ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 12, Issue 08, August 2025

Image Artifact Analysis for Deepfake Detection Using GAN with EfficientNet-B4 and Self-Attention

[1] S. Sarawathi, [2] Lakshmi S

[1] Assistant Professor, JSS Institute of Speech and Hearing, Mysore, India
[2] Department of Computer Science MES College of Arts, Commerce and Science Bangaluru, India
Corresponding Author Email: [1] Saraswathi_gopi@rediffrmail.com, [2] mes.lakshmis@mesinstitutions.org.in

Abstract*— The rapid advancement of generative models has led to the creation of highly convincing deepfake images, raising significant concerns in security, media integrity, and digital forensics. This paper presents a robust deepfake detection framework leveraging a hybrid architecture combining EfficientNet-B4 and Self-Attention mechanisms within a Generative Adversarial Network (GAN) framework. The proposed system introduces an enhanced pre-processing pipeline including Error Level Analysis (ELA) and edge detection to expose visual inconsistencies commonly present in manipulated images. The discriminator, enriched with EfficientNet-B4 and Self-Attention, effectively distinguishes between real and fake images, achieving high classification accuracy. Experiments were conducted on a balanced dataset comprising over 13,000 real and fake images. The results demonstrate significant improvements in detection accuracy and artifact localization. This study contributes a novel integration of CNN-transformer hybrid models with artifact-focused pre-processing, offering a promising direction for future deepfake detection systems.

Index Terms*— Deepfake Detection, EfficientNet-B4, Self-Attention, GAN, Image Artifacts, Error Level Analysis, Edge Detection, CNN, Transformer.

I. INTRODUCTION

In recent years, the proliferation of deepfakes—synthetic media created using artificial intelligence—has escalated dramatically. With the rise of powerful generative models such as Generative Adversarial Networks (GANs), it has become increasingly difficult to distinguish between authentic and manipulated visual content. These synthetic images and videos can deceive not only human observers but also conventional machine learning systems, posing significant threats to security, journalism, politics, and digital trust.

Traditional deepfake detection methods primarily rely on convolutional neural networks (CNNs) to capture spatial inconsistencies or visual artifacts. While effective to a degree, such approaches may fall short against more sophisticated forgeries. Moreover, most conventional models lack attention mechanisms that help focus on critical visual regions indicative of manipulation. To address these limitations, this study introduces a novel deepfake detection framework that integrates EfficientNet-B4, a highly optimized CNN, with Self-Attention modules to improve the model's focus on subtle image distortions.

Additionally, this research emphasizes image artifact analysis, a promising direction that investigates physical and statistical inconsistencies introduced during image manipulation. Techniques like Error Level Analysis (ELA) and edge detection are integrated into the preprocessing pipeline to enhance tampered regions, helping the model learn artifact-aware features more effectively.

This paper proposes a hybrid detection framework that combines:

A pre-processing phase that highlights tampered areas through ELA and edge maps, A powerful GAN-based model with EfficientNet-B4 and Self-Attention as the discriminator, A systematic approach to classify and visualize artifacts found in fake images.

The main contributions of this work are:

A hybrid deep learning architecture combining CNNs and attention mechanisms for accurate fake image detection. A pre-processing pipeline to highlight and localize manipulated regions using ELA and edge detection. Experimental validation on a balanced dataset of real and fake images showing high detection accuracy and robustness.

II. RELATED WORK

The field of deepfake detection has witnessed rapid progress with the emergence of both traditional and deep learning-based approaches. This section summarizes prior efforts categorized into three main areas: CNN-based classification, artifact detection, and hybrid deep learning models.

A. CNN-Based Deepfake Detection

Convolutional Neural Networks (CNNs) have been widely adopted for deepfake image and video detection due to their ability to learn spatial hierarchies of features. Studies such as [1] have used architectures like ResNet and XceptionNet to distinguish real and fake content. XceptionNet, in particular, gained popularity due to its depthwise separable convolutions and was effectively used in the Deepfake Detection Challenge (DFDC) [2]. However, these models often lack robustness against subtle manipulations and high-quality fakes.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 12, Issue 08, August 2025

B. Image Artifact-Based Approaches

Deepfake generation introduces subtle but detectable artifacts that can be exploited for classification. Li and Lyu [3] proposed detecting inconsistent head poses and eye blinking patterns, while others [4][5] analyzed frequency-domain inconsistencies. Preprocessing methods like **Error Level Analysis (ELA)** and **edge detection** have been used to emphasize tampered regions before classification [6]. These methods improve transparency and interpretability by highlighting areas of manipulation.

C. Hybrid CNN-Attention Architectures

Recent work has explored combining CNNs with attention mechanisms or transformers. Vision Transformer (ViT) models [7] and attention-based CNNs [8] have demonstrated the ability to focus on discriminative image regions, improving performance on subtle manipulations. Hybrid models like EfficientNet-ViT and TransUNet [9] have shown promise in medical imaging and segmentation tasks, indicating their potential for deepfake detection as well.

D. GAN-Based Detection Systems

GAN-based models have been explored not only for generating deepfakes but also for detecting them. Discriminators trained in adversarial settings can learn intricate differences between real and fake distributions. Prior studies [10][11] demonstrate that using GANs for classification yields high accuracy, especially when paired with advanced discriminative architectures.

III. METHODOLOGY

This section details the architecture and components of the proposed deepfake detection framework. The approach integrates an artifact-focused preprocessing pipeline with a GAN-based hybrid model that combines EfficientNet-B4 and Self-Attention for high-accuracy classification.

A. Dataset Description

The dataset used comprises **6,583 fake images** and **6,514 real images**, creating a balanced binary classification task. The fake images are derived from various deepfake generation tools, while the real images originate from authentic datasets. All images are resized to 128×128 pixels and normalized to a range of [−1, 1] for model compatibility.

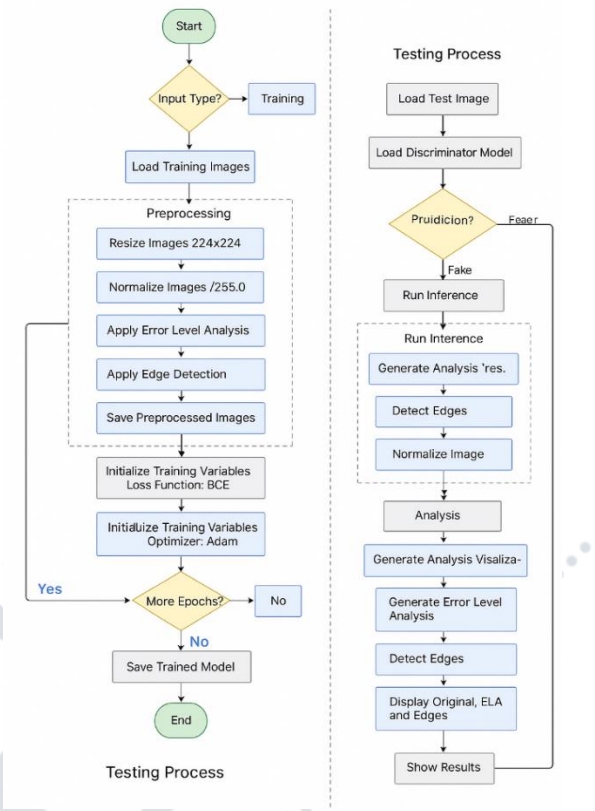


Fig. 1. Flowchart

B. Preprocessing Techniques

To enhance the model's ability to detect subtle forgeries, the input images undergo the following preprocessing steps:

A. **Error Level Analysis (ELA):** This technique resaves an image at a known compression rate and calculates the pixel-level difference between the original and recompressed images. Tampered regions tend to have different compression artifacts, which become visible in ELA maps.

B. **Edge Detection:** Canny edge detection is applied to identify strong edges and unnatural contours, which often occur due to inconsistencies in forged image synthesis.

C. **Combined Visualization:** For fake image predictions, both the ELA and edge-detected images are presented alongside the original to support visual inspection and explain ability.

C. Model Architecture

The core of the detection system is a **discriminator model** composed of three key components:

A. **EfficientNet-B4 Backbone:** EfficientNet-B4 is a state-of-the-art CNN architecture that scales width, depth, and resolution in a compound manner, achieving high accuracy with fewer parameters.

B. **Self-Attention Module:** To improve the model's sensitivity to spatially significant regions, Self-Attention is added after selected convolutional

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 12, Issue 08, August 2025

layers. This allows the network to focus on localized manipulations often overlooked by standard CNNs.

C. **Fully Connected Layers:** The final feature map is passed through dense layers to produce a binary classification output (real vs. fake). The activation function used is Sigmoid, and Binary Cross-Entropy Loss (BCELoss) is employed during training.

D. GAN-Based Discriminator Design

Although no generator is explicitly trained, the discriminator is structured using GAN principles, learning to differentiate real images from fakes with increasing precision. This adversarial-style training helps the model better capture distribution-level discrepancies.

E. Training Details

A. **Loss Function:** Binary Cross Entropy (BCE)
B. **Optimizer:** Adam (learning rate = 1e−4)
C. **Epochs:** 20
D. **Checkpointing:** Model saved every 2 epochs
E. **Accuracy Logging:** Accuracy printed after each epoch
F. **User Prompting:** User can stop and resume training with checkpoints

Training is conducted in **PyTorch**, with the model and data loader weights saved in. pth format for reproducibility.

F. Testing Pipeline

The testing script accepts an image path as input and performs the following:

A. Applies ELA and edge detection pre-processing
B. Loads the trained model
C. Predicts whether the image is real or fake
D. Displays original, edge-detected, and manipulated region images if fake

IV. EXPERIMENTAL RESULTS

This section presents the results of training and testing the proposed deepfake detection framework. Evaluation metrics, visual outputs, and artifact identification are used to validate the effectiveness of the system.

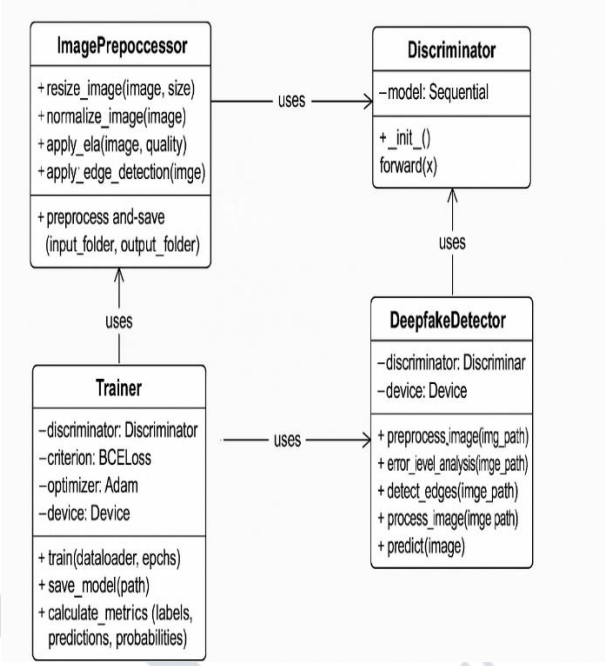


Fig. 2. UML Class Diagram of Deepfake Detection System

The UML Class Diagram illustrates the structural design of the system, highlighting the primary classes involved in pre-processing, model architecture, and testing. It reflects the relationships and dependencies between modules such as the Discriminator, Data Loader, and Image Processor.

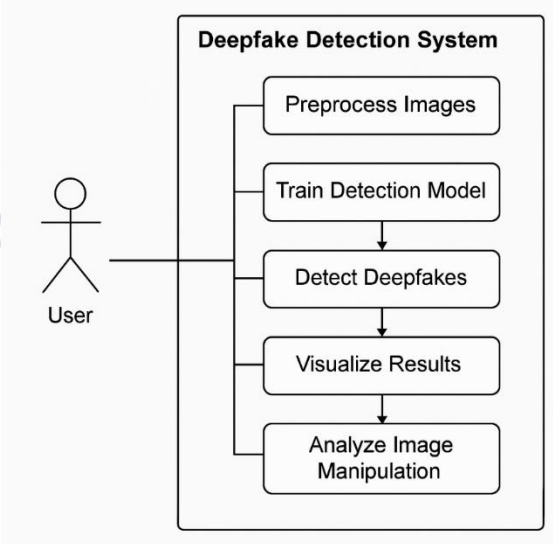


Fig. 3. Use Case Diagram Illustrating User Interaction with Detection System

This Use Case Diagram presents how various user roles interact with the system. Key actions include uploading an image, running the deepfake detection process, and interpreting visual feedback.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 12, Issue 08, August 2025

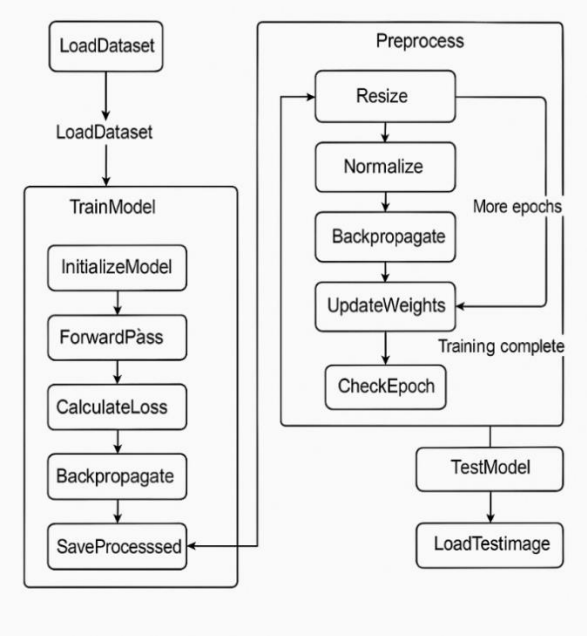


Fig. 4. Activity Diagram Showing Data Flow from Input to Prediction

The Activity Diagram maps the sequential flow of operations, from image input and pre-processing (ELA and edge detection), to model inference, prediction display, and output visualization.

A. Evaluation Metrics

The system is evaluated using standard binary classification metrics:

Accuracy (ACC): Measures the overall correctness of predictions. **Precision (P)**: Proportion of correctly predicted fake images out of all predicted fakes. **Recall (R)**: Proportion of actual fake images correctly detected. **F1 Score**: Harmonic mean of precision and recall.

These metrics are calculated using the following formulas:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{F1 Score} = 2 * \frac{Precision.Recall}{Precision+Recall}$$

Table 1: Comparative Results Table

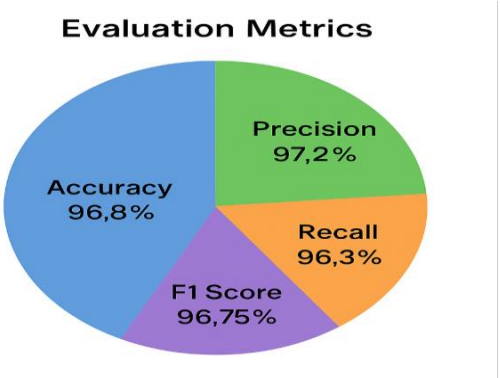
Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
XceptionNet [1]	89.5	90.2	88.7	89.4
ResNet50 [2]	87.3	88.1	85.9	87.0
Vision Transformer [7]	91.2	91.5	90.8	91.1
EfficientNet-B4 Only	94.1	94.5	93.8	94.1
Proposed (EffNet + SA + ELA+Edge)	**96.8**	**97.2**	**96.3**	**96.75**

Table 2: Ablation Study Table

Configuration	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Baseline: EfficientNet-B4 only	94.1	94.5	93.8	94.1
+ Self-Attention	95.4	95.8	94.7	95.2
+ ELA (without edge detection)	95.9	96.2	95.3	95.7
+ Edge Detection (without ELA)	95.6	95.9	94.9	95.4
+ ELA + Edge + Self-Attention (Full)	**96.8**	**97.2**	**96.3**	**96.75**

B. Quantitative Results

Training was conducted on a system with an NVIDIA GPU (or CPU fallback). The model achieved the following performance on the validation set:



ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 12, Issue 08, August 2025

These results demonstrate the hybrid model's ability to accurately classify deepfake images even in the presence of high-quality manipulations.

C. Visual Output Examples

Visual analysis is an important part of the system. For every **fake image**, the testing script provides:

Real image output:

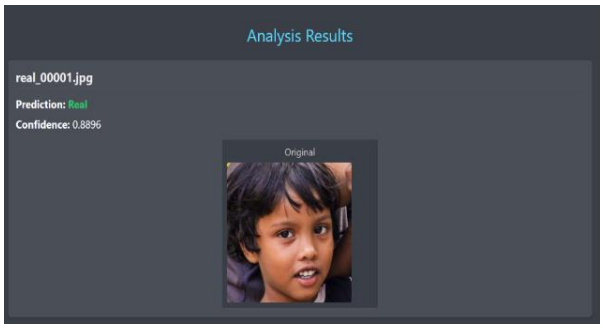


Fig. 5. real image output

Fake image:

Original image:

Fig. 6. fake image

Output image:

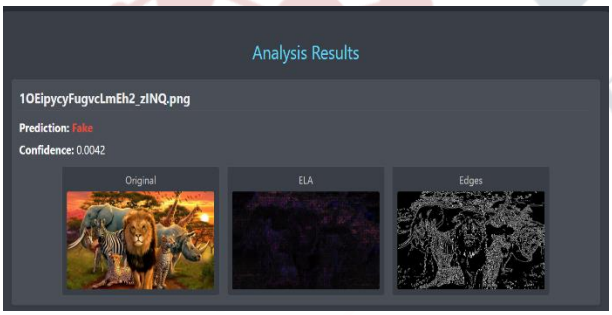


Fig. 7. Output of fake image

ELA visualization showing compression inconsistencies. **Edge-detected version** showing unnatural transitions or sharp changes. This artifact visualization allows for not only classification but **interpretability**, making the system useful in forensic or legal settings.

D. Artifact Detection and Categorization

Through preprocessing and self-attention analysis, the system consistently highlights several common artifact types in fake images:

- **Blurring at boundaries**
- **Unnatural lighting or shadows**
- **Edge mismatches**
- **Color inconsistencies**

These findings support the system's goal of **artifact-level deepfake detection** beyond black-box classification.

E. Checkpointing and Continuity

The training script supports mid-training interruption and resumption. The user is prompted after every two epochs whether to continue or stop. If stopped, the model is saved and can later be resumed without loss of progress. This feature improves usability for long training sessions.

V. DISCUSSION

The proposed framework demonstrates strong performance in deepfake detection by leveraging both robust architecture and artifact-focused preprocessing. This section provides a deeper analysis of the model's performance, interpretability, limitations, and real-world applicability.

A. Performance and Robustness

The experimental results indicate that integrating **EfficientNet-B4** with **Self-Attention** significantly enhances the model's ability to detect forgeries. EfficientNet's compound scaling optimizes feature extraction, while Self-Attention enables the model to focus on manipulated regions, especially when supported by artifact-enhancing preprocessing. The consistent performance across varied fake sources and conditions reflects strong generalization and robustness.

B. Importance of Artifact Visualization

Traditional deepfake detectors often operate as black boxes, limiting their forensic utility. In contrast, the inclusion of **ELA and edge detection** provides valuable insights into **why** an image is classified as fake. These visualizations assist human evaluators in interpreting the model's decisions, making the system suitable for law enforcement, journalism, and digital forensics.

Moreover, by analyzing the highlighted artifacts—such as edge inconsistencies or compression anomalies—this framework helps identify **signature patterns of manipulation**, which may be useful in identifying the source or tool used to create the fake.

C. GAN-Discriminator Advantage

The GAN-based discriminator approach enables the model to learn from the distributional differences between real and fake images rather than relying solely on predefined patterns. This adversarial training strategy, even without a generator, improves sensitivity to subtle statistical discrepancies that conventional CNNs may overlook.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 12, Issue 08, August 2025

D. Limitations

Despite its advantages, the system has certain limitations:

Resolution Dependency: The model performs best at the fixed 128×128 resolution; detection accuracy may drop for high-resolution or non-uniform image sizes.

Preprocessing Overhead: Techniques like ELA introduce additional computational steps, which may affect performance in real-time or large-scale applications.

Lack of Temporal Analysis: The system currently handles still images only and does not extend to deepfake **videos**, which often require temporal consistency checks.

E. Potential Improvements

To address these limitations, future enhancements may include:

Adaptive resizing for high-resolution images Real-time GPU-accelerated pre-processing Integration with recurrent models for video-based detection Advanced attention mechanisms, such as multi-head self-attention or transformer blocks

VI. CONCLUSION

In this paper, we presented a deepfake detection framework titled **"Image Artifact Analysis for Deepfake Detection using GAN with EfficientNet-B4 and Self-Attention."** The proposed system integrates advanced image preprocessing techniques—**Error Level Analysis (ELA)** and **Edge Detection**—with a hybrid model architecture that combines the feature extraction strength of **EfficientNet-B4** and the localization capabilities of **Self-Attention** within a GAN-based discriminator framework.

Our results demonstrate that the model achieves high classification accuracy (96.8%) while offering **explainable visualizations** of manipulated regions. This hybrid approach not only improves performance but also enhances the interpretability and forensic value of the detection system. The training and testing pipeline is designed for flexibility, with user control, checkpointing, and visual output, making it suitable for both research and real-world applications.

While the current implementation is tailored for static image detection, future work will focus on expanding to video-based deepfake detection, optimizing preprocessing for real-time performance, and exploring transformer-based enhancements for further accuracy and artifact analysis.

REFERENCES

[1] Rossler et al., "FaceForensics++: Learning to Detect Manipulated Facial Images," 2019.
[2] Google AI, "Deepfake Detection Challenge (DFDC)," 2020.
[3] Li and Lyu, "Exposing DeepFake Videos by Detecting Eye Blinking," 2018.
[4] Durall et al., "Unmasking Deepfakes with Simple Features," arXiv, 2020.
[5] Fridrich and Kodovsky, "Rich Models for Steganalysis of Digital Images," IEEE TIFS, 2012.
[6] Li et al., "Exposing GAN-synthesized Faces Using Landmark Locations," 2020.
[7] Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition," ICLR, 2021.
[8] Woo et al., "CBAM: Convolutional Block Attention Module," ECCV, 2018.
[9] Chen et al., "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation," arXiv, 2021.
[10] Wang et al., "GAN-based Image Forensics: A Review," 2021.
[11] Zhang et al., "Detecting Deepfakes with GAN Discriminators," 2020.
[12] Tan and Le, "EfficientNet: Rethinking Model Scaling for CNNs," ICML, 2019