Robust Detection Of Deepfake Videos Using Deep Learning

[1]V. Prathyusha, [2]Dr. J. Praveen Kumar

[1]M.Tech Scholar in Department of CSE Teegala Krishna Reddy Engineering College
panduvistarakula@gmail.com
[2]Associate Professor in Department of IT Teegala Krishna Reddy Engineering College
praveentkrecit@gmail.com

Abstract
Recent improvements in computational capabilities have significantly propelled the development of deep learning models, making it easier than ever to create synthetic videos that closely resemble real human speech and facial expressions—widely referred to as deepfakes. These convincingly fabricated videos can be exploited for malicious purposes, including political manipulation, staged acts of terrorism, revenge-based pornography, and various forms of digital extortion. To address these threats, this work proposes a deep learning-based system capable of distinguishing real videos from AI-generated deepfakes. The proposed method leverages artificial intelligence to fight against AI-driven deception. Frame-level features are initially extracted from the input video and analyzed using a ResNeXt Convolutional Neural Network. These spatial features are then forwarded to a Recurrent Neural Network architecture powered by Long Short-Term Memory (LSTM) units, which enables the detection of temporal distortions commonly found in manipulated videos. Although the model currently does not support the identification of specific deepfake subtypes such as reenactment or face replacement, it establishes a robust baseline for further innovation in automated deepfake detection.

Keywords: Deepfake Detection, Synthetic Media, ResNeXt CNN, LSTM, Recurrent Neural Network (RNN), Video Forensics, AI-generated Videos, Temporal Feature Analysis, Deep Learning Framework.

I INTRODUCTION

In the evolving landscape of social media, one of the most pressing concerns surrounding artificial intelligence is the emergence of deepfakes—hyper-realistic synthetic videos that manipulate a person's face or speech in a convincing manner. These fabricated videos pose significant threats, ranging from inciting political instability and orchestrating fake terrorist incidents to disseminating non-consensual explicit content or even engaging in financial extortion. Notable examples have included fabricated nude videos featuring celebrities such as Brad Pitt and Angelina Jolie, underscoring the urgency of effective deepfake detection mechanisms.Given the sophisticated nature of these manipulations, distinguishing between authentic content and deepfakes has become critically important. In our research, we address this challenge by deploying artificial intelligence against its own misuse. Popular applications like Face Swap [12] and FaceApp [11] use pretrained models such as Generative Adversarial Networks (GANs) and Autoencoders to generate deepfake media. In contrast, our method extracts spatial features frame by frame using a pretrained ResNeXt Convolutional Neural Network (CNN), and then analyzes temporal dynamics using a Long Short-Term Memory (LSTM)-based Recurrent Neural Network (RNN).

The ResNeXt CNN helps capture detailed visual cues at the frame level, while the LSTM component processes sequential dependencies across frames to determine whether the video has been manipulated. To

improve generalization and robustness, our model was trained on a large and balanced dataset derived from several publicly available sources, including Face Forensics++ [1], the Deepfake Detection Challenge dataset [2], and Celeb-DF [3].a user-friendly front-end interface that allows individuals to upload videos for analysis. Once processed, the system delivers a classification result—indicating whether the video is real or a deepfake—along with a confidence score, providing transparency and ease of interpretation for end users.

II LITERATURE SURVEY

One of the foundational works in deepfake detection is the Face Forensics++ benchmark dataset, introduced to provide a comprehensive platform for evaluating detection models. It includes thousands of manipulated and pristine videos created using various face-swapping techniques, enabling researchers to train and test models under realistic conditions. The study behind Face Forensics++ highlights the difficulty of detecting subtle artifacts, especially as manipulation methods evolve and improve over time. The dataset serves as a cornerstone for many deepfake detection models due to its diversity and high-quality annotations [1].

The Deepfake Detection Challenge (DFDC) initiated by Facebook and other partners aimed to crowdsource solutions for deepfake detection by providing an extensive, diverse video dataset. The challenge helped reveal the limitations of existing models and emphasized the importance of generalization to unseen manipulations. Many approaches from this challenge employed ensemble methods combining CNNs, attention mechanisms, and audio-visual consistency analysis to boost accuracy [2].

The Celeb-DF dataset addressed a major limitation in previous datasets by providing high-quality deepfake videos with fewer compression artifacts. The work accompanying this dataset introduced new evaluation protocols and revealed that models trained on older datasets failed to generalize well to Celeb-DF, underlining the importance of training on realistic and high-resolution data [3].

In the work by Rossler et al., researchers emphasized the vulnerability of detection models to adversarial perturbations. Their findings demonstrate how even minimal pixel-level modifications can drastically reduce a model's accuracy. This led to the development of robust training methods incorporating adversarial learning to harden models against such manipulations [4].

A unique approach by Nguyen et al. focused on head pose estimation and geometric inconsistencies between manipulated faces and original scenes. By analyzing inconsistencies in the 3D head pose estimations, they were able to effectively identify synthetic face insertions that failed to conform to natural motion dynamics [5].

Sabir et al. explored spatio-temporal modeling using CNN-LSTM hybrid architectures, which allowed their model to not only detect local frame-level inconsistencies but also understand temporal patterns across frames. This combination outperformed models based purely on spatial analysis, especially in detecting subtle temporal artifacts introduced during deepfake generation [6].

In their work, Li et al. introduced a method based on lip-sync inconsistency detection. Their model compared the alignment of phonemes and visemes in speech and mouth movements. As deepfake generation often fails to precisely synchronize these two modalities, this approach proved to be effective in multimodal detection [7].

A study by Zhao et al. utilized frequency domain analysis to detect anomalies introduced during compression and synthesis. Their work showed that deepfakes exhibit specific frequency patterns absent in real videos, and models trained in the frequency domain performed better in scenarios involving heavy post-processing [8].

Korshunov and Marcel addressed the cross-dataset generalization problem. They demonstrated that models trained on a specific dataset often performed poorly on others due to overfitting to particular manipulation

techniques or compression artifacts. They proposed dataset-agnostic features and transfer learning to improve robustness [9].

III EXISTING SYSTEM

Deepfake detection systems today mainly depend on deep learning models, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), to spot irregularities in facial expressions, mismatched audio-visual cues, and visual inconsistencies within video frames. These systems are typically trained on publicly available datasets that contain both real and manipulated videos. Many approaches focus on detecting visual artifacts, such as blurry or low-resolution areas that reveal signs of tampering. Others emphasize analyzing the temporal behavior of faces in videos, using models that track how facial features change over time to spot unnatural movements or patterns.

Some techniques explore specific cues like mismatched lip-sync, incorrect head positioning, or unexpected frequency patterns in the video signal. Although these systems often achieve good results under controlled conditions, they tend to struggle when applied to videos from different sources or when faced with newer, more sophisticated deepfake generation methods. To address this, some researchers have experimented with advanced training techniques and multi-modal detection methods that incorporate video, audio, and even physiological signals like subtle changes in skin tone. Despite these efforts, most current systems are limited in scope, focusing on just one type of manipulation or working best on the datasets they were trained on. As a result, there is a strong need for more flexible, accurate, and real-time detection systems that can generalize well across various deepfake types and real-world scenarios.

IV PROBLEM STATEMENT

With the rapid advancement of deep learning techniques, generating highly realistic deepfake videos has become increasingly accessible, raising serious concerns about misinformation, identity misuse, and privacy violations. Although several deepfake detection systems have been developed using Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and hybrid models, they still face significant limitations. Most existing approaches are tailored to detect specific types of manipulations or are trained on limited datasets, leading to poor generalization when tested on new or unseen deepfake formats.current systems often rely heavily on spatial or temporal features in isolation and are not optimized for real-time detection or deployment across diverse platforms. Attempts to enhance performance through advanced techniques like adversarial training and multimodal analysis have shown promise but remain largely experimental and computationally intensive. As deepfake generation methods continue to evolve, there is a growing need for a more robust, adaptable, and scalable detection framework that can accurately distinguish between real and manipulated videos across various datasets and manipulation types in real-world scenarios.

Objectives

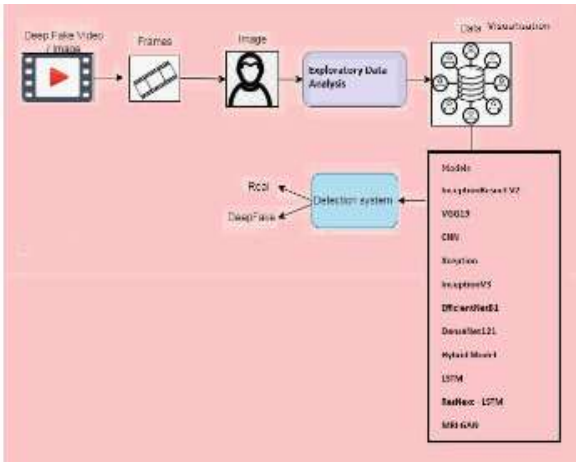
this research is to explore how deepfake technology distorts reality by manipulating visual and audio elements in a way that deceives viewers. By studying the underlying techniques and artifacts used in deepfake generation, the project aims to better understand how these manipulations affect the perception of truth in digital media. Through the development of an effective detection system, this research seeks to reduce the number of individuals who fall victim to online deception, harassment, or identity misuse caused by deepfake content. By providing a tool that can reliably flag manipulated videos, the project aims to support safer online environments.

Another core objective is to create a deep learning-based framework that can accurately differentiate between real and artificially generated (deepfake) videos. The system will analyze both spatial and temporal features of the video to make informed classifications.

V PROPOSED SYSTEM

The proposed system aims to tackle the rising concern of deepfake media by developing a smart, reliable, and deep learning-driven detection framework. As deepfakes become more realistic and harder to detect with the naked eye, existing detection methods are often unable to keep up. To address these challenges, this system brings together a variety of powerful deep learning models, each chosen for their strengths in recognizing the subtle visual and temporal distortions that often occur in manipulated media.The core of the system is a hybrid architecture that combines several high-performing models such as InceptionResNetV2, VGG19, Xception, InceptionV3, EfficientNetB1, DenseNet121, and other custom CNNs. These models are used to extract spatial features from individual video frames or images, helping to detect irregularities that would typically go unnoticed. To capture changes over time—like unnatural blinking or out-of-sync lip movements—the system also employs Long Short-Term Memory (LSTM) networks and ResNeXt-LSTM combinations. This allows it to understand patterns across video sequences rather than just isolated frames.To make the system more adaptive and robust, advanced methods like MRI-GAN are also included. These techniques help the model learn from challenging or deceptive examples, improving its ability to catch newer and more complex types of deepfakes. By using multiple models together, the system can cross-check its results, minimizing false alarms and increasing accuracy for various kinds of deepfake content, such as face swaps, synthetic speech, or fully AI-generated videos.The entire system is designed with modularity and scalability in mind, making it easier to update with newer models or integrate additional features in the future. Real-time detection is a key goal, enabling quick analysis and classification of uploaded videos. To make this technology accessible, a simple and user-friendly interface will also be developed.

VI SYSTEM ARCHITECTURE



System Architecture

The architecture of the proposed deepfake detection system follows a structured and systematic pipeline designed to ensure accurate and efficient identification of manipulated content. The process begins by importing both genuine and deepfake videos, which are then divided into individual frames to facilitate detailed frame-by-frame analysis. Before feeding these frames into the deep learning models, they undergo preprocessing steps including resizing, normalization, and augmentation to standardize input data and improve model generalization.

To better understand the data characteristics, the system incorporates Exploratory Data Analysis (EDA) with visualizations, helping uncover patterns, outliers, and distribution differences between real and manipulated frames. An image data generator is then used to supply the neural networks with transformed image batches during training, which helps prevent overfitting and enhances learning efficiency.

The core detection mechanism is powered by multiple deep learning architectures such as InceptionResNetV2, VGG19, CNN, and Xception. These models work in parallel to process each frame independently, extracting spatial features and identifying subtle inconsistencies that may signal forgery. By leveraging the individual strengths of these networks, the system can capture a wide range of manipulation artifacts across different types of deepfake videos.

To evaluate performance, the system uses a comprehensive set of metrics including accuracy, precision, recall, F1-score, specificity, sensitivity, Mean Absolute Error (MAE), and Mean Squared Error (MSE). These metrics ensure a balanced assessment of both detection effectiveness and error rates. Altogether, this architecture integrates preprocessing, visual analysis, and deep learning to deliver a robust and scalable solution for detecting deepfakes with high reliability.

VI IMPLEMENTATION

The proposed deepfake detection system integrates a variety of advanced deep learning models, each chosen for their specific strengths in feature extraction, classification, and sequence analysis. These algorithms work together to analyze spatial and temporal aspects of video frames, enabling accurate identification of manipulated content.

Inception-ResNetV2 is a powerful convolutional neural network that blends the strengths of two well-established architectures: Inception and ResNet. By combining inception modules with residual connections, it enhances both the accuracy and efficiency of feature extraction, making it suitable for identifying subtle deepfake artifacts. Similarly, **VGG19**, a 19-layer CNN developed by the Visual Geometry Group at the University of Oxford, is widely recognized for its simplicity and effectiveness in image classification tasks. It uses small convolutional filters and deep layers to extract high-level visual features.

Convolutional Neural Networks (CNNs) form the backbone of many image analysis tasks. They consist of stacked layers of convolution, pooling, and fully connected layers, allowing the system to detect patterns such as facial features, lighting inconsistencies, and texture irregularities in each frame. **Xception**, an extension of the Inception architecture, introduces depthwise separable convolutions to capture both spatial and channel-wise information with fewer parameters and improved accuracy. **InceptionV3**, another evolution of the Inception family, adds efficient factorized convolutions and dimensionality reduction techniques, optimizing performance while reducing computation.

To improve computational efficiency without compromising accuracy, the system incorporates **EfficientNetB1**, part of a model family known for balancing network depth, width, and resolution. It achieves excellent performance with fewer parameters compared to conventional CNNs. Another high-performing model included is **DenseNet121**, which establishes dense connections between all layers within each block. This architecture improves feature propagation and gradient flow, resulting in better learning and generalization.

A **Hybrid Model** is also included, combining elements from multiple architectures to tailor the network to the specific needs of deepfake detection. This fusion allows the system to leverage the strengths of each model for more robust performance across varied input types.

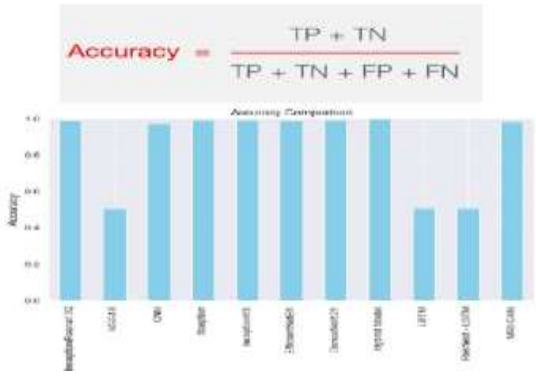
For analyzing video content over time, **Long Short-Term Memory (LSTM)** networks are employed. LSTMs are a form of Recurrent Neural Network (RNN) designed to capture long-term dependencies in sequential data. This makes them well-suited for detecting inconsistencies in facial expressions, blinking, and lip movements across multiple frames. The system also uses a **ResNeXt-LSTM** hybrid, which combines the spatial feature extraction power of ResNeXt with the sequential learning capabilities of LSTM. This integration enables effective processing of both visual and temporal aspects of video content.

VII RESULTS


Deepfake Detection

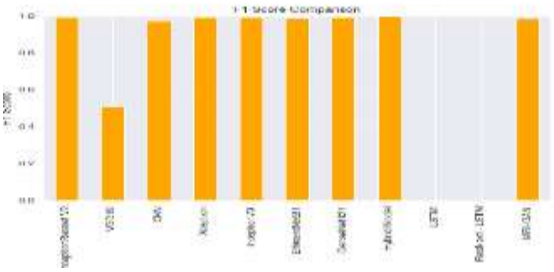

Deepfake Detection


Deepfake Detection
Frames Split
Face Cropped Frames
Play to see Result


Accuracy = TP + TN / TP + TN + FP + FN

$$\text{F1 Score} = \frac{2}{\left(\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}\right)}$$

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



VIII CONCLUSION

we demonstrated that the proposed deepfake detection model, which leverages neural networks for classification, offers a reliable approach to distinguishing between genuine and manipulated video content. The system effectively analyzes short video sequences and is capable of delivering predictions in near real-time—processing at a rate of 10 frames per second, which equates to providing results within just one second of video playback. To build this system, we employed a pre-trained ResNeXt Convolutional Neural Network (CNN) for frame-level feature extraction. This model captures detailed spatial information from each video frame. We then integrated a Long Short-Term Memory (LSTM) network to process temporal sequences, enabling the detection of subtle transitions and inconsistencies between consecutive frames (t and t-1). Our model is equipped to handle input sequences of various lengths, including 10, 20, 40, 60, 80, and 100 frames, making it adaptable to different video durations and use cases.the results validate the effectiveness of our hybrid deep learning approach and highlight its potential for practical deployment in combating the spread of deepfake media.

References

[1] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies,Matthias Nießner, “FaceForensics++: Learning to Detect Manipulated Facial Images” in arXiv:1901.08971.
[2] Deepfake detection challenge dataset :https://www.kaggle.com/c/deepfake-detection- challenge/data Accessed on 26 March, 2020
[3] YuezunLi,XinYang,PuSun,HonggangQiandSiweiLyu“Celeb-DF:A Large-scale Challenging Dataset for DeepFake Forensics” in arXiv:1909.12962
[4] Deepfake Video of Mark Zuckerberg Goes Viral on Eve of House A.I. Hearing :
https://fortune.com/2019/06/12/deepfake-mark-zuckerberg/ Accessed on 26 March, 2020
[5] 10 deep fake examples that terrified and amused the internet :
https://www.creativebloq.com/features/deepfake-examples Accessed on 26 March,2020
[6] TensorFlow:https://www.tensorflow.org/(Accessedon26March,2020)
[7] Keras:https://keras.io/(Accessedon26March,2020)
[8] PyTorch:https://pytorch.org/(Accessedon26March,2020)
[9] G. Antipov, M. Baccouche, and J.-L. Dugelay. Face aging with conditional gen- erative adversarial networks. arXiv:1702.01983, Feb. 2017.
[10] J. Thies et al.Face2Face: Real-time face capture and reenactment of rgb videos. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2387–2395, June 2016. Las Vegas, NV.
[11] Faceapp:https://www.faceapp.com/(Accessedon26March,2020)
[12] FaceSwap:https://faceswaponline.com/(Accessedon26March,2020)
[13] Deep fakes, Revenge Porn, And The Impact On Women :
https://www.forbes.com/sites/chenxiwang/2019/11/01/deepfakes-revenge-porn-and-the-impact-on-women/

[14] The rise of the deep fake and the threat to democracy: https://www.theguardian.com/technology/ng-interactive/2019/jun/22/the-rise-of-the-deepfake-and-the-threat-to-democracy(Accessed on 26 March, 2020).