WWW.IJRPR.COM

**International Journal of Research Publication and Reviews**

Journal Homepage: www.ijrpr.com ISSN 2582-7421

# AI Image Forensics Detection System: A Multi-Model Framework for AI-Generated Image and Deepfake Detection

***Tejaswini K P[1], Dr. Santoshkumar M[2]***

*[1]Department of CSE, GM Institute of Technology, Davanagere - 577006, Affiliated to VTU, Belagavi, Karnataka, India*
*[2]Associate Professor, Department of CSE, GM Institute of Technology, Davanagere - 577006, Affiliated to VTU, Belagavi, Karnataka, India*
*Email: [1]tejaswinikp2003@gmail.com,[2] santoshkumarm@gmit.ac.in*

**ABSTRACT**

The increasing availability of artificial intelligence techniques for generating synthetic images and manipulating facial content has created a need for effective image-analysis methods. This paper presents an AI Image Forensics Detection System that integrates three deep-learning models: a ResNet-18-based AI-generated image detector associated with the CIFAKE dataset, a ResNet-18-based human-face deepfake detector associated with the 140K Real and Fake Faces dataset, and an EfficientNet-B3-based face deepfake detector associated with the FaceForensics++ dataset. The system preprocesses input images according to each model's requirements and provides REAL/FAKE predictions with confidence scores. Gradient-weighted Class Activation Mapping (Grad-CAM) generates heatmaps and overlays to help visualize image regions influencing individual model predictions. In addition, a separate image-similarity module compares two images using the Structural Similarity Index (SSIM), Sobel and Canny edge analysis, Local Binary Patterns (LBP), Fast Fourier Transform (FFT), Discrete Cosine Transform (DCT), and deep-feature similarity, producing similarity scores and a difference heatmap. A weighted decision engine combines the predictions and confidence values of the three detectors to produce a final assessment of LIKELY REAL, LIKELY FAKE, or UNCERTAIN, together with a trust score reflecting model confidence and agreement. Grad-CAM visualizations and image-similarity results are separate analysis outputs and are not inputs to the weighted final decision. The system is intended to support preliminary image-forensics analysis rather than provide definitive authentication of digital images.

**Keywords:** AI Image Forensics, AI-Generated Image Detection, Deepfake Detection, ResNet18, EfficientNet-B3, Grad-CAM, Image Similarity Analysis

## 1. INTRODUCTION

The rapid development of artificial intelligence (AI) and deep learning has enabled the generation of realistic synthetic images and the manipulation of facial images. These developments have created challenges in determining whether digital images are genuine, AI-generated, or manipulated. Deep learning models, particularly Convolutional Neural Networks (CNNs), can learn visual features to support image classification and deepfake detection. However, different image-analysis tasks may require different model architectures and preprocessing procedures, motivating the development of systems that combine multiple detection approaches and provide interpretable results.

This paper presents an AI Image Forensics Detection System that integrates three deep-learning models: a ResNet18-based AI-generated image detector associated with the CIFAKE dataset, a ResNet18-based human-face deepfake detector associated with the 140K Real and Fake Faces dataset, and an EfficientNet-B3-based face deepfake detector associated with the FaceForensics++. The system combines the predictions and confidence values of these three detectors through a weighted decision mechanism to produce an overall assessment of LIKELY REAL, LIKELY FAKE, or UNCERTAIN, together with a trust score reflecting model confidence and agreement.

To support the interpretation of individual model predictions, the system uses Gradient-weighted Class Activation Mapping (Grad-CAM) to generate heatmaps and overlays highlighting image regions that influence the models' predictions. In addition, a separate image-similarity module compares two images using the Structural Similarity Index (SSIM), Sobel and Canny edge analysis, Local Binary Patterns (LBP), Fast Fourier Transform (FFT), Discrete Cosine Transform (DCT), and deep-feature similarity. This module produces similarity scores and a difference heatmap, providing additional information about similarities and differences between the images. The image-similarity module operates independently, and its results are not included in the final decision generated by the weighted decision mechanism. The final decision relies on the predictions and confidence values of the three detection models. The main contribution of this work is the implementation of these complementary capabilities in a unified desktop application for the preliminary examination of potentially AI-generated or manipulated images. The system is intended to support image-forensics analysis rather than provide definitive authentication.

## 2. PROBLEM STATEMENT

The increasing availability of AI-generated images and manipulated facial images makes it difficult to distinguish authentic images from synthetic or altered content through visual inspection alone. Different image types may require different detection models and preprocessing procedures, while individual model predictions may not provide sufficient information for a broader forensic assessment. Therefore, a system is needed to support multiple image-detection tasks and provide additional information for examining potentially manipulated images.

To address this need, the proposed AI Image Forensics Detection System integrates three deep-learning models: ResNet18 for AI-generated image detection, ResNet18 for face deepfake detection, and EfficientNet-B3 for face deepfake detection. The system also provides Grad-CAM visualizations to help examine image regions influencing individual model predictions, a separate image-similarity module for comparing two images, and a weighted decision mechanism that combines model predictions and confidence values to produce an overall assessment of LIKELY REAL, LIKELY FAKE, or UNCERTAIN. The system is intended to support preliminary image-forensics examination; its outputs should not be treated as definitive proof of image authenticity.

## 3. LITERATURE REVIEW

Several studies have explored deep-learning techniques for AI-generated image detection and DeepFake analysis. Bird and Lotfi introduced the CIFAKE dataset for distinguishing real images from AI-generated synthetic images and investigated explainability techniques for interpreting model predictions [1]. Naskar et al. proposed a DeepFake detection approach that combines deep features extracted using Xception and EfficientNet-B7 through feature stacking and meta-learning to classify real and fake video sequences [2]. Jhun et al. introduced an image waveform representation to capture texture information and proposed a two-stream Convolutional Neural Network (CNN) for DeepFake detection using the original image and its waveform representation [3]. Lau et al. integrated a Convolutional Block Attention Module (CBAM) with ResNet-18 to improve explainability in fake-image detection [4]. Deepa et al. investigated an EfficientNet-B3-based DeepFake detection framework and compared deep-learning models with machine-learning techniques for image classification [5].

These studies provide background on synthetic-image classification, deep-feature combination, texture analysis, attention-based fake-image detection, and EfficientNet-B3-based detection. The proposed AI Image Forensics Detection System integrates three deep-learning image detectors into a single desktop application with a weighted decision mechanism that combines their predictions and confidence values to produce an overall assessment. The system also provides Grad-CAM visualizations to help interpret individual model predictions, a separate image-similarity module for comparing two images, and a Full Forensic Analysis feature that presents the detector outputs, overall assessment, and supporting forensic information when available. The image-similarity module operates separately from the weighted decision mechanism, while Full Forensic Analysis presents the combined assessment alongside the available supporting information. Together, these capabilities support preliminary image-forensics analysis.

## 4. PROPOSED SYSTEM AND METHODOLOGY

### A. Proposed System

The proposed AI Image Forensics Detection System is a desktop application developed using Python and the CustomTkinter graphical user interface. It integrates three deep-learning models to analyse AI-generated images and manipulated facial images. The system consists of a ResNet18-based AI-generated image detector associated with the CIFAKE dataset, a ResNet18-based human-face deepfake detector associated with the 140K Real and Fake Faces dataset, and an EfficientNet-B3-based face deepfake detector associated with FaceForensics++.

Each model processes the input image according to its specific preprocessing requirements and produces a REAL or FAKE prediction with a confidence score. Grad-CAM is used to generate heatmaps and overlays that visually explain the regions influencing each model's prediction. In addition, a separate image similarity analysis module compares two uploaded images using multiple image-analysis techniques. The predictions and confidence values from the three detection models are combined by a multi-model decision mechanism to generate a final assessment of LIKELY REAL, LIKELY FAKE, or UNCERTAIN, together with a trust score.

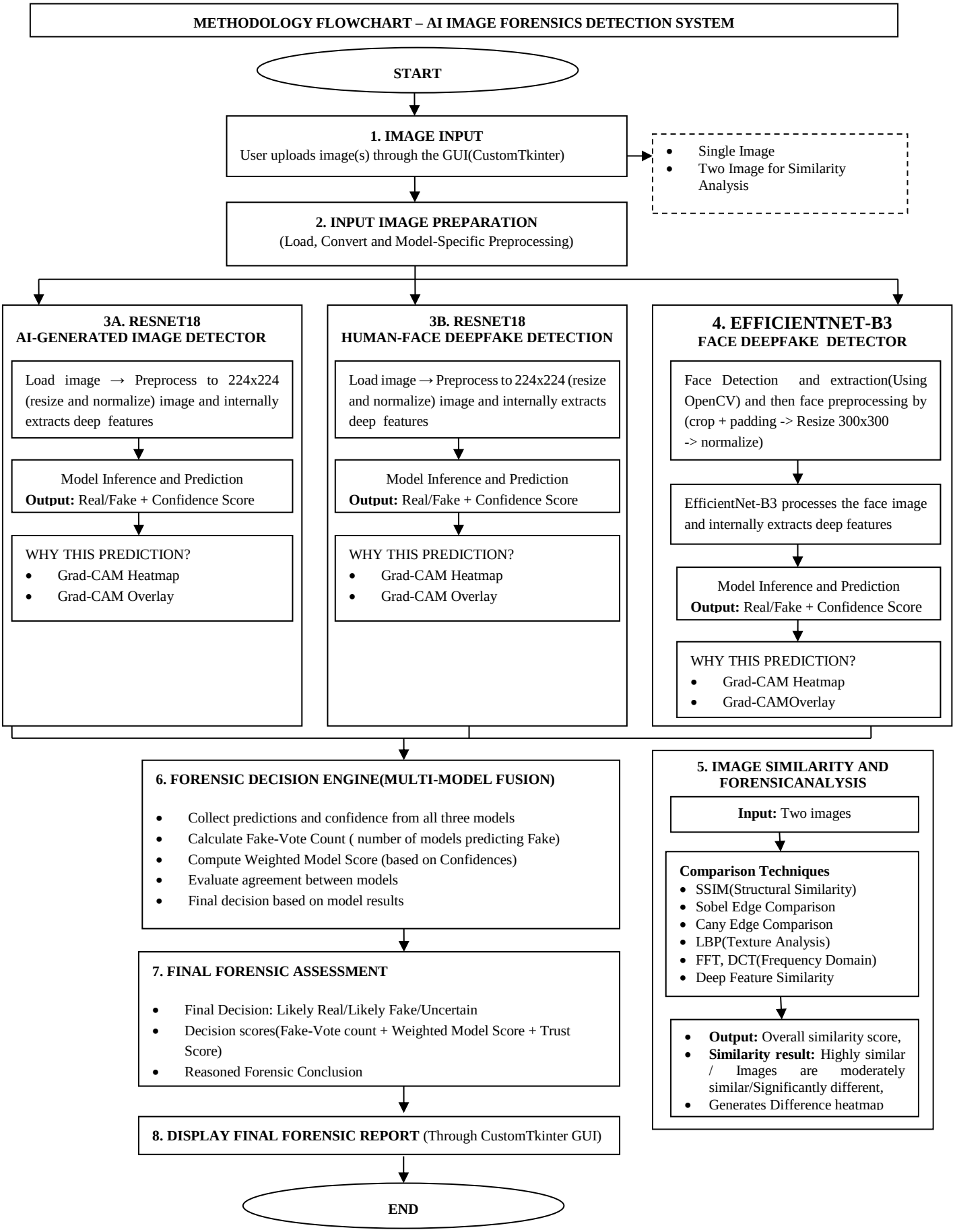



**Figure 4.1: Methodology flowchart of the proposed AI Image Forensics Detection System**

**B. System Workflow**

The system begins when the user uploads an image through the graphical user interface (GUI) developed using CustomTkinter. The uploaded image is loaded and preprocessed according to the input requirements of each detection model. The three deep-learning models independently perform their respective classifications and produce predictions with confidence values. These outputs are passed to the forensic decision engine, which applies the implemented weighted scoring and voting rules to determine the final assessment. The application displays the individual model predictions, confidence values, final assessment, and trust score. When requested, Grad-CAM visualizations are generated separately to help explain individual model predictions. If two images are provided for comparison, the separate image similarity module analyses their similarities and differences and generates similarity measurements and a difference heatmap. Grad-CAM visualizations and image similarity results are separate analysis outputs and are not used as inputs to the current forensic decision engine.

**C. Image Loading and Preprocessing**

The input images are loaded using image-processing libraries such as OpenCV and PIL. Preprocessing is performed separately for each model to match its expected input dimensions and normalization requirements. The two ResNet18 models receive RGB images resized to 224 × 224 pixels. Pixel values are normalized using the ImageNet mean and standard deviation. For the EfficientNet-B3 detector, OpenCV's Haar cascade is used to detect faces. The largest detected face is cropped with additional padding and resized to 300 × 300 pixels before normalization. If a valid face crop cannot be obtained, the system falls back to processing the full image.

**D. Deep-Learning Detection Models**

**i) ResNet18-Based AI-Generated Image Detection:** This model analyses input images to distinguish between REAL and FAKE categories. The image is resized to 224 × 224 pixels and normalized before inference. The trained ResNet18 model produces a prediction and confidence score. This detector is associated with the CIFAKE dataset.

**ii) ResNet18-Based Human-Face Deepfake Detection:** The second ResNet18 model analyses facial images to classify them as REAL or FAKE. It uses the same 224 × 224 input size and ImageNet normalization procedure. The model produces an individual prediction and confidence score and is associated with the 140K Real and Fake Faces dataset.

**iii) EfficientNet-B3-Based Face Deepfake Detection:** The third model uses EfficientNet-B3 to analyse the detected facial region. The largest face is cropped with padding, resized to 300 × 300 pixels, and normalized before inference. The model produces a REAL or FAKE prediction and confidence score. This detector is associated with the FaceForensics++ dataset.

**E. Grad-CAM-Based Explainability**

Gradient-weighted Class Activation Mapping (Grad-CAM) is implemented to provide visual explanations for the predictions of all three deep-learning models. When the user selects the “WHY THIS PREDICTION?” option, the system generates a Grad-CAM heatmap and an overlay on the input image. These visualizations highlight image regions that influence the model's prediction and help users inspect the model's response. Grad-CAM serves as an interpretability aid and does not independently establish whether an image is authentic or manipulated. The resulting visualizations are presented separately and are not used as inputs to the forensic decision engine.

**F. Image Similarity Analysis**

The separate image similarity module compares two user-provided images using multiple techniques. These include Structural Similarity Index (SSIM), Sobel and Canny edge analysis, Local Binary Patterns (LBP), Fast Fourier Transform (FFT), Discrete Cosine Transform (DCT), and deep-feature similarity. The module produces individual similarity scores, an overall similarity score, and a difference heatmap. The results support the visual examination of similarities and differences between the images. This module operates independently of the final three-model decision mechanism.

**G. Multi-Model Decision Mechanism**

The decision engine combines the predictions and confidence values from the three detection models. The models are assigned weights of 30% to the ResNet18 AI-generated image detector, 30% to the ResNet18 human-face deepfake detector, and 40% to the EfficientNet-B3 face deepfake detector.

The engine calculates the fake-vote count, weighted model score, agreement between model predictions, and trust score. The final assessment is determined using the implemented decision rules. The system classifies an image as LIKELY FAKE when at least two models predict FAKE and the weighted model score meets the implemented threshold of 0.50. It classifies an image as LIKELY REAL when at least two models predict REAL and the weighted model score is below 0.40. Other combinations are classified as UNCERTAIN. The trust score summarizes model confidence and agreement according to the implemented scoring procedure. It should not be interpreted as a calibrated probability that an image is authentic. Grad-CAM visualizations and image similarity measurements are not inputs to this decision mechanism.

**H. Implementation Environment**

The AI Image Forensics Detection System was implemented in Python using CustomTkinter to develop the desktop graphical user interface. PyTorch and Torchvision were used for loading the trained models, performing inference, and preprocessing images, while the timm library was used to implement the EfficientNet-B3 architecture. OpenCV was used for face detection and image processing, PIL/Pillow for image loading, and NumPy for numerical

operations. Model training and evaluation were performed in Google Colab, while the desktop application was developed and integrated in Visual Studio Code.

The application integrates three deep-learning detection models: two ResNet18-based models and one EfficientNet-B3 model. Grad-CAM was implemented for all three models to generate visual explanations of their predictions. A separate Image Similarity Comparison module compares two uploaded images and provides similarity measurements and a difference heatmap. The Full Forensic Analysis module combines the predictions and confidence values of the three detection models using a weighted decision mechanism to produce the final assessment and trust score.

## 5. IMPLEMENTATION RESULTS AND DISCUSSION

The proposed AI Image Forensics Detection System was implemented using three deep-learning models: ResNet18 for AI-generated image detection using the CIFAKE dataset, ResNet18 for face deepfake detection using the 140K Real and Fake Faces dataset, and EfficientNet-B3 for facial deepfake detection using FaceForensics++. Each model was trained and evaluated according to its respective dataset configuration. Table 5.1 summarizes the dataset sample counts and the recorded training, validation, and testing accuracies.

| **Model/ dataset** | **Total Samples** | **Training accuracy (in %)** | **Validation accuracy (in %)** | **Testing Accuracy (in %)** |
|---|---|---|---|---|
| ResNet18– CIFAKE | 120,000 | 93.46 | 96.48 | 96.48 |
| ResNet18–140K Real and Fake Faces | 140,000 | 91.26 | 94.16 | 92.09 |
| EfficientNet-B3 FaceForensics++ | 182,246 | 97.94 | 94.63 (peak at epoch 3) | 95.20 |

**Table 5.1 summarizes the dataset sample counts and the recorded training, validation, and testing accuracies.**

### A. ResNet18-Based AI-Generated Image Detection

The training and validation accuracy graph illustrates the learning performance of the ResNet18 model across eight epochs. The confusion matrix evaluates 20,000 samples, of which 19,296 were correctly classified and 704 were misclassified, resulting in an accuracy of 96.48%. The classification report presents the precision, recall, and F1-score for both REAL and FAKE image classes. The application was also manually tested by uploading different images and observing the predicted class and confidence score. However, since the same evaluation data was used for checkpoint selection, the reported 96.48% accuracy should not be interpreted as an independent test result.

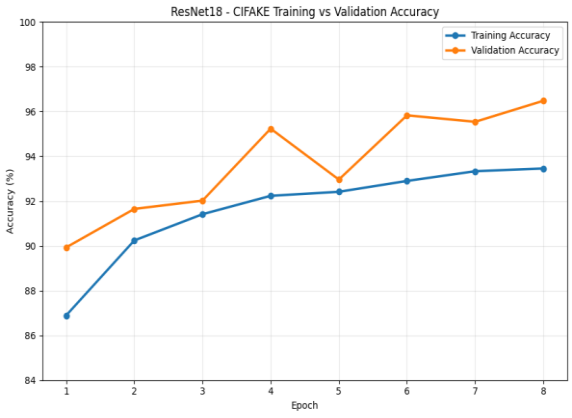



**Fig. 5.1. Training and validation accuracy of the ResNet18 CIFAKE-associated model**

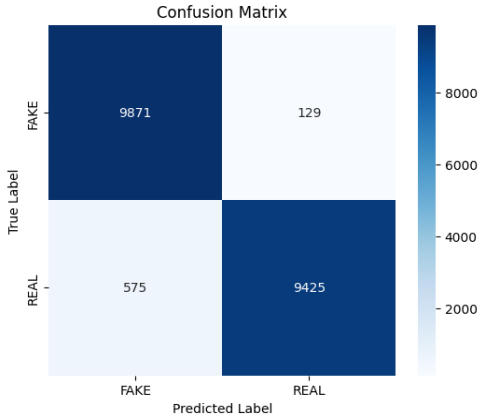



**Fig.5.2. Confusion matrix of the ResNet18 CIFAKE associated model**

**ResNet18 - CIFAKE Classification Report**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| FAKE | 0.94 | 0.99 | 0.97 | 10000 |
| REAL | 0.99 | 0.94 | 0.96 | 10000 |
| accuracy | | | 0.96 | 20000 |
| macro avg | 0.97 | 0.96 | 0.96 | 20000 |
| weighted avg | 0.97 | 0.96 | 0.96 | 20000 |

**Fig. 5.3. Classification report of the ResNet18 CIFAKE-associated model**

**B. ResNet18-Based Human-Face Deepfake Detection**

The training and validation accuracy graph illustrates the learning performance of the ResNet18 model for human-face deepfake detection. The confusion matrix evaluates the model on 20,000 samples, of which 18,418 were correctly classified and 1,582 were misclassified, resulting in an accuracy of 92.09%. The classification report provides the precision, recall, and F1-score for the real and fake face classes, offering further insight into the model's classification performance. In addition to this numerical evaluation, the implemented application was manually tested by uploading different face images and observing the predicted class and confidence score. These tests demonstrated the application's ability to generate predictions for uploaded images. However, manual testing of individual images does not independently establish the overall model accuracy. Therefore, the reported 92.09% accuracy describes the evaluated samples and should be interpreted in the context of the model's evaluation procedure

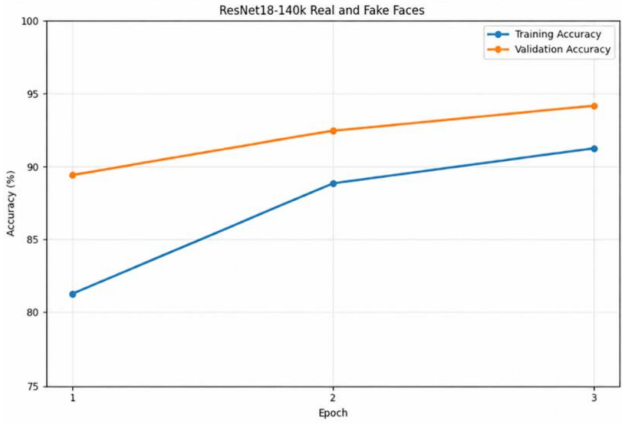



**Fig.5. 4. Training and validation accuracy of the ResNet18 human-face associated model**

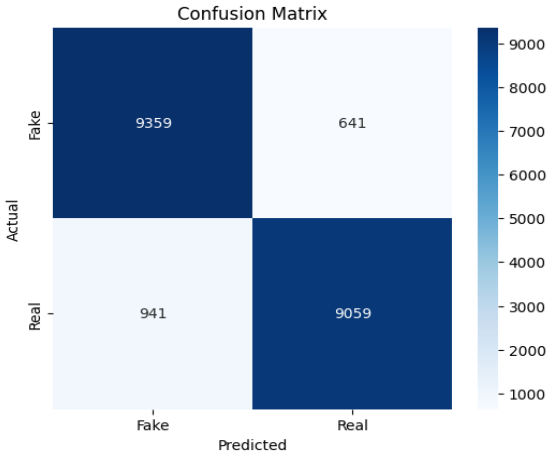



**Fig.5.5. Confusion matrix of the ResNet18 human-face associated model**

**ResNet18 - 140K Real and Fake Faces Classification Report**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Fake | 0.91 | 0.94 | 0.92 | 10000 |
| Real | 0.93 | 0.91 | 0.92 | 10000 |
| accuracy | | | 0.92 | 20000 |
| macro avg | 0.92 | 0.92 | 0.92 | 20000 |
| weighted avg | 0.92 | 0.92 | 0.92 | 20000 |

**Fig. 5.6. Classification report of the human-face ResNet18 model**

**C. EfficientNet-B3-Based Face Deepfake Detection**

The third model uses EfficientNet-B3 for face deepfake detection and is associated with the FaceForensics++ dataset. The training and validation accuracy graph illustrates the model's learning behaviour across epochs. The confusion matrix contains 26,276 samples, of which 25,016 were correctly classified and 1,260 were misclassified, resulting in a calculated accuracy of approximately 95.20%. The classification report provides class-wise precision, recall,

and F1-score for the REAL and FAKE classes. In the implemented dataset loader, REAL images are assigned label 0, while FAKE images are assigned label 1.

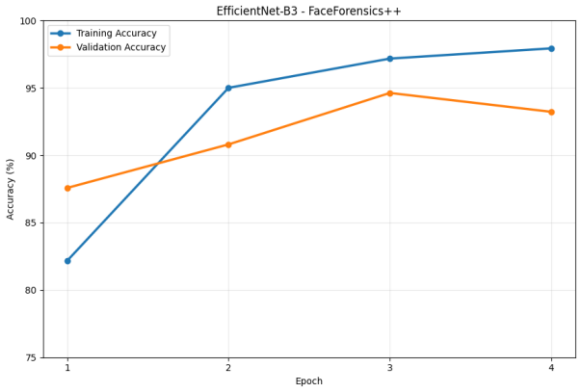



**Fig. 5.7. Training and validation accuracy of the EfficientNet-B3 model**

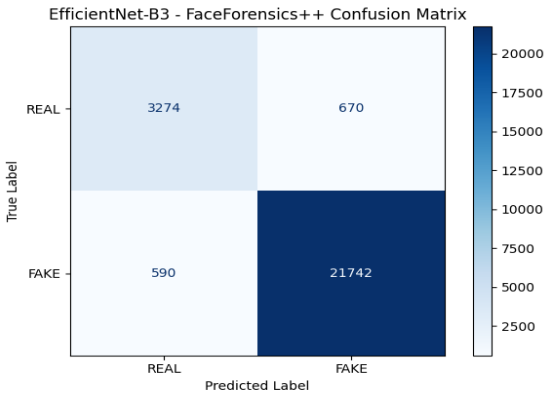



**Fig.5.8. Confusion matrix of the EfficientNet-B3 model**

**EfficientNet-B3 - FaceForensics++ Classification Report**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.85 | 0.83 | 0.84 | 3944 |
| 1 | 0.97 | 0.97 | 0.97 | 22332 |
| accuracy | | | 0.95 | 26276 |
| macro avg | 0.91 | 0.90 | 0.91 | 26276 |
| weighted avg | 0.95 | 0.95 | 0.95 | 26276 |

**Fig. 5.9. Classification report of the EfficientNet-B3 model**

Overall, the proposed system integrates three deep-learning models to provide individual image predictions, confidence scores, and Grad-CAM heatmaps and overlays for each model. The desktop application also includes a separate image similarity analysis module and a multi-model decision mechanism that combines the three models predictions to generate a final assessment of LIKELY REAL, LIKELY FAKE, or UNCERTAIN.

## 6. CONCLUSION

This paper presents an AI Image Forensics Detection System that integrates three deep-learning models for AI-generated image detection and face deepfake detection. The system provides individual predictions, confidence scores, and Grad-CAM heatmaps and overlays for each model to visualize the image regions influencing their predictions. It also includes a separate image similarity analysis module that compares two images using multiple structural, edge-based, texture-based, frequency-based, and deep-feature techniques. Furthermore, the multi-model decision mechanism combines the predictions of the three detectors to generate a final assessment of LIKELY REAL, LIKELY FAKE, or UNCERTAIN. The developed desktop application brings these features together to support the preliminary examination of potentially AI-generated and manipulated images.

## 7. FUTURE SCOPE

The proposed system establishes a multi-model approach to AI-generated and manipulated image analysis by integrating three deep-learning models, Grad-CAM visual explanations, image similarity analysis, and a weighted decision mechanism within a desktop application. Future work may focus on extending the system to a web-based platform while retaining its existing detection and forensic analysis capabilities. Further studies may investigate the system's performance on additional image datasets and explore improvements in detection reliability and generalization to different image manipulation techniques. These extensions could broaden the practical application of the proposed system in digital image forensics.

## REFERENCES

[1] J. J. Bird and A. Lotfi (2024), “CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images,” *IEEE Access*, vol. 12, pp. 15642–15650.

[2] G. Naskar, S. Mohiuddin, S. Malakar, E. Cuevas, and R. Sarkar (2024), "Deepfake Detection Using Deep Feature Stacking and Meta-Learning," *Heliyon*, vol. 10, no. 4, Art. no. e25933.

[3] G. S. Jhun, P. M. Hong, K. Lee, J. H. Ahn, Y. S. Kim, D. Kang, and Y. K. Lee (2024), "A New Wave of Texture Feature: Enhancing Deepfake Detection via Image Waveform," in *Proceedings of the 15th International Conference on Information and Communication Technology Convergence (ICTC)*, pp. 234–238.

[4] S. K. Lau, S.-L. Chua, and L. K. Foo (2025), "Towards Explainable Deep Learning with Attention Mechanisms for Fake Image Detection," in *Artificial Intelligence and Human-Computer Interaction, Frontiers in Artificial Intelligence and Applications*, vol. 404, pp. 55–61.

[5] B. G. Deepa, C. K. Lokesh, D. Umamaheswari, B. Ayshwarya, P. V. Yethish, and K. P. Suhaas (2026), "An Enhanced Deep Learning Framework for DeepFake Detection Using EfficientNet-B3: Comparative Evaluation of Deep and Machine Learning Techniques," *Discover Computing*, vol. 29, Art. no. 18.

[6] I. Goodfellow, Y. Bengio, and A. Courville (2016), *Deep Learning*. Cambridge, MA, USA: MIT Press.

[7] K. He, X. Zhang, S. Ren, and J. Sun (2016), "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778.

[8] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner (2019), "FaceForensics++: Learning to Detect Manipulated Facial Images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1–11.

[9] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. Canton Ferrer (2020), "The DeepFake Detection Challenge (DFDC) Dataset," *arXiv preprint*, arXiv:2006.07397.

[10] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu (2020), "Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3207–3216.