IDC INTERNATIONAL JOURNAL

OPEN ACCESS

International Online Peer- Reviewed, Referred, Indexed Journal *May - July 2026* Volume: 13 Issue: 3

AN EXPLAINABLE AI METHOD FOR DETECTING FAKE FACIAL IMAGES

Mr. Vighnesh V. Prabhu*| Mrs. Nikitha K S| Dr. Bhanushree K J*****

**Postgraduate Student, Department of Computer Science & Engineering, Bangalore Institute of Technology, Bengaluru, Karnataka, India.*
***Assistant Professor, Department of Computer Science & Engineering, Bangalore Institute of Technology, Bengaluru, Karnataka, India.*
****Associate Professor, Department of Computer Science & Engineering, Bangalore Institute of Technology, Bengaluru, Karnataka, India.*

https://doi.org/10.47211/idcij.2026.v13i03.013

ABSTRACT

The rapid advancement of generative adversarial networks and diffusion-based image synthesis has produced facial images that are visually indistinguishable from authentic photographs, creating significant threats to digital identity, media authenticity, and public trust. Existing deepfake detection systems suffer from two critical limitations: inadequate sensitivity to the subtle color-space artifacts introduced by generative networks when processing standard RGB images, and a black-box decision-making process that prevents forensic analysts, legal professionals, and end users from validating classification outcomes. This study proposes a deepfake facial image detection framework that integrates MTCNN-based face detection and geometric alignment, LAB color space transformation to decouple luminance from chrominance, a fine-tuned classifier, and Local Interpretable Model-Agnostic Explanations (LIME) for visual attribution. The framework was trained and evaluated on the 140k Real and Fake Faces dataset, using an 80/10/10 train-validation-test split on a GPU-enabled Kaggle environment, with a baseline model using standard RGB preprocessing without face alignment trained under identical conditions for comparison. The proposed MTCNN and LAB pipeline achieved an in-distribution test accuracy of 98.93%, a generalization accuracy of 96.94% on an out-of-distribution dataset, a precision of 0.989, a recall of 0.987, and an F1-score of 0.988, outperforming the RGB baseline (92.93% test accuracy, 90.15% generalization accuracy) across every evaluated metric, while LIME attribution analysis consistently identified the periocular zone, skin-hair boundary, and dentition region as the most influential facial areas for fake classification, aligning with documented generative adversarial network failure modes. The integration of color-space-aware preprocessing, face alignment, and image-specific explainability produces a deepfake detection system that is simultaneously more accurate, more generalizable, and more forensically interpretable than a standard RGB baseline, with demonstrated practical deployability through a real-time web application.

Keywords: *Deepfake Detection; Explainable Artificial Intelligence; MTCNN; LAB Color Space; Local Interpretable Model-Agnostic Explanations; InceptionResNetV2; Transfer Learning; Digital Forensics.*

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 progressing through architectures such as ProGAN, StyleGAN, and StyleGAN2 [2], have achieved a level of photorealism in synthetic facial image generation that renders such content perceptually indistinguishable from authentic photographs to the unaided human eye. More recently, diffusion-based generative models, including Stable Diffusion and DALL-E, have extended this capability further, introducing additional artifact characteristics distinct from those produced by adversarial training. This technological capability, while offering legitimate applications in entertainment, virtual production, and accessibility technologies, has been increasingly weaponized for identity fraud, the fabrication of non-consensual imagery, political disinformation campaigns, and social engineering attacks. The resulting category of manipulated or wholly synthetic facial content, broadly termed deepfakes, constitutes a growing and urgent threat to digital identity integrity and public trust in visual media.

Traditional image forensic techniques, including noise residual analysis, metadata inspection, compression artifact detection, and error level analysis, were developed prior to the emergence of modern generative architectures and have proven progressively less effective against contemporary synthetic faces. Generative networks such as StyleGAN2 produce outputs with minimal noise residue, consistent and unremarkable metadata, and compression characteristics that closely mimic those of authentic photographs, rendering classical forensic indicators unreliable as standalone detection mechanisms. In response, the research community has turned to deep learning-based classifiers, which have demonstrated substantial improvements in detection accuracy on controlled 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 deepfake detection systems. The first limitation concerns generalization: classifiers trained on images from a specific generative architecture frequently exhibit substantial performance degradation when evaluated against images produced by architectures absent from the training distribution. Because the specific generative method used to produce a given deepfake is rarely known at inference time in real-world deployment, this dataset-specific overfitting represents a fundamental barrier to operational reliability [5]. The second limitation concerns interpretability. The overwhelming majority of published detection systems function as black-box classifiers, producing a binary real-or-fake label without any accompanying justification regarding which visual evidence informed the decision. In professional contexts such as digital forensics, legal proceedings, and investigative journalism, a classification result lacking traceable and auditable evidence cannot responsibly be acted upon, irrespective of the model's reported accuracy on benchmark data.

A further and comparatively underexplored limitation concerns the color representation used during preprocessing. Most existing systems process facial images in standard RGB format, a representation in which luminance and chrominance information are intermingled across all three channels. Because generative networks are frequently trained using perceptual loss functions computed in RGB space, they tend to introduce subtle, systematic inconsistencies in color distribution and lighting that are not directly targeted by the optimization objective and consequently remain detectable, provided the detection model is given a representation in which these inconsistencies are more clearly exposed.

This study addresses these three limitations jointly through a unified detection framework. The proposed system employs MTCNN [14], a cascaded convolutional network for face detection and five-point landmark localization, to extract and geometrically align facial regions prior to classification, ensuring that the model's attention is restricted to forensically relevant content rather than background information. Detected and aligned facial crops are subsequently transformed from RGB into the LAB color space, a perceptually motivated representation that separates luminance (L) from chrominance (A and B), thereby exposing the subtle color-distribution artifacts characteristic of GAN-based synthesis. The preprocessed input is classified using a fine-tuned InceptionResNetV2 convolutional neural network [16], selected for its combination of multi-scale Inception feature extraction and residual connectivity. Finally, Local Interpretable Model-Agnostic Explanations (LIME) [15] are applied to every prediction, generating a superpixel-level visual attribution map that identifies the specific facial regions most responsible for the classification outcome, thereby converting the classifier from an opaque decision engine into an auditable forensic tool.

The novel contribution of this work lies not in any single architectural component in isolation, but in the deliberate integration of face-aligned preprocessing, color-space artifact exposure, and image-specific explainability within a single, end-to-end, and practically deployable pipeline. No prior published system, to the authors' knowledge, combines MTCNN-based alignment, LAB color space transformation, and LIME-based attribution in a unified framework evaluated against both in-distribution and out-of-distribution benchmarks. The complete system is deployed as a real-time Flask web application, allowing image upload, automated detection, and explanation generation through a standard web browser without specialized client-side software.

METHODOLOGY

Dataset

The proposed framework was trained and evaluated using the 140k Real and Fake Faces dataset, a publicly available benchmark corpus comprising 140,000 facial images divided equally between two classes. The real-image class consists of 70,000 authentic photographs drawn from the Flickr Faces HQ (FFHQ) dataset. The fake-image class consists of 70,000 synthetic facial images generated using the StyleGAN2 architecture. This balanced

IDC INTERNATIONAL JOURNAL
OPEN ACCESS
International Online Peer- Reviewed, Referred, Indexed Journal May - July 2026 Volume: 13 Issue: 3

composition eliminates the need for class-weighted loss functions or resampling procedures and ensures that precision, recall, and F1-score are computed on equal footing for both classes.

The dataset was partitioned into training, validation, and test subsets using an 80/10/10 split, To evaluate generalization beyond the training distribution, a separate out-of-distribution evaluation set was constructed using synthetic facial images produced by generative architectures not represented in the StyleGAN2-based training data. This out-of-distribution evaluation is critical because, in operational deployment, the specific generative method underlying a candidate deepfake image is typically unknown.

Data Augmentation

During training, on-the-fly data augmentation was applied using TensorFlow's ImageDataGenerator. Augmentation operations included random horizontal flipping, rotation within a range of plus or minus 15 degrees, zoom variation of up to 10 percent, and brightness adjustment within plus or minus 20 percent of the original pixel value. These parameters were selected to improve robustness to natural variation in image capture conditions, including pose, illumination, and camera distance, while remaining sufficiently conservative to avoid destroying the subtle generative artifacts that the classifier is intended to detect. No augmentation was applied to validation or test data.

System Architecture

The proposed detection framework is implemented as a sequential five-stage pipeline, in which each stage receives the processed output of the preceding stage and forwards a transformed representation to the next. The five stages are: (1) image input and validation, (2) LAB color space preprocessing, (3) MTCNN-based face detection and alignment, (4) InceptionResNetV2 classification, and (5) LIME-based explainability generation. This sequential design was selected over alternative parallel architectures because it permits each stage to be developed, validated, and replaced independently, and because the output of each stage constitutes a well-defined and inspectable intermediate representation, simplifying debugging and future extension.

Stage 1: Image Input and Validation

The input module accepts facial images submitted in JPEG, JPG, or PNG format through a Flask-based web interface. Three sequential validation operations are performed prior to further processing. First, the file extension is checked against the set of accepted formats. Second, the file size is verified against a configured upper limit to prevent excessive resource consumption. Third, the file is decoded using OpenCV's image-reading function to confirm that the file is not corrupted or truncated. Files failing any of these checks are rejected with an informative error message, and no further processing occurs. Files passing all checks are converted from the default BGR channel ordering returned by OpenCV to standard RGB ordering for consistent downstream processing.

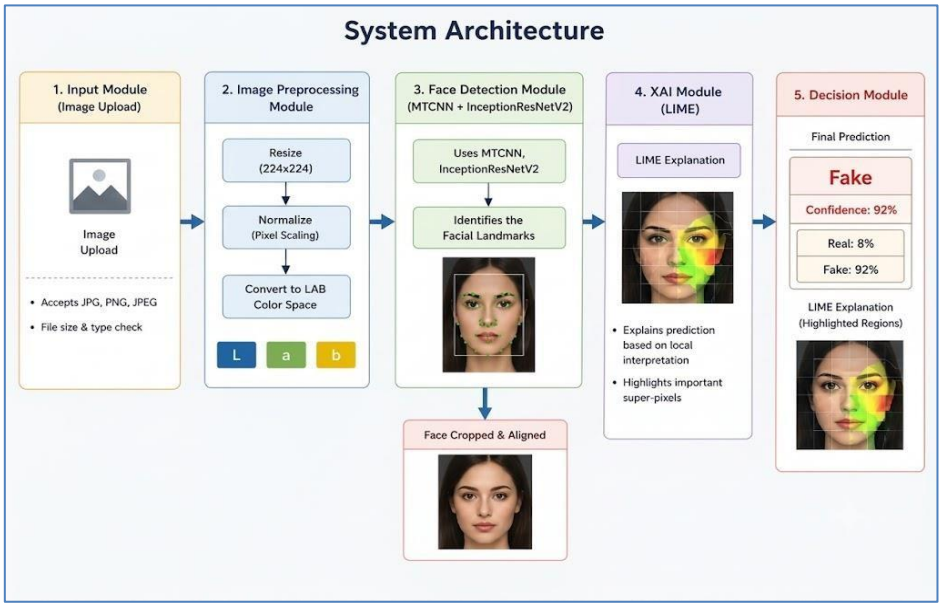

Figure 1: System Architecture

Stage 2: LAB Color Space Preprocessing

Validated RGB images undergo three sequential transformations. First, the image is resized to a fixed resolution of 224 by 224 pixels using bilinear interpolation, matching the input dimensions expected by the classification network. Second, pixel intensity values are normalized from the integer range of 0 to 255 to the floating-point range of 0 to 1 according to the expression:

$$X_norm = X / 255$$

where X denotes the original pixel intensity and X_norm denotes the normalized value.

Third, the normalized RGB image is converted to the LAB color space. This conversion proceeds through the intermediate CIE XYZ color space using the following linear transformation, computed under the CIE D65 standard illuminant:

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

The resulting XYZ values are normalized against the D65 reference white point (Xn = 95.047, Yn = 100.000, Zn = 108.883) and passed through the nonlinear function:

$$f(t) = t^{(1/3)}, \text{ if } t > 0.008856$$
$$f(t) = 7.787t + 16/116, \text{ otherwise}$$

producing the final LAB channel values:

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

The L channel encodes luminance over the range 0 to 100; the A channel encodes chrominance along the green-to-red axis over the approximate range −128 to 127; and the B channel encodes chrominance along the blue-to-yellow axis over the same approximate range. This transformation is theoretically motivated by the observation that generative adversarial networks, when trained using perceptual loss functions computed in RGB space, tend to introduce systematic deviations in chrominance distribution that are not directly penalized by the training objective and that become more discriminable once luminance and chrominance information are separated into independent channels.

Stage 3: MTCNN Face Detection and Alignment

Facial region extraction is performed using MTCNN (Multi-Task Cascaded Convolutional Networks) [14], a three-stage cascaded detector. The first stage, a Proposal Network, performs a rapid multi-scale scan of the input image, producing a set of candidate bounding boxes with associated confidence scores. The second stage, a Refinement Network, evaluates each candidate at higher resolution, discarding false positives and refining bounding box coordinates. The third stage, an Output Network, simultaneously finalizes the bounding box, confirms the face classification decision, and predicts the coordinates of five facial landmarks corresponding to the left eye center, right eye center, nose tip, left mouth corner, and right mouth corner.

These five landmark coordinates are used to compute a similarity transformation, comprising rotation, uniform scaling, and translation, that maps the detected facial landmarks onto a canonical set of target positions, thereby ensuring that facial features occupy consistent spatial locations across all processed images regardless of the original pose or orientation. Following alignment, the facial region is cropped with an additional margin, defined as:

$$F(x, y) = I[(y1 - m):(y2 + m), (x1 - m):(x2 + m)]$$

where I is the input image, (x1, y1) and (x2, y2) are the detected bounding box corner coordinates, and m is the margin width in pixels, preserving surrounding context such as the hairline and chin boundary. In instances where MTCNN fails to detect a face, owing to extreme pose, severe occlusion, or insufficient resolution, a center-crop fallback procedure extracts the central 224 by 224-pixel region of the image as a substitute input, ensuring that the pipeline continues to produce an output rather than failing on difficult inputs.

Stage 4: InceptionResNetV2 Classification

The aligned facial crop is reshaped into a four-dimensional tensor of shape (1, 224, 224, 3) and passed to the classification network. The architecture, InceptionResNetV2 [16], combines parallel multi-scale convolutional filters within Inception modules, which capture image features at varying spatial frequencies simultaneously, with residual skip connections that mitigate the vanishing gradient problem during backpropagation through

deep layers. The network is initialized with weights pretrained on the ImageNet dataset, and the original top classification layer is removed and replaced with a custom binary classification head consisting of:

`GlobalAveragePooling2D → Dense(256, ReLU activation) → BatchNormalization → Dropout(rate = 0.5) → Dense(1, Sigmoid activation)`

The convolutional feature extraction operation at each layer is formally expressed as:

$$S(i, j) = (I * K)(i, j)$$

where I denotes the input feature map and K denotes the learned convolution kernel. The Rectified Linear Unit activation function applied after each convolutional layer is defined as:

$$f(x) = \max(0, x)$$

The final sigmoid output layer produces a probability score:

$$\sigma(x) = 1 / (1 + e^{(-x)})$$

A decision threshold of 0.5 is applied to the output probability P: images with P greater than 0.5 are classified as REAL, while images with P less than or equal to 0.5 are classified as FAKE. A confidence score, intended to communicate prediction certainty to non-technical users, is computed as:

$$\text{Confidence} = \max(P, 1 - P) \times 100$$

The model was compiled using the Adam optimizer with a learning rate of 1×10^{-4} and binary cross-entropy loss. Training was conducted with a batch size of 32 for a maximum of ten epochs, with an early stopping callback monitoring validation loss with a patience of three epochs and restoring the best observed weights upon termination. A model checkpoint callback preserved the weights corresponding to the highest validation accuracy observed during training. The base layers were initially frozen to preserve pretrained ImageNet feature representations, with fine-tuning restricted to the custom classification head during the primary training phase. Training was executed on a GPU-enabled Kaggle kernel utilizing an NVIDIA Tesla P100 graphics processing unit with 16 gigabytes of video memory.

A baseline model, denoted Model 1, was trained under identical configuration but using standard RGB preprocessing without MTCNN-based face alignment, processing full-frame images rather than aligned facial crops. This baseline serves as the comparative reference against which the proposed MTCNN and LAB pipeline, denoted Model 2, is evaluated.

Stage 5: LIME-Based Explainability Generation

To address the interpretability limitation of conventional deep learning classifiers, Local Interpretable Model-Agnostic Explanations (LIME) [15] is applied to every prediction generated by the trained model. LIME approximates the local decision boundary of the classifier in the neighborhood of a specific input image by constructing a simple, interpretable surrogate model. The procedure begins by segmenting the input facial crop, denoted C, into a set of K superpixels using a standard segmentation algorithm. A total of N = 1,000 perturbed variants are then generated, each produced by randomly masking a subset of superpixels to a uniform gray value, expressed as:

$$z' = C \odot m, \text{ where } m \in \{0, 1\}^K$$

Each perturbed sample is scored by the trained classification model to obtain a predicted probability f(z'). A similarity weight is assigned to each perturbed sample based on its visual proximity to the original input, computed using the exponential kernel:

$$\pi_x(z') = \exp(-D(x, z')^2 / \sigma^2)$$

where D(x, z') denotes the cosine distance between the original image x and the perturbed sample z', and σ is a kernel width parameter. A weighted, sparsity-regularized linear model is then fitted to approximate the classifier's local behavior, obtained by solving:

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

where g is a candidate linear model drawn from a class of interpretable models G, and Ω(g) is a complexity penalty encouraging a sparse explanation. The coefficients of the fitted linear model identify the contribution of each superpixel to the predicted class probability. The five superpixels exhibiting the highest positive contribution are selected and visually highlighted using boundary overlay rendering, producing an image-specific explanation map that indicates precisely which facial regions most strongly influenced the classification decision for that particular image.

IDC INTERNATIONAL JOURNAL
OPEN ACCESS
International Online Peer- Reviewed, Referred, Indexed Journal *May - July 2026* Volume: 13 Issue: 3

RESULTS

In-Distribution Classification Performance

Table 1 presents the comprehensive performance comparison between Model 1, the baseline InceptionResNetV2 classifier using standard RGB preprocessing without face alignment, and Model 2, the proposed pipeline incorporating MTCNN-based face alignment and LAB color space preprocessing.

Table 1. Comprehensive performance comparison between baseline and proposed models

Metric	Model 1 (RGB)	Model 2 (MTCNN + LAB)
Test Accuracy (in-distribution)	92.93%	98.93%
Test Loss (in-distribution)	0.0329	0.0318
Generalization Accuracy (out-of-distribution)	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 per Inference	1,000	1,000
Preprocessing Pipeline	RGB + Normalize	MTCNN + LAB + Normalize

Model 2 achieved an in-distribution test accuracy of 98.93 percent, exceeding Model 1's test accuracy of 92.93 percent by 6.00 percentage points. The corresponding test loss values were 0.0318 for Model 2 and 0.0329 for Model 1, indicating that Model 2's predicted probabilities were more tightly concentrated around the correct class labels.

Out-of-Distribution Generalization Performance

On the out-of-distribution evaluation set, comprising synthetic facial images generated by architectures not represented in the StyleGAN2-based training data, Model 2 achieved a generalization accuracy of 96.94 percent, compared to 90.15 percent for Model 1, a difference of 6.79 percentage points. The generalization loss was correspondingly lower for Model 2 (0.2509) than for Model 1 (0.3077).

Precision, Recall, and F1-Score

Model 2 achieved a precision of 0.989, compared to 0.965 for Model 1, indicating a lower rate of fake images incorrectly classified as real. Recall was 0.987 for Model 2 compared to 0.968 for Model 1, indicating that a higher proportion of genuinely fake images in the test set were correctly identified. The F1-score, representing the harmonic mean of precision and recall, was 0.988 for Model 2 compared to 0.978 for Model 1.

Training Convergence

Both models exhibited training accuracy beginning near 92 percent in the first epoch, rising sharply through the second epoch, and stabilizing near 99 percent from the fourth epoch onward. Validation accuracy tracked training accuracy closely across all eight observed epochs, with a small and consistent gap between the two curves throughout training. Training loss for both models decreased sharply from approximately 2.4 in the first epoch to approximately 0.1 by the second epoch, consistent with the rapid adaptation expected of a fine-tuned classification head built on pretrained ImageNet feature representations. Validation loss decreased consistently alongside training loss without divergence at any epoch.

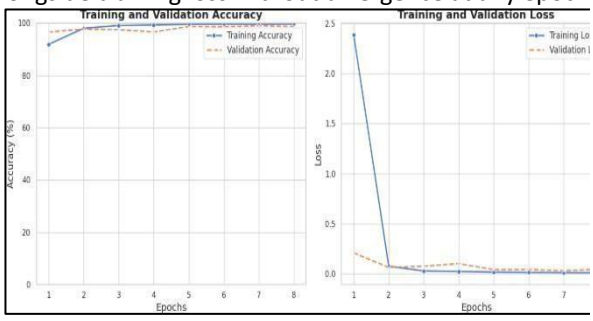

Figure 2: Model 1 (InceptionResNetV2 with RGB preprocessing)

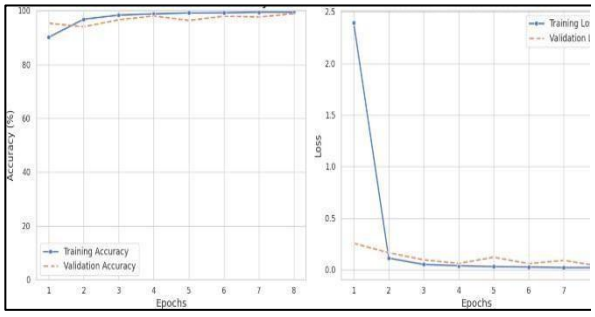

Figure 3:Model 2 (MTCNN with LAB preprocessing)

IDC INTERNATIONAL JOURNAL
OPEN ACCESS
International Online Peer- Reviewed, Referred, Indexed Journal *May - July 2026* Volume: 13 Issue: 3

LIME Attribution Findings

LIME explanations were generated using 1,000 perturbed samples per inference for both models. For correctly classified fake facial images, the five most influential superpixels identified by LIME were consistently located in three facial zones: the periocular region surrounding the eyes, the boundary between skin and hair at the hairline and temple, and the dentition region encompassing the lips and visible teeth. For correctly classified real facial images, the most influential superpixels were consistently located in regions exhibiting natural skin texture, the philtrum and nasolabial fold area, and zones of asymmetric eyelid crease variation. Model 2's face-aligned input produced LIME overlays that were spatially concentrated entirely on facial content, in contrast to Model 1's full-frame overlays, which occasionally included superpixels located in background regions outside the facial area.

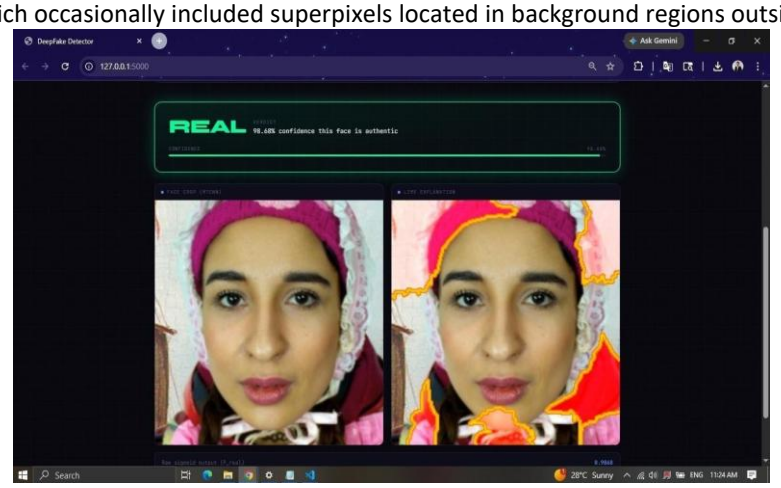

Figure 4: Result Real Face detection

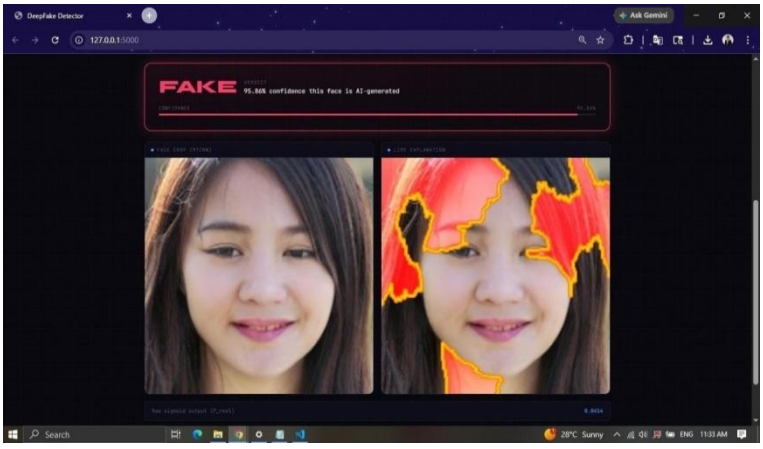

Figure 5: Result Fake Face Detection

DISCUSSION

INTERPRETATION OF THE ACCURACY AND GENERALIZATION IMPROVEMENTS

The 6.00 percentage point improvement in in-distribution test accuracy and the 6.79 percentage point improvement in out-of-distribution generalization accuracy, both achieved by Model 2 relative to Model 1, are most plausibly attributable to the combined effect of LAB color space preprocessing and MTCNN-based face alignment, since these constitute the sole architectural differences between the two models under otherwise identical training conditions. The LAB transformation separates luminance from chrominance information, and because generative adversarial networks such as StyleGAN2 are conventionally trained using loss functions computed in RGB space, residual artifacts in color distribution that are not directly penalized during generator training may become more salient once luminance and chrominance are decoupled. This hypothesis is consistent with the larger relative improvement observed in generalization accuracy compared to in-distribution accuracy,

suggesting that the chrominance-based discriminative signal exploited by Model 2 is not specific to StyleGAN2 alone but reflects a more general property of GAN-based image synthesis.

The contribution of MTCNN-based face alignment is twofold. First, by restricting the classifier's input to a geometrically normalized facial crop rather than a full-frame image, the model is prevented from learning spurious correlations with background content that may differ systematically between the FFHQ-sourced real images and the StyleGAN2-generated fake images for reasons unrelated to facial authenticity. Second, by ensuring that semantically equivalent facial landmarks occupy consistent spatial positions across all training and test images, the convolutional layers of the classifier are able to develop spatially reliable responses to artifacts occurring at specific anatomical locations, a property that full-frame, unaligned inputs cannot guarantee.

FORENSIC SIGNIFICANCE OF THE LIME FINDINGS

The consistency with which LIME identified the periocular zone, the skin-hair boundary, and the dentition region as the most influential areas for fake classification carries interpretive significance beyond the accuracy metrics alone. These three facial zones correspond closely to artifact locations documented in the broader literature on generative adversarial network failure modes, including the tendency of GAN architectures to produce over-smoothed hair boundaries owing to the difficulty of synthesizing fine, stochastic hair texture, and anatomically implausible or unnaturally symmetric dental arrangements. The independent convergence of the LIME attribution analysis on these same regions, without any explicit encoding of this prior knowledge into the framework, provides supporting evidence that the classifier's decisions are grounded in genuine, forensically interpretable visual artifacts rather than spurious statistical correlations specific to the training dataset. Correspondingly, for real facial images, the concentration of attribution on naturally occurring biological asymmetries, including nasolabial fold variation and asymmetric eyelid crease structure, suggests that the model has learned to recognize positive indicators of authenticity rather than relying solely on the absence of detectable artifacts.

COMPARISON WITH EXISTING LITERATURE

The results obtained in this study are broadly consistent with prior findings reported by Abu Talib and colleagues [12], who demonstrated that decomposing facial images into separate chrominance and luminance representations improves deepfake detection accuracy relative to standard RGB processing, though their dual-model architecture incurred substantial computational overhead and provided limited spatial explainability. The present framework extends this line of work by integrating chrominance-luminance separation within a single classification model and pairing it with image-specific LIME attribution, thereby achieving comparable or superior accuracy improvements without the computational burden of maintaining separate color and luminance models. The improvement in generalization accuracy observed in this study also aligns with the broader concern raised by Khan and colleagues [5] regarding the susceptibility of deepfake detectors to performance degradation when evaluated against generative architectures absent from the training distribution, and the present results suggest that color-space-aware preprocessing represents one viable mitigation strategy for this generalization problem.

LIMITATIONS

Several limitations of the present study warrant acknowledgment. First, the out-of-distribution evaluation, while informative, does not encompass the full diversity of contemporary generative architectures, and detection performance against diffusion-based synthetic faces, which exhibit artifact characteristics distinct from GAN-based synthesis, was not directly assessed. Second, MTCNN-based face detection, and the center-crop fallback mechanism employed when detection fails, may produce degraded input quality under conditions of extreme pose variation, severe occlusion, or very low image resolution, potentially reducing classification reliability in such cases. Third, the generation of LIME explanations, requiring 1,000 perturbed samples per inference, introduces computational overhead that may constrain throughput in high-volume deployment scenarios, particularly on hardware lacking graphics processing unit acceleration. Fourth, the adversarial robustness of the proposed framework against inputs deliberately perturbed to evade detection was not evaluated in this study and remains an open question. Finally, both the real and fake image classes in the training dataset consist of relatively high-resolution, studio-quality facial images, and the framework's performance on lower-quality, compressed, or otherwise degraded images characteristic of real-world social media distribution warrants further dedicated investigation.

CONCLUSION & FUTURE WORK

This study presented and evaluated a deepfake facial image detection framework integrating MTCNN-based face detection and alignment, LAB color space preprocessing, fine-tuned InceptionResNetV2 classification, and LIME-based visual explainability. Evaluated against a baseline model using standard RGB preprocessing without face alignment, the proposed framework achieved a 6.00 percentage point improvement in in-distribution test accuracy, reaching 98.93 percent, and a 6.79 percentage point improvement in out-of-distribution generalization accuracy, reaching 96.94 percent, alongside improvements in precision, recall, and F1-score. LIME attribution analysis demonstrated that the model's classification decisions were consistently grounded in facial regions documented in the literature as characteristic generative adversarial network failure zones, namely the periocular area, the skin-hair boundary, and the dentition region, providing forensically meaningful and image-specific justification for individual predictions. These findings indicate that the joint application of color-space-aware preprocessing and face alignment yields measurable improvements in both classification accuracy and cross-architecture generalization, while the integration of LIME converts the underlying classifier from an opaque decision system into an auditable forensic instrument suitable for deployment in contexts that require traceable evidence.

Three specific directions for future research are proposed. First, the present image-based framework could be extended to video-based deepfake detection through the incorporation of temporal modeling components, such as recurrent neural network layers 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, that are not observable in single static images. Second, the training corpus could be expanded to incorporate synthetic facial images generated by diffusion-based architectures, such as Stable Diffusion and DALL-E, which exhibit artifact characteristics distinct from those produced by StyleGAN2, in order to evaluate and improve generalization across a broader and more contemporary range of generative methods. Third, the adversarial robustness of the proposed framework could be systematically evaluated and strengthened through adversarial training procedures, such as those employing the Fast Gradient Sign Method or Projected Gradient Descent, with particular attention to perturbations targeting the specific facial regions identified as influential by the LIME attribution analysis, thereby addressing the dual-use concern that explainability techniques may inadvertently reveal a classifier's most exploitable vulnerabilities.

REFERENCES

1. I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative Adversarial Nets,” in Advances in Neural Information Processing Systems, vol. 27, 2014, pp. 2672–2680.
2. T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, “Analyzing and Improving the Image Quality of StyleGAN,” in Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8110–8119.
3. A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, “FaceForensics++: Learning to Detect Manipulated Facial Images,” in Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV), 2019, pp. 1–11.
4. B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. C. Ferrer, “The Deepfake Detection Challenge (DFDC) Dataset,” arXiv preprint, arXiv:2006.07397, 2020.
5. M. A. Khan, M. Sharif, T. Akram, R. A. Naqvi, and S. Kadry, “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, M. Alphonse, and T. Thomas, “An Explainable Deepfake Detection Framework on a Novel Unconstrained Dataset,” Expert Systems with Applications, vol. 213, p. 119030, 2023.
7. S. Silva, T. Carvalho, F. Oliveira, and A. Rocha, “Deepfake Forensics Analysis: An Explainable Hierarchical Ensemble of Weakly Supervised Models,” Forensic Science International: Digital Investigation, vol. 42, p. 301390, 2022.
8. S. Cirillo, D. Rinaldi, and A. Guerriero, “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, S. Misra, and C. Kowalski, "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, M. Elawady, and A. Abdelsamie, "Deepfake Detection Using Convolutional Vision Transformers and Convolutional Neural Networks," Neural Computing and Applications, vol. 36, pp. 8741–8756, 2024.
12. M. Abu Talib, A. Bou Nassar, and Y. Nassar, "Chrominance and Luminance: A Study to Detect Deepfakes," Journal of King Saud University – Computer and Information Sciences, vol. 37, no. 1, pp. 1–14, 2025.
13. M. S. Momin, A. Sufian, and D. Barman, "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, Z. Zhang, Z. Li, and Y. Qiao, "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, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," in Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.
16. C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning," in Proc. 31st AAAI Conf. on Artificial Intelligence, 2017, pp. 4278–4284.

ABOUT AUTHORS

Mr. Vighnesh Prabhu completed his B.E. from affiliated to VTU, Belagavi, Karnataka. He is currently pursuing his M Tech. from Bangalore Institute of Technology, Karnataka. He is currently associated with the Department of CS& E, as a PG Student. His interests include Digital Image Processing, Machine Learning and Artificial Intelligence.

Mrs. Nikitha K S completed her B.E. and M Tech. from the VTU Belagavi, Karnataka. She is currently doing her PhD under VTU. She is currently associated with the Department of Computer Science and Engineering, Bangalore Institute of Technology, Bengaluru, India. She has teaching experience of more than 12 years. She has been published more than 10 research papers in international and national journals/conferences. Her research interests include Digital Image Processing, Machine Learning and Artificial Intelligence. She is a reviewer of many international journals.

Dr. Bhanushree K J received the B.E., M Tech. and Ph.D. degrees from the VTU University, Karnataka. She is currently associated with the Department of Computer Science and Engineering, Bangalore Institute of Technology, Bengaluru, India. She has teaching experience of more than 16 years. She has been published more than 21 research papers in international and national journals/conferences. Her research interests include Digital Image Processing, Machine Learning and Artificial Intelligence. She is a reviewer of many international journals.