Detecting and Watermarking Fake Videos using AI Model

Prince Kumar* , Lav Raman Sinha, Subrojeet Das, Anuranjana

Amity School of Engineering and Technology, Amity University Uttar Pradesh, Noida, India
*sprincekumar316@gmail.com

* Corresponding author

doi: https://doi.org/10.21467/proceedings.7.6.35

Abstract

Detecting deepfake videos has become crucial for ensuring media authenticity in the digital age. This work presented an AI-driven system that utilized a custom Convolutional Neural Networks (CNNs) for spatial analysis to detect manipulated media. The system also integrated a novel watermarking approach to mark identified deepfakes, thereby enhancing traceability and accountability. To improve detection accuracy, the custom CNN model was trained on diverse datasets to generalize across various deepfake techniques, including FaceSwap. The proposed method addressed key challenges such as computational complexity, adversarial robustness, and real-time processing efficiency. Additionally, the watermarking solution embedded digital signatures within detected deepfake media, enabling their identification even after post-processing modifications. By combining deep learning-based detection with watermarking, the system provided a comprehensive approach to mitigate misinformation and preserve media integrity. This research contributes to the ongoing fight against digital deception and offers a scalable, effective solution for deepfake detection.

Keywords— Deepfake detection, AI models, CNN, ResNet-50 Watermarking.

I. INTRODUCTION

Deepfake technology refers to the use of artificial intelligence (AI) to create manipulated or synthetic media, typically videos, that convincingly mimic real people. This technology has rapidly advanced, leading to the creation of highly realistic videos where people appear to say or do things they never actually did. While deepfakes can be used for creative or entertainment purposes, they also pose significant risks, including spreading misinformation, identity theft, and damaging reputations [1]. Detecting deepfake videos has become an essential task in maintaining the integrity of digital media. AI algorithms play a critical role in this effort by analyzing subtle inconsistencies in video content that are difficult for the human eye to detect [2]. These algorithms often leverage machine learning techniques such as Convolutional Neural Networks and Generative Adversarial Networks to identify features like unnatural facial movements, visual artifacts, or anomalies in video frames [3]. By employing advanced AI methods, it is possible to develop robust systems capable of distinguishing between real and fake videos, helping to mitigate the harmful effects of deepfakes in various domains, including media, politics, and cybersecurity. By continuously evolving detection methods, AI algorithms help safeguard the authenticity of digital media, providing tools to counteract the growing threat posed by deepfakes across social media, news platforms, and online communication.

II. LITERATURE SURVEY

The detection of deepfake videos has garnered significant attention from the research community, with numerous studies focusing on developing effective AI-based techniques to address this growing challenge. This literature review outlines key contributions and methods used in the domain of deepfake detection, highlighting advancements and the evolving landscape of the field.

- Generative Adversarial Networks (GANs) and Deepfake Creation: Deepfake technology primarily relies on Generative Adversarial Networks. GANs consist of two neural networks: the generator, which creates fake data, and the discriminator, which attempts to differentiate between real and fake content. This adversarial training process leads to the production of highly realistic media. While GANs are powerful tools for content creation, they also pose significant challenges for detection, as the generated media becomes more difficult to distinguish from authentic material[3].
- CNN-based Detection Approaches: Convolutional Neural Networks have been extensively studied for image and video analysis tasks, making them a natural choice for detecting deepfake videos. Afchar et al. proposed the MesoNet architecture, specifically designed to detect deepfake artifacts in video frames. Their model focused on identifying mesoscopic features, such as pixel inconsistencies and blurred regions around the face, which are often overlooked by the human eye. CNNs have also been used to detect spatial anomalies in deepfake videos, analyzing each frame independently for visual cube like unnatural facial expressions or lighting inconsistencies[5].

Proceedings DOI: 10.21467/proceedings.7.6; Series: AIJR Proceedings; ISSN: 2582-3922; ISBN: 978-81-989164-9-5

- Temporal Analysis with RNNs and LSTMs: Since videos consist of sequences of images, some researchers have focused on temporal dynamics to enhance deepfake detection. Guera et al. combined
- CNNs with Long Short-Term Memory (LSTM) networks to capture both spatial and temporal inconsistencies [4]. Their model analyses not only individual frames but also the progression of frames over time, detecting subtle differences in motion, facial movements, and eye blinking patterns that are commonly mishandled by deepfake algorithms.

TABLE 1: Comparison of AI models

Model	Type	Key Strengths	Key Weaknesses	Typical Use in Deepfake Detection	Performance Notes
ResNet-50 [6]	CNN (Convolutional)	Deep architecture with skip connections avoids vanishing gradient, strong spatial feature extraction.	High computational cost; limited temporal analysis without additional layers.	Extracts frame-level spatial features (e.g., facial inconsistencies).	Often achieves 95-97% accuracy when paired with other models; excels in image-based tasks.
GAN [7]	Generative Model	Generates realistic synthetic data; can be adapted for detection via discriminators.	Prone to mode collapse; training instability; not inherently a detector.	Used to train detectors by generating deepfakes for comparison.	Detection accuracy varies (e.g., 88% in some studies) ; strong in feature learning.
CNN [5]	Convolutional Neural Network.	Excellent at identifying spatial patterns (e.g., artifacts, distortions).	Lacks temporal sequence analysis; may miss subtle frame-to-frame inconsistencies.	Frame-by-frame analysis for visual artifacts.	High accuracy (up to 97%) on static frames; less effective alone for video sequences.
RNN [8]	Recurrent Neural Network.	Captures temporal dependencies across sequences, good for video analysis.	Struggles with long-term dependencies; prone to vanishing gradient problems.	Analyzes temporal inconsistencies in video sequences.	Accuracy around 85% standalone; improved when combined with CNNs.

III. PROPOSED MODEL

- Data Collection and Pre-processing: The design began with gathering a large dataset of both real and deepfake videos from sources like DeepFake_Faces[9], FaceForensics++, Celeb-DF, or the Deepfake Detection Challenge dataset. Pre-processing steps included frame extraction, resizing, and normalization to ensure uniformity. Data augmentation techniques, such as flipping and rotating frames, may also be applied to increase variability and improve model robustness.

- Dataset Utilized: In this study, the model utilized a dataset named Deepfake Faces, comprising approximately 96,000 images, each with a resolution of 224 × 224 pixels. The images are labeled as either 'REAL' or 'FAKE,' where the FAKE images are annotated with additional metadata regarding their source of origin, while the REAL images lack such source information.

- Feature Extraction: The system focused on identifying key features that distinguish deepfake videos from real ones. These included irregularities in facial movements, inconsistencies in lighting, unnatural eye blinking, and visual artifacts like blurriness or pixel-level distortions. AI model was designed to extract these features automatically from video frames.

Proceedings DOI: 10.21467/proceedings.7.6
ISBN: 978-81-989164-9-5

Series: AIJR Proceedings
ISSN: 2582-3922

- Detection and Classification: Once trained, the model was deployed to detect and classify videos as real or fake. Each video was analyzed frame-by-frame, with the AI algorithm identifying anomalies or unnatural elements. Based on the analysis, the system generated a probability score or binary classification indicating whether the video was genuine or a deepfake.
- Selection of AI Model: Convolutional Neural Networks were typically employed for image analysis, as they excel at detecting spatial features within individual frames. These models could detect inconsistencies over time, such as subtle changes in facial expressions that might signal a deepfake [4],[5].
- Evaluation and Testing: The performance of the detection system was evaluated using standard metrics like accuracy, precision, recall, F1-score, and the Area Under the Curve (AUC). The system was tested on unseen data to ensure it generalizes well to different types of deepfake videos. To further enhance the model's resilience, adversarial training or continuous updating may be integrated into the design, enabling the system to adapt to evolving deepfake techniques.

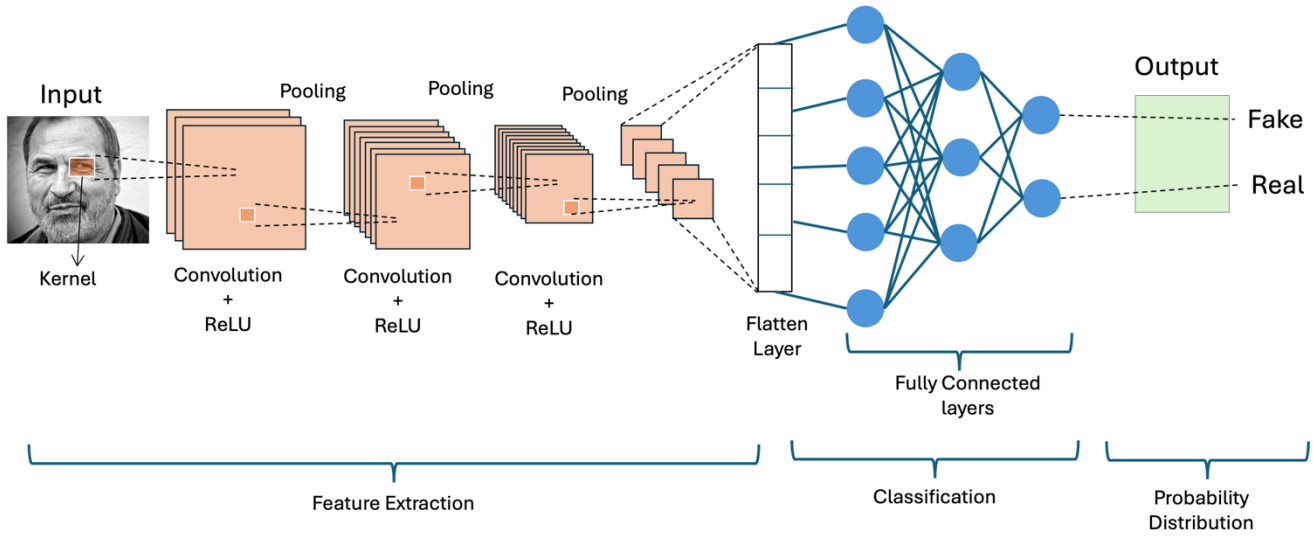

FIGURE 1: CONVOLUTIONAL NEURAL NETWORK DIAGRAM

A convolutional neural network (CNN) architecture for deepfake detection is depicted in the image. A kernel first pulls localized features from an input image. To add non-linearity, the network uses ReLU activation functions after several convolutional layers. Pooling layers preserve key properties while gradually reducing the spatial dimensions. After being flattened into a one-dimensional vector, the collected features are then run through fully connected layers for classification. In the end, a probability distribution is produced that categorizes the input as "Fake" or "Real." This architecture makes use of CNNs' capacity to recognize complex patterns in images and is frequently employed in deepfake detection tasks.

IV. METHODOLOGY

1. Data Collection

The first step in deepfake detection involved gathering relevant data, primarily videos containing both genuine and manipulated faces. A widely used dataset for this purpose is **DeepFake Faces**, which consisted of labeled real and fake deepfake videos. Instead of analyzing entire videos, frames were extracted, focusing on facial regions to detect inconsistencies. Each frame was categorized as real or fake, forming a structured dataset essential for model training. The quality and diversity of the collected data significantly impacted the accuracy of the detection model, ensuring it could recognize a wide range of deepfake manipulations across different scenarios.

2. Data Preprocessing

Preprocessing was a critical step to ensure consistency and enhance the model's efficiency. The **OpenCV** library was used to extract and crop facial regions from video frames, ensuring a uniform focus on relevant features. All images were resized to a fixed input shape of **[224, 224]**, making them compatible with Convolutional Neural Networks (CNNs). To improve generalization, techniques like **normalization and data augmentation** (rotation, flipping, and brightness adjustments) were applied. Additionally, noise reduction enhances image

Proceedings DOI: 10.21467/proceedings.7.6
ISBN: 978-81-989164-9-5

Series: AIJR Proceedings
ISSN: 2582-3922

clarity, enabling the model to better distinguish between real and manipulated faces based on structural and texture inconsistencies.

3. Feature Extraction

Feature extraction helped identify key characteristics that differentiate real images from deepfakes. CNNs automatically learnt essential patterns from images, detecting irregularities such as unnatural lighting, distorted facial regions, and texture mismatches. Deepfake images often contain noticeable artifacts around the **eyes, mouth, and hair**, where blending imperfections were more apparent. Convolutional filters captured these inconsistencies by extracting edges, textures, and structural details from images. These extracted features form the basis for classification, allowing the model to recognize even subtle alterations in facial features, making deepfake detection more effective and accurate.

4. Deepfake Detection Using CNN

A **Convolutional Neural Network (CNN)** was employed to classify images as real or fake based on extracted features. The basic architecture includes:

1. **Convolutional Layers** – Extract important visual features.
2. **Pooling Layers** – Reduce dimensionality while preserving key information.
3. **Fully Connected Layers** – Process features for classification.
4. **Sigmoid Activation** – Outputs probabilities for each class. For multi-class classification, softmax activation is generally used.

The CNN learnt to identify deepfake characteristics such as unnatural textures, pixel inconsistencies, and abrupt transitions in facial regions. Advanced models like XceptionNet or EfficientNet may further improve accuracy by leveraging deeper layers and optimized feature extraction techniques.

5. Evaluation Metrics

To measure the performance of the deepfake detection model, various evaluation metrics are used:

- **True Positive (TP):** Fake images correctly identified as fake.
- **True Negative (TN):** Real images correctly classified as real.
- **False Positive (FP):** Real images mistakenly marked as fake.
- **False Negative (FN):** Fake images incorrectly classified as real.

V. RESULT AND ANALYSIS

The suggested Deepfake Detector's performance was assessed using common classification measures, such as F1-Score, Accuracy, Precision, and Recall. These metrics offer a thorough evaluation of the model's ability to identify videos that have been altered.The model regularly produced values between 90% and 95%, demonstrating good performance across all evaluation criteria. The model's capacity to accurately detect deepfake movies while reducing false positives and false negatives is demonstrated by the F1-Score, which strikes a compromise between precision and recall.

- Accuracy: Indicates how accurate the model's predictions are overall. With an accuracy of 90–95%, the suggested system demonstrated a high degree of dependability in differentiating between authentic and fraudulent films (1).

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \qquad (1)$$

- Precision: Indicates the percentage of accurately detected deepfake videos among all videos that have been flagged as fraudulent. The model's accuracy ranged from 80 to 85%, guaranteeing that few real films were misclassified (2).

$$\text{Precision} = \frac{TP}{TP+FP} \qquad (2)$$

- Recall: Indicates the percentage of real deepfake videos that the model properly detected. The algorithm successfully identifies modified content with a recall rate between 80 and 85 percent (3).

$$\text{Recall} = \frac{TP+TN}{TP+TN+FP+FN} \qquad (3)$$

Proceedings DOI: 10.21467/proceedings.7.6
ISBN: 978-81-989164-9-5

Series: AIJR Proceedings
ISSN: 2582-3922

- F1-Score: A single indicator of the model's equilibrium between identifying phoney films and preventing misclassifications, calculated as the harmonic mean of precision and recall. The detection algorithm's resilience was confirmed by the F1-Score, which stayed within the 80–85% range (4).

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

Table 2: Performance Metrics

Parameters	Values
Accuracy	95.4
Precision	80.87
F1 Score	79.4
Recall	77.4
Validation accuracy	75.4

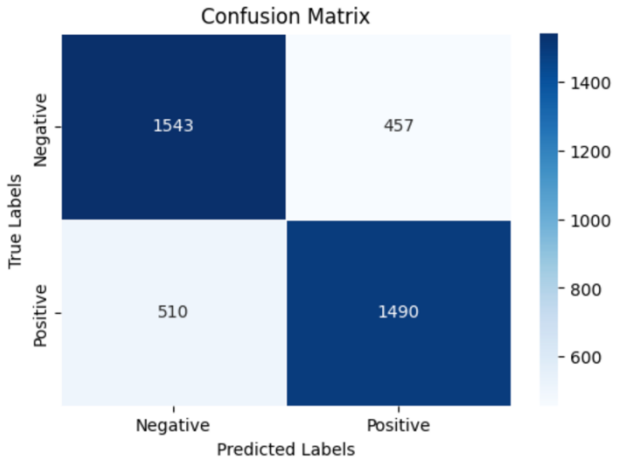

FIGURE 2: Convolutional Neural network Diagram

We had used confusion matrix for evaluation of the True Positive(TP), True Negative (TN) , false Positive (FN) , false Negative (FN).

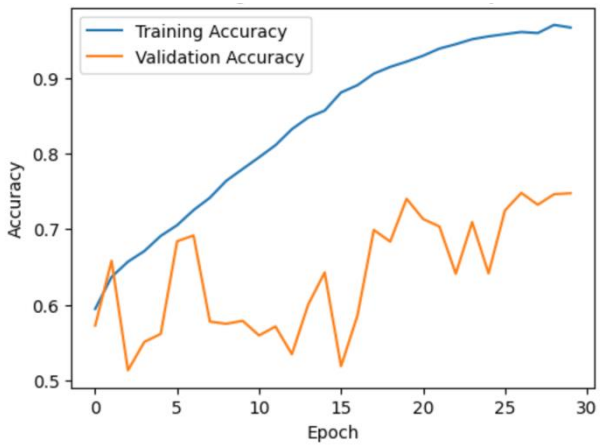

FIGURE 3: Training and validation accuracy

Proceedings DOI: 10.21467/proceedings.7.6
ISBN: 978-81-989164-9-5

Series: AIJR Proceedings
ISSN: 2582-3922

A machine learning model's training and validation accuracy over 30 epochs is depicted in the graph. Training accuracy was represented by the blue line, which rose gradually and surpasses 90%, signifying that the model was successfully extracting patterns from the training data. The validation accuracy indicated by the orange line, however, varies greatly and stayed substantially lower, at most 70%. This disparity rose the possibility of overfitting, in which the model did well on training data but found it difficult to transfer to unobserved validation data. Techniques like regularization, dropout, or expanding the dataset size could be used to solve the variations in validation accuracy, which could be a sign of problems like inadequate data, a complex model, or unsuitable hyperparameters.

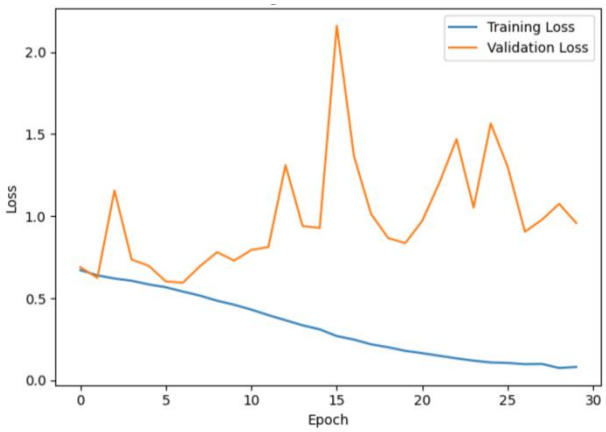

FIGURE 4: TRAINING AND VALIDATION LOSS

A machine learning model's training and validation losses across 30 epochs were shown in the graph. Training loss, represented by the blue line, steadily declines, showing that the model was picking up new information and getting better at using the training set. The orange line, which stood for validation loss, varied a lot yet stayed high. This implied that even while the model fitted the training data quite well, overfitting was probably the cause of its difficulties generalizing to unseen validation data. Further evidence of instability was provided by the steep increases in validation loss, which could be brought on by an improper learning rate, a too complex model, or a lack of data. Regularization strategies that reduced overfitting and enhanced validation performance include dropout, early halting, and data augmentation.

VI. CONCLUSION

In conclusion, the deepfake detection project demonstrated a robust and effective approach to tackling the growing challenges posed by manipulated media in the digital era. By leveraging advanced machine learning techniques, including custom convolutional neural networks and optimized architectures, the system successfully identified and differentiated between authentic and deepfake video content. The model showed strong performance during training and evaluation, highlighting its potential to reduce risks associated with misinformation and media manipulation. The system provided accurate detection capabilities and served as a practical tool for applications in media integrity, security enforcement, and digital trust. While the outcomes were promising, the project recognized limitations such as the dependency on high-quality training data and the rapidly evolving nature of deepfake techniques. Looking ahead, continuous learning and adaptation will be essential to ensure the system evolves alongside emerging threats. Integration of technologies like blockchain can enhance content verification, while user-friendly interfaces and real-time tools will broaden accessibility and adoption. Furthermore, public education and awareness initiatives remain vital to help users better understand and detect manipulated content. Watermarking emerged as a valuable addition to reinforce digital media authenticity. Future developments may focus on creating resilient and adaptive watermarking methods capable of withstanding compression, editing, or adversarial attacks. Combining watermarking with blockchain may further ensure content traceability and integrity. Optimizing these systems for real-time use on digital platforms and social media will enhance their practical deployment. Overall, this research contributed to the ongoing effort to safeguard the integrity of digital information and emphasized the critical role of technology in protecting society against the growing threat of deepfakes.

DECLARATIONS

We, the undersigned authors, hereby declare that the research work presented in this paper entitled **"Detecting and Watermarking Fake Videos Using AI Model"** is our original and authentic work, carried out under academic guidance. This paper has not been submitted, published, or presented elsewhere in part or in full

Proceedings DOI: 10.21467/proceedings.7.6
ISBN: 978-81-989164-9-5

Series: AIJR Proceedings
ISSN: 2582-3922

for the award of any degree, diploma, or certificate. All sources of information, data, images, and research used in this paper have been duly acknowledged. The implementation of AI-based detection techniques and watermarking systems has been conducted ethically and responsibly, in line with academic and professional research standards. We affirm that the work is free from plagiarism and that each contributor has participated significantly in the research and preparation of this manuscript.

REFERENCES

[1] B. U. Mahmud and A. Sharmin, "Deep insights of deepfake technology: A review," *arXiv preprint arXiv:2105.00192*, 2023.

[2] Qureshi SM, Saeed A, Almotiri SH, Ahmad F, Al Ghamdi MA. 2024. Deepfake forensics: a survey of digital forensic methods for multimodal deepfake identification on social media. *PeerJ Computer Science* 10:e2037 **https://doi.org/10.7717/peerj-cs.2037**.

[3] Almars, A. (2021) Deepfakes Detection Techniques Using Deep Learning: A Survey. *Journal of Computer and Communications*, **9**, 20-35. doi: 10.4236/jcc.2021.95003

[4] Tiwari, Aniruddha, Rushit Dave, and Mounika Vanamala. "Leveraging deep learning approaches for deepfake detection: A review." *Proceedings of the 2023 7th International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence*. 2023.

[5] S. Suratkar, E. Johnson, K. Variyambat, M. Panchal and F. Kazi, "Employing Transfer-Learning based CNN architectures to Enhance the Generalizability of Deepfake Detection," *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, Kharagpur, India, 2020, pp. 1-9, doi: 10.1109/ICCCNT49239.2020.9225400.

[6] Devvi Sarwinda, Radifa Hilya Paradisa, Alhadi Bustamam, Pinkie Anggia,Deep Learning in Image Classification using Residual Network (ResNet)Variants for Detection of Colorectal Cancer,Procedia Computer Science,Volume 179,2021,Pages 423-431,ISSN1877-0509,https://doi.org/10.1016/j.procs.2021.01.025.

[7] Goodfellow, Ian J., et al. "Generative adversarial nets." *Advances in neural information processing systems* 27 (2014).

[8] Marhon, S.A., Cameron, C.J.F., Kremer, S.C. (2013). Recurrent Neural Networks. In: Bianchini, M., Maggini, M., Jain, L. (eds) Handbook on Neural Information Processing. Intelligent Systems Reference Library, vol 49. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-36657-4_2.

[9] DeepFake_Faces:https://www.kaggle.com/datasets/dagnelies/deepfake-faces?resource=download

[10] Dunya Ahmed Alkurdi, Mesut Cevik, Abdurrahim Akgundogdu,Advancing Deepfake Detection Using Xception Architecture: A Robust Approach for Safeguarding against Fabricated News on Social Media,Computers, Materials and Continua, Volume 81, Issue 3, 2024, Pages 4285-4305, ISSN 1546-2218, https://doi.org/10.32604/cmc.2024.057029.

[11] A. Qureshi, D. Megías and M. Kuribayashi, "Detecting Deepfake Videos using Digital Watermarking," 2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Tokyo, Japan, 2021, pp. 1786-1793.