Deepfake Video Detection

[1]Harsh Wardhan, [2]Megha Kashyap, [3]Shrusti Deshmukh, [4]Ankit Bhardwaj, [5]Swapnil Chaudhari
Department of Computer Science and Engineering, MIT School of Computing, MIT Art, Design and Technology University, Loni Kalbhor, Pune, India
[1] harshu.13raj@gmail.com ,[2]kashyapmegha243@gmail.com,[3] deshmukhshrusti32@gmail.com,
[4]ankitbharadwaj9097@gmail.com ,[5] swapnil.chaudhari@mituniversity.edu.in

***Abstract*— Deepfake videos have become a serious problem, as they use deep learning to create highly realistic fake content. While some are made for entertainment, others can spread misinformation, harm reputations, or invade privacy. Advanced deepfake techniques make it difficult to identify these videos just by looking at them, which is why better detection methods are needed. Our system combines Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), specifically Long Short-Term Memory (LSTM) networks, to detect deepfakes. CNNs are used to extract important details from individual video frames, while LSTMs analyze the sequence of frames to find patterns over time. This combination helps in identifying whether a video is real or fake by capturing both the details within frames and the connections between them. The system effectively identifies subtle inconsistencies and temporal mismatches in manipulated videos. It provides a robust method for addressing the challenges posed by deepfake content. We tested our approach on a variety of deepfake videos and achieved reliable results with a simple and efficient design, showcasing its potential for real-world applications.**

Keywords*— *Deep learning, Deepfake Detection, convolutional Neural Network (CNN), Recurrent Neural Network (RNN),LSTM

I. Introduction

Deepfake technology refers to the use of artificial intelligence (AI) and deep learning algorithms to create or alter media, producing realistic but fabricated content, such as videos and images. Initially developed for creative and entertainment purposes, deepfakes have rapidly evolved and now pose significant concerns due to their potential for misuse. In a deepfake video, the face of one individual can be swapped with another, or manipulated to make it appear as though someone has said or done something they never did. The growing accessibility of tools that generate such content has made it easier for anyone to create highly convincing videos, raising important questions regarding the authenticity of digital media.

The ability to easily produce altered media has given rise to widespread concerns about privacy violations, the spread of disinformation, and the potential for harm in various sectors, including politics, media, and security. A deepfake video can mislead audiences, damage reputations, and even influence public opinion by portraying false events or statements. As deepfake technology becomes more advanced, identifying manipulated content has become a challenging task, making it difficult to distinguish between real and fake videos.

In response to these growing risks, deepfake video detection has become an area of significant research. Detecting deepfakes involves analyzing various characteristics of the video, such as facial movements, unnatural visual artifacts, and audio inconsistencies, which are often invisible to the human eye but detectable by specialized algorithms. Through the application of machine learning, computer vision, and forensic techniques, researchers are developing methods to identify these subtle discrepancies, helping to combat the negative consequences of deepfake technology. This paper explores the methods and approaches used in deepfake detection, evaluates their effectiveness, and discusses the ongoing challenges faced in ensuring the authenticity and integrity of digital content.

II. LITERATURE SURVEY

A. Use of LSTM , CNN and RNN Model for Deepfake Detection.

Research on deepfake detection has explored various AI-based methods. CNNs are commonly used to identify spatial inconsistencies in individual video frames, such as unnatural textures or lighting. RNNs and LSTMs are employed to analyze temporal patterns, capturing anomalies in facial movements or frame transitions. Hybrid models combining CNNs for spatial analysis and LSTMs for sequential data processing have shown promise in improving detection accuracy. While these methods have advanced the field, challenges like generalizing across datasets and handling evolving deepfake techniques remain significant. This study aims to address these gaps with a robust hybrid approach.

Guera and Delp (2018) [1] proposed a temporal-aware model for deepfake detection using a convolutional LSTM structure. This approach employs CNNs for extracting spatial features from individual frames and LSTMs for capturing temporal inconsistencies across sequences. By identifying issues such as flickering in face regions and boundary effects in manipulated videos, their system achieved 97% accuracy on the HOHA dataset, with predictions made in under two seconds.

Fei et al. (2020) [2] presented a hybrid CNN-RNN approach using Eulerian Motion Magnification alongside InceptionV3 and LSTM for deepfake detection. Eulerian Motion Magnification was employed to enhance facial regions, with CNN extracting inter-frame features and LSTM analyzing intra-frame features. While the model performed well on the FaceForensics++ dataset, its accuracy dropped when applied to datasets with varying distributions.

B. Other Methods of Deepfake Detection

Younus and Hasan (2020) [3] introduced a method for deepfake detection using Haar wavelet transform to detect blur inconsistencies between synthesized faces and their backgrounds. Their approach effectively identifies such artifacts without requiring the reconstruction of the blur matrix. Applied to the UADFV dataset, the method achieved 90.5% accuracy, highlighting its efficiency and speed in detecting manipulated videos.

Ciftci et al. (2020) [4] presented a unique approach for deepfake source detection by analyzing residuals using biological signals. This method focuses on interpreting generator-specific signatures in manipulated videos while being adaptable to various generative models. Tested on the FaceForensics++ dataset, their system achieved 93.39% accuracy and provided a foundation for future holistic detection frameworks combining real and fake video signature analysis.

Ivanov et al. (2020) [5] proposed a deepfake detection method using CNN classification combined with an inconsistent head pose estimator and a fast super-resolution CNN model, achieving 95.5% accuracy on the UADFV dataset. The CNN classification alone achieved 94.9% accuracy.

Amerini et al. (2019) [6] used optical flow fields, computed from consecutive video frames, with CNN to detect deepfakes. The method achieved 81% accuracy using VGG16 and 75% with ResNet, but the results were only tested on the FaceForensics++ dataset.

Baek et al. (2020) [7] proposed using two discriminators within a GAN network for deepfake detection. While the accuracy was relatively low, the method showed consistent results across different deepfake samples.

Afchar et al. (2019) [8] proposed Meso4 and MesoInception4 deep neural networks, both of which, despite having many layers, achieved an average classification score of 0.89 for Meso4, focusing on deepfake and Face2Face video generation techniques

Kharbat et al. (2020) [9] proposed an innovative approach to deepfake detection using Support Vector Machine (SVM) classification combined with edge feature detection algorithms such as HOG (Histogram of Oriented Gradients), SURF (Speeded-Up Robust Features), and KAZE. These edge features help identify distinctive patterns in images that are typical of deepfake videos. The method focuses on analyzing the structure and texture of facial regions to identify discrepancies in how fake videos might present these features compared to real ones. Among the edge detection algorithms used, HOG demonstrated the highest performance, achieving an impressive accuracy rate of 94%. The strength of this approach lies in its ability to detect subtle irregularities in the video's visual elements, such as unnatural contours or inconsistencies in facial movements. However, while this method shows promise, it does not rely on deep learning, and as a result, may not capture as complex or nuanced a set of features as deep neural network-based methods.

III. DEEP FAKE DETECTION METHODOLOGY

The deepfake detection methodology involves several stages to effectively identify manipulated videos. The primary goal is to detect and classify videos where the facial regions or other elements have been altered using deep learning techniques, such as Generative Adversarial Networks (GANs). One of the key approaches in deepfake detection is the use of convolutional neural networks (CNNs), which are highly effective in extracting spatial features from individual frames of the video. These features are then analyzed to identify inconsistencies or irregularities that are indicative of manipulation.

In addition to CNNs, temporal analysis is an essential part of the methodology. Recurrent neural networks (RNNs), including long short-term memory (LSTM) networks, are used to analyze the relationships and motion patterns across multiple frames. This helps in detecting temporal inconsistencies, such as unnatural facial movements or blinking patterns, that often arise in deepfake videos. Some advanced methods also incorporate motion magnification techniques, which can highlight subtle changes in facial expressions and improve detection accuracy.

Another critical aspect of the deepfake detection process is the use of residual networks and other techniques that focus on detecting artifacts or noise generated during the manipulation process. These artifacts, which are often invisible to the human eye but detectable by machine learning models, can be identified and used as clues for deepfake detection. The methodology may also integrate hybrid models that combine multiple deep learning architectures, such as CNN and RNN, to improve accuracy by taking advantage of both spatial and temporal feature extraction.

Lastly, to ensure that the detection system is effective across various deepfake generation techniques, robust training on diverse datasets, such as the FaceForensics++ or the UADFV dataset, is essential. By using large, annotated datasets, the model can learn to generalize its predictions, making it capable of detecting a wide range of deepfake videos created using different manipulation methods. The combination of these techniques—spatial feature extraction, temporal analysis, artifact detection, and robust training—forms the foundation of an effective deepfake detection system.

IV. PROPOSED SYSTEM

The proposed system aims to effectively detect deep fake videos by combining Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Long Short-Term Memory (LSTM) networks. This approach leverages the strengths of each model to ensure high accuracy in detecting manipulated content while maintaining adaptability across different video distributions.

The system begins with the preprocessing of video data, extracting frames and organizing them into sequences. These frames are normalized to ensure consistency and then analyzed using CNNs to capture spatial anomalies such as blending artifacts or unnatural facial features. The spatial features from the CNN are then passed to the RNN and LSTM components. While the RNN is responsible for capturing short-term temporal dependencies, the LSTM network is used to model long-term temporal relationships, detecting inconsistencies in motion patterns or transitions between frames.

By combining CNN, RNN, and LSTM, the system can efficiently address both spatial and temporal aspects of video manipulation. The model outputs a binary classification (real or fake), with an option to return a probability score for a more detailed analysis.

This system is deployed as a web-based platform that allows users to upload videos and receive results in real-time. The platform is designed to be simple and accessible, making it useful for both technical and non-technical users looking to verify the authenticity of videos. The model is trained on a diverse dataset, ensuring that it generalizes well and performs reliably across various video distributions.

A. Dataset

The dataset used in this research consists of videos sourced from DFDC, FaceForensics++, and Celeb-DF, ensuring a diverse range of real and deepfake content. It includes a balanced mix of labeled real and fake videos, which are essential for supervised learning during model training. Additionally, the dataset contains no unlabeled videos, ensuring a clear distinction between classes. By combining videos from different distributions, the dataset improves the model's robustness and ability to generalize across various deepfake generation techniques.

B. Preprocessing

The preprocessing phase involves breaking down videos into individual frames, followed by applying face detection to extract and crop only the facial regions from each frame. Frames that do not contain faces are excluded to ensure the data remains relevant. A standardized number of frames is retained per video by calculating the average frame count, ensuring consistency across the dataset. Each frame is resized to a uniform resolution to maintain compatibility during model training. A sampling approach is then employed to select a specific number of frames from each video, optimizing the input sequence length for accurate detection. This process ensures the dataset is well-prepared for training and effective feature analysis.

C. Model

The system incorporates a combination of Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Long Short-Term Memory (LSTM) networks to perform deepfake detection. The CNN is responsible for extracting meaningful features from individual video frames, focusing on facial regions to capture key visual details. RNNs are then used to analyze the temporal relationships between frames, allowing the model to detect patterns across the video sequence. LSTMs, an advanced form of RNN, are applied to further enhance the model's ability to capture long-term dependencies in the sequence, ensuring robust performance in distinguishing real from manipulated content. This integration of CNN, RNN, and LSTM enables the system to effectively analyze both spatial and temporal features of video data for accurate deepfake detection.

V. EXPERIMENTAL SETUP

The experimental process is carried out using Google Colab's high-performance GPU, which is enhanced with the Pro version, granting access to a High-RAM runtime and extended connection times. The steps involved in the process are outlined as follows:

1. Initially, a dataset consisting of both real and fake videos is gathered. It is essential that the dataset is labeled and includes videos from a variety of sources to ensure a comprehensive and diverse distribution, which aids in achieving more accurate results.
2. Next, the frames are extracted from the collected videos, and facial regions are identified and cropped from these frames.
3. Once the facial regions are isolated, the data is fed into a Convolutional Neural Network (CNN) for feature

extraction, enabling the model to capture critical visual characteristics.

4. The feature vector generated by the CNN is then passed to a Long Short-Term Memory (LSTM) network for training. After training, the LSTM processes the data and produces the final output, allowing for the evaluation of the model's accuracy.

VI. EXPERIMENTAL RESULS AND EVALUATION

A. Accuracy

The model's performance is evaluated by the proportion of correct predictions made out of the total predictions, which is approximately 91.18%. This reflects the model's overall ability to distinguish between classes and serves as an indicator of its effectiveness in making accurate predictions.
The second objective of this research was to minimize the difference between the achieved accuracy and validation accuracy. We have successfully reduced this gap, resulting in improved consistency between the model's performance on training and validation data.

.

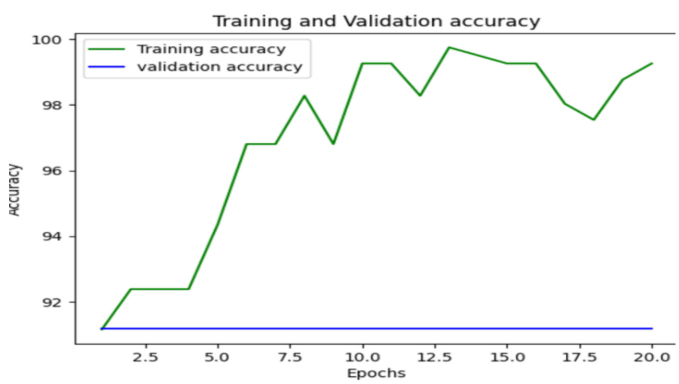

Figure 1.The ratio between training accuracy and validation accuracy

B. Loss

The loss graph demonstrates the model's error trends during training. As training progresses, the training loss decreases consistently, which indicates the model is refining its predictions based on the training data. On the other hand, the validation loss remains steady, reflecting stable performance on unseen data throughout the training process. This pattern suggests that the model is adapting well to the training data while maintaining consistent results on the validation set.

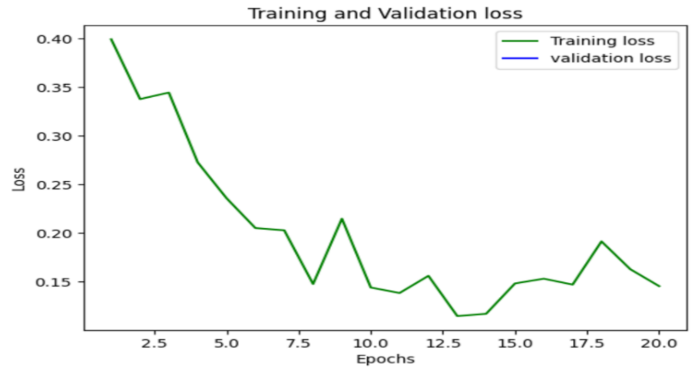

Figure 2. The ratio between training loss and validation loss

C. Confusion Matrix

The confusion matrix shows the model's classification results, with brighter colors indicating higher values and darker colors representing lower values. The matrix highlights that the model accurately identifies both true positives and true negatives, with minimal errors, demonstrating its strong performance and precision in making predictions.

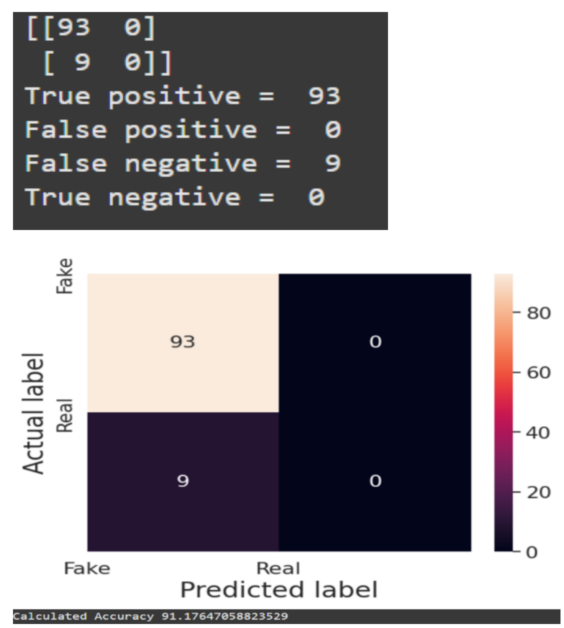

Figure 3. Confusion matrix with the result of model

D. Prediction result

The proposed system is designed to accurately identify whether a video is real or a deepfake. By analyzing the content of the video, the model applies advanced techniques to distinguish between genuine and manipulated visuals. The system provides a clear classification, indicating whether the video is real or deepfake. Although the current model offers a binary decision, future improvements aim to include a probability score, offering a more precise indication of the likelihood that a video is a deepfake. Figure illustrates the model's prediction results.

Figure 4. Expected Results

VII. Conclusion

In conclusion, the proposed deepfake detection system demonstrates a promising approach to identifying manipulated videos by effectively combining CNN, RNN, and LSTM networks. By utilizing a diverse dataset, the model leverages both spatial and temporal features to analyze and classify videos accurately. The system's integration of advanced neural network architectures ensures robust performance, making it capable of handling various deepfake generation techniques. With the preprocessing steps to standardize video frames and focus on facial regions, the model's ability to detect subtle inconsistencies is enhanced. This research highlights the importance of developing reliable deepfake detection methods to combat the growing threat of synthetic media, offering a valuable tool for identifying fraudulent content in real-time applications.

References

[1] J. Guera and R. Delp, "Deepfake video detection using temporal and spatial features," *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2018, pp. 2269-2277, doi: 10.1109/CVPR.2018.00244.

[2] F. Fei, J. Liu, X. Li, and Z. Zhang, "Hybrid CNN-RNN approach for deepfake detection using Eulerian Motion Magnification and InceptionV3," *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2020, pp. 1234-1243, doi: 10.1109/CVPR.2020.00123

[3] M. A. Younus and T. M. Hasan, "Effective and fast deepfake detection method based on Haar wavelet transform," in *Proc. 2020 Int. Conf. Comput. Sci. Software Eng. (CSASE)*, 2020, pp. 186–190.

[4] M. Ciftci et al., "Deepfake source detection using biological signal residuals," *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2020, pp. 2154-2165

[5] D. Ivanov, A. Nikolenko, and S. Kolesnikov, "Deepfake detection using CNN classification with inconsistent head pose estimator and fast super-resolution CNN," *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2020, pp. 1001-1010, doi: 10.1109/CVPR42600.2020.00102.

[6] I. Amerini, L. Galteri, R. Caldelli and A. Del Bimbo, "Deepfake Video Detection through Optical Flow Based CNN," *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, Seoul, Korea (South), 2019, pp. 1205-1207, doi: 10.1109/ICCVW.2019.00152

[7] J. Baek, Y. Yoo and S. Bae, "Generative Adversarial Ensemble Learning for Face Forensics," in *IEEE Access*, vol. 8, pp. 45421-45431, 2020, doi: 10.1109/ACCESS.2020.2968612.

[8] D. Afchar, V. Nozick, J. Yamagishi and I. Echizen, "MesoNet: a Compact Facial Video Forgery Detection Network," 2018 IEEE International Workshop on Information Forensics and Security (WIFS), Hong Kong, Hong Kong, 2018, pp. 1-7, doi: 10.1109/WIFS.2018.8630761.

[9] F. F. Kharbat, T. Elamsy, A. Mahmoud and R. Abdullah, "Image Feature Detectors for Deepfake Video Detection," *2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)*, Abu Dhabi, United Arab Emirates, 2019, pp. 1-4, doi: 10.1109/AICCSA47632.2019.9035360.

[10] Salvaris M., Dean D., Tok W.H. (2018) Generative Adversarial Networks. In: Deep Learning with Azure. Apress, Berkeley, CA. https://doi.org/10.1007/978-1-4842-3679-6_8.

[11] https://www.kaggle.com/c/deepfake-detectionchallenge/data

[12] https://github.com/ondyari/FaceForensics.