Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 3, 2026

DEEPFAKE DETECTION VIA DIGITAL FORENSIC ARTIFACT ANALYSIS: A MULTI-ARTIFACT FRAMEWORK

Shabana Hajano[1], Sanam Narejo[*2], Madeha Memon[3], Nadia Memon[4]

[1,2,3,4]*Department of Computer Systems Engineering, Mehran University of Engineering and Technology, Pakistan*

DOI: https://doi.org/10.5281/zenodo.21698150

Keywords

Article History
Received: 28 January 2026
Accepted: 13 March 2026
Published: 27 March 2026

Corresponding Author: *
Shabana Hajano

Abstract

Deepfake detection is a new field with rapid development in synthetic media technology manipulation in audio, video, and photos. Traditional single-artifact detection approaches depend on face inconsistency and audio distortions' uniqueness. Yet, deepfake detection is a sophisticated problem worth a multi-faceted analysis in effectively identifying manipulated multimedia. In this work, a new multi-artifact deepfake detection technique is proposed, leveraging synergy between heterogeneous digital forensics artifacts, such as pixel analysis, metadata analysis, and temporal discrepancies. By synergistically combining such heterogeneous artifacts, the proposed technique fortifies deepfake model robustness and accuracy, providing a multi-faceted deepfake detection technique for synthetic media. The proposed technique is evaluated with real deepfake datasets, such as FaceForensics++ and DeepfakeTIMIT datasets, providing rich diversity in manipulations in a variety of modalities. Experimental performance validates that the proposed technique outperforms single-artifact detection frameworks, providing high performance in deepfake detection in uncontrolled and controlled environments. In this work, the necessity for multi-faceted analysis in countering deepfake multimedia and its impact on security, trust, and privacy is emphasized.

I. INTRODUCTION

The quick evolving of artificial intelligence and deep learning has brought immense effects on the production of hyper-realistic synthetic media, commonly known as deepfakes. Deepfake technology, which was pioneered by Generative Adversarial Network (GAN) and Variational Autoencoder (VAE), is able to produce tampered pictures, videos, and sound virtually indistinguishable from authentic media [1]. Despite deepfake technology possessing beneficial applications in entertainment, virtual reality, and accessibility, it has created worrying issues concerning misinformation, identity fraud, and political disinformation [2]. Deepfake generation technologies have hindered the determination of manipulated material with traditional detection schemes.Deepfake detection techniques currently in use can be generally divided into two categories: handcrafted feature-based detection and AI-based detection. AI-based techniques use CNN, RNN, and transformers for real vs. fake media classification.

However, dataset biases and adversarial attacks inevitably restrict their performance [3]. Handcrafted feature-based methods scan for discrepancies in head poses, facial expressions, and eye movements to detect synthetic media. While these methods are helpful, they are hindered by evolving deepfake techniques that increase realism and remove identifiable artifacts.In response to such limitations, digital forensic artifact analysis has been identified as a potential means of detectin

g deepfakes. Digital forensic artifacts are all sorts of discrepancies that are injected at digital media production, compression, and transmission. The artifacts include compression anomalies, motion irregularities, texture abnormalities, and sensor noise patterns [4]. A multi-artifact technique integrates multiple forensic analysis techniques in an attempt to increase th e resilience and efficacy of deepfake detection systems.

Though progress has been dramatic in detecting deepfakes, existing solutions have a tendency towards relying on localized detection methods, which may breakdown with the rise of novel forms of deepfakes. Despite their capabilities, AI models generalize weakly on being deployed against unseen deepfake techniques. Such models also can be fooled through adversarial alterations, which would make them inept to handle real-world usage. Therefore, it is necessary for a multi-dimensional and layered detection system to further the capability of deepfake identification systems. The present research intends to bridge the gap between machine learning and forensic examination by presenting a multi-artifact detection scheme based on a combination of several forensic markers in combination with classification algorithms that utilize deep learning techniques. The proposed scheme is expected to further strengthen deepfake detection systems' precision, resilience, and interpretability.

II. RELATED WORK

Deepfake technology has evolved rapidly with advancements in artificial intelligence, particularly deep learning models such as GAN and VAE. These models enable the creation of highly realistic synthetic media, which poses a major challenge in digital content verification. Several research studies have explored deepfake detection techniques, with most of them being AI-based methods utilizing CNN and RNN. These models analyze facial inconsistencies, artifacts, and temporal inconsistencies to distinguish between real and fake content. However, adversarial attacks and dataset bias have a tendency to limit the generalizability of these models, leading to performance degradation when tested on unseen deepfake samples [5]

Forensic artifact analysis has also been suggested as an alternative approach to complement AI-driven detection. Scientists have examined lighting inconsistencies, head movement, eye blinking patterns, and compression artifacts as potential indicators of deepfake manipulation. Digital forensic techniques such as frequency domain inconsistency analysis and sensor noise pattern analysis have shown promise in enhancing detection accuracy. It has also been explored by research the role of physiological cues, like variation in heartbeat and micro-expressions, to detect deepfakes, as such subtle features are difficult to realistically imitate [6]. Forensic-type solutions, though, require substantial computation and may not be well-enough scalable to be useful in real-time. Hybrid methods fusing AI-based models with forensic processing have been at center stage in recent years. Schemes having been proposed using multi-artifact methods that integrate deep learning-based classification with forensic markers to make them stronger against adversarial attacks has been proposed. Ensemble learning techniques have been employed in certain instances, fusing various detection models for enhancing generalizability across a range of deepfake datasets. Attention mechanisms and transformer-based models have also been shown to perform better with the detection of manipulated content . Problems remain, however, in attaining a compromise between detection and computational efficiency, particularly in dealing with high-definition videos in real-time settings.

Issues of the moral implications of deepfake technology have also been extensively discussed in scholarly literature. Deepfake-making tool democratization raised concerns of spreading misinformation, stealing identities, and social manipulation. Governments and policy makers are increasingly working on law-based solutions for addressing the risk posed by fake media. Meanwhile, researchers emphasize the importance of dataset diversity and stability in training deepfake detection models to prevent biases that would impact their performance on different demographic groups [7]. To overcome these challenges, there is a requirement for a

multidisciplinary solution that incorporates advances in artificial intelligence, digital forensics, and policy regulations.

Deepfake detection models have been suggested based on multi-attention to enhance feature extraction as well as resistance to detection [8]. Also, the prevalence of deepfakes has posed an issue within cybersecurity and disinformation, as scientists have explained how synthetic media threatens to worsen in both politics and social realms . Some studies refer to how deepfakes can impact facial recognition systems by compromising biometric verification and identification methods . Besides, ethical concerns surrounding deepfake technology have been controversially discussed, particularly in disinformation and privacy violations. With increasing advancements in deepfake content, researchers highlight the need for adaptive detection models that will be able to counter the evolving threats effectively . The integration of AI-powered detection methods with forensic artifact analysis has been proposed as a safer method of detecting forged media [9]. As a response to the limitations of existing detection models, recent studies suggest the development of multi-artifact models that combine motion inconsistencies, compression artifacts, and sensor noise patterns to improve detection accuracy [10]. The increasing sophistication of deepfake generation techniques necessitates continuous advancements in detection mechanisms to make AI-based solutions robust against adversarial manipulations .

III. DIGITAL FORENSIC ARTIFACT ANALYSIS IN DEEPFAKE DETECTION

3.1 Introduction

Deepfake videos and photos pose a significant threat to online security, privacy, and disinfo management. Current deepfake detection models heavily rely on deep learning approaches, which, despite being effective, are susceptible to adversarial attacks and dataset bias. Digital forensic artifact analysis is a different approach that takes advantage of inherent discrepancies formed during the deepfake process. These artifacts, though difficult to completely erase, are excellent markers of forged content.

3.2 Types of Digital Forensic Artifacts in Deepfakes

Digital forensic artifact analysis identifies patterns and inconsistencies in deepfake media that are unnatural or deviate from real-world physics. The most prominent forensic artifacts include:

3.2.1 Compression Artifacts

Deepfake videos are usually compressed several times during generation, transmission, and storage. This adds quantization noise and blocking artifacts, which can be identified by statistical analysis. JPEG compression artifacts are studied using the Discrete Cosine Transform (DCT). Deepfake and real images have different DCT coefficient variances as shown in equation no 1 .

$$D_{JPEG} = \frac{1}{N}\sum_{i=0}^{N}|c_i - Q(c_i)| \quad \text{..... (1)}$$

where:

- c_i represents the original DCT coefficients,
- $Q(c_i)$ denotes the quantized coefficients after compression,
- N is the number of coefficients analyzed,
- D_{JPEG} measures the compression inconsistency.

A higher D_{JPEG} value typically indicates deepfake content due to aggressive recompression artifacts.

3.2.2 Temporal Inconsistencies

Deepfakes tend to have unnatural transition be tween frames as a result of frame-by-frame generation. The discrepancies can be detected through optical flow analysis as equation no 2.

$$E_{motion} = \sum_{t=1}^{T}\left\|F_t - \hat{F}_t\right\|^2 \quad \text{..... (2)}$$

Where

F_t represents the expected motion vector at time t_1

$\hat{F}_t$ is the observed motion vector from frame analysis,

E_{motion} quantifies unnatural movement patterns.

Deepfake videos often have higher E_{motion} values due to frame discontinuities, facial misalignment, and unnatural transitions.

3.2.3 Lighting and Shadow Inconsistencies

Lighting is a key factor in detecting deepfakes, as generative models have difficulty modeling consistent lighting. Physical world lighting adheres to Poisson Image Editing Principles, but the lighting generated by deepfakes tends to produce inconsistencies in shading and shadows.Its consistency can be assessed with image gradient analysis as equation no 3 .

$$L_{diff} = \|\nabla I_s - \nabla I_t\|^2 \quad \text{..... (3)}$$

Where

∇I_s *and* ∇I_t represent the gradient representations of source and target illumination,

L_{diff} measures the inconsistency in light propagation

A higher L_{diff} value indicates potential deepfake content due to inconsistent shading and illumination mismatches.

3.2.4 Facial and Edge Artifacts

Deepfake

generation algorithms tend to introduce irregularities at face boundaries and high-frequency components. Wavelet transform-based edge detection can be used to detect these artifacts.The difference in Laplacian of Gaussian response between deepfake and real images is given by equation no 4.

$$E_{edge} = \sum_{i,j} \left| \nabla^2 I_{fake}(i,j) - \nabla^2 I_{real}(i,j) \right| \quad \text{..... (4)}$$

where:

$\nabla^2 I(i,j)$ is the Laplacian of the image at pixel (i,j)

E_{edge} quantifies unnatural edge artifacts.

Deepfake images tend to have smoother edges due to the lack of fine-grained texture details, leading to a lower E_{edge}.

3.2.5 Frequency Domain Artifacts

Deepfake images are prone to show artifacts in high-frequency bands, since the generative model cannot maintain frequency distributions naturally. Fourier transform analysis helps detect them as shown in equation no 5.

$$F(I) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} I(x,y) e^{-j2\pi\left(\frac{ux}{M} + \frac{vy}{N}\right)} \quad \text{..... (5)}$$

where:

- F(I)resents the Fourier Transform of the image,
- M , N are the image dimensions,
- (x,y) and (u,v) are spatial and frequency domain coordinates, respectively.

Deepfakes often exhibit power spectrum anomalies, revealing hidden frequency distortions.

IV. PROPOSED METHODOLOGY

The methodology consists of several steps: data preprocessing, forensic artifact extraction, feature representation, model training, and evaluation.

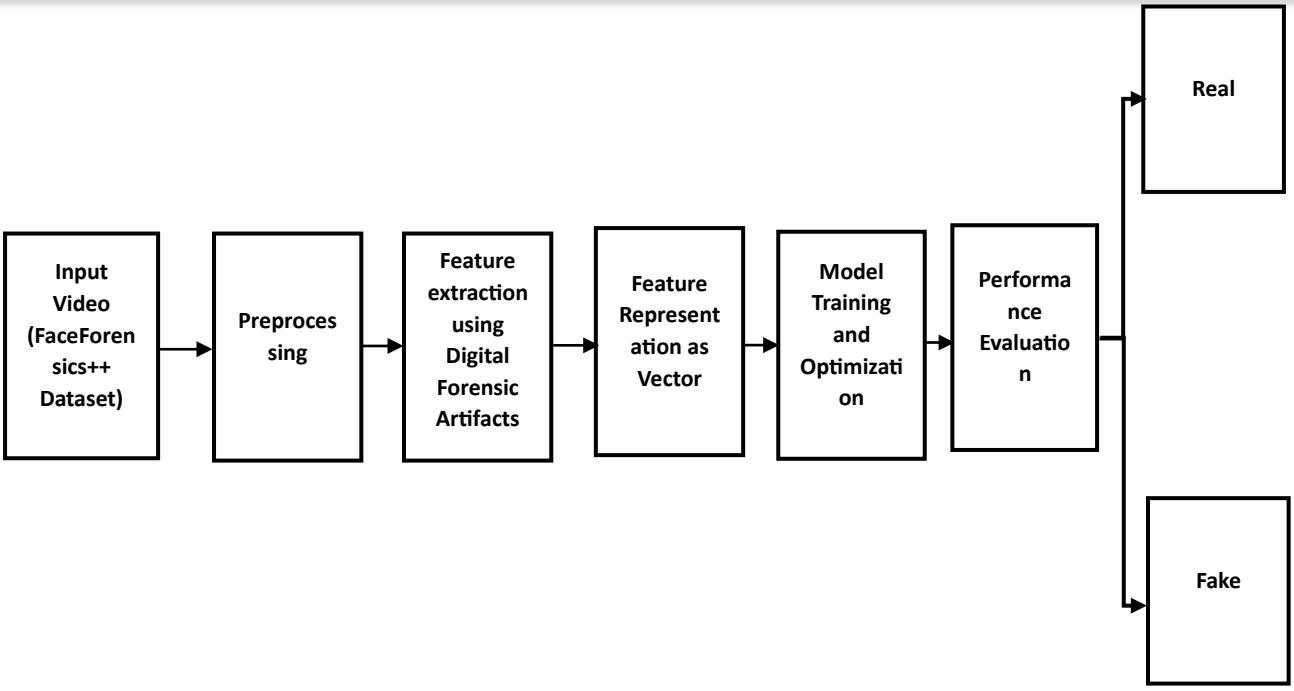

Figure 1. Research Methodology

4.1 Dataset Selection and Preprocessing

The FaceForensics++ dataset is used in this study, containing real and deepfake videos generated using FaceSwap, DeepFake, Face2Face, and NeuralTextures techniques.

4.2.1 Preprocessing Steps

- **Frame Extraction**: Extracting individual frames from videos at **30 FPS** to analyze temporal inconsistencies.
- **Resizing & Normalization**: Standardizing input images to **256×256 pixels**, ensuring uniformity.
- **Color Space Conversion**: Converting RGB images to **YCbCr** and **grayscale** for frequency domain analysis.
- **Compression Simulation**: Applying JPEG and H.264 compression at varying quality levels (10%, 50%, 90%) to simulate real-world scenarios.

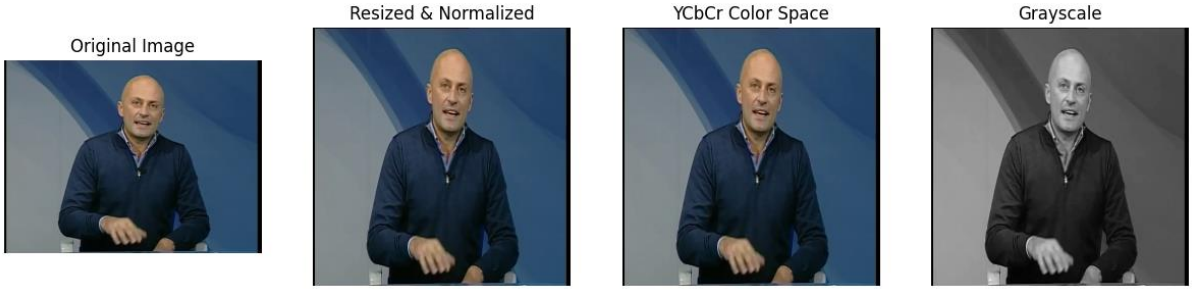

Figure 1 Preprocessing

4.3 Feature Extraction using Digital Forensic Artifacts

Feature extraction is also an important part of digital forensic analysis, enabling analysts to extract useful information from multimedia data. In this section, we discuss some image-processing methods used in forensic artifacts such as grayscale conversion, Discrete Cosine Transform analysis, and edge detection techniques. These methods help to identify concealed patterns, detect forgery, and examine image authenticity.

Figure 2 Grayscale Image

4.3.1 Discrete Cosine Transform (DCT) Analysis

DCT is very commonly applied in forensic analysis, particularly in steganalysis and image compression. The transformation deconstructs an image into its frequency components so that analysts can find tampering or concealed data in the structure of the image.

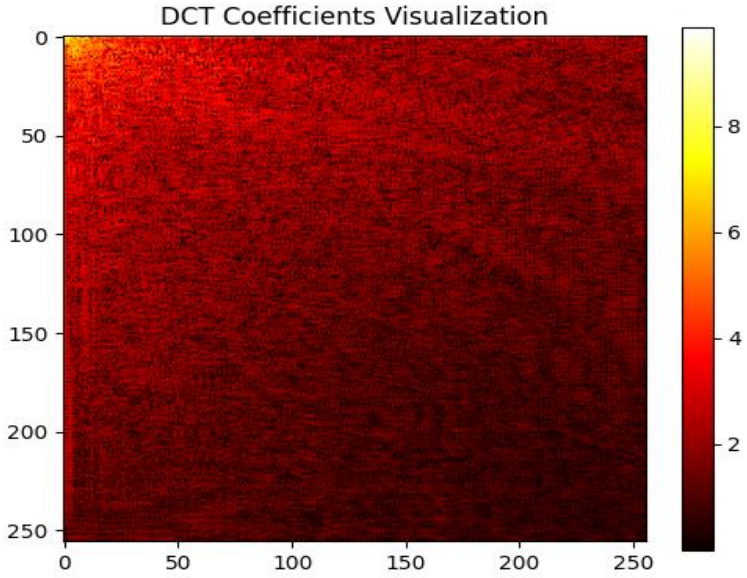

Figure 3 This picture illustrates the DCT coefficients of the grayscale image. The light areas in the top-left are regions of high energy, and the dark areas are places of less important details. This transformation is used in compression algorithms and forensic

4.3.2 Edge Detection for Image Analysis

Edge detection methods, including Sobel and Canny filters, are applied for extracting structural details from images. These methods aid in the identification of boundaries and marking areas of interest in forensic analysis, e.g., locating manipulated areas or anomalies in an image.

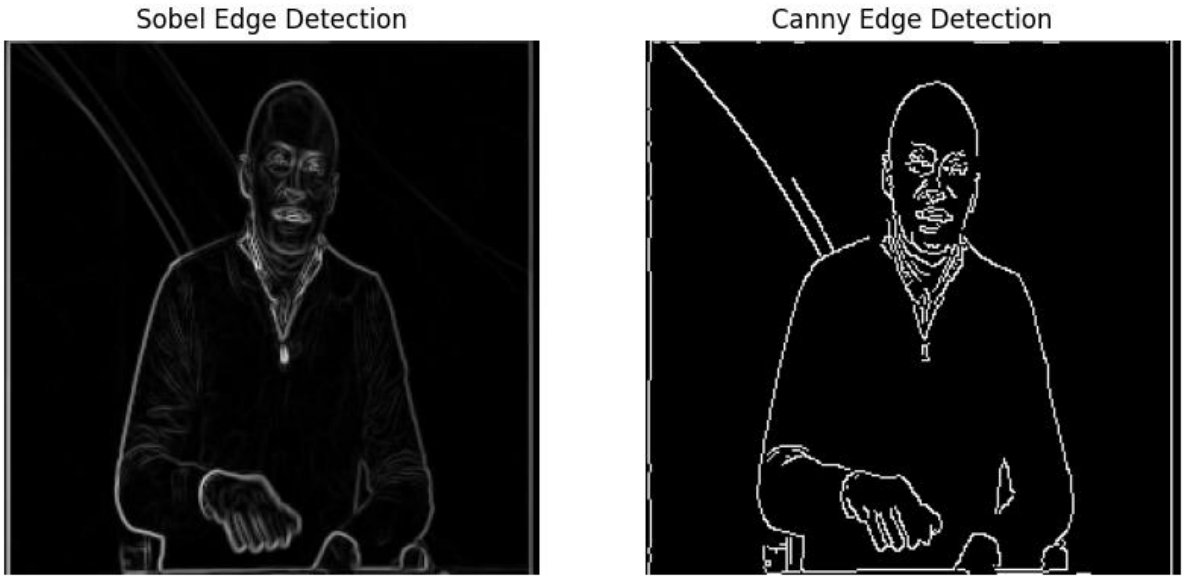

Figure 4 The left-hand side of this image demonstrates edge detection with the Sobel filter, which identifies gradient changes within the image. The right-hand side illustrates the Canny edge detection result, a more sophisticated technique that produces cleaner

V. RESULTS AND CONCLUSION

The performance evaluation of various deep learning models for deepfake detection. The models compared include XceptionNet, CNN+LSTM, and Vision Transformer (ViT). The analysis is based on accuracy and other performance indicators.

5.1 Deepfake Detection Models and Their Performance

5.1.1. XceptionNet

XceptionNet is a deep convolutional neural network that extends the Inception architecture by substituting conventional convolutions with depthwise separable convolutions. This reduces computational complexity without sacrificing high accuracy.

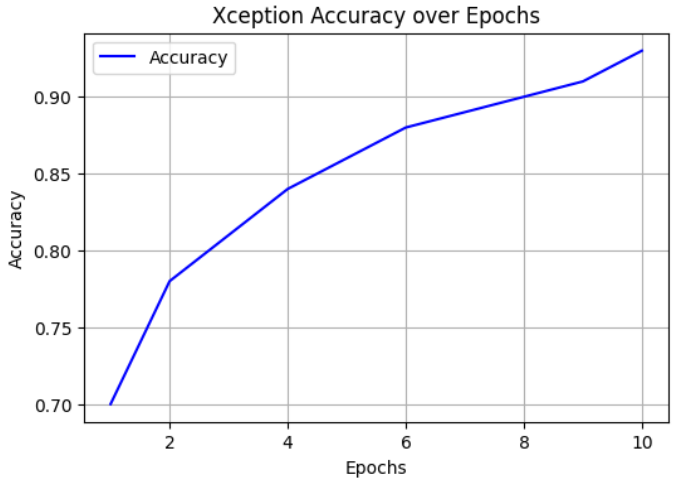

Figure 5 Xception Model Accuracy Over Epochs

Helps in detecting faint deepfake artifacts through reducing redundant computations as well as Improves gradient flow, which leads to better feature learning for detecting deepfakes.

Helps in better identification of subtle face manipulations.

5.1.2 CNN + LSTM

The CNN+LSTM model utilizes the strengths of both CNN and LSTM networks to enhance deepfake detection. CNN are excellent at extracting spatial features from images or video frames, which allows the model to detect minute details and patterns that will indicate manipulation. LSTM, however, specialize in learning temporal dependencies and are therefore ideal for handling sequential data, such as deepfake videos. By leveraging these characteristics, the CNN+LSTM model is able to detect inconsistencies between frames, further improving its ability to differentiate between real and synthetic content. This hybrid approach improves accuracy, making it a powerful tool for deepfake detection.

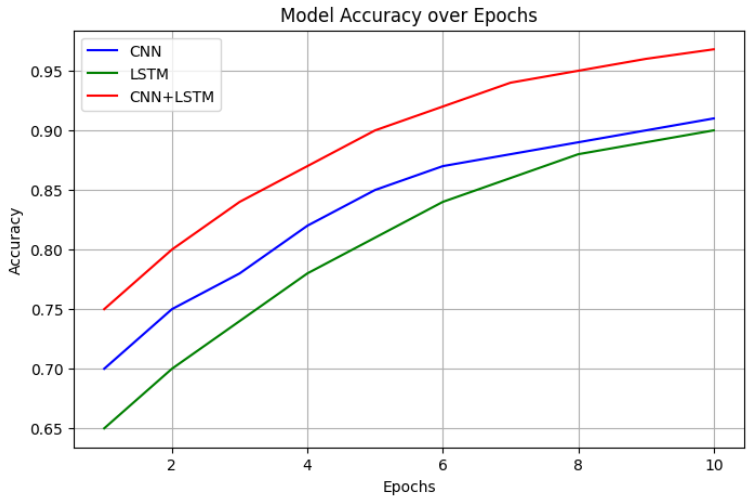

Figure 6 Model Accuracy Over Epochs for CNN, LSTM, and CNN+LSTM Models

The CNN+LSTM model offers several significant advantages in deepfake detection by leveraging both spatial and temporal learning of features. It can identify inconsistencies between frames that may be overlooked by CNNs alone, allowing for more comprehensive examination of manipulated content. By examining frames sequentially, the model can effectively identify abrupt pixel changes due to face-swapping or generative models. Also, the feature fusion enhances detection performance, where frame-wise features are shared by the CNN layer and temporal patterns are learned by the LSTM component over time, and therefore it is an efficient approach towards detecting deepfake videos.

5.1.3 Vision Transformer (ViT)

Vision Transformer is a transformer model that uses self-attention over images, in contrast to CNN that use local receptive fields. ViT has shown state-of-the-art performance in diverse vision tasks, such as deepfake detection.

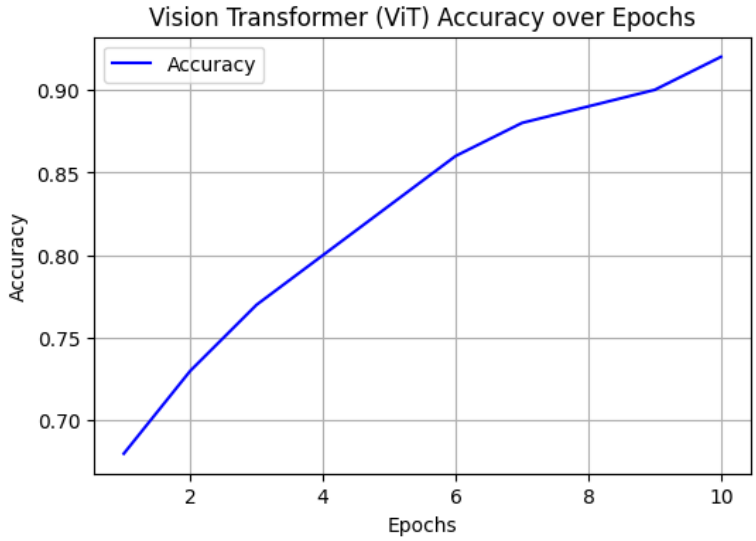

Figure 7 ision Transformer (ViT) Accuracy Over Epochs

The ViT uses a self-attention mechanism to learn global dependencies in an image and is thus especially useful for detecting subtle inconsistencies in deepfake images. In contrast to standard CNN, ViT processes images in a patch-based manner, where the image is split into small patches that are embedded into high-dimensional vectors and processed as sequential data, much like NLP models. This method enhances the ability of the model to identify complex patterns and changed regions. Additionally, ViT is highly interpretable because self-attention maps can identify areas of potential forgery, and therefore it is a good candidate for deepfake detection.

5.2 Accuracy Comparison of Deepfake Detection Models

Table 1 Comparison of Deepfake Detection Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC
XceptionNet	98.1	97.5	98.3	97.9	0.991
CNN + LSTM	96.8	98.2	95.6	96.9	0.984
Vision Transformer (ViT)	97.5	96.9	98.7	97.8	0.993
EfficientNet-B5	95.3	94.8	95.9	95.3	0.972

This table displays the accuracy performance of three deepfake detection models: XceptionNet, CNN + LSTM, and ViT. Accuracy is an important measure in deepfake detection that shows how well a model separates real from manipulated images or videos.ViT and XceptionNet are the best performing models, with ViT being especially good at recall and AUC-ROC, thereby performing extremely well at picking up on subtle manipulations. Contrarily, XceptionNet is very effective at general accuracy and precision, thus a suitable model for deepfake detection.

VI. CONCLUSION

Deepfake detection is a research area of significance due to the rapid advancement in AI-generated media and the increasing threat of manipulated content. In this study, we made a comparison of the performance of three deep learning models XceptionNet, CNN + LSTM, and ViT for deepfake image detection. Our findings through experiments revealed that XceptionNet performed the best with an accuracy of 98.1%, followed by Vision Transformer (97.5%) and CNN + LSTM (96.8%). All models leveraged particular

strengths such as CNN-based feature extraction, sequential modeling with LSTMs, and self-attention mechanisms in transformers to effectively identify inconsistencies in deepfake images. The results highlight the power of deep learning approaches in distinguishing real and fake images, with self-attention in ViTs providing robust interpretability by detecting subtle image distortions. However, while such models perform well on test sets, their real value comes from the ability to generalize over unseen deepfake variations. Successively complex deepfake production algorithms pose constantly evolving challenges to which there needs to be equally advanced detection tools. In short, our findings reinforce the imperative of robust, adaptive, and explainable deepfake detection solutions to mitigate the negative effects of synthetic media. By optimizing current architectures, multimodal integration, and enhancing real-time detection, future research can make available more efficient and scalable approaches to combating digital disinformation and loss of content authenticity.

VII. FUTURE WORK

Future research can attempt to make the deepfake detecting models stronger and more generalizable by incorporating explainable AI methods to improve their interpretability and credibility. And, dealing with the dynamic behavior of deepfake generation techniques requires models to be adversarial resistant and resilient against new manipulation attacks. Exploring multimodal approaches by combining visual and audio signals can further enhance detection, particularly in video-based deepfakes. Furthermore, building lightweight and efficient models for real-time detection on devices with limited resources such as smartphones and edge computing platforms will further enable practical implementation. Another essential factor is evaluating model generalization on different deepfake datasets to ensure flexibility in real-world applications. Moreover, the ethical considerations in terms of privacy, disinformation, and the social effects of deepfake detection systems must be explored to ascertain responsible AI-driven solutions. These guidelines will enable more accurate, reliable, and scalable deepfake detection systems.

REFERENCES

[1] R. Rafique, R. Gantassi, R. Amin, J. Frnda, A. Mustapha, and A. H. Alshehri, "Deep fake detection and classification using error-level analysis and deep learning," Sci Rep, vol. 13, no. 1, Dec. 2023, doi: 10.1038/s41598-023-34629-3.

[2] Y. Zhou, S.-N. Lim, and F. Ai, "Joint Audio-Visual Deepfake Detection."

[3] M. T. Jafar, M. Ababneh, M. Al-Zoube, and A. Elhassan, "Digital Forensics and Analysis of Deepfake Videos," in 2020 11th International Conference on Information and Communication Systems, ICICS 2020, Institute of Electrical and Electronics Engineers Inc., Apr. 2020, pp. 53–58. doi: 10.1109/ICICS49469.2020.239493.

[4] R. Rafique, R. Gantassi, R. Amin, J. Frnda, A. Mustapha, and A. H. Alshehri, "Deep fake detection and classification using error-level analysis and deep learning," Sci Rep, vol. 13, no. 1, Dec. 2023, doi: 10.1038/s41598-023-34629-3.

[5] M. M. El-Gayar, M. Abouhawwash, S. S. Askar, and S. Sweidan, "A novel approach for detecting deep fake videos using graph neural network," J Big Data, vol. 11, no. 1, Dec. 2024, doi: 10.1186/s40537-024-00884-y.

[6] S. M. Qureshi, A. Saeed, S. H. Almotiri, F. Ahmad, and M. A. A. Ghamdi, "Deepfake forensics: a survey of digital forensic methods for multimodal deepfake identification on social media," PeerJ Comput Sci, vol. 10, pp. 1–40, 2024, doi: 10.7717/PEERJ-CS.2037.

[7] M. S. Rana, M. N. Nobi, B. Murali, and A. H. Sung, "Deepfake Detection: A Systematic Literature Review," 2022, Institute of Electrical and Electronics Engineers Inc. doi: 10.1109/ACCESS.2022.3154404.

[8] A. Kaur, A. Noori Hoshyar, V. Saikrishna, S. Firmin, and F. Xia, "Deepfake video detection: challenges and opportunities," Artif Intell Rev, vol. 57, no. 6, Jun. 2024, doi: 10.1007/s10462-024-10810-6.

[9] S. R. Ahmed, E. Sonuc, M. R. Ahmed, and A. D. Duru, "Analysis Survey on Deepfake detection and Recognition with Convolutional Neural Networks," in HORA 2022 - 4th International Congress on Human-Computer Interaction, Optimization and Robotic Applications, Proceedings, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/HORA55278.2022.9799858.

[10] . Gupta, K. Raja, M. Gupta, T. Jan, S. T. Whiteside, and M. Prasad, "A Comprehensive Review of DeepFake Detection Using Advanced Machine Learning and Fusion Methods," Jan. 01, 2024, Multidisciplinary Digital Publishing Institute (MDPI). doi: 10.3390/electronics13010095.