quality should not be judged by in-dataset accuracy alone. Generalization to unseen manipulation methods, robustness to compression and quality changes, reproducibility, explainability, and computational cost are equally important for practical deployment. Benchmark work such as FaceForensics++ and DeepfakeBench demonstrates the value of standardized evaluation, while recent studies on domain adaptation and frequency debiasing show that generalization remains an active research challenge [1], [6], [8].

For future implementation, a layered evaluation strategy beginning with a strong CNN baseline and extending to complementary frequency and temporal evidence is recommended. Such a design can provide a practical balance between detection capability, robustness, and computational requirements. The study concludes that reliable deepfake detection is best viewed as an evolving forensic problem requiring continuous evaluation against new forms of synthetic media.

ACKNOWLEDGMENTS

The authors express their sincere gratitude to Prof. Rana Afreen Sheikh for her guidance, valuable suggestions, and academic support during the preparation of this research work.

REFERENCES

- [1] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: Learning to Detect Manipulated Facial Images," in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), 2019, pp. 1–11, doi: 10.1109/ICCV.2019.00009.
- [2] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: a Compact Facial Video Forgery Detection Network," in Proc. IEEE Int. Workshop on Information Forensics and Security (WIFS), 2018, pp. 1–7, doi: 10.1109/WIFS.2018.8630761.
- [3] L. Verdoliva, "Media Forensics and DeepFakes: An Overview," IEEE Journal of Selected Topics in Signal Processing, vol. 14, no. 5, pp. 910–932, 2020.
- [4] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo, "Face X-Ray for More General Face Forgery Detection," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5001–5010, doi: 10.1109/CVPR42600.2020.00505.
- [5] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in Frequency: Face Forgery Detection by Mining Frequency-Aware Clues," in Computer Vision – ECCV 2020, 2020, pp. 86–103, doi: 10.1007/978-3-030-58610-2_6.
- [6] H. Kashiani, N. Alipour Talemi, and F. Afghah, "FreqDebias: Towards Generalizable Deepfake Detection via Consistency-Driven Frequency Debiasing," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2025, pp. 8775–8785, doi: 10.1109/CVPR52734.2025.00820.
- [7] L. Y. Gong and X. J. Li, "A Contemporary Survey on Deepfake Detection: Datasets, Algorithms, and Challenges," Electronics, vol. 13, no. 3, Art. no. 585, 2024, doi: 10.3390/electronics13030585.



- [8] Z. Yan, Y. Zhang, X. Yuan, S. Lyu, and B. Wu, "DeepfakeBench: A Comprehensive Benchmark of Deepfake Detection," in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), 2023.
- [9] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A New Dataset for DeepFake Forensics," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2020.
- [10] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. Ferrer, "The DeepFake Detection Challenge (DFDC) Dataset," arXiv:2006.07397, 2020.
- [11] K. Zhang, Z. Hou, Z. Hua, Y. Zheng, and L. Y. Zhang, "Boosting Deepfake Detection Generalizability via Expansive Learning and Confidence Judgement," IEEE Transactions on Circuits and Systems for Video Technology, vol. 35, no. 1, pp. 953–966, 2025, doi: 10.1109/TCSVT.2024.3462985.
- [12] Y. Li, M. Chang, and S. Lyu, "In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking," in Proc. IEEE Int. Workshop on Information Forensics and Security (WIFS), 2018.
- [13] Y. Li and S. Lyu, "Exposing DeepFake Videos by Detecting Face Warping Artifacts," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops (CVPRW), 2019.
- [14] D. Güera and E. J. Delp, "Deepfake Video Detection Using Recurrent Neural Networks," in Proc. IEEE Int. Conf. Advanced Video and Signal Based Surveillance (AVSS), 2018.
- [15] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1251–1258.

