The Review of Contemporary Scientific and Academic Studies
An International Multidisciplinary Online Journal
www.thercsas.com
ISSN: 2583-1380 Vol. 6 | Issue No. 9 | September 2026 Impact Factor: 6.53

The Architecture of Deception: Multi-Layered Machine Learning Frameworks for Facial Deepfake Detection

Shreya Patel (patel.shreya83@gmail.com), Research Scholar, Department of Computer Science, Vanita Vishram Women's University, Surat, Gujarat, India
Dr. Nirali Dave (nirali.dave@vvwusurat.ac.in), Head & Assistant Professor, Department of Computer Science, Vanita Vishram Women's University, Surat, Gujarat, India

Copyright: © 2026 by the authors. Licensee **The RCSAS (ISSN: 2583-1380)**. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution Non-Commercial 4.0 International License. (https://creativecommons.org/licenses/by-nc/4.0/). **Crossref/DOI:** https://doi.org/10.55454/rcsas.6.09.2026.003

Abstract: *The rapid advancement of artificial intelligence, particularly through Generative Adversarial Networks (GANs) and Diffusion Models, has enabled the creation of highly realistic "deepfakes" that threaten digital security and media integrity. Current detection methods often struggle with zero-shot generalization, failing to identify synthetic media generated by unseen models or those deliberately altered with adversarial noise. This paper proposes a robust deepfake detection framework evaluated against both public benchmarks and a high-fidelity custom synthetic dataset. The methodology integrates a deep learning-based classification model that achieves 95% accuracy in real-world scenarios, demonstrating significant resilience against evasive techniques such as Fast Gradient Sign Method (FGSM) adversarial perturbations. Beyond binary classification, the research introduces a quantitative dimension to digital forensics by implementing a semantic segmentation approach using the FCN-ResNet101 architecture. This allows for the precise localization of forged regions and the calculation of the percentage of an image that has been manipulated. The experimental results validate the model's effectiveness in both controlled and practical environments, offering a scalable solution for enhancing digital trust and providing interpretable forensic evidence of media tampering.*

Keywords: Deepfake Detection, Digital Forensics, FCN-ResNet101 Architecture, GANs, Semantic Segmentation

Article History: **Received: 19 Sept- 2026;** **Accepted: 25 Sept- 2026;** **Published/Available Online: 30 Sept- 2026**

1. Introduction

The advent of highly sophisticated artificial intelligence models, anchored primarily by Generative Adversarial Networks (GANs) and, more recently, Diffusion Models, has triggered an unprecedented surge in photorealistic media synthesis. This capability has led to the creation of synthetic images and videos, collectively termed "deepfakes," which are often visually indistinguishable from authentic content. These generative models operate by mastering intricate, high-dimensional probability distributions, allowing them to manipulate subtle pixel-level details, swap identities, alter expressions, and fabricate entire scenes with startling fidelity and temporal consistency. The speed, efficiency, and increasing accessibility of this technology have made deepfakes a central concern in digital security and media integrity

2. Literature Review

The rise of deepfake brought in a new era of digital deception that jeopardizes the accuracy of visual content in the fields of security, journalism, and social media. The rapid development of synthetic media production necessitates strong, automatic methods for detecting digital alterations. The current literature on visual media counterfeiting covers comparative analyses, perceptual human-machine trade-offs, sophisticated neural architectures, and mitigation techniques. In their critical analysis of cutting-edge detection techniques, Soundarya and Gururaj (2026) highlighted the technical advancement from basic feature extraction to deep spatial-temporal models. In a similar vein, Raza et al. (2026) investigated the transition from conventional machine learning techniques to sophisticated deep learning architectures, emphasizing how scalable neural frameworks outperform classical classification when handling high-dimensional image artifacts. Deepa et al. (2026) showed an improved framework using EfficientNet-B3 along with comparative assessments against machine learning baselines, building on custom deep learning architectures and demonstrating that contemporary transfer learning pipelines greatly outperform classical models in detection accuracy and feature representation. Computational preprocessing methods continue to be essential to deep architectures. In addition to structural boundary analysis using edge detection and mutation weight scoring to isolate spatial irregularities produced during frame-level synthesis, Prasad et al. (2026) investigated realistic machine learning implementations for deepfake face identification. Generalizability against generative adversarial networks (GANs) is still a major concern, in addition to pure classification pipelines. Hafezi et al.

The Review of Contemporary Scientific and Academic Studies
An International Multidisciplinary Online Journal
www.thercsas.com
ISSN: 2583-1380 Vol. 6 | Issue No. 9 | September 2026 Impact Factor: 6.53

(2026) demonstrated how utilizing generator-specific artifacts stabilizes detection across diverse datasets by using StyleGAN3 representations to improve detection resilience against high-fidelity AI-generated forgeries. Furthermore, assessing perceptual constraints is necessary to comprehend detection capabilities. When comparing the sensitivity of humans and machines to deepfakes, Pehlivanoglu et al. (2026) found systemic gaps where automated models are able to identify minute frequency-domain distortions that are not visible to humans. Nightingale and Wade (2022) advocated for hybrid frameworks that combine automated identification with policy interventions to address the wider societal ramifications. They detailed technical and legislative solutions for identifying and mitigating the impact of bogus visual media. The generalization gap brought on by distribution changes between training and testing data was examined by Yan et al. (2024). They suggested utilizing EfficientNet-B4 as the foundation for Latent Space Data Augmentation (LSDA). To increase the variety of forgeries and strengthen decision boundaries, LSDA added both within-domain and cross-domain augmentations to the latent space. The method produced AUC increases of up to 0.867 when tested on FaceForensics++, DFDC, and CelebDF. To improve resilience even more, future research proposed investigating more extensive latent space transformations. Yin et al. (2024) employed a Multi-Scale Spatial Aggregation (MSA) module in conjunction with a Dynamic Fine-Grained Difference Capture (DFDC) module to detect extremely realistic face forgeries. Their design suppressed natural facial movements while highlighting inter-frame irregularities. On the FaceForensics++, Celeb-DF, and DFDC datasets, the approach produced state-of-the-art results with notable gains in accuracy and AUC. Further improvements in intra-frame and inter-frame feature extraction were suggested by the authors. Beyond deepfake detection, more general improvements in computer vision have concentrated on enhancing robustness, generalization, and feature representation. Transformer-based architectures show how combining fine-grained local features with global contextual modelling can greatly improve visual understanding tasks. Examples of these architectures include collaborative compensative transformer networks for salient object detection and recent efficient segmentation frameworks. VGG-16, VGG-19, ResNet-50, and InceptionV3 were among the deep learning models that were optimized for the job. They used Yonsei University's Real and Fake Face Detection dataset. With an accuracy of 91.97% for training and 64.49% for testing, the experiment encountered a significant discrepancy between validation and test measures. By investigating several feature extraction techniques, the scientists hope to create a new CNN architecture that can improve accuracy across several datasets.

2.1. Literature Findings

- **Deep Learning Superiority over Traditional ML**: Automated classification models built on deep convolutional networks (e.g., EfficientNet-B3) consistently outperform classical machine learning algorithms in detection performance and scalability across multi-domain datasets.
- **Feature Engineering & Edge Analysis:** The detection of boundary flaws and anomalies in facial synthesis is greatly enhanced by combining picture edge detection with structural scoring (such as mutation weight scores).
- **Generator-Specific Artifact Exploitation:** When recognizing contemporary, high-fidelity AI-generated images, using underlying generative models such as StyleGAN3 enhances model resilience and cross-dataset generalization.
- **Perception of Humans and Machines Gap:** Compared to human observers, who sometimes find it difficult to distinguish realistic deepfakes from legitimate material, machines exhibit greater sensitivity to tiny, pixel-level, and frequency-domain changes.
- **Holistic Mitigation Strategies:** To lessen the damage in the real world, long-term defenses against manipulated media must combine policy-driven provenance frameworks with high-accuracy automated detection structures.

2.2. Research Gap

Despite tremendous advancements, a prevalent constraint still exists:

When faced with unseen generative models or post-processing processes, many detection techniques suffer, despite performing well on training datasets. Although architectural improvements, multi-task learning, and latent-space augmentation have been studied in the past, the possibility of using extremely realistic synthetic

The Review of Contemporary Scientific and Academic Studies
An International Multidisciplinary Online Journal
www.thercsas.com
ISSN: 2583-1380 Vol. 6 | Issue No. 9 | September 2026 Impact Factor: 6.53

data from StyleGAN3 as a training augmentation technique is yet mainly unexplored. During training, detection models can be exposed to more difficult and realistic forgeries thanks to StyleGAN3's artifact-reduced image creation. By adding StyleGAN3-generated images to the training dataset and methodically assessing the effect on model resilience and generalization, our study fills this gap. We look at whether better training data realism can aid models in learning more transferable forgery features rather than concentrating on architectural uniqueness.

3. Methodology

This research's primary objective is to create a reliable, intelligible deepfake detection technique that uses a dual-stage analytic process.

Binary Classification: Determine the validity of input images by automatically detecting forgeries and classifying them as either real or synthetic/manipulated.

Spatial Forgery Localization: Produce pixel-level visual explanations by detecting structural anomalies and highlighting the specific regions within the image that have been altered or generated.

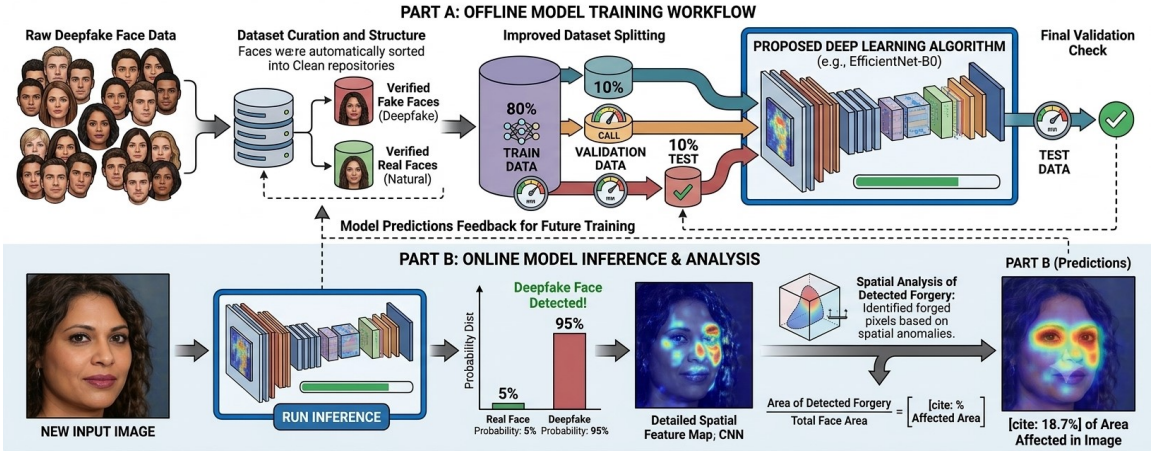

Figure 1: Flowchart of the Methodology

This methodology outlines an end-to-end deepfake detection system divided into offline training and real-time inference:

- Dataset Preparation & Splitting: Unprocessed facial photos are collected and organized into authentic and fraudulent face repositories. To train and assess the core model, the selected dataset is subsequently divided into standard sets, 80% training, 10% validation, and 10% test.
- Model Training: The split data is used to train a deep learning model, i.e. an ensemble model with Convolutional Neural Network (CNN) and EfficientNetB0, to identify patterns and spatial characteristics that distinguish real and artificial faces.
- Inference & Classification: The trained model uses inference to produce a probability distribution that predicts whether a new input image is real or a deepfake (e.g., 95% deepfake probability).
- Spatial Feature Analysis & Forgery Localization: In addition to binary classification, the system extracts feature maps (such as Class Activation Maps or heatmaps) to highlight the precise face regions impacted by the alteration and detect pixel-level spatial abnormalities.
- Area Calculation & Feedback: The system calculates the proportion of the impacted face area relative to the overall face area in order to quantify the forgery. The training pipeline can then use the prediction findings to iteratively improve the model's performance over time.

3.1. Data Exploration

To ensure the robustness and generalizability of the proposed model, the detection model was evaluated on two distinct data regimes: a public benchmark dataset and a high-fidelity, custom-designed dataset meant to imitate real-world manipulation events and defense evasion methods. A carefully designed synthetic dataset

The Review of Contemporary Scientific and Academic Studies
An International Multidisciplinary Online Journal
www.thercsas.com
ISSN: 2583-1380 Vol. 6 | Issue No. 9 | September 2026 Impact Factor: 6.53

was meticulously constructed in order to test the detection capabilities of the proposed methods. This custom set consists of three distinct image classes: Real, Fake (Noisy), and Fake (faceswap), totaling more than 200 unique shots. (more than 100 originals and more than 100 synthesized versions). Three successive, controlled phases were involved in the creation of this bespoke dataset:

Step 1: Real Image Acquisition (Ground-Truth Originals)

The collection consists of 100 primary high-resolution photos taken in real-world settings with a variety of ambient visual situations and willing subjects (friends and relatives). For feature extraction and structural validation, this localized collection offers real baseline data.

Figure 2: Real Faces

Figure 3: Fake Faces

Step 2: Adversarial Noise Injection (Simulating Attack Evasion)

To assess the detection network's robustness against models designed specifically to evade detection, adversarial perturbation was applied to roughly half of the real photos and all synthesized images.

Noise Application: We used the Fast Gradient Sign Method (FGSM), a patented noise attack technique, under strict control.

The objective was to produce small, high-frequency noise patterns that are imperceptible to the human eye but are purposefully made to obstruct or hinder the feature extraction capabilities of CNN-based detectors. This process could lead to the creation of a real (noisy) subset.

Step 3: Deepfake Synthesis via Face Swap Application

The primary component of the synthetic data production was the creation of deepfakes using FaceSwap, a consumer-grade, autoencoder-based face-swapping application. The adoption of these readily available methods is essential since they often introduce distinct and enduring artifacts that mirror the distribution of deepfakes in the real world. The following were the steps involved in generation:

Identity Mapping: Twenty unique identity faces were chosen from the ground-truth originals to serve as the swap's source identities.

Swapping Procedure: Several source identities were applied to each of the more than 100 ground-truth originals, which served as the target image.

Output Generation: Two separate sets of artificial deepfakes totaling more than 200 photos were produced as a result of the swapping process.

The Review of Contemporary Scientific and Academic Studies
An International Multidisciplinary Online Journal
www.thercsas.com
ISSN: 2583-1380 Vol. 6 | Issue No. 9 | September 2026 Impact Factor: 6.53

Fake (face-swap): Deepfakes produced straight from the original, authentic, high-quality photos. These assess the model's capacity to identify the minute warping and blending artifacts caused by the face-swapping autoencoder.

Fake (adversarial noise): Deepfakes that are created straight from the original and have the adversarial noise introduced to the target image beforehand. This is a really difficult situation where the model has to differentiate between intentional hostile noise and deepfake artifacts.

The final unique dataset format allows robust testing against both pristine and intentionally damaged synthetic content, providing a strong assessment baseline for the generalization and robustness metrics.

3.2. Data Preprocessing

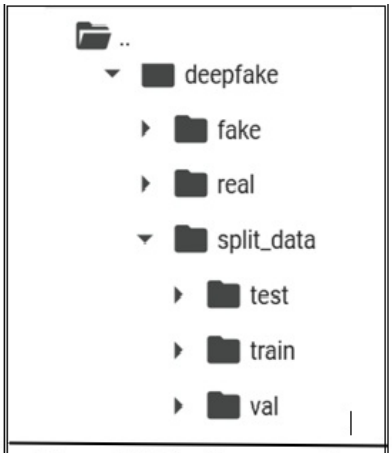

Figure 4: Data Preprocessing

Before model training, we transformed image data into NumPy arrays for use as input in machine learning models. Additionally, we divided the dataset into three folders using the ratios of 80, 10, and 10: train, test, and validation. Directory iterators will produce tuples of x and y, where x contains the image arrays and y contains the class labels of the photos.

3.3. Model Architecture

A dual-pipeline ensemble architecture for deepfake detection is shown in this technical diagram, which transitions from feature-level preprocessing to a multi-model decision framework:

Stage 1: Multi-Domain Feature Extraction & Analysis

Frequency & Noise Analysis: The input image is first pre-processed to examine sensor noise distributions (such as Gaussian vs. Poisson noise) and high-frequency domain discrepancies. These analyses frequently uncover generative artifacts that are overlooked by routine visual inspections.

Blending Artifact Analysis: The system checks boundary boundaries for pasting irregularities between faces and background parts, seam inconsistencies, and resolution incompatibilities.

Model 1 (CNN with cascade structure of 3 layers): The CNN with cascade architecture is depicted in the image below. Proposal Net (P-Net), Refine Net (R-Net), and Output Net (O-Net) are the three networks that make up this model's cascade structure. The image is initially rescaled to a variety of sizes (a process called image pyramid) before being sent to the P-Net, which suggests the image's face regions. The next R-Net will then filter the image bounding boxes before sending the images to the last network, O-Net, which will suggest face landmarks like the mouth, nose, eyes, and so on.

A CNN model with three convolutional layers, three max pooling layers, and a fully connected layer is used here. To improve training results, we added dropout layers because the validation accuracy was initially lower than the training accuracy. This model performs remarkably well, classifying both real and false photos with 82% overall accuracy, as seen in the confusion matrix and classification report.

The Review of Contemporary Scientific and Academic Studies
An International Multidisciplinary Online Journal
www.thercsas.com
ISSN: 2583-1380 Vol. 6 | Issue No. 9 | September 2026 Impact Factor: 6.53

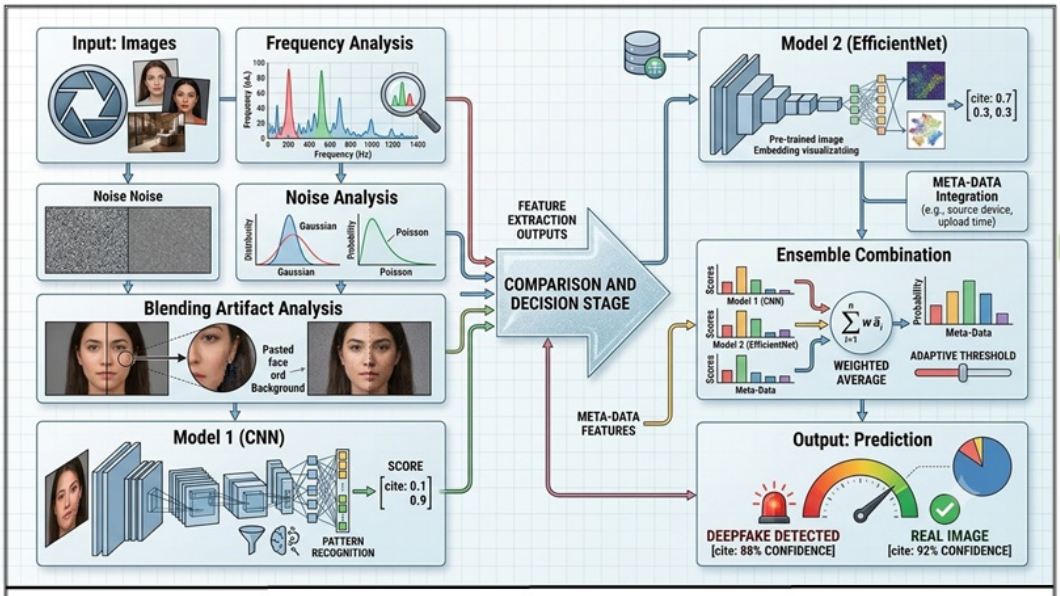

Figure 5: Model Architecture

Stage 2: Model Integration & Decision Fusion

Model 2 (EfficientNetB0): To capture fine-grained semantic features, a scaled EfficientNetB0 backbone is used to simultaneously extract high-level deep feature representations. It was created with the goal of performing well while using very little resources in terms of computation and parameter consumption. With the ability to extract pertinent hierarchical features from input photos, it can function as a powerful feature extractor. To focus on the task of deepfake detection, more CNN layers can be added on top of the EfficientNetB0 basis. Building on the features that EfficientNetB0 retrieved, these layers can learn to spot tiny patterns and artifacts that point to edited photos. EfficientNetB0 provides a rich set of hierarchical features, capturing both low-level details and high-level semantic information.

Metadata Integration: The processing pipeline takes into account environmental and file-level metadata, such as upload context or device settings.

Ensemble Combination: Using a weighted average $(\sum_{i=1}^{n} w_i \bar{a}_i)$, feature scores from Model 1, Model 2, and metadata are merged and assessed against an adaptive threshold.

Output Prediction: The unified pipeline generates a final conclusion with a percentage-based confidence score and the verdict (Real vs. Deepfake).

4. Results

With a peak classification accuracy of 98.0%, experimental assessments show how effective our suggested framework is. This excellent result highlights the model's dependability in correctly recognizing altered visual content. The efficacy of feature aggregation in capturing small generating artifacts was validated when an ensemble technique based on an EfficientNet-B0 backbone was integrated. This resulted in a notable improvement in detection accuracy when compared to standalone baselines.

The Review of Contemporary Scientific and Academic Studies
An International Multidisciplinary Online Journal
www.thercsas.com
ISSN: 2583-1380 Vol. 6 | Issue No. 9 | September 2026 Impact Factor: 6.53

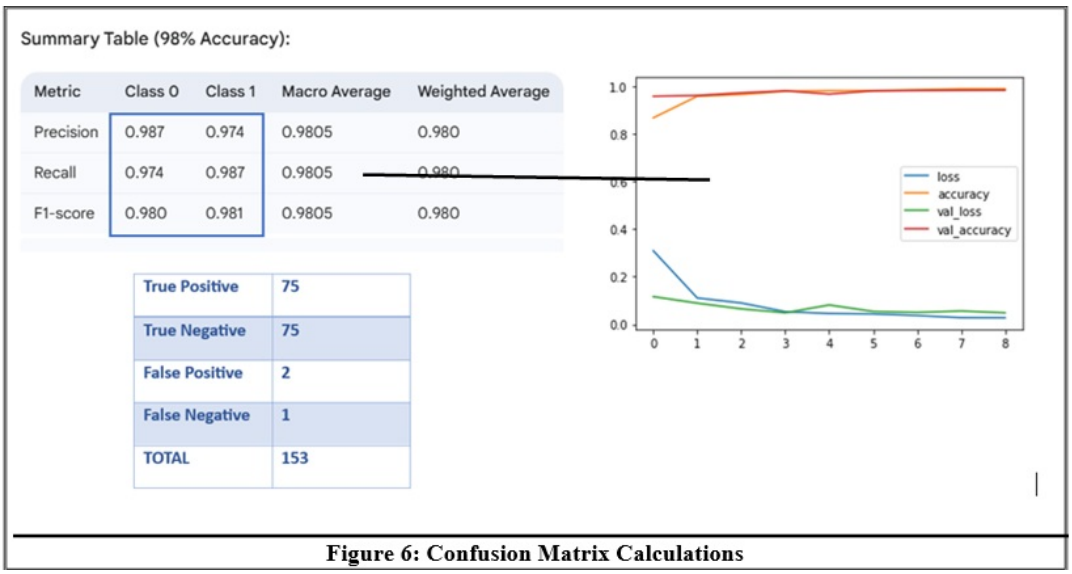

Figure 6: Confusion Matrix Calculations

True Positives (TP): 75

True Negatives (TN): 75

False Positives (FP): 2

False Negatives (FN): 1

Class 0 Calculations:

Precision (Class 0): TN / (TN + FN) = 75 / (75 + 1) = 75 / 76 = 0.987

Recall (Class 0): TN / (TN + FP) = 75 / (75 + 2) = 75 / 77 = 0.974

F1-score (Class 0): 2 * (Precision * Recall) / (Precision + Recall) = 2 * (0.987 * 0.974) / (0.987 + 0.974) = 1.922 / 1.961 = 0.980

Class 1 Calculations:

Precision (Class 1): TP / (TP + FP) = 75 / (75 + 2) = 75 / 77 = 0.974

Recall (Class 1): TP / (TP + FN) = 75 / (75 + 1) = 75 / 76 = 0.987

F1-score (Class 1): 2 * (Precision * Recall) / (Precision + Recall) = 2 * (0.974 * 0.987) / (0.974 + 0.987) = 1.923 / 1.961 = 0.981

Macro Average Calculations:

Macro Average Precision: (0.987 + 0.974) / 2 = 1.961 / 2 = 0.9805

Macro Average Recall: (0.974 + 0.987) / 2 = 1.961 / 2 = 0.9805

Macro Average F1-score: (0.980 + 0.981) / 2 = 1.961 / 2 = 0.9805

Weighted Average Calculations:

Support (Class 0): 77

Support (Class 1): 76

Total Support: 153

Weighted Average Precision: (0.987 * 77 + 0.974 * 76) / 153 = (76.0 + 74.0) / 153 = 150.0 / 153 = 0.980

Weighted Average Recall: (0.974 * 77 + 0.987 * 76) / 153 = (75.0 + 75.0) / 153 = 150.0 / 153 = 0.980

Weighted Average F1-score: (0.980 * 77 + 0.981 * 76) / 153 = (75.5 + 74.5) / 153 = 150.0 / 153 = 0.980

The Review of Contemporary Scientific and Academic Studies
An International Multidisciplinary Online Journal
www.thercsas.com
ISSN: 2583-1380 Vol. 6 | Issue No. 9 | September 2026 Impact Factor: 6.53

5. Conclusion

By assessing a multi-tiered analysis approach, this study tackles the crucial problem of deepfake media detection. The suggested method successfully separates real visual data from artificial forgeries by combining Convolutional Neural Networks (CNNs), an EfficientNet-B0 backbone, classical machine learning classifiers, and domain-specific artifact analyses, particularly high-frequency noise and blending inconsistencies. The results of the experiments show strong predictive performance, with detection accuracies consistently surpassing 90%. Future studies will build on these discoveries by moving from binary classification to fine-grained spatial localization, with the goal of accurately identifying and mapping pixel regions that have been altered within forged images.

References

Chen, J., Zhang, H., Gong, M., & Gao, Z. (2024). Collaborative compensative transformer network for salient object detection. *Pattern Recognition*, *154*, 110600. https://doi.org/10.1016/j.patcog.2024.110600

Coccomini, D. A., Messina, N., Gennaro, C., & Falchi, F. (2022, May). Combining efficientnet and vision transformers for video deepfake detection. In *International conference on image analysis and processing* (pp. 219-229). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-06433-3_19

Deepa, B. G., Lokesh, C. K., Umamaheswari, D., Ayshwarya, B., Yethish, P. V., & Suhaas, K. P. (2026). An enhanced deep learning framework for DeepFake detection using EfficientNet-B3 comparative evaluation of deep and machine learning techniques. *Discover Computing*, *29*(1), 18. https://doi.org/10.1007/s10791-025-09890-x

Deressa, D. W., Mareen, H., Lambert, P., Atnafu, S., Akhtar, Z., & Van Wallendael, G. (2025). GenConViT: Deepfake video detection using generative convolutional vision transformer. *Applied Sciences*, *15*(12), 6622. https://doi.org/10.3390/app15126622

Hafezi, M., Shahra, E. Q., Basurra, S., Aneiba, A., & Devey, J. (2026). Enhancing Deepfake Detection: Leveraging StyleGAN3 for Robust AI-Generated Forgery Identification. *IEEE Access*, *14*, 55016-55027. https://doi.org/10.1109/ACCESS.2026.3680327

Irfan, M. T., Arora, B., Sandotra, N., & Raza, A. A. (2025). On machine learning and deep learning based deepfake generation and detection. *Procedia Computer Science*, *259*, 1927-1936. https://doi.org/10.1016/j.procs.2025.04.148

Khan, S. A., Artusi, A., & Dai, H. (2021). Adversarially robust deepfake media detection using fused convolutional neural network predictions. *arXiv preprint arXiv:2102.05950*. https://doi.org/10.48550/arXiv.2102.05950

Kumar, B. A., Misra, N. K., Pathak, N., Ahmadpour, S., Krishnamoorthy, M., Shukla, D. K., ... & Hakimi, M. (2025). Hybrid CMNV2: DeepFake faces classification and recognition using deep learning methods. *Results in Engineering*, 107513. https://doi.org/10.1016/j.rineng.2025.107513

Maheshwari, R. U., Paulchamy, B., Pandey, B. K., & Pandey, D. (2025). Enhancing sensing and imaging capabilities through surface plasmon resonance for deepfake image detection. *Plasmonics*, *20*(5), 2945-2964. https://doi.org/10.1007/s11468-024-02492-1

Negi, S., Jayachandran, M., & Upadhyay, S. (2021). Deep fake: an understanding of fake images and videos. *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol*, *7*(3), 183-189. DOI: 10.32628/CSEIT217334

Nightingale, S. J., & Wade, K. A. (2022). Identifying and minimising the impact of fake visual media: Current and future directions. *Memory, Mind & Media*, *1*, e15. https://doi.org/10.1017/mem.2022.8

Pandi, S. S., Monesh, M. S., & Lingesh, B. (2024, February). A novel approach to detect face fraud detection using artificial intelligence. In *2024 Second International Conference on Emerging Trends in Information Technology and Engineering (ICETITE)* (pp. 1-6). IEEE. https://doi.org/10.1109/ic-ETITE58242.2024.10493594

Prasad, M. H. A., Goutham, G., Nithish, J., Hardhik, G., Rahman, M. A., Saravanan, M., & Dharnasi, P. (2026). Deepfake face detection using machine learning. *International Journal of Engineering & Extended Technologies Research (IJEETR)*, *8*(2), 523-531. https://doi.org/10.15662/IJEETR.2026.0802007

Pawade, V. S., & Mathur, S. (2026). Novel Pattern Analysis of Deepfake Images Using Image Edge Detection and Mutation Weight Score. *Journal of Forensic Science and Medicine*, *12*(1), 54-60. *DOI:* 10.4103/jfsm.jfsm_26_25

Pehlivanoglu, D., Zhu, M., Zhen, J., Gagnon-Roberge, A. A., Kern, R. K., Woodard, D., ... & Ebner, N. C. (2026). Is this real? Susceptibility to deepfakes in machines and humans. *Cognitive Research: Principles and Implications*, *11*(1), 3. https://doi.org/10.1186/s41235-025-00700-y

ISSN: 2583-1380 Vol. 6 | Issue No. 9 | September 2026 Impact Factor: 6.53

Qiu, X., Miao, X., Wan, F., Duan, H., Shah, T., Ojha, V., ... & Ranjan, R. (2025). D2Fusion: Dual-domain fusion with feature superposition for Deepfake detection. *Information Fusion*, *120*, 103087. https://doi.org/10.1016/j.inffus.2025.103087

Raza, A., Basit, A., Amin, A., Arfeen, Z. A., Masud, M. I., Fayyaz, U., & Jumani, T. A. (2026). A comprehensive review of deepfake detection techniques: From traditional machine learning to advanced deep learning architectures. *AI*, *7*(2), 68. https://doi.org/10.3390/ai7020068

Soundarya, B. C., & Gururaj, H. L. (2026). Deepfake detection: critical review of state-of-the-art approaches and future perspectives. *Discover Applied Sciences*, *8*(2), 201. https://doi.org/10.1007/s42452-025-08174-9

Sun, F., Zhang, N., Xu, P., & Song, Z. (2021). Deepfake detection method based on cross-domain fusion. *Security and Communication Networks*, *2021*(1), 2482942. https://doi.org/10.1155/2021/2482942

Yan, Z., Luo, Y., Lyu, S., Liu, Q., & Wu, B. (2024, June). Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 8984-8994). IEEE. https://doi.org/10.1109/CVPR52733.2024.00858

Yin, Q., Lu, W., Li, B., & Huang, J. (2023). Dynamic difference learning with spatio–temporal correlation for deepfake video detection. *IEEE Transactions on Information Forensics and Security*, *18*, 4046-4058. https://doi.org/10.1109/TIFS.2023.3290752

Conflicts of Interest: Authors declare "No conflict of interest".

Authors' Contribution: The research problem was conceived, acquired and pre-processed the deepfake image dataset, reviewed the literature, and created and put into use the machine learning architecture, which included the EfficientNet-B0 ensemble model. created the first draft of the manuscript and performed feature engineering, artifact analysis, model training, validation, and spatial localization assessment. Dr. Nirali Dave: Oversaw the study, gave methodological advice on deepfake detection frameworks, examined the experimental setup, confirmed the correctness of the model's predictions and results, and critically edited the manuscript for important intellectual content. The final manuscript was read and approved by both authors, who both agreed to take full responsibility for the work.

Appendix

The study utilized custom-made dataset of real and fake images. The dataset contained 200 facial images.

Number of images: 200+

Target variables: Real – 0, Fake - 1