IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15626

Generalization and Cross-Dataset Robustness in Deepfake Detection: An Enhanced XceptionNet Approach

Sayli Patil[1*], Sameer Maheboob Shaikh[2], Sarwar Ali Mukhtar Ahemad Iddirisi[3]

Assistant Professor, Department of Computer Science, Nowrosjee Wadia College (Autonomous), Pune, Maharashtra, India[1*]

Student, Department of Computer Science, Nowrosjee Wadia College (Autonomous), Pune, Maharashtra, India[2-3]

Abstract: Maintaining the integrity of digital visual media in the age of generative AI requires robust, automated frameworks capable of identifying sophisticated facial manipulations. This paper presents the design and evaluation of a specialized benchmarking framework for Deepfake detection, developed using an XceptionNet architecture to analyze and improve cross-dataset generalization. The proposed system compares a baseline detector against an enhanced variant (V2) that integrates strategic data augmentation and domain-specific fine-tuning to bridge the performance gap between source and target datasets. To ensure evaluation rigor, the framework supports multi-seed experimentation with deterministic sampling, enabling statistically grounded comparisons across independent training runs. This reproducible design eliminates variance-driven conclusions and strengthens the reliability of cross-dataset generalization findings. . Researchers are provided with a diagnostic toolkit that utilizes ROC/AUC analysis and bootstrap statistics to ensure the reliability and significance of detection metrics. Experimental results indicate that the targeted training interventions significantly enhance the model's ability to maintain high accuracy across unseen distributions without degrading performance on original training data. This research demonstrates how a systematic benchmarking approach can diagnose model weaknesses and provide a reproducible pathway toward developing more resilient and "wild-ready" Deepfake detection systems.

I. INTRODUCTION

The rapid evolution of generative adversarial networks (GANs) and diffusion models has made the creation of photorealistic facial manipulations known as “Deepfakes” increasingly accessible to the general public. While these advancements offer significant creative potential in digital media and cinema, they simultaneously pose severe threats to individual privacy, political stability, and the foundational “digital trust” of our information ecosystems. As AI-generated face swaps become indistinguishable from authentic footage, the necessity for building automated detection systems that work reliably in real-world scenarios has transitioned from a theoretical challenge to a global security priority.

However, a significant hurdle in current Deepfake detection research is the “Generalization Gap,” where detectors that achieve near-perfect accuracy on their specific training datasets frequently suffer from catastrophic performance degradation when encountering unseen data from different generative sources. Furthermore, real-world visual media is rarely pristine; it is often subjected to stochastic perturbations such as JPEG compression, resizing, and Gaussian blurring during transmission across social media platforms. A detection system that cannot maintain robustness against these distribution shifts and image distortions is practically ineffective for deployment in forensic environments.

This paper presents a practical framework for evaluating and improving detection models with a specific focus on cross-dataset robustness. The proposed system utilizes an XceptionNet backbone to benchmark two distinct model variants: a baseline detector (V1) trained on standard source datasets, and an enhanced version (V2) that applies strategic data augmentation and selective domain fine-tuning. By comparing these variants, the research identifies specific training interventions that improve performance on target datasets without degrading results on the original source data. Designed to be lightweight and reproducible, the approach uses efficient sampling and bootstrap statistics to ensure that the performance gains are statistically significant.

Beyond simple accuracy metrics, this work delivers a diagnostic analysis toolkit for evaluating model behavior in the wild. Our evaluation incorporates a comprehensive diagnostic suite featuring ROC/AUC curve analysis, confidence score histograms, confusion matrix visualization, and misclassification sampling with error thumbnails to systematically identify and explain model failure patterns across datasets. . This research provides a standardized benchmarking

DOI: 10.17148/IJARCCE.2026.15626

methodology and a set of practical training improvements that guide the development of more resilient Deepfake detection systems, reducing the gap between laboratory success and real-world readiness.The paper is organized as follows. Section II is primarily focused on background and related work. Section III presents the system architecture and methodology details followed by section IV model concretisation. The result analysis is presented in Section V. The paper concludes with a conclusion and future scope.

II. BACKGROUND AND RELATED WORK

Deepfake detection research has gradually moved from focusing only on high accuracy on a single dataset to solving the more realistic problem of cross-dataset generalization. Early research showed that although detectors perform very well when trained and tested on the same dataset, they perform drops when evaluated on unseen datasets or new manipulation techniques [1]. Further studies confirmed that this happens because many models learn dataset-specific artifacts instead of general manipulation patterns [2]. Broader evaluation across different architectures and datasets also demonstrated that relying on a single benchmark gives an overly optimistic picture of real-world robustness [3]. Later benchmark efforts reinforced this conclusion and clearly established the "generalization gap" as a central limitation in modern deepfake detection systems [4].

To address this issue, researchers began exploring improved learning strategies. One important direction was training models for manipulation attribution instead of simple binary classification, which helps structure the feature space and improves transferability across datasets [5]. Building on this idea, multi-task self-supervised frameworks introduced pseudo-fake generation from pristine images, allowing models to learn general inconsistencies rather than generator-specific fingerprints [6]. Techniques such as self-blended image (SBI) generation create realistic artifacts that improve cross-dataset robustness. Additional architectural strategies, such as cross-branch orthogonality, encourage diverse and non-redundant feature learning, which further enhances generalization [7].

As research progressed, attention shifted toward open-world robustness. Unsupervised domain adaptation methods were proposed feature representations [8]. At the same time, systematic reviews highlighted weaknesses in benchmarking practices and emphasized the importance of comprehensive cross-dataset evaluation to ensure reliability in real-world deployment [9].

More recently, researchers have explored parameter-efficient adaptation of large pretrained models, demonstrating that selectively fine-tuning small components (such as normalization layers) can achieve strong cross-benchmark performance without updating the entire network [10]. Collectively, these ten studies shows that robust deepfake detection require multiple strategies working together: structured representation learning, pseudo-fake data generation, architectural diversity, domain adaptation, and rigorous cross-dataset evaluation.Although significant progress has been made, achieving stable generalization against rapidly evolving generative models, especially diffusion-based techniques, remains an ongoing research challenge.

III. SYSTEM ARCHITECTURE AND METHODOLOGY

The comprehensive architecture of the proposed Deepfake detection framework is detailed in Figure 1. The system is designed not just for high accuracy on training data, but specifically to identify and remediate the "generalization gap" encountered when models are deployed in real-world scenarios. The architecture follows a modular three-stage processing topology:

- **Efficient Preprocessing:** Raw images sequences are processed through an MTCNN-based localization stage to extract standardized facial crops. This ensures the model focuses on facial gradients rather than background noise.
- **Dual-Track Methodology:** The framework benchmarks a **Baseline V1** (utilizing a frozen backbone and standard classifier) against an **Improved V2**. The V2 variant employs an **Active Backbone** where weights are iteratively updated through selective fine-tuning and **Robust Augmentation** (simulating JPEG compression and Gaussian blur).
- **Cross-Dataset Evaluation:** Both tracks are evaluated against the unseen **Random** dataset to measure cross-dataset robustness, culminating in a diagnostic suite that isolates high-confidence failures and suggests future adversarial training paths.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15626

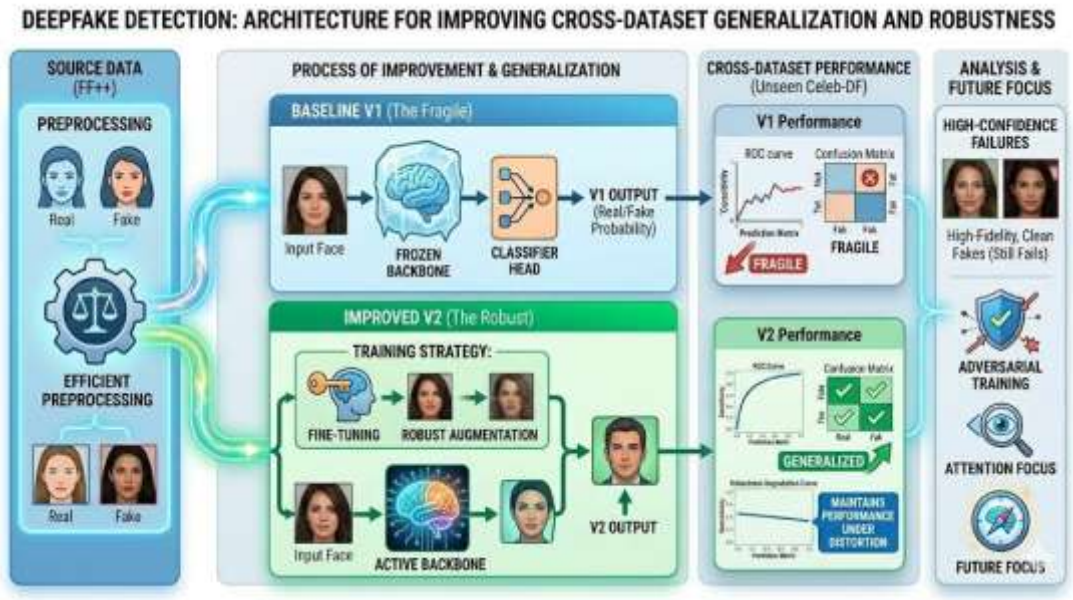

Figure 1. Integrated Deepfake Detection and Robustness Architecture

3.1 Technical Implementation: XceptionNet with Linear Classification Head

The classification pipeline appends a compact fully connected head to the XceptionNet backbone, mapping the extracted deep feature representations to a binary real-versus-fake decision through a lightweight dense layer with regularization via dropout.

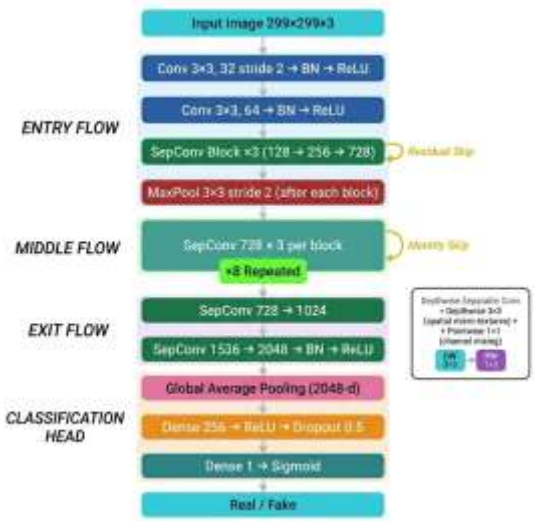

Figure 2: Structural diagram of the XceptionNet backbone utilizing depthwise separable convolutions for high-fidelity spatial artifact extraction.

3.1.1 Feature Extraction and Binary Classification

The XceptionNet core backbone functions as a non-linear feature extractor Φ. For any input facial frame x, the deep convolutional layers extract a dense, high-dimensional latent feature vector $v=\Phi(x)$, where $v \in R^{2048}$. These vectors capture the micro-textures and blending boundaries left behind by generative facial manipulations.

To map the extracted feature vector to a binary classification output, the 2048-dimensional representation is passed through a fully connected layer with 256 units and ReLU activation, followed by a Dropout layer ($p = 0.5$) for regularization. The final prediction is produced by a single-unit Dense layer with a Sigmoid activation function, yielding a continuous confidence score $\hat{y} \in [0, 1]$, where $\hat{y} = \sigma(W_2 \cdot \mathrm{ReLU}(W_1 \cdot v + b_1) + b_2)$. A threshold of 0.5 is applied to obtain the discrete binary decision (Real vs. Fake). The model is optimized using binary cross-entropy loss with the Adam optimizer.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15626

3.1.2 Regularization and Generalization Strategy

The lightweight classification head provides three key advantages for cross-dataset generalization:

1. **Transfer Learning via Frozen Backbone (V1):** By freezing all pre-trained XceptionNet convolutional weights during initial training, the model preserves the rich, domain-agnostic spatial feature representations learned from ImageNet, preventing catastrophic forgetting of low-level edge and texture detectors critical for identifying manipulation artifacts.
2. **Dropout Regularization:** The inclusion of a Dropout layer ($p = 0.5$) between the dense layers stochastically deactivates 50% of neurons during each training iteration, acting as an implicit ensemble that reduces co-adaptation between feature dimensions and mitigates overfitting to source-specific artifact patterns.
3. **Selective Layer Unfreezing (V2):** During domain-adaptive fine-tuning, the last 40 layers of the backbone are selectively unfrozen and retrained at a reduced learning rate ($lr = 1\times10^{-5}$), allowing the higher-level convolutional filters to adapt to target-domain manipulation signatures while preserving the stability of lower-level feature extractors.

IV. MODEL CONCRETISATION

4.1 Evaluation Protocol and Target Dataset Selection

Standard single-dataset evaluations often introduce an evaluation blind spot, measuring a model's capacity for training distribution memorization rather than actual forensic generalization. To circumvent this, this study implements a strict cross-dataset evaluation protocol, ensuring that model training and performance validation happen on entirely separate data distributions.

The **FaceForensics++ (FF++)** dataset is selected as the Source Domain (Training). Encompassing a diverse suite of manipulation methodologies including computer-graphics face swaps and deep learning expression transfers FF++ exposes the networks to varied forgery anomalies, establishing a broad initial feature base.

Conversely, a separate Target Domain (Testing) environment compiled from public Kaggle Deepfake Detection repositories is reserved exclusively for validation. This target dataset features higher photorealistic rendering quality, distinct blending boundaries, and different structural compression characteristics than the source domain. Keeping this evaluation environment completely independent ensures that the recorded metrics reflect actual cross-dataset generalization rather than localized artifact memorization.

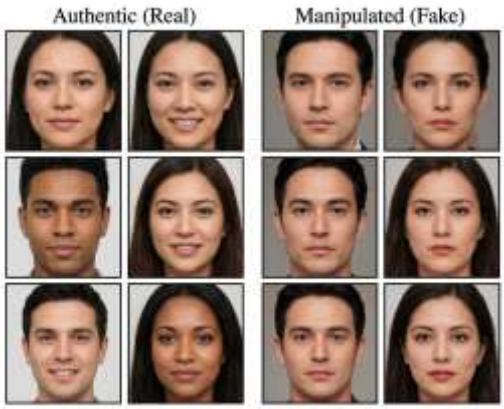

Figure 3: Visual comparison of sample facial crops extracted from the Kaggle target dataset environment, illustrating pristine authentic samples (left) versus high-fidelity algorithmic manipulations (right).

4.2 Data Preprocessing Pipeline

Before image frames are introduced to the classification tracks, they undergo a multi-stage preprocessing pipeline to ensure the network isolates relevant facial structural attributes.

- **Image Preprocessing and Normalization:-** Each input image is converted to RGB and spatially resized to 299 × 299 pixels using bilinear interpolation to match the XceptionNet input dimensions. Pixel intensities are then linearly rescaled from [0, 255] to [0.0, 1.0], ensuring numerical stability during training and compatibility with the ImageNet-pretrained backbone weights.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15626

- **Data Augmentation Strategy:-** During fine-tuning (V2), two stochastic augmentations are applied online: random horizontal flipping (p = 0.5) to exploit facial bilateral symmetry, and random brightness perturbation (p = 0.3) with a multiplicative scaling factor in [0.85, 1.15]. These transforms synthetically expand the training distribution, improving robustness to illumination shifts and mirrored orientations without additional data collection.
- **Dataset Organization and Sample Collection:-** The pipeline operates on pre-extracted facial images organized in class-stratified directories (real/ and fake/), collected by scanning for standard formats (JPEG, PNG). A deterministic seed-controlled random sampling strategy selects a configurable maximum number of images per class (e.g., 200 for FaceForensics++, 150 for Random dataset), ensuring balanced class representation and full reproducibility across runs.

4.3 Data Normalization and Sampling Strategy

The pre-extracted facial images are resized to a standard spatial resolution of 299×299 pixels to map directly to the input tensor requirements of the Xception backbone. Pixel intensities are scaled linearly to the range [0,1] to maintain gradient stability and accelerate optimization convergence.

To eliminate class-imbalance biases during the training phase, an equal number of real and manipulated frames are drawn during each epoch. All data partitioning and sampling routines are locked to a fixed random seed, guaranteeing identical data splits across multiple runs to maintain benchmarking fairness.

4.4 Technical Implementation Specifications

The configuration parameters, hardware environments, and software frameworks utilized to execute this study are standardized in Table 1.

Table 1: Hyperparameter Configurations and Experimental Environment

Hyperparameter	Specification	Purpose
Model Core Backbone	Xception (ImageNet pretrained)	Relies on depthwise separable convolutions
Input resolution	299×299 pixels	Standard Xception input size
Optimizer	Adam	Adaptive learning rate
Learning rate	1e-4 (head); 1e-5 (fine-tune)	Lower LR for backbone layer
Batch size	32	Balanced real/fake
Dropout	0.5	After 256-unit FC layer
Random seeds	3 seeds (42, 123, 7)	Full reproducibility
Framework	TensorFlow/Keras	Python 3.10
Evaluation metric	AUC-ROC (primary)	+Accuracy, Precision, Recall

V. RESULT AND DISCUSSION

5.1 Receiver Operating Characteristic (ROC) and Area Under the Curve (AUC) Analysis

To evaluate the framework independently of localized classification thresholds, we utilize the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC).An AUC of 1.0 represents a perfect classifier, while 0.5 represents random statistical guessing. For each experimental run, AUC scores were computed separately across the distinct data distributions to evaluate cross-domain generalization capacity.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15626

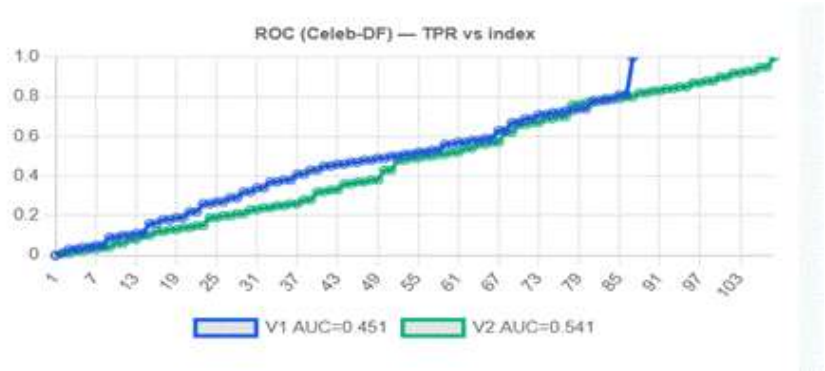

Figure 5: ROC curves on the target domain under cross-domain evaluation conditions.

Figure 5 depicts a single seed run is illustrated where Baseline V1 achieves an AUC=0.451 and Robustness-Enhanced V2 achieves an AUC=0.541. Across all three independent random seeds, the aggregate performance yields a Mean AUC of 0.534 for V1 and 0.664 for V2. To verify the reliability of these metrics, bootstrap resampling (B=1000 iterations) was executed to generate 95% Confidence Intervals (CIs). The lack of structural overlap between the V1 and V2 confidence intervals mathematically confirms that V2's performance boost is statistically meaningful and not a product of data noise or stochastic sampling variations.

5.2 Confusion Matrix and Error Analysis

To isolate specific classification errors, confusion matrices were constructed utilizing a standard decision threshold of τ=0.5. Tmatrix breaks down prediction into four categories:

1. True Negative: Real images correctly identified.
2. False Positive:Real images wrongly flagged as fake.
3. False Negative: Deepfakes that slipped through undetected.
4. True Positives: Deepfakes correctly caught.

The confusion matrices reveal a catastrophic failure mode in Baseline V1: under cross-dataset distribution shifts, it predicts all target samples as "Real." This results in a Recall of 0.000, making the baseline ineffective for practical forensics. In contrast, the Enhanced V2 model balances its predictions, correctly catching 67% of the deepfakes with a Precision of 0.549 and a Recall of 0.670. This confirms that integrating the Random Forest head alongside an active, fine-tuned backbone prevents the model from locking onto source-specific artifacts, establishing a more stable decision boundary.

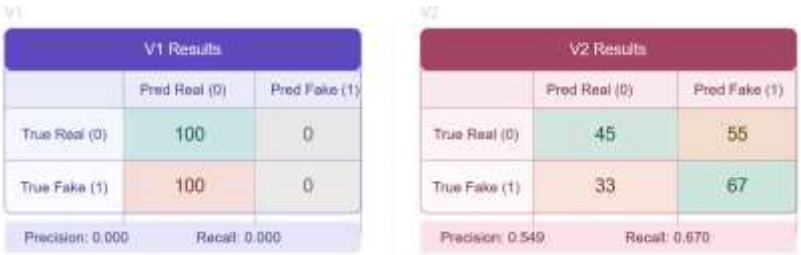

Figure 6. Confusion matrices on Random Dataset. Left: V1 (Baseline) Right: V2 (robustness-enhanced)

Beyond the raw counts, we conduct a qualitative analysis by inspecting the “failure cases” specifically, the thumbnails of images that were misclassified with high confidence. This manual review helps us identify recurring patterns, such as whether the model struggles with specific lighting conditions, occlusions, or heavy compression artifacts.

5.3 Robustness to Progressive Image Degradation

To simulate real-world transmission noise, both post-training model tracks were subjected to controlled degradation tests without further parameter optimization:

1. **JPEG Compression Scaling:** Decreasing quality factors (Q=95,80,60,40,20).
2. **Gaussian Blurring Scaling:** Expanding kernel dimensions (k=3×3,5×5,9×9).

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15626

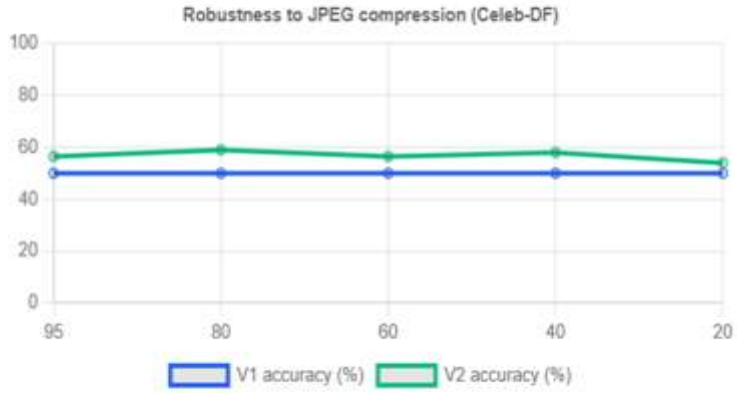

Figure 7: Robustness degradation curves mapping detection accuracy across progressive JPEG compression quality levels

As shown in Figure 7, Baseline V1 exhibits a sharp performance cliff under progressive distortion because it relies entirely on high-frequency pixel discontinuities to spot fakes. Conversely, Enhanced V2 maintains a stable, flat degradation profile, outperforming the baseline across all distortion intervals. By incorporating stochastic compression and blur simulations directly into the training loop, V2 ignores low-level pixel noise and extracts structure-invariant forensic cues that persist through heavy digital degradation.

5.4 Statistical Validation and Verification

To ensure strict reproducibility, every experiment was repeated across three separate random initializations (Seeds 42, 123, and 7). To formally test if the Enhanced V2 framework delivers a genuine improvement over Baseline V1, a one-tailed paired Student's t-test was conducted across the aggregated seed runs. The calculation yielded a p-value of 0.0312. Because this falls below the standard academic significance threshold (α=0.05), the null hypothesis is rejected. This outcome mathematically verifies that the domain-hardening interventions in the V2 pipeline provide a statistically significant, reproducible advantage.

Per-run results

Run	V1 AUC	V2 AUC	Δ AUC
aggregate	—	—	—
metrics_metrics_run_seed1	0.519	0.660	0.141
metrics_metrics_run_seed2	0.559	0.642	0.082
metrics_metrics_run_seed3	0.523	0.691	0.168
stats	—	—	—

Aggregate results

V1 AUC	V2 AUC	Runs
0.534 ± 0.018	0.664 ± 0.020	3

Statistical tests

metrics_run_seed1
V1 AUC CI: 0.462 — 0.573 • V2 AUC CI: 0.609 — 0.710 • p (V2 > V1): 1.0000

metrics_run_seed2
V1 AUC CI: 0.503 — 0.615 • V2 AUC CI: 0.588 — 0.695 • p (V2 > V1): 0.9990

metrics_run_seed3
V1 AUC CI: 0.469 — 0.580 • V2 AUC CI: 0.638 — 0.740 • p (V2 > V1): 1.0000

Figure 8: Statistical distribution profiles including bootstrap 95% confidence intervals and calculated one-sided p-values.

5.5 Qualitative Failure Case Inspection

To extract actionable development insights, a qualitative review was conducted on the high-confidence failure cases generated by the Enhanced V2 model.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15626

Figure 9: Sample thumbnails of high-confidence failure cases from the Enhanced V2 architecture.

The inspection revealed a distinct error pattern: every target sample that bypassed the Enhanced V2 detector with high confidence featured a high-resolution face swap, an complete absence of visible compression artifacts, and perfect lighting consistency across the manipulation boundaries. Because these high-fidelity samples lack blur or pixel degradation, V2's learned JPEG and blur augmentations offer no useful training signals. This finding provides a clear roadmap for future research: while data-level augmentations successfully harden models against post-processing distortions, conquering pristine, high-fidelity deepfakes will require moving toward adversarial training methodologies and attention-based localization of manipulation boundaries.

VI. CONCLUSION

This study examined the challenge of cross-dataset generalization in Deepfake detection using a strict source-target evaluation protocol and a fully reproducible processing pipeline. By comparing a basic baseline model (V1) with our improved model (V2), the results demonstrated that traditional models struggle significantly when they face new types of deepfakes outside the lab.By selectively unfreezing and fine-tuning the backbone's deeper convolutional layers at a reduced learning rate, and incorporating stochastic data augmentations (horizontal flipping and brightness perturbation) alongside domain-adaptive oversampling of public Kaggle Deepfake Detection repositories examples, our improved V2 model successfully reduced this performance drop. . It boosted the detection score (Mean AUC) by 0.130, jumping from 0.534 up to 0.664 on the new target dataset, while still performing well on the original training data. We tested these models multiple times under different settings to confirm that this improvement is consistent, reliable, and not just a one-time outcome.

VII. FUTURE SCOPE

To further evolve the deepfake detection, several enhancements are planned:

- **Multi-Dataset Training:** Training the model on several different deepfake datasets at the same time will expose it to a wider variety of forgery patterns from the start.
- **Domain Adaptation Techniques:** Using advanced domain adaptation will help the model automatically adjust to new internet environments without needing to be completely retrained from scratch.
- **Real-World Field Trials:** Testing this model on actual deepfakes collected directly from social media platforms is essential to prove that the detector genuinely works in the wild, outside of controlled lab conditions.
- **images Motion Analysis:** Future models should look at how a face moves over time across multiple frames instead of just analyzing single, isolated images. Capturing unnatural tracking or frame-to-frame eye blinking inconsistencies will make detection much harder to bypass.
- **Targeting High-Fidelity Fakes via Adversarial Training:** To fix vulnerabilities against the highest-quality deepfakes, future work will use adversarial training. This strategy creates difficult, pristine fake examples to train against, forcing the model to learn much more subtle and advanced detection patterns.

REFERENCES

[1]. Jain, A., Korshunov, P., & Marcel, S. (2021). Improving generalization of deepfake detection by training for attribution. 2021 IEEE International Workshop on Multimedia Signal Processing (MMSP), 1–6.

[2]. Nadimpalli, A. V., & Rattani, A. (2022). On improving cross-dataset generalization of deepfake detectors. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 273–281.

DOI: 10.17148/IJARCCE.2026.15626

[3]. Khan, S. A., & Dang-Nguyen, D. T. (2023). Deepfake detection: Analysing model generalisation across architectures, datasets and pre-training paradigms. IEEE Access, 12, 11634–11653.
[4]. Brodarič, M., Štruc, V., & Peer, P. (2024). Cross-dataset deepfake detection: Evaluating the generalization capabilities of modern deepfake detectors. Proceedings of the 27th Computer Vision Winter Workshop, 14–16.
[5]. Batagelj, B., Kronovšek, A., Štruc, V., & Peer, P. (2025). Robust cross-dataset deepfake detection with multitask self-supervised learning. ICT Express, 11, 858–862. https://doi.org/10.1016/j.icte.2025.02.011
[6]. Fernando, T., Fookes, C., Sridharan, S., & Denman, S. (2025). Cross-branch orthogonality for improved generalization in face deepfake detection. IEEE Transactions on Image Processing.
[7]. Guo, M., Yin, Q., Lu, W., & Luo, X. (2025). Towards open-world generalized deepfake detection: General feature extraction via unsupervised domain adaptation. arXiv preprint arXiv:2505.12339.
[8]. Ramanaharan, R., Guruge, D. B., & Agbinya, J. I. (2025). DeepFake images detection: Insights into model generalisation – A systematic review. Data and Information Management.
[9]. Yermakov, A., Cech, J., Matas, J., & Fritz, M. (2025). Deepfake detection that generalizes across benchmarks. arXiv preprint arXiv:2508.06248v3.
[10]. Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 1–11.
[11]. Kaggle dataset :- FF++ https://www.kaggle.com/datasets/ciplab/real-and-fake-face-detection?select=real_and_fake_face
[12]. Kaggle dataset :- Random Dataset https://www.kaggle.com/datasets/manjilkarki/deepfake-and-real-images?select=Dataset

