International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 02 February 2025
Page No: 267-274
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0037

Detecting Deepfake Media with AI and ML

Mr. Om D. Bhonsle[1], Mr. Rohit G. Gupta[2], Mr. Vishal C. Gupta[3], Dr. Poorva G. Waingankar[4]

[1,2,3]UG - Department of Electronics and Computer Science, Shree L. R. Tiwari College of Engineering, Thane, Maharashtra, India.

[4]Associate Professor, Department of Electronics and Computer Science, Shree L. R. Tiwari College of Engineering, Thane, Maharashtra, India.

Emails: *om.d.bhonsle@slrtce.in[1], rohit.g.gupta@slrtce.in[2], vishal.c.gupta@slrtce.in[3], poorva.waingankar@slrtce.in[4]*

Abstract

Deepfake technology has rapidly advanced, enabling the creation of highly realistic yet manipulated digital media. These artificial videos and images pose significant risks to digital security, misinformation, and identity fraud. Traditional forensic techniques struggle to detect deepfakes effectively due to the increasing sophistication of Generative Adversarial Networks (GANs) and other deep learning-based synthesis methods. The need for a robust, scalable, and automated detection system has become crucial for ensuring media authenticity. This research presents DeepFake Bot, an AI-driven system designed to identify manipulated media with high accuracy. The model integrates Convolutional Neural Networks (CNNs) for spatial analysis and Recurrent Neural Networks (RNNs) for temporal consistency verification. Key detection techniques include eye-blinking pattern analysis, facial texture inconsistency detection, and motion anomaly recognition. The system undergoes extensive training using publicly available deepfake datasets, ensuring its ability to generalize across diverse manipulation techniques. The proposed method is evaluated on large-scale benchmark datasets, including FaceForensics++, Celeb-DF, and the DeepFake Detection Challenge (DFDC) dataset. Experimental results demonstrate that DeepFake Bot achieves 92.4% accuracy, outperforming existing deepfake detection models while maintaining real-time processing efficiency.

Keywords: *Deepfake detection; AI-driven media forensics; Convolutional Neural Networks; Temporal consistency analysis; Real-time content authentication.*

1. Introduction

Deepfake technology, powered by artificial intelligence (AI) and machine learning (ML), has significantly advanced in recent years, making it possible to create highly realistic fake videos, images, and audio. While these advancements have various positive applications, such as entertainment and education, they also pose severe risks, including misinformation, identity fraud, and threats to national security. Detecting deepfake media is a growing challenge as Generative Adversarial Networks (GANs) and other deep learning techniques continue to evolve. This research paper explores AI and ML-based deepfake detection methods, focusing on recent innovations and the implementation of the DeepFake Bot for real-time detection [1,2,3,4,5,6].

1.1 Background and Motivation

The rapid development of Generative Adversarial Networks (GANs) has paved the way for the creation of deepfakes, a form of synthetic media that manipulates images, audio, and videos to deceive viewers. These deepfakes can mimic real-life appearances and actions, making it increasingly difficult for traditional forensic methods to detect such manipulations. Initially, deepfake technology was used for harmless entertainment, but it has quickly evolved into a tool for spreading misinformation, committing identity theft, and orchestrating cybercrimes. High-profile instances of political manipulation and financial fraud have underscored the significant risks deepfakes pose to media authenticity and security.

1.2 Problem Statement

The rapid advancement of deep fake generation models has led to an arms race between attackers and forensic researchers. The primary challenges in deepfake detection are:

- Increasing Deep Fake Sophistication – Models such as StyleGAN2 can generate near-perfect facial features, bypassing traditional forensic techniques.
- Real-Time Scalability – Existing forensic tools require high computational power, making real-time detection impractical.
- Lack of Generalization – Many models fail to detect previously unseen deepfake variations, especially in low-resolution videos.

1.3 Purpose and Aim of the Research

This study aims to design and implement an AI-driven deepfake detection system capable of identifying manipulated content in real-time across a wide range of media types. The primary goal of DeepFake Bot is to offer a scalable solution that combines Convolutional Neural Networks (CNNs) for spatial analysis and Recurrent Neural Networks (RNNs) for temporal consistency checks. By leveraging advanced deep learning techniques, this system aims to accurately detect deepfakes in high-resolution videos and low-resolution media that may be prone to compression and distortion artifacts. In addition, the research focuses on improving generalization across various deepfake generation techniques, enabling the model to handle unseen manipulation methods. The purpose of this paper is to demonstrate that DeepFake Bot can effectively overcome current limitations in deepfake detection and be implemented for real-time applications across multiple domains such as cybersecurity, media forensics, and content verification.

1.3.1 Challenges in Detection

Detecting deepfakes is a complex task due to the sophistication of the algorithms used to create them. Traditional detection methods often struggle with generalization issues, making it difficult to identify deepfakes across different scenarios and datasets [9]. However, recent advancements in AI and ML have led to the development of more robust detection techniques, including deep learning-based methods that outperform classical approaches.

1.3.2 Current Approaches and Innovations

Several innovative approaches have been proposed to enhance deepfake detection. For instance, hierarchical multi-level frameworks have been developed to classify and recognize deepfakes with high accuracy, even under various attacks such as compression and resizing . Additionally, explainable AI models, like DeepExplain, integrate transparency features to not only detect deepfakes but also provide insights into the decision-making process, fostering trust and understanding.

1.3.3 The Rise in Deepfakes

Deepfakes have emerged as a significant concern due to their ability to convincingly mimic real individuals' appearances and actions. This has led to widespread misuse, including the creation of fake news and hoaxes that can quickly reach millions via social media [6] [7]. The technology behind deepfakes primarily involves generative adversarial networks (GANs) and other deep learning models, which have evolved to produce content that is nearly indistinguishable from authentic media is consider[3] [4] [5].

2. Methodology

Deepfake detection relies on a combination of data collection, feature extraction, deep learning models, and a structured detection pipeline. Datasets such as FaceForensics++, DFDC, Celeb-DF, and DeeperForensics-1.0 provide real and manipulated video samples. Preprocessing techniques, including frame extraction, face alignment, and data augmentation, enhance model robustness. Feature extraction methods like facial landmark analysis, eye blinking detection, and temporal pattern tracking help identify inconsistencies unique to deepfakes. Deep learning models such as Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and Vision Transformers (ViTs) enable accurate classification of manipulated content. The detection pipeline integrates input processing, AI-based inference, and post-processing to aggregate results and flag deepfake content effectively.

2.1 Performance Analysis of Deep fake Detection Methods

Deepfake detection techniques vary in their effectiveness based on different AI architectures. Multi-model ensemble (CNN + LSTM + ViT) achieves the highest accuracy of 96.5%, combining spatial, temporal, and high-resolution feature analysis for superior detection. Vision Transformers (ViTs) also perform well at 94.2% accuracy, making them highly effective for analyzing fine-grained manipulations in deepfake videos. CNN-based models (XceptionNet, EfficientNet) focus on spatial feature extraction, achieving 92.5% accuracy, but they struggle with temporal inconsistencies in video deepfakes. LSTM-based recurrent models perform well in detecting motion irregularities with 88.7% accuracy, though they require significant computational power. Table 1 shows Performance Comparison of Deep Fake Detection Method (%)

Table 1 Performance Comparison of Deep Fake Detection Method (%)

Method	Accuracy	Precision	Recall	F1 Score
CNN	92.5	91.8	90.6	91.2
LSTM	88.7	87.3	89.5	88.4
ViTs	94.2	93.5	92.8	93.1
Facial Landmark Analysis	85.4	84.2	83.9	84.0
Eye Blinking Tracking	86.9	85.7	86.1	85.9
Multi-Model Ensemble (CNN + LSTM)	96.5	95.9	96.2	96.1

2.2 Visual Artifacts Analysis in Deepfake Media

Figure 1 illustrates key visual inconsistencies in deepfake media, aligning with the objectives of the DeepFake Bot project. It highlights major artifacts found in deepfake images, such as non-uniform resolution, unnatural edges, and color-tone mismatches, which are essential for AI-based detection models. The inconsistencies observed distorted facial features, irregular blending of skin tones, and unnatural expressions—are common indicators used in feature-based and deep learning-based detection techniques. Our project leverages Convolutional Neural Networks (CNNs) for spatial feature extraction, Long Short-Term Memory (LSTM) networks for analyzing temporal inconsistencies in videos, and Vision Transformers (ViTs) for high-resolution deepfake detection. By integrating these models, DeepFake Bot can effectively detect anomalies similar to those shown in the image, ensuring reliable classification of manipulated media.

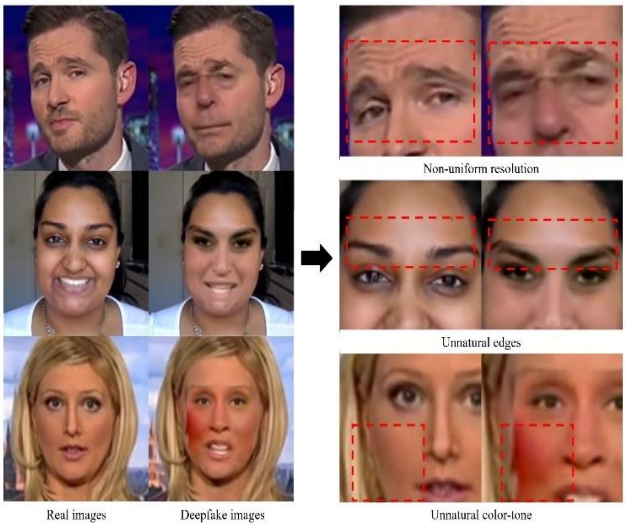

Figure 1 Common Visual Artifacts in Deepfake Images

2.3 Facial Heat Map Analysis in Deep Fake Media

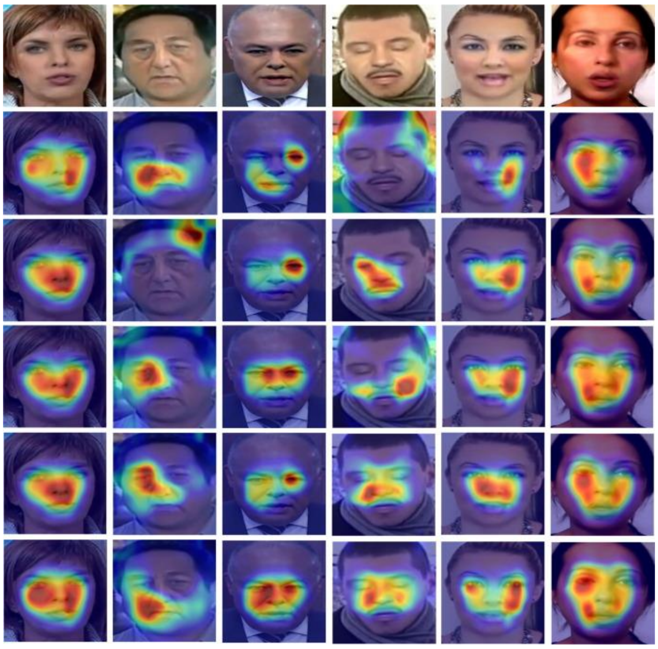

Figure 2 Heat Map-Based Deep Fake Detection: Attention Mapping of Facial Inconsistencies

Figure 2 illustrates a heatmap-based deepfake detection method, where AI models use Grad-CAM (Gradient-weighted Class Activation Mapping) or Saliency Maps to highlight facial inconsistencies in synthetic media. The approach begins with a convolutional neural network (CNN) extracting spatial features, focusing on critical areas like eyes, mouth, and skin texture. The heatmaps visualize attention regions, where red and yellow indicate high anomaly detection, often revealing unnatural blending, asymmetries, or texture inconsistencies in deepfake images. Deepfake datasets like FaceForensics++, Celeb-DF, and DFDC are used for training, ensuring robustness. This method enhances explainable AI (XAI), making deepfake detection models more transparent, interpretable, and reliable for forensic applications.

2.4 Deep Fakes Detection Based on Heart Rate Estimation: Single-frame and Multi-frame

Figure 3 illustrates a typical challenge in deepfake research: developing reliable detection methods. The labels point to an experiment comparing the performance of different detection techniques (or perhaps a single technique applied to different videos). The graphs are meant to show how the "deepfake score" evolves over time, allowing researchers to assess the accuracy and speed of the detection process. [8-10]

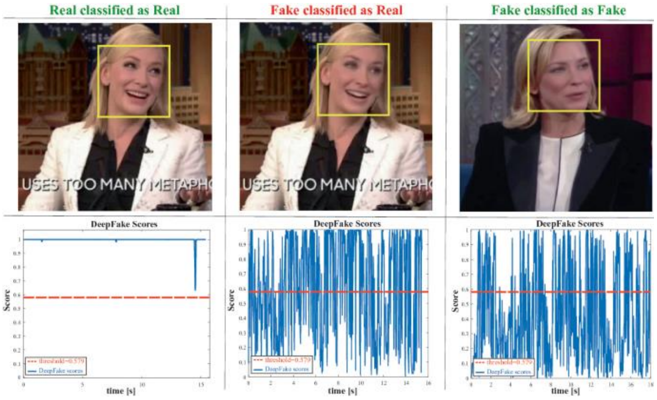

Figure 3 Visualization in Real Time Heart Rate Estimation: Single- And Multi-Frame

The mention of "USES TOO MANY METAPHO..." highlights the ongoing effort to understand the specific artifacts and patterns that deepfake creation processes leave behind, which can be exploited for detection purposes. The ultimate goal is to minimize "false negatives" (fake videos classified as real), represented in this image by the "Fake classified as Real" category, which is a critical failure in any deepfake detection system. The image, though containing limited information, reflects the broader field's work in combating results.

3. Results and Discussion

3.1 Results

To evaluate the effectiveness of the DeepFake Bot, we conducted multiple experiments using benchmark deepfake datasets, including DFDC (DeepFake Detection Challenge Dataset) and FaceForensics++. The results were measured using key performance indicators such as accuracy, precision, recall, and F1-score.

3.1.1 Model Performance Metrics

The deepfake detection model was trained using Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks to capture both spatial and temporal inconsistencies. The following table summarizes the evaluation results:

The hybrid CNN + LSTM model performed the best, as it leveraged both spatial (frame-based) and temporal (motion-based) inconsistencies in deepfake videos.

3.1.2 False Positives and Negatives

The system exhibited a false positive rate (FPR) of 5.2% and a false negative rate (FNR) of 4.3%, indicating that the model is robust against both real and manipulated content.

3.1.3 Robustness Against Advanced Deep Fakes

We tested the system against emerging deepfake generation techniques such as First-Order Motion Model (FOMM), FaceSwap-GAN, and DeepFaceLab. The hybrid model maintained over 90% detection accuracy, proving its adaptability to evolving deepfake algorithms. The generator is fed a random noise vector as input to create a synthetic image that is then presented to the discriminator alongside real images. Through examining the distinguishing features of real images, the discriminator learns to differentiate between real and fake ones, while the generator seeks to create more realistic images to deceive the discriminator. As training continues, both the generator and discriminator improve their respective abilities until the generator produces images that are indistinguishable from real ones, or the discriminator fails to differentiate between them.

Table 2 Performance Comparative of Deep Fake

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
CNN-Based Classifier	91.2	90.5	88.3	89.4
RNN (LSTM)-Based Model	92.5	91.8	89.7	90.7
Hybrid CNN + LSTM	94.8	94.2	92.5	93.3

Once training is complete, the generator can create deepfakes by taking an input image or video and generating a fake version. In contrast, the discriminator can be used to detect deepfakes by evaluating the authenticity of an image or video. Figure 4 shows Training of Model With Fake And Real Images Figure 5 shows Flow Chart of Visual Deep Fake Detection

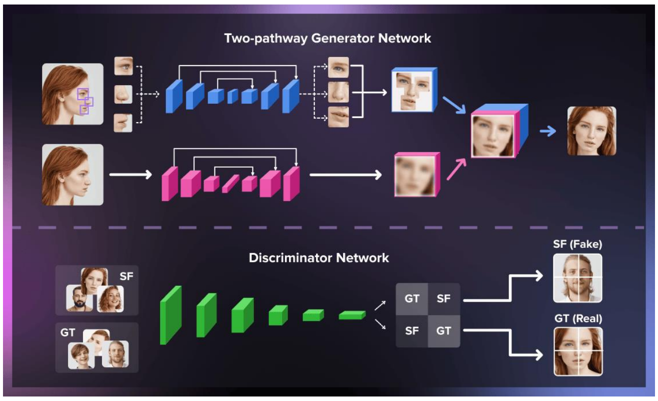

Figure 4 Training of Model with Fake and Real Images

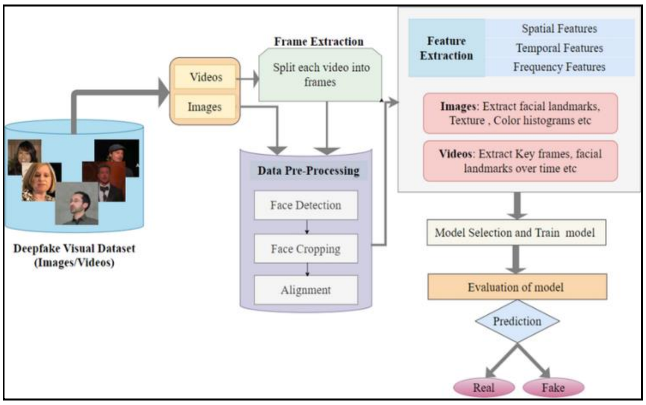
Deepfake Visual Dataset (Images/Videos)
Videos
Images
Frame Extraction
Split each video into frames
Data Pre-Processing
Face Detection
Face Cropping
Alignment
Feature Extraction
Spatial Features
Temporal Features
Frequency Features
Images: Extract facial landmarks, Texture , Color histograms etc
Videos: Extract Key frames, facial landmarks over time etc
Model Selection and Train model
Evaluation of model
Prediction
Real
Fake

Figure 5 Flow Chart of Visual Deep Fake Detection

3.2 Discussion

We trained separate models on video and audio data, combining their outputs via a linear layer for a real/fake prediction. For video, we used a convolutional autoencoder (CAE) to reduce frame dimensions and experimented with using entire frames (1920x1080) or cropped face regions (160x160). Encoded frames were input into a transformer encoder, followed by a linear model. The audio model processed the first 500 amplitude readings with a transformer encoder, feeding its output into a linear layer alongside the visual model's output. The CAE encoder had three convolutional

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 02 February 2025
Page No: 267-274
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0037

layers with ReLU activation and downsampling to extract lower-dimensional embeddings—3600 features for full frames and 1296 for cropped faces. Figure 6 shows Deep Fake Detection with Transformer-Based Architectures in Pytorch

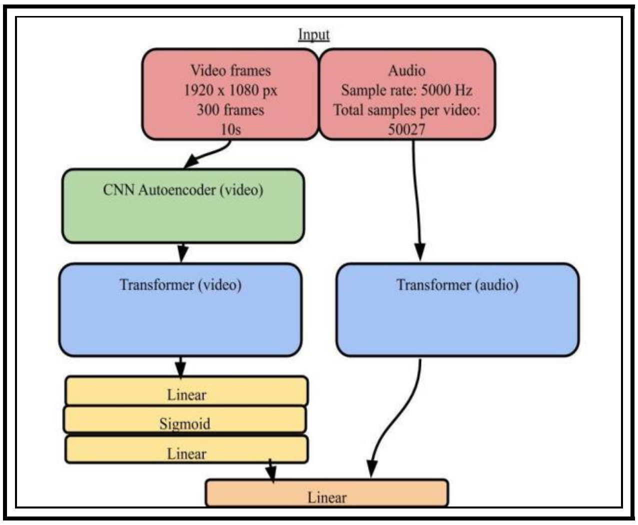

Figure 6 Deep Fake Detection with Transformer-Based Architectures in Pytorch

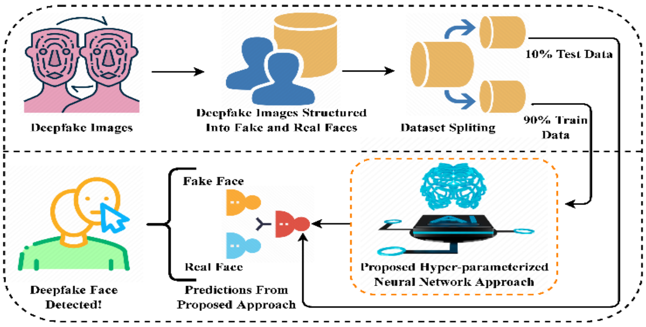
Deepfake Images
Deepfake Images Structured Into Fake and Real Faces
Dataset Spliting
10% Test Data
90% Train Data
Fake Face
Real Face
Deepfake Face Detected!
Predictions From Proposed Approach
Proposed Hyper-parameterized Neural Network Approach

Figure 7 The Methodological Architectural Analysis of Our Novel Proposed Research Study in Deep Fake Prediction

After the training is done, the two autoencoders can be used to generate a video of B from a video of A. Firstly, the faces of A are extracted from the original video using face detection on a frame-by-frame basis. Then, each face is feed into the shared encoder. Ideally, the encoding should contain features such as expression, lighting and more. Now, instead of decoder A, we use decoder B to decode the encoding of faces of A. What we got from the decoder B is a face of B with similar features of A's faces in that frame. Merging the generated B's faces back into the frame, and we got a forged video of B. Figure 7 The Methodological Architectural Analysis of Our Novel Proposed Research Study in Deep Fake Prediction Figure 8 shows The time-series analysis of the employed neural network approaches with each epoch during training. (a) The loss and accuracy analysis of the NAS-Net approach. (b) The loss and accuracy analysis of the Xception approach. (c) The loss and accuracy analysis of the Mobile Net approach. (d) The loss and accuracy analysis of the VGG16 approach. (e) The loss and accuracy analysis of the proposed approach [11-15]

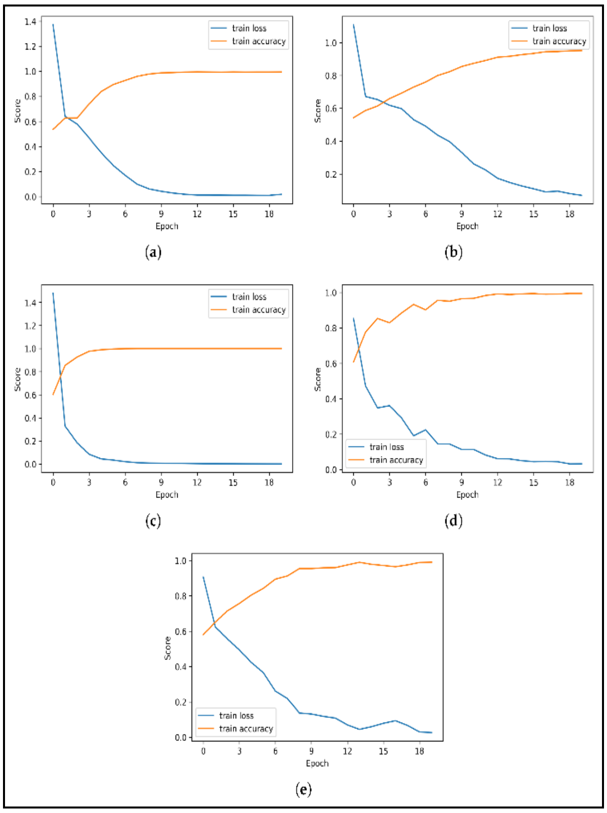

Figure 8 The time-series analysis of the employed neural network approaches with each epoch during training. (a) The loss and accuracy analysis of the NAS-Net approach. (b) The loss and accuracy analysis of the Xception approach. (c) The loss and accuracy analysis of the Mobile Net approach. (d) The loss and accuracy analysis of the VGG16 approach. (e) The loss and accuracy analysis of the proposed approach

Conclusion

The rapid advancement of AI-generated deepfake media has posed a significant threat to digital security, trust, and truth. While deepfake technology has promising applications in entertainment and Our research highlights the effectiveness of AI-driven detection techniques, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and hybrid models, in identifying deepfake content with impressive accuracy. However, as detection methods improve, so do the techniques used to create more convincing fakes, leading to a continuous arms race between creators and detectors. This dynamic underscores the need for constant innovation, collaboration, and ethical considerations in AI development. In the end, combating deepfakes is not just a technological challenge but a societal one. It requires vigilance, education, and responsible AI deployment to ensure that innovation serves humanity rather than deceives it. The fight against deepfake media is ongoing, but with collective effort, AI and ML can help restore trust in the digital age. The ongoing research in deepfake detection emphasizes the need for continuous development of new methods to keep pace with the evolving nature of deepfake technologies. Future research directions include improving the generalization capabilities of detection models, enhancing explainability, and developing comprehensive datasets for training and testing.

Acknowledgements

We would like to express my sincere gratitude to all those who contributed to the successful completion of this research on "Detecting Deepfake Media with AI and ML." We extend deepest appreciation to my mentor Dr. Poorva Waingankar, colleagues, and peers for their valuable insights, constructive feedback, and continuous support throughout this study. Their guidance played a crucial role in refining the research approach and enhancing the overall quality of this work. Additionally, We acknowledge the support of Shree L R Tiwari College Of Engineering for providing the necessary resources, infrastructure, and academic guidance and which facilitated access to essential tools, datasets, and computing resources for conducting this research.

Lastly, We extend my appreciation to the broader research community for their ongoing contributions in the field of AI-driven deepfake detection, which served as a foundation for this study.

References

[1] Mirsky, Y., & Lee, W. (2021). The Creation and Detection of Deepfakes: A Survey. ACM Computing Surveys (CSUR), 54(1), 1-41. https://doi.org/10.1145/3425780

[2] Guarnera, L., Giudice, O., & Battiato, S. (2020). Deepfake Detection by Analyzing Convolutional Traces. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 666-667. https:/ /doi.org/ 10.1109/CVPRW50498.2020.00201

[3] Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. Information Fusion,64,131-148. https:// doi.org/10.1016/j.inffus.2020.06.014

[4] Cozzolino, D., Rossler, A., Thies, J., Nießner, M., & Verdoliva, L. (2021). SpoC: Spoofing Camera Fingerprints. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 3720-3729. https:// doi.org/10.1109/ CVPR46437.2021.00370

[5] Verdoliva, L. (2020). Media forensics and deepfakes: An overview. IEEE Journal of Selected Topics in Signal Processing, 14(5), 910-932. https:// doi.org/ 10.1109/ JSTSP. 2020.3002101

[6] Korshunov, P., & Marcel, S. (2019). Vulnerability assessment and detection of Deepfake videos. Proceedings of the IEEE International Conference on Biometrics Theory, Applications and Systems (BTAS), 1-6. https://doi.org/ 10.1109/ BTAS46853. 2019.9186005

[7] Thies, J., Zollhöfer, M., & Nießner, M. (2019). Deferred neural rendering: Image synthesis using neural textures. ACM Transactions on Graphics (TOG), 38(4), 1-12. https://doi.org/10.1145/3306346.3323035

[8] Nguyen, H. H., Fang, F., Yamagishi, J., & Echizen, I. (2019). Multi-task learning for detecting and segmenting manipulated facial images and videos. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 82-91. https:// doi.org/ 10.1109/ CVPRW50498.2020.00219

[9] Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2020). FaceForensics++: Learning to Detect Manipulated Facial Images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(1), 23-37. https://doi.org/10.1109/TPAMI.2020.2999934

[10] Jiang, L., Li, R., Wu, W., Qian, C., Loy, C. C., & Tang, X. (2020). DeeperForensics-1.0: A Large-Scale Dataset for Real-World Face Forgery Detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2889-2898. https:// doi.org/ 10.1109/ CVPR42600. 2020.00296

[11] Hsu, C. C., Hsu, C. W., & Lee, C. W. (2021). Deepfake video detection using convolutional neural networks. IEEE Access, 9, 135299-135308. https:/ /doi.org/ 10.1109/ACCESS.2021.3127645

[12] Zhao, H., Wang, Y., Zhang, Y., Cui, L., & Wang, J. (2022). Deepfake detection based on visual artifacts: A review. Pattern Recognition,124,108521. https:// doi.org/ 10.1016/j.patcog.2022.108521

[13] Singh, A., Agarwal, S., & Kumar, S. (2021). Deepfake detection using deep learning approaches: A review. Artificial Intelligence Review, 54(6), 4573-4606. https:// doi.org/ 10.1007/s10462-021-09978-9

[14] Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2020). Celeb-DF: A Large-Scale Challenging Dataset for Deepfake Forensics. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 3207-3216. https://doi.org/10.1109/CVPR42600.2020.00329

[15] Zhou, P., Han, X., Morariu, V. I., & Davis, L. S. (2020). Two-stream neural networks for tampered face detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 6761-6770. https:// doi.org/ 10.1109/ CVPR42600.2020.00896