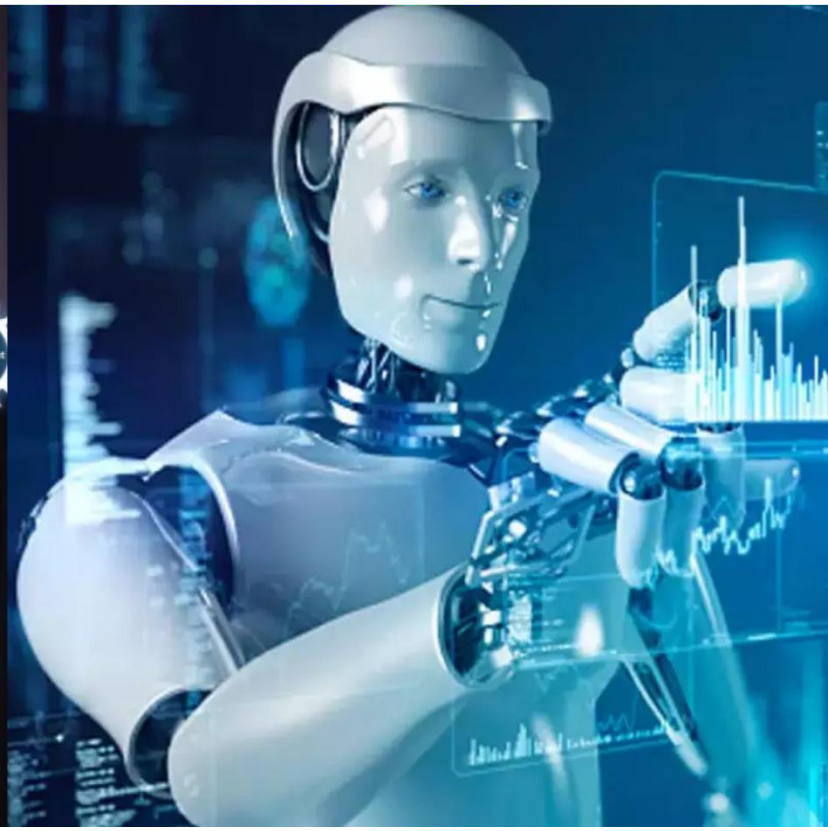
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 15, Issue 4, April 2026

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 4, April 2026

|DOI: 10.15680/IJIRSET.2026.1504128|

Cross-Dataset Evaluation of Deepfake Detection Models: Assessing Reliability and Robustness

Koteshwar Shankar Katkuri

Department of Data Science, SIES College of Arts, Science & Commerce (Empowered Autonomous), Mumbai, India

ABSTRACT: Rapid progress in generative deep learning has made fabricated facial media increasingly convincing, creating serious threats to digital authenticity. While many published deepfake detection models report outstanding performance on standard benchmarks, their capacity to generalise beyond the training distribution is seldom examined. This paper presents a controlled cross-dataset experimental study evaluating four architectures—ResNet-50, Xception, EfficientNet-B4, and Vision Transformer (ViT)—trained solely on one facial image dataset and tested on a completely separate dataset. Under intra-dataset conditions every model converged to near-perfect accuracy; however, cross-dataset accuracy for all four architectures collapsed to approximately 50%, equivalent to random guessing. Per-class analysis further revealed extreme recall asymmetry, with some models almost entirely ignoring the fake class. These findings establish that source-specific artefact overfitting, dataset bias, and the absence of standardised cross-dataset evaluation protocols are the primary obstacles to deploying reliable deepfake detectors in real-world contexts.

KEYWORDS: Cross-dataset generalisation, deepfake detection, face forgery forensics, model robustness, transfer learning.

I. INTRODUCTION

Generative adversarial networks and neural rendering have fundamentally altered the landscape of synthetic media production [1]. Among the most consequential outputs of these technologies is the deepfake, a digitally manufactured video or image in which a person's face is convincingly replaced or re-enacted with another. The social risks are substantial: fraudulent content has been weaponised in disinformation campaigns, non-consensual imagery, identity theft, and political manipulation [2].

Automated detection is the primary defence against deepfake misuse. First-generation detectors relied on hand-crafted physiological markers—irregular eye-blink rates, implausible head-pose trajectories, and facial landmark distortions [6][12]. As generation pipelines matured, these rule-based approaches became insufficient. Convolutional neural networks (CNNs) then took centre stage, learning discriminative spatial features directly from pixel data [3][2]. Frequency-domain analysis further extended this capability by capturing spectral artefacts that survive lossy compression [17][20]. Most recently, Vision Transformers (ViTs) have been explored for their ability to model long-range dependencies across facial regions through self-attention mechanisms [4][5].

Despite steady architectural progress, a critical methodological blind spot persists. The overwhelming majority of published evaluation protocols operate entirely within a single dataset—models are trained and tested on different partitions of the same source. This intra-dataset paradigm yields impressive accuracy numbers yet fails to approximate operational conditions, where manipulated content arises from diverse, previously unseen generative processes [22][24].This paper directly addresses that gap. Four architectures spanning convolutional and attention-based paradigms are each trained on one dataset (Dataset 1) and tested on a completely disjoint dataset (Dataset 2). The resulting performance gap quantifies generalisation failure and provides actionable insight into the root causes of model brittleness.

II. LITERATURE SURVEY

A. Early Hand-Crafted Feature Approaches

The earliest deepfake detection methods were grounded in observable physiological and geometric cues. Li et al. [6] observed that GAN-based face synthesis rarely captures complete eye-closure sequences during training, leading to abnormal blink patterns in generated videos. Their detector exploited this statistical gap by monitoring temporal blink

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 4, April 2026

|DOI: 10.15680/IJIRSET.2026.1504128|

frequency and duration, achieving reliable classification on first-generation deepfake content. While effective at the time of publication, this approach became obsolete once generative pipelines began incorporating full eye-motion dynamics.

Yang et al. [12] proposed a complementary approach based on head-pose estimation. By fitting a 3D face model to predicted facial landmarks, their method detected inconsistencies between the inner and outer face regions in swapped content. These geometric inconsistencies arise because face-swapping pipelines manipulate the central facial region independently of the surrounding head and neck, introducing subtle but measurable pose misalignments. Although informative, landmark-based approaches are susceptible to misdetection when high-quality blending masks are applied.

Li and Lyu [7] extended physiological analysis to face warping artefacts, demonstrating that affine transformation boundaries introduce warping distortions detectable through resolution mismatch analysis. Their work highlighted the importance of examining boundary regions between the pasted face and the original background, a cue that deep learning approaches would later learn implicitly. Thies et al. [13], as creators of the Face2Face reenactment system, demonstrated how expression transfer could be accomplished in real time using tracked facial parameters, simultaneously illustrating the difficulty of the detection problem. Collectively, these early methods established valuable detection benchmarks but exhibited narrow applicability, failing to generalise across manipulation techniques and quickly becoming ineffective as synthesis quality improved.

B. Deep Learning and CNN-Based Detection

The limitations of hand-crafted feature detectors motivated a shift toward data-driven deep learning approaches. Afchar et al. [1] introduced MesoNet, a compact CNN architecture designed specifically for facial forgery detection. By operating at the mesoscopic level—between pixel-level and semantic-level analysis—MesoNet captured blending artefacts invisible to hand-crafted methods. Despite its lightweight design, MesoNet achieved competitive performance on early benchmark datasets, demonstrating that even shallow CNNs could learn meaningful manipulation signatures. However, as the authors themselves acknowledged, even compact models tend to overfit rapidly to source-specific cues when training data lacks diversity.

The publication of FaceForensics++ by Rössler et al. [2] represented a major milestone for the field. This large-scale benchmark encompassed four manipulation categories—Deepfakes, Face2Face, Face Swap, and Neural Textures—across over one thousand video sequences, enabling systematic evaluation of detection methods. The benchmark established XceptionNet [3] as the reference CNN architecture for deepfake detection, owing to its use of depthwise separable convolutions that efficiently capture fine-grained spatial artefacts within blended facial regions. FaceForensics++ became the de facto standard evaluation protocol, though its controlled laboratory conditions and limited manipulation diversity would later be identified as a source of overfitting bias.

Subsequent CNN research pursued two parallel goals: improving detection accuracy and reducing computational cost. EfficientNet-based detectors [15] applied compound scaling to simultaneously optimise network depth, width, and input resolution, yielding models with favourable accuracy-to-parameter ratios well suited for practical deployment. Nguyen et al. [8] proposed a capsule-network architecture that leveraged part-whole relationship encoding to detect spatial inconsistencies more robustly than standard convolution operations. Capsule networks demonstrated improved resistance to spatial transformations, suggesting that encoding explicit geometric relationships between facial components could complement standard feature learning. Güera and Delp [9] extended temporal analysis by applying long short-term memory networks to sequences of CNN-extracted frame features, capturing inter-frame inconsistencies in facial motion that single-frame detectors inevitably miss. These temporal approaches underlined the importance of video-level analysis for detecting manipulation techniques that maintain per-frame realism while introducing cross-frame artefacts.

More recent CNN advances have explored multi-stream architectures and attention mechanisms. Wang et al. [18] proposed a dual-stream framework that simultaneously processed RGB spatial features and edge-map representations, with a cross-attention module fusing complementary information from both streams. Zhao et al. [19] introduced multi-attentional deepfake detection, using multiple spatial attention maps to direct the model toward the most informative facial sub-regions simultaneously, improving sensitivity to localised manipulation traces. Cozzolino et al. [21] demonstrated that camera model fingerprints, extracted through a CNN-based Noiseprint technique, carry latent

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 4, April 2026

|DOI: 10.15680/IJIRSET.2026.1504128|

signatures that differentiate genuine and synthesised imagery, suggesting that even apparently noise-like signals encode actionable forensic information. These developments collectively raised the intra-dataset accuracy ceiling but did not substantially address the fundamental challenge of cross-dataset generalisation.

C. Frequency-Domain and Signal-Level Analysis

A parallel line of investigation examined whether deepfake detection could benefit from frequency-domain representations. Wang et al. [16] made the notable observation that CNN-generated images exhibit spectral anomalies in their high-frequency components, arising from the upsampling operations intrinsic to most GAN decoder architectures. These spectral fingerprints proved surprisingly persistent even after JPEG compression and resizing operations commonly applied to shared media, enabling reliable detection using simple frequency-domain classifiers. Their findings suggested that spatial-domain feature learning alone leaves significant discriminative information unexploited.

Qian et al. [17] formalised frequency-aware detection through a dedicated face forgery detection network that mines frequency-domain clues across multiple frequency bands. By decomposing images into sub-band frequency maps and learning cross-band correlations alongside spatial CNN features, their F3-Net achieved improved robustness against post-processing degradations. Frank et al. [20] reinforced these findings by analysing the phase spectrum of GAN-generated images, demonstrating that phase-level periodicity artefacts complement amplitude-based spectral signatures. Together, these works established frequency-domain analysis as a valuable auxiliary modality for deepfake detection, though cross-dataset evaluations revealed that spectral fingerprints are also generation-method-specific, limiting their transferability across different GAN architectures.

D. Transformer-Based Architectures

The introduction of the Vision Transformer (ViT) by Dosovitskiy et al. [4] offered a fundamentally different architectural paradigm. By dividing input images into fixed-size non-overlapping patches and processing them as sequential tokens through multi-head self-attention layers, ViT discarded the locality inductive bias of convolution entirely. This allowed the model to capture interactions between spatially distant facial regions, potentially identifying globally distributed manipulation inconsistencies that local CNN filters cannot detect. Empirical studies confirmed that ViT models achieve strong recognition performance when pre-trained on large-scale datasets such as ImageNet-21k, though their data efficiency is lower than that of CNNs on limited data.

Liu et al. [5] addressed ViT's computational limitations through the Swin Transformer, which computes self-attention within local non-overlapping windows that shift between layers to enable cross-window information exchange. The resulting hierarchical feature maps enable multi-scale analysis comparable to CNN feature pyramids while retaining the global contextual modelling advantage of attention mechanisms. When applied to deepfake detection, Swin-based models can simultaneously analyse localised blending boundaries and globally distributed texture inconsistencies across an entire face image. However, the computational overhead of large transformer models presents practical deployment challenges, particularly in resource-constrained environments.

E. Benchmark Datasets and Evaluation Protocols

The development of large-scale benchmark datasets has been essential for advancing the field. Li et al. [10] introduced Celeb-DF, a high-quality deepfake dataset constructed from celebrity interview videos using an improved face-swapping algorithm that reduces visible blending artefacts compared to earlier benchmarks. Their analysis demonstrated that Celeb-DF poses substantially greater detection difficulty than FaceForensics++, motivating more rigorous evaluation standards. Dolhansky et al. [11] compiled the DeepFake Detection Challenge (DFDC) dataset on behalf of Facebook AI, encompassing over 100,000 video clips from paid actors using multiple generation methods and encompassing demographic diversity. The DFDC dataset highlighted the gap between laboratory detection performance and real-world conditions involving diverse individuals and recording environments. Korshunov and Marcel [14] conducted a human perceptual study comparing detection accuracy between human observers and automated systems, finding that humans perform poorly on high-quality deepfakes, underlining the severity of the societal threat.

F. Generalisation Challenges and Cross-Dataset Limitations

Despite the significant progress documented across all architectural families, a persistent and critical limitation recurs throughout the literature: detection models trained and evaluated within a single dataset consistently fail to maintain

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 4, April 2026

|DOI: 10.15680/IJIRSET.2026.1504128|

their performance when applied to data from a different source. Vakili and Escobar [22] conducted a systematic analysis of this generalisation gap, cataloguing multiple published studies in which accuracy on an intra-dataset test partition exceeded 95% yet dropped below 60% when the same model was applied to an alternative benchmark. They attributed this degradation primarily to three sources of dataset bias: differences in the GAN pipeline used for manipulation, variation in post-processing steps such as compression and resizing, and inconsistencies in the preprocessing conventions applied during dataset preparation.

Mirsky and Lee [24] provided a comprehensive taxonomy of deepfake creation and detection methods, identifying cross-dataset generalisation as one of the central open problems in the field. Their survey observed that many proposed methods are evaluated only on the dataset used for training, rendering their reported accuracies uninformative with respect to real-world deployment. Zhang et al. [23] reinforced this perspective by demonstrating that local feature statistics effective for discrimination within one dataset—such as patch-level noise residuals and gradient magnitude distributions—lose discriminative power when distribution shifts occur, as the statistical properties of those features differ fundamentally across datasets produced by different generation pipelines. Westerlund [25] situates these technical challenges within the broader societal context, arguing that the rapid democratisation of deepfake technology has outpaced the maturation of detection capabilities, with the generalisation gap representing the most urgent technical barrier to effective deployment.

Dang-Nguyen et al. [15] provided a broad overview of media forensics and deepfake detection, emphasising that evaluation methodologies must move beyond single-dataset benchmarks to provide meaningful assessments of robustness. Their review documented multiple cases in which architectures achieving state-of-the-art accuracy on FaceForensics++ exhibited substantially degraded performance on Celeb-DF or DFDC, despite being evaluated under the same binary classification setting. Wang et al. [18] and Zhao et al. [19] separately confirmed that even architecturally sophisticated detectors incorporating attention mechanisms and multi-stream representations remain susceptible to cross-dataset performance degradation when trained without explicit domain generalisation objectives. These converging observations across independent studies constitute the principal motivation for the present work, which applies a controlled cross-dataset evaluation framework to four representative architectures in order to quantify and characterise the generalisation failure systematically.

III. PROBLEM STATEMENT

A fundamental tension exists between the accuracy figures reported in deepfake detection literature and the functional reliability of those models outside controlled laboratory settings. When training and evaluation data share the same generative source, compression scheme, and acquisition conditions, inflated estimates of detection capability result. Operational environments present manipulated media drawn from heterogeneous pipelines; models that have memorised source-specific artefacts produce unreliable decisions in these settings. Moreover, the field lacks agreed-upon cross-dataset evaluation protocols, making it impossible to objectively compare or rank the real-world robustness of competing proposals.

IV. OBJECTIVES

The study pursues five specific objectives:
(1) Measure the intra-dataset accuracy ceiling of each selected architecture under controlled training conditions.
(2) Quantify cross-dataset performance degradation without any target-domain adaptation.
(3) Characterise per-class error patterns to distinguish symmetric accuracy decline from recall asymmetry.
(4) Contrast the generalisation behaviour of CNN-based and transformer-based architectures.
(5) Derive practical recommendations for evaluation methodology and future model design.

V. METHODOLOGY

A. Dataset Configuration

Two independent facial image datasets were selected to enable rigorous cross-domain testing. Dataset 1, containing labelled real and manipulated images, served exclusively as the training corpus and the source for intra-dataset evaluation. Dataset 2—comprising 10,905 test samples across real and fake sub-directories—was held completely

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 4, April 2026

|DOI: 10.15680/IJIRSET.2026.1504128|

separate and used solely for cross-dataset testing. This strict separation eliminates distribution leakage and ensures that any detection success on Dataset 2 reflects genuine transferability rather than memorisation.

B. Preprocessing

All images from both datasets were uniformly resized to 224 × 224 pixels, matching the canonical input resolution for ImageNet-pretrained models and balancing spatial detail against GPU memory constraints (NVIDIA RTX 3050, 4 GB VRAM). Pixel values were normalised with ImageNet mean and standard deviation. A folder-based labelling scheme compatible with the PyTorch Image Folder loader provided binary class assignments.

C. Model Architectures

Four architectures were selected to span the convolutional and attention-based design paradigms: (i) ResNet-50—a residual network providing a strong intra-dataset baseline; (ii) Xception—employing depthwise separable convolutions shown effective on FaceForensics++ [3]; (iii) EfficientNet-B4—a compound-scaled network balancing accuracy and efficiency; and (iv) ViT-Base/16—a pure transformer partitioning images into 16 × 16 patch tokens [4]. All models were initialised from ImageNet pre-trained weights; classification heads were replaced with two-class linear layers.

D. Training Protocol

Each model was trained exclusively on Dataset 1, simulating the realistic constraint of having labelled data from a single source. Cross-entropy loss was minimised using the Adam optimiser with a learning rate of 1×10^{-4}. Batch sizes were tuned to hardware capacity. Training ran for two to three epochs, limiting the opportunity for overfitting while allowing loss convergence. No data augmentation beyond normalisation was applied so that all architectural differences in cross-dataset generalisation could be attributed to model design rather than augmentation strategy.

E. Evaluation Protocol

Each trained model was assessed under two complementary conditions. Intra-dataset evaluation used held-out samples from Dataset 1 to establish a performance baseline. Cross-dataset evaluation applied the identical trained model—without fine-tuning—to all samples in Dataset 2. Reported metrics include classification accuracy, per-class precision, recall, and F1-score, plus confusion matrices. The magnitude of the accuracy gap between the two conditions serves as the principal measure of generalisation capability.

VI. EXPERIMENTAL RESULTS

A. Intra-Dataset Performance

All architectures converged rapidly and reliably on Dataset 1. ResNet-50 reached training accuracy of 98.84% across three epochs and achieved 100% on the Dataset 1 test set. Xception similarly recorded 99.22% training accuracy in two epochs and 100% test accuracy. EfficientNet-B4 attained 96.80% training accuracy and 93.79% test accuracy. The ViT achieved 88.85% training accuracy in a single epoch pass. Steadily decreasing loss curves across all models confirmed stable optimisation. These results are consistent with published intra-dataset benchmarks and demonstrate that each architecture successfully captured the manipulation signatures present in Dataset 1.

B. Cross-Dataset Performance

When deployed on Dataset 2 without domain adaptation, all four models exhibited precipitous accuracy declines. Table I summarises the complete quantitative outcomes. Xception achieved the highest cross-dataset accuracy at 51.94%, with ResNet-50 at 49.89%, EfficientNet-B4 at 49.78%, and the ViT at 49.68%. Every figure clusters near random-chance performance (50%), confirming that no architecture transferred discriminative knowledge across the dataset boundary.

TABLE I: Performance Summary — Intra-Dataset vs. Cross-Dataset

Model	Intra Acc. (%)	Cross Acc. (%)	Fake Recall	Real Recall	F1 Score
ResNet-50	100.00	49.89	0.07	0.93	0.12

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 4, April 2026

|DOI: 10.15680/IJIRSET.2026.1504128|

Model	Intra Acc. (%)	Cross Acc. (%)	Fake Recall	Real Recall	F1 Score
Xception	100.00	51.94	0.47	0.57	0.50
EfficientNet-B4	93.79	49.78	0.34	0.66	0.40
Vision Transformer	88.85	49.68	0.00	1.00	0.01

Per-class recall patterns expose contrasting failure modes. ResNet-50 achieved 0.93 real-class recall but only 0.07 fake-class recall—classifying nearly all samples as real. The ViT took this tendency to its extreme, attaining a perfect 1.00 real recall with 0.00 fake recall and effectively operating as a single-class classifier. Xception exhibited more balanced but still degraded performance (fake recall 0.47, real recall 0.57). EfficientNet-B4 fell between these extremes with 0.34 fake recall and 0.66 real recall. Confusion matrix analysis corroborates these patterns: the ViT confusion matrix shows [25, 5467; 20, 5393], indicating an overwhelming preference for the real label on Dataset 2.

VII. DISCUSSION

The experimental outcomes present a consistent narrative: architectures that are individually capable of near-perfect intra-dataset discrimination share a universal weakness—they encode source-specific manipulation fingerprints rather than manipulation-invariant features. When those fingerprints are absent in an unseen dataset, the learned decision boundary becomes arbitrary, and accuracy converges on chance.

The observed recall asymmetry illuminates the nature of this overfitting. Models trained on Dataset 1 appear to have learned that the majority of Dataset 2 imagery superficially resembles the real class of their training distribution. Without access to the particular compression artefacts, GAN upsampling patterns, or blending boundaries that distinguished fake faces in Dataset 1, the models default to predicting real. This behaviour is especially pronounced in the ViT, whose global self-attention mechanism may reinforce holistic face representations at the expense of local manipulation cues.

Xception's comparatively balanced—though still poor—cross-dataset performance is notable. Depthwise separable convolutions explicitly decouple spatial and channel-wise feature extraction; this architectural inductive bias may discourage over-concentration on dataset-specific spectral signatures, yielding marginally broader generalisation. The implication for future work is that architecture choices influence not only intra-dataset accuracy but also the type of generalisation failure.

The findings reinforce the research gap identified in the Problem Statement: intra-dataset accuracy is an insufficient proxy for real-world utility. Any detection model intended for operational deployment should be evaluated against at least one held-out dataset originating from a different generative pipeline before accuracy claims are accepted.

VIII. FUTURE SCOPE

Closing the generalisation gap identified in this study requires advances in training strategy, model architecture, and evaluation methodology. On the training side, multi-source training—simultaneously ingesting data from FaceForensics++ [2], Celeb-DF [10], and DFDC [11]—forces models to discover manipulation cues that transcend any individual pipeline. Combining this with domain generalisation objectives, such as inter-domain contrastive alignment, can further reduce sensitivity to source-specific artefact distributions.

Self-supervised and contrastive pre-training offer a complementary path. By constructing pretext tasks that require distinguishing genuine facial dynamics from synthesised ones across diverse unlabelled data, models can acquire transferable representations before supervised fine-tuning. This reduces dependence on large manually annotated corpora, which are expensive to produce as generative techniques proliferate.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 4, April 2026

|DOI: 10.15680/IJIRSET.2026.1504128|

Temporal modelling represents an orthogonal extension. Image-based detectors discard frame-to-frame consistency information that is often retained even by high-fidelity face-swapping pipelines. Recurrent architectures [9] and video transformers operating on clip-level representations may exploit these temporal artefacts, improving robustness against spatially realistic but temporally inconsistent forgeries.

Finally, the community needs standardised cross-dataset benchmarks analogous to those in speaker verification and machine translation. A shared evaluation framework—specifying which dataset pairs are used for training and testing, and mandating reporting of per-class metrics—would enable objective comparison of detection methods and provide an honest measure of deployment readiness.

IX. CONCLUSION

This paper presented a systematic cross-dataset evaluation of four deepfake detection architectures. Controlled experiments confirmed that ResNet-50, Xception, EfficientNet-B4, and the Vision Transformer each achieve near-perfect performance on the dataset used for training while collectively failing to exceed random-chance accuracy on an unseen target dataset. The magnitude and consistency of this degradation provide strong empirical evidence that dataset-specific overfitting—rather than inadequate architectural capacity—is the dominant obstacle to robust deepfake detection.

Per-class recall analysis further revealed systematic biases in failure mode: most architectures collapse toward classifying unseen images as real, reflecting the dominance of real-class features in the cross-dataset distribution. Xception's marginally more balanced cross-dataset recall suggests that architectural inductive biases influence generalisation behaviour, motivating future design choices that prioritise manipulation-invariant feature extraction.

The study's principal contribution is empirical and methodological: it demonstrates through reproducible experiments that cross-dataset evaluation must become a standard reporting requirement in deepfake forensics research. Architectures and training strategies that perform well only under intra-dataset conditions cannot be considered deployment-ready. Achieving true reliability will require multi-source training, domain-robust learning objectives, and the community-wide adoption of standardised cross-dataset benchmarks.

REFERENCES

[1] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A compact facial video forgery detection network," in Proc. IEEE Int. Workshop Inf. Forensics Security (WIFS), 2018.
[2] 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.
[3] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in Proc. IEEE Conf. Computer Vision Pattern Recognition (CVPR), 2017.
[4] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in Proc. Int. Conf. Learning Representations (ICLR), 2021.
[5] Z. Liu et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), 2021.
[6] Y. Li, M.-C. Chang, and S. Lyu, "In ictu oculi: Exposing AI generated fake face videos by detecting eye blinking," in Proc. IEEE Int. Workshop Inf. Forensics Security (WIFS), 2018.
[7] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," in Proc. IEEE Conf. Computer Vision Pattern Recognition Workshops (CVPRW), 2019.
[8] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-forensics: Using capsule networks to detect forged images and videos," in Proc. IEEE Int. Conf. Acoustics, Speech Signal Process. (ICASSP), 2019.
[9] D. Güera and E. J. Delp, "Deepfake video detection using recurrent neural networks," in Proc. IEEE Int. Conf. Advanced Video Signal Based Surveillance (AVSS), 2018.
[10] Y. Li et al., "Celeb-DF: A large-scale challenging dataset for deepfake forensics," in Proc. IEEE/CVF Conf. Computer Vision Pattern Recognition (CVPR), 2020.
[11] B. Dolhansky et al., "The deepfake detection challenge (DFDC) dataset," arXiv preprint arXiv:2006.07397, 2020.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 4, April 2026

|DOI: 10.15680/IJIRSET.2026.1504128|

[12] X. Yang, Y. Li, and S. Lyu, "Exposing deep fakes using inconsistent head poses," in Proc. IEEE Int. Conf. Acoustics, Speech Signal Process. (ICASSP), 2019.
[13] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2Face: Real-time face capture and reenactment of RGB videos," in Proc. IEEE Conf. Computer Vision Pattern Recognition (CVPR), 2016.
[14] P. Korshunov and S. Marcel, "Deepfake detection: Humans vs. machines," arXiv preprint arXiv:2009.03186, 2020.
[15] K. Dang-Nguyen et al., "Media forensics and deepfakes: An overview," IEEE J. Sel. Topics Signal Process., vol. 14, no. 5, pp. 969–984, 2020.
[16] S. Wang et al., "CNN-generated images are surprisingly easy to spot... for now," in Proc. IEEE/CVF Conf. Computer Vision Pattern Recognition (CVPR), 2020.
[17] Y. Qian et al., "Thinking in frequency: Face forgery detection by mining frequency-aware clues," in Proc. European Conf. Computer Vision (ECCV), 2020.
[18] H. Wang, Z. Wang, and X. Xie, "Dual-stream CNN for deepfake detection," in Proc. IEEE Int. Conf. Multimedia Expo (ICME), 2020.
[19] Y. Zhao et al., "Multi-attentional deepfake detection," in Proc. IEEE/CVF Conf. Computer Vision Pattern Recognition (CVPR), 2021.
[20] J. Frank et al., "Leveraging frequency analysis for deep fake image recognition," in Proc. Int. Conf. Machine Learning (ICML), 2020.
[21] D. Cozzolino, G. Poggi, and L. Verdoliva, "Noiseprint: A CNN-based camera model fingerprint," IEEE Trans. Inf. Forensics Security, vol. 15, pp. 144–159, 2020.
[22] A. Vakili and M. J. Escobar, "Generalisation challenges in deepfake detection," IEEE Access, vol. 9, pp. 151–162, 2021.
[23] Y. Zhang, F. Ding, and S. Kwong, "Detecting deepfake images via local feature statistics," Pattern Recognition, vol. 116, 2021.
[24] S. Mirsky and W. Lee, "The creation and detection of deepfakes: A survey," ACM Comput. Surv., vol. 54, no. 1, pp. 1–41, 2021.
[25] M. Westerlund, "The emergence of deepfake technology: A review," Technol. Innov. Manage. Rev., vol. 9, no. 11, pp. 40–53, 2019.

ABOUT THE AUTHOR

Koteshwar Shankar Katkuri (RollNo.SMSC2526125) is a postgraduate student in the Department of Data Science at SIES College of Arts, Science & Commerce (Empowered Autonomous), Mumbai, India. His research interests centre on trustworthy artificial intelligence, media forensics, and the design of robust deep learning evaluation frameworks. He conducted this study as part of his capstone research, implementing and benchmarking deepfake detection models across diverse datasets using PyTorch on an NVIDIA RTX 3050 GPU. He can be contacted at koteshwarsmsc2526125@ascs.sies.edu.in

Dr. Abuzar Ansari is Professor and Head of the Department of Data Science at S.I.E.S. College of Arts, Science & Commerce (Autonomous), Mumbai. His research spans applied machine learning, educational data mining, and natural language processing. He serves as Principal Investigator for multiple funded data science research projects within the institution.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details