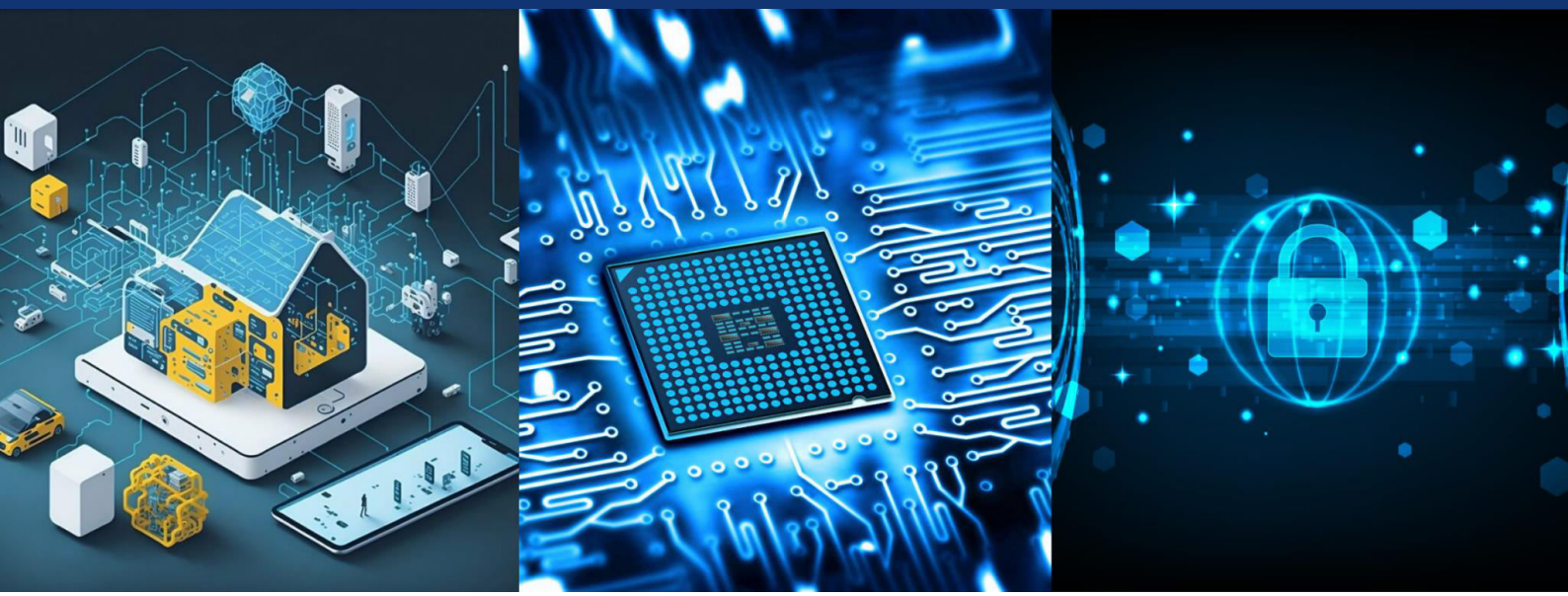
IJIRCCE

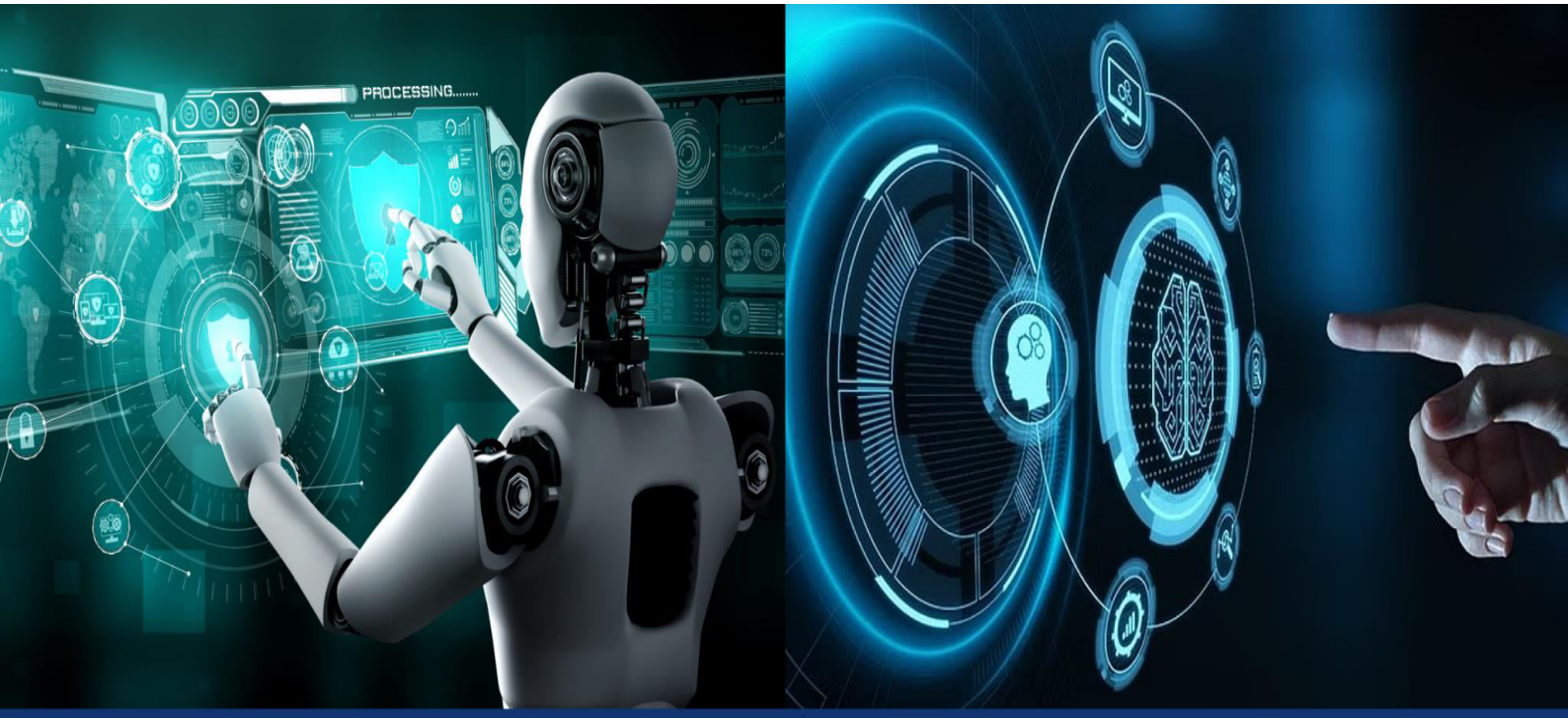
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

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



Impact Factor: 9.927

Volume 14, Issue 7, July 2026

© 2026 IJIRCCE | Volume 14, Issue 7, July 2026 DOI: 10.15680/IJIRCCE.2026.1407064

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

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

An Explainable AI Method for Detecting Fake Facial Images

Vighnesh V Prabhu[1], Nikitha K S[2]

Department of CS& E, Bangalore Institute of Technology, Bengaluru, India[1]

Assistant Professor, Department of CS& E, Bangalore Institute of Technology, Bengaluru, India[2]

ABSTRACT: The rapid advancement of generative adversarial networks and diffusion-based image synthesis has produced synthetic facial images visually indistinguishable from authentic photographs, creating serious threats to digital identity, media integrity, and public trust. Existing detection approaches suffer from two critical limitations: inadequate sensitivity to color-space artifacts when processing standard RGB images, and a black-box decision-making process that prevents forensic analysts and legal professionals from validating classification outcomes. This study addresses both limitations by proposing a deepfake facial image detection framework integrating MTCNN-based face detection and geometric alignment to ensure the classifier focuses exclusively on forensically relevant facial content rather than background noise, LAB color space transformation to decouple luminance from chrominance and surface subtle generation artifacts characteristic of synthesized faces, fine-tuned InceptionResNetV2 classification, and LIME-based superpixel attribution maps highlighting suspicious facial zones including skin-hair boundaries, periocular regions, and dentition areas. The framework was trained on the 140k Real and Fake Faces dataset using an 80/10/10 split on a GPU-enabled environment. The proposed pipeline achieved a test accuracy of 98.93%, generalization accuracy of 96.94%, with good precision, recall and F1-score, outperforming the baseline RGB model across every evaluated metric, with demonstrated real-time deployability through a Flask web application.

KEYWORDS: Deepfake Detection; Explainable AI; MTCNN; LAB Color Space; LIME; InceptionResNetV2; Transfer Learning; Digital Forensics

I. INTRODUCTION

The proliferation of generative artificial intelligence has fundamentally altered the landscape of digital image creation. Generative Adversarial Networks (GANs), beginning with the original formulation by Goodfellow et al. [1] and advancing through ProGAN, StyleGAN, and StyleGAN2 [2], have achieved a level of photorealism in synthetic facial generation that renders such content perceptually indistinguishable from authentic photographs. More recently, diffusion-based models including Stable Diffusion and DALL-E have extended this capability further. While offering legitimate creative applications, these technologies have been weaponized for identity fraud, political disinformation, non-consensual imagery creation, and social engineering attacks, constituting a growing threat to digital identity integrity and public trust in visual media [3].

Traditional forensic techniques including noise residual analysis, metadata inspection, error level analysis, and compression artifact detection were developed prior to modern generative architectures and have proven progressively ineffective against contemporary synthetic faces. StyleGAN2 outputs exhibit minimal noise residue, unremarkable metadata, and compression characteristics closely mimicking authentic photographs, rendering classical forensic indicators unreliable. In response, the research community has developed deep learning-based classifiers demonstrating substantial detection accuracy improvements on benchmark datasets such as FaceForensics++ [3] and the Deepfake Detection Challenge corpus [4].

Despite these advances, two persistent limitations constrain the practical utility of existing detection systems. First, generalization failure: classifiers trained on images from a specific generative architecture exhibit substantial performance degradation when evaluated against images from architectures absent from the training distribution [5], a critical barrier in operational contexts where the generative source is unknown. Second, interpretability absence: the overwhelming majority of deployed systems produce binary labels without identifying which visual evidence informed

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

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

the classification. In forensic and legal contexts, an unexplained classification result cannot responsibly be acted upon irrespective of reported accuracy.

A further underexplored limitation concerns color representation during preprocessing. Standard RGB processing conflates luminance and chrominance across all three channels, reducing sensitivity to the subtle color-distribution artifacts that GANs characteristically introduce through perceptual loss optimization in RGB space. This study addresses all three limitations through a unified framework integrating MTCNN-based face alignment [14] to restrict classifier input to forensically relevant facial content, LAB color space preprocessing to decouple chrominance from luminance and expose generative color artifacts, fine-tuned InceptionResNetV2 classification [16], and LIME-based superpixel attribution [15] converting the classifier from an opaque decision engine into an auditable forensic instrument. The complete system is deployed as a real-time Flask web application accessible through standard browsers.

The principal contributions of this work are: (1) demonstration that MTCNN face alignment combined with LAB preprocessing significantly improves both in-distribution accuracy and cross-architecture generalization over standard RGB full-frame processing; (2) integration of LIME attribution providing image-specific forensic justification for every prediction; (3) comprehensive empirical evaluation under both in-distribution and out-of-distribution conditions; and (4) practical real-time web deployment accessible to non-technical users in forensic and media verification contexts.

II. RELATED WORK

A. GAN-Based Deepfake Generation

The original GAN formulation [1] established the adversarial training paradigm that underlies most contemporary deepfake synthesis. Karras et al. progressively refined this through ProGAN, StyleGAN, and StyleGAN2 [2], the last of which introduced path length regularization and weight demodulation that eliminated the characteristic blob artifacts of earlier GAN outputs, substantially raising the perceptual realism threshold. More recent diffusion-based models, including Stable Diffusion and Midjourney, have further complicated the detection landscape by producing synthetic faces with artifact profiles qualitatively distinct from GAN-based images. Khan et al. [5] conducted a comprehensive survey of detection techniques across image, video, audio, and multimodal domains, identifying cross-architecture generalization, adversarial vulnerability, and the absence of standardized evaluation benchmarks as the three primary unresolved challenges in the field.

B. Deep Learning Detection Approaches

Rossler et al. [3] introduced the FaceForensics++ benchmark and demonstrated that XceptionNet achieves strong detection performance on compressed deepfake video frames, establishing an important baseline for subsequent comparative evaluation. Barik et al. [10] evaluated MTCNN, InceptionResNetV1, and FaceNet for identity-based deepfake detection through facial embedding consistency analysis, demonstrating that face-aligned embeddings provide an effective forensic signal. Soudy et al. [11] proposed a hybrid CNN and Convolutional Vision Transformer architecture for video-based detection, demonstrating that transformers capture global spatial relationships that convolutional-only models tend to miss.

C. Color Space Preprocessing in Detection

Abu Talib et al. [12] developed ColorDense and LightDense, two DenseNet-based models trained separately on chrominance and luminance features extracted from alternative color space representations. Their results provide strong empirical evidence that color-space decomposition improves deepfake detection accuracy relative to standard RGB processing, since GAN perceptual loss functions operating in RGB space leave detectable residual inconsistencies in chrominance channels that become more salient when luminance and color are separated. However, their dual-model architecture incurred substantial computational overhead and provided no spatial explanation of which facial regions contained the detected artifacts. The present work builds on this insight by incorporating LAB preprocessing within a single classification model, paired with LIME-based attribution for spatial interpretability.

D. Explainable AI in Deepfake Detection

Mathews et al. [6] demonstrated that XAI-integrated detection and high accuracy can be achieved simultaneously using an Inception-based classifier with visual explanation maps on an unconstrained dataset. Silva et al. [7] developed a hierarchical ensemble of weakly supervised CNNs with Grad-CAM visualization for forensic-level interpretability.

© 2026 IJIRCCE | Volume 14, Issue 7, July 2026 DOI: 10.15680/IJIRCCE.2026.1407064

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

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

Cirillo et al. [8] demonstrated that explainability-aware adversarial attacks targeting LIME- and Grad-CAM-identified facial regions are substantially more effective than generic perturbations, revealing a dual-use concern. Mansoor and Iliev [9] applied Network Dissection techniques to associate internal neuron activations with semantic facial concepts, including periocular features and skin texture. Momin et al. [13] reviewed explainable detection methods across modalities, identifying computational overhead, lack of standardized explanation quality benchmarks, and limited real-time capability as the primary barriers to adoption. The present work specifically selects LIME over Grad-CAM because LIME is model-agnostic, image-specific, and operates at superpixel resolution, making its outputs directly usable as spatially precise forensic evidence.

III. METHODOLOGY

A. Dataset and Experimental Setup

The 140k Real and Fake Faces dataset comprises 70,000 authentic photographs from the Flickr Faces HQ (FFHQ) corpus and 70,000 synthetic images generated using StyleGAN2, providing a perfectly balanced binary classification benchmark. An 80/10/10 train-validation-test split yielded approximately 112,000 training, 14,000 validation, and 14,000 test samples. For generalization evaluation, a separate out-of-distribution dataset was constructed using synthetic images from architectures not represented in the StyleGAN2-based training data. On-the-fly augmentation applied during training included random horizontal flipping, rotation within ±15 degrees, zoom variation up to 10%, and brightness adjustment within ±20%, selected to improve robustness to capture condition variation without destroying generative artifacts. Training was executed on a GPU-enabled Kaggle kernel utilizing an NVIDIA Tesla P100 with 16GB VRAM.

B. System Architecture Overview

The proposed framework implements a sequential five-stage pipeline: (1) image input and validation, (2) LAB color space preprocessing, (3) MTCNN face detection and alignment, (4) InceptionResNetV2 classification, and (5) LIME explainability generation. Each stage receives the processed output of the preceding stage, forwarding a transformed representation to the next. This sequential design was selected to permit each stage to be independently developed, validated, and replaced, and because each stage constitutes an inspectable intermediate representation, simplifying diagnostic analysis and future extension.

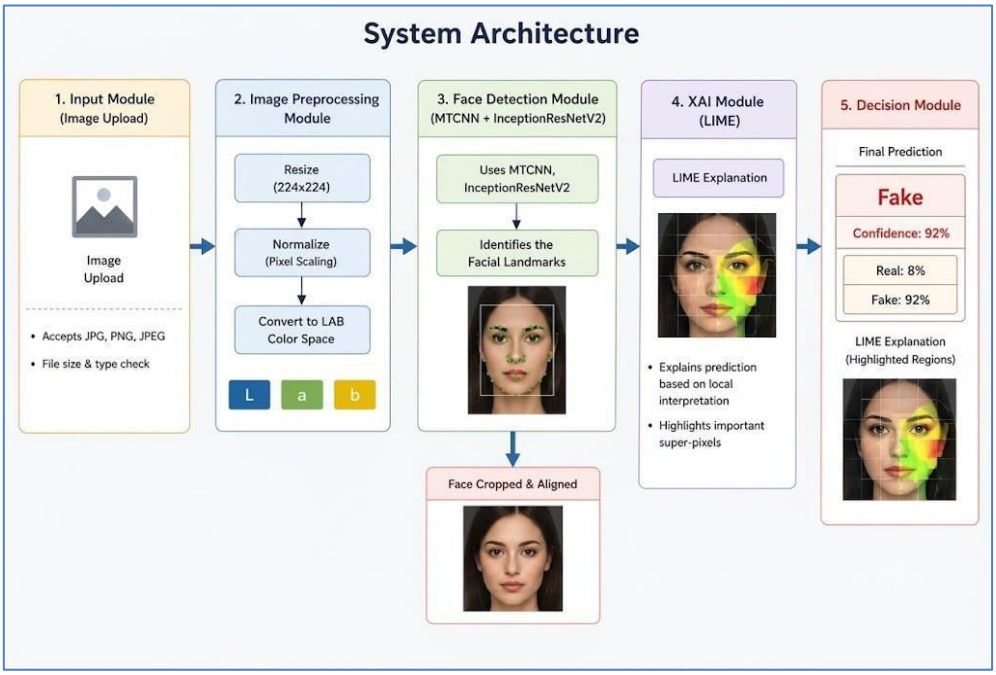


Figure 1: System Architecture

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

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

C. LAB Color Space Preprocessing

Validated RGB images are resized to 224×224 pixels and normalized using X_norm = X/255. They are then converted from RGB to the CIE LAB color space through the intermediate XYZ space using the D65 standard illuminant transformation matrix:

$$X = 0.412453R + 0.357580G + 0.180423B$$
$$Y = 0.212671R + 0.715160G + 0.072169B$$
$$Z = 0.019334R + 0.119193G + 0.950227B$$

Normalized against the D65 reference white point (Xn=95.047, Yn=100.000, Zn=108.883) and passed through f(t)=t^(1/3) for t>0.008856, or f(t)=7.787t+16/116 otherwise, producing the final LAB channels:

$$L = 116 \cdot f(Y/Yn) - 16, \quad A = 500 \cdot [f(X/Xn) - f(Y/Yn)], \quad B = 200 \cdot [f(Y/Yn) - f(Z/Zn)]$$

The L channel encodes luminance [0,100] while A and B encode chrominance axes [−128,127]. This transformation is theoretically motivated by the property that GANs trained on RGB-space perceptual loss functions introduce systematic chrominance deviations that are not directly penalized during generator optimization and become more discriminable once luminance and chrominance are separated into independent channels.

D. MTCNN Face Detection and Alignment

MTCNN operates through a three-stage cascaded architecture. The Proposal Network performs a rapid multi-scale scan producing candidate bounding boxes. The Refinement Network evaluates candidates at higher resolution, discarding false positives and refining coordinates. The Output Network simultaneously finalizes the bounding box and predicts five facial landmark coordinates (left eye, right eye, nose tip, left mouth corner, right mouth corner). These landmarks enable a similarity transformation mapping detected positions to canonical target coordinates, ensuring consistent spatial positioning of facial features across all inputs. The aligned face is cropped with margin m: F(x,y) = I[(y1−m):(y2+m), (x1−m):(x2+m)], preserving hairline and chin context. When detection fails due to extreme pose, occlusion, or low resolution, a center-crop fallback ensures pipeline continuity.

E. InceptionResNetV2 Classification

The aligned face crop is reshaped to a tensor of shape (1,224,224,3) and passed to the InceptionResNetV2 architecture, combining parallel multi-scale Inception modules with residual skip connections to mitigate vanishing gradients in deep networks. The base model is initialized with ImageNet weights, and the original classification layer is replaced with a custom head: GlobalAveragePooling2D → Dense(256, ReLU) → BatchNormalization → Dropout(0.5) → Dense(1, Sigmoid). The sigmoid output produces probability P∈[0,1], with the decision rule: Label=REAL if P>0.5, FAKE otherwise, and confidence S=max(P,1−P)×100. The model was trained using Adam (lr=1×10^{-4}), binary cross-entropy loss, batch size 32, and early stopping with patience=3 monitoring validation loss.

F. LIME Explainability Generation

LIME approximates the classifier's local decision boundary by segmenting the input face crop C into K superpixels and generating N=1,000 perturbed variants z'=C⊙m (m∈{0,1}^K), each masking a random superpixel subset to uniform gray. Each perturbed sample is scored by the trained model. Sample weights are computed via the exponential similarity kernel $\pi_x(z')=\exp(-D(x,z')^2/\sigma^2)$, and a weighted sparse linear surrogate model is fitted by minimizing:

$$\xi(x) = \text{argmin}_g \Sigma \pi_x(z') \cdot (f(z') - g(z'))^2 + \Omega(g)$$

The five superpixels with the highest positive linear model coefficients are highlighted using boundary overlay rendering, producing an image-specific attribution map identifying the facial regions most responsible for the classification outcome.

IV. RESULTS AND PERFORMANCE ANALYSIS

A. Quantitative Performance

Table I presents the comprehensive performance comparison between Model 1 (baseline InceptionResNetV2 with standard RGB preprocessing) and Model 2 (proposed pipeline with MTCNN face alignment and LAB preprocessing) under identical training conditions.

© 2026 IJIRCCE | Volume 14, Issue 7, July 2026 DOI: 10.15680/IJIRCCE.2026.1407064

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

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

Table I. Comprehensive Performance Comparison: Baseline vs. Proposed Model

Metric	Model 1 — RGB	Model 2 — MTCNN+LAB
Test Accuracy (in-dist.)	92.93%	98.93%
Test Loss (in-dist.)	0.0329	0.0318
Generalization Accuracy	90.15%	96.94%
Generalization Loss	0.3077	0.2509
Precision	0.965	0.989
Recall	0.968	0.987
F1-Score	0.978	0.988
LIME Samples / Inference	1,000	1,000

Model 2 achieved an in-distribution test accuracy of 98.93%, exceeding the baseline by 6.00 percentage points. The test loss of 0.0318 (vs. 0.0329) confirms tighter probability calibration around correct class labels. On the out-of-distribution evaluation set, Model 2 achieved 96.94% generalization accuracy compared to 90.15% for Model 1, a 6.79 percentage point improvement, with generalization loss reduced from 0.3077 to 0.2509. Model 2 achieved precision of 0.989 versus 0.965, indicating substantially fewer fake images incorrectly classified as real — a critical property for forensic deployment where false negatives carry severe consequences. Recall improved from 0.968 to 0.987, confirming higher sensitivity to genuine fake images without sacrificing specificity. The F1-score of 0.988 versus 0.978 confirms more balanced and reliable classification across both classes.

B. Training Convergence

Both models exhibited training accuracy beginning near 92% in the first epoch, rising sharply by the second epoch, and stabilizing above 98% from the fourth epoch onward. Validation accuracy tracked training accuracy closely throughout all eight observed epochs, with a consistently small gap between curves confirming the absence of overfitting. Training loss decreased sharply from approximately 2.4 in epoch one to 0.1 by epoch two, consistent with the rapid classification head adaptation expected of transfer learning from pretrained ImageNet representations. Validation loss decreased consistently alongside training loss without divergence, confirming appropriate learning rate configuration and stable gradient dynamics.

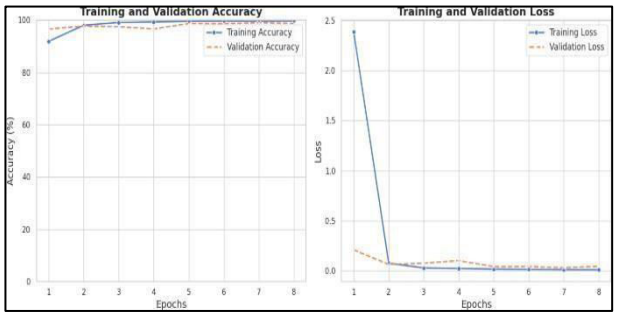


Figure 2: Model 1 (InceptionResNetV2 with RGB preprocessing)

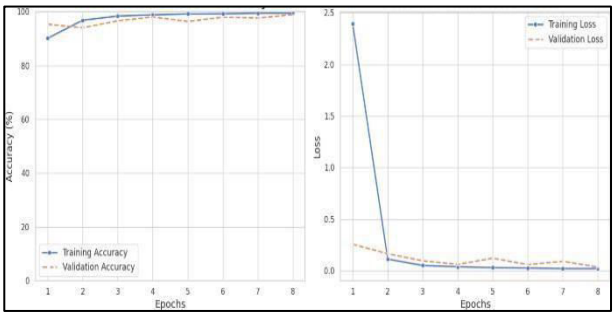


Figure 3: Model 2 (MTCNN with LAB preprocessing)

C. LIME Attribution Analysis

LIME explanations were generated using 1,000 perturbed samples per inference for all correctly classified test samples. For correctly classified fake facial images, the five most influential superpixels identified by LIME were consistently located in three zones: the periocular region surrounding the eyes, the skin-hair boundary at the hairline and temple, and the dentition region encompassing the lips and visible teeth. For correctly classified real facial images, attribution

© 2026 IJIRCCE | Volume 14, Issue 7, July 2026 DOI: 10.15680/IJIRCCE.2026.1407064

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

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

maps consistently highlighted natural skin texture regions, and asymmetric eyelid crease variation. Model 2's face-aligned inputs produced spatially concentrated LIME overlays restricted entirely to facial content, whereas Model 1's full-frame overlays occasionally included superpixels located in background regions outside the facial boundary, reducing forensic precision.

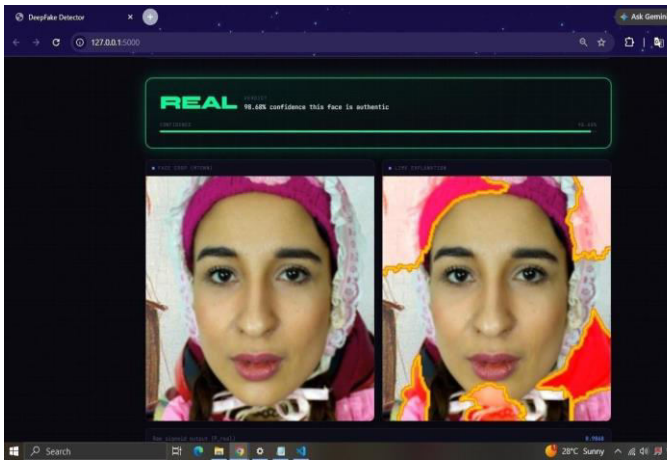


Figure 4: Result Real Face detection

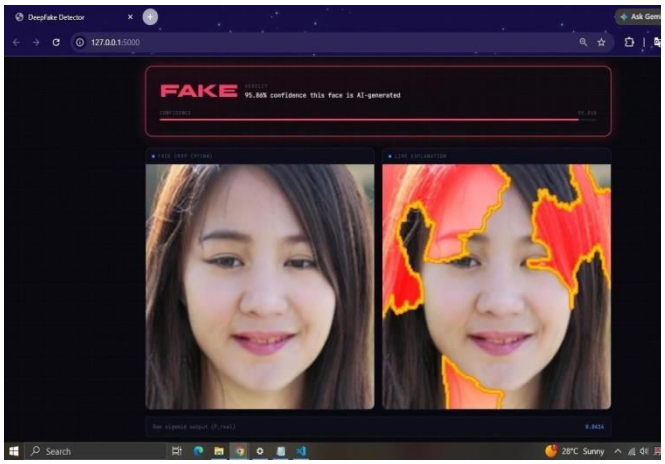


Figure 5: Result Fake Face Detection

V. DISCUSSION

A. Mechanistic Interpretation of LAB Preprocessing

The 6.00 and 6.79 percentage point improvements in in-distribution and out-of-distribution accuracy respectively are most plausibly attributable to the combined effect of LAB preprocessing and MTCNN alignment, since these constitute the sole architectural differences between the two models. GANs trained using perceptual loss functions computed in RGB space tend to introduce systematic deviations in chrominance distribution that are not directly penalized during generator optimization, as the perceptual loss primarily captures luminance-driven texture features. These residual chrominance artifacts become more discriminable as isolated signals in the A and B channels of LAB space, providing the classifier with a more informative feature representation at each spatial location. The larger relative improvement in generalization accuracy compared to in-distribution accuracy supports the interpretation that the chrominance-based discriminative signal captured by Model 2 reflects a general property of GAN synthesis rather than StyleGAN2-specific characteristics, explaining the improved cross-architecture generalizability.

B. Role of Face Alignment

MTCNN face alignment contributes through two complementary mechanisms. The background isolation mechanism prevents the classifier from learning spurious correlations with image background content that may differ systematically between the FFHQ real images and StyleGAN2 fake images for reasons unrelated to facial authenticity. The spatial consistency mechanism ensures that semantically equivalent facial landmarks — eyes, nose, mouth — occupy consistent pixel positions across all inputs, enabling convolutional filters to develop spatially reliable artifact detection responses at known anatomical locations. Without alignment, equivalent generative artifacts appearing at different pixel positions across different training images would require larger spatial pooling operations to generalize, effectively reducing the precision with which the model can localize artifact zones.

C. Forensic Validity of LIME Explanations

The independent convergence of the LIME attribution analysis on the periocular zone, skin-hair boundary, and dentition region — without any explicit encoding of GAN failure mode literature into the framework — provides supporting evidence that the classifier's decisions are grounded in genuine forensic artifacts rather than spurious statistical correlations. These three zones correspond precisely to the regions documented in the generative AI literature as most difficult for GAN architectures to synthesize authentically: fine stochastic hair strand texture at the hairline, natural dental variation and anatomical asymmetry in the dentition, and subtle periocular eyelid crease variation that

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

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

GANs tend to render with unnatural bilateral symmetry. The forensic alignment of LIME-identified regions with known GAN failure zones is particularly significant for legal and investigative deployment, where an explanation must be grounded in observable visual evidence rather than statistical artifact to be admissible or actionable.

D. Comparison with Existing Literature

The results are consistent with Abu Talib et al. [12], who demonstrated that chrominance-luminance decomposition improves detection accuracy, though their dual-model architecture incurred substantial computational overhead and provided no superpixel-level spatial attribution. The present framework achieves comparable accuracy improvement within a single unified model while additionally providing per-prediction LIME attribution. The generalization improvements align with concerns raised by Khan et al. [5] regarding dataset-specific overfitting, suggesting that color-space-aware preprocessing represents a practical mitigation strategy. Compared with Mathews et al. [6] and Silva et al. [7], the LIME explanations generated by the proposed system are spatially more precise due to the face-aligned input, operating entirely on facial content rather than full-frame images where attribution may be diluted by background superpixels.

E. Limitations

Several limitations warrant acknowledgment. The out-of-distribution evaluation does not encompass diffusion-based synthetic faces, which exhibit qualitatively distinct artifact profiles from GAN-generated images. MTCNN detection degrades under extreme pose variation, severe occlusion, and very low image resolution, with center-crop fallback potentially reducing classification quality in such cases. LIME explanation generation, requiring 1,000 perturbed samples per inference, introduces computational overhead that may constrain throughput in high-volume CPU-only deployment scenarios. The adversarial robustness of the framework against inputs deliberately perturbed to evade detection was not evaluated. Both real and fake training images are high-resolution studio-quality, and performance on lower-quality, socially distributed images warrants further investigation.

VI. CONCLUSION AND FUTURE WORK

This paper presented an explainable deepfake facial image detection framework integrating MTCNN face alignment, LAB color space preprocessing, fine-tuned InceptionResNetV2 classification, and LIME visual attribution in a unified, practically deployable pipeline. Against a baseline RGB model trained under identical conditions, the proposed framework achieved a 6.00 percentage point improvement in in-distribution test accuracy (98.93%), a 6.79 percentage point improvement in out-of-distribution generalization accuracy (96.94%), precision of 0.989, recall of 0.987, and F1-score of 0.988. LIME attribution analysis consistently identified the periocular zone, skin-hair boundary, and dentition region as the most influential areas for fake classification, and the forensic alignment of these regions with documented GAN failure modes provides evidence that the classifier performs genuine forensic reasoning rather than spurious pattern matching. The integration of face-aligned, color-space-aware preprocessing with image-specific explainability converts a deep learning classifier into a forensically auditable detection instrument suitable for deployment in digital forensics, media verification, and cybersecurity applications.

Three specific future research directions are proposed. First, extension to video-based detection through recurrent or transformer-based temporal attention mechanisms capable of exploiting physiological consistency cues across consecutive frames, including eye-blink frequency, head pose trajectory, and lip-movement synchronization. Second, expansion of the training corpus to include diffusion-model-generated synthetic faces from Stable Diffusion and DALL-E to evaluate and improve generalization across contemporary non-GAN generative architectures. Third, systematic evaluation and enhancement of adversarial robustness through adversarial training procedures targeting the specific facial regions identified as most influential by LIME attribution, addressing the dual-use concern that published explainability results may reveal exploitable classifier vulnerabilities to adversarial actors.

REFERENCES

[1] I. J. Goodfellow et al., “Generative Adversarial Nets,” in Proc. NeurIPS, vol. 27, pp. 2672–2680, 2014.
[2] T. Karras et al., “Analyzing and Improving the Image Quality of StyleGAN,” in Proc. IEEE/CVF CVPR, pp. 8110–8119, 2020.
[3] A. Rossler et al., “FaceForensics++: Learning to Detect Manipulated Facial Images,” in Proc. IEEE/CVF ICCV, pp. 1–11, 2019.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

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

[4] B. Dolhansky et al., "The Deepfake Detection Challenge (DFDC) Dataset," arXiv:2006.07397, 2020.
[5] M. A. Khan et al., "A Survey on Multimedia-Enabled Deepfake Detection: State-of-the-Art Tools and Techniques, Emerging Trends, Current Challenges and Future Directions," IEEE Access, vol. 13, pp. 1–28, 2025.
[6] A. Mathews et al., "An Explainable Deepfake Detection Framework on a Novel Unconstrained Dataset," Expert Systems with Applications, vol. 213, p. 119030, 2023.
[7] S. Silva et al., "Deepfake Forensics Analysis: An Explainable Hierarchical Ensemble of Weakly Supervised Models," Forensic Science International: Digital Investigation, vol. 42, p. 301390, 2022.
[8] S. Cirillo et al., "Explainability-Driven Adversarial Robustness Assessment for Generalized Deepfake Detectors," Pattern Recognition Letters, vol. 178, pp. 45–52, 2025.
[9] U. Mansoor and A. Iliev, "Explainable AI for DeepFake Detection," Journal of Imaging, vol. 11, no. 2, pp. 1–18, 2025.
[10] K. Barik et al., "Practical Evaluation and Performance Analysis for Deepfake Detection Using Advanced AI Models," Multimedia Tools and Applications, vol. 83, pp. 12541–12563, 2024.
[11] A. Soudy et al., "Deepfake Detection Using Convolutional Vision Transformers and CNNs," Neural Computing and Applications, vol. 36, pp. 8741–8756, 2024.
[12] M. Abu Talib et al., "Chrominance and Luminance: A Study to Detect Deepfakes," Journal of King Saud University – CIS, vol. 37, no. 1, pp. 1–14, 2025.
[13] M. S. Momin et al., "Explainable Deepfake Detection Across Different Modalities: An Overview of Methods and Challenges," Artificial Intelligence Review, vol. 58, no. 3, pp. 1–42, 2025.
[14] K. Zhang et al., "Joint Face Detection and Alignment Using Multi-Task Cascaded Convolutional Networks," IEEE Signal Processing Letters, vol. 23, no. 10, pp. 1499–1503, 2016.
[15] M. T. Ribeiro et al., "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," in Proc. 22nd ACM SIGKDD, pp. 1135–1144, 2016.
[16] C. Szegedy et al., "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning," in Proc. AAAI, pp. 4278–4284, 2017.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details