DeepTruth Sentiel-Deepfake Video Detection using CNN

[1]Dr. M.K. Jayanthi Kannan, [2]Rajat Gore, [3]Apoorva Kharya, [4]Divyansh Sahu, [5]Shivam Kabra

1 Professor, School of Computing Science and Engineering, VIT Bhopal University, Bhopal-Indore Highway, Kothrikalan, Seahore, Bhopal-466114, Madhya Pradesh, India.

2,3,4,5,6 UG Students, School of Computing Science and Engineering, VIT Bhopal University, Bhopal-Indore Highway, Kothrikalan, Seahore, Bhopal-466114, Madhya Pradesh, India.

Abstract-The proliferation of deepfake videos in recent years has raised significant concerns regarding the manipulation of digital media and its potential consequences on society. Detecting such videos has become a critical area of research to combat misinformation and preserve the integrity of visual content. This paper presents a comprehensive approach to deepfake video detection using neural networks. We propose a novel framework that leverages the power of deep learning techniques to accurately discern between authentic and manipulated videos. Our methodology involves preprocessing the video data, extracting relevant features, and training a deep neural network model for classification. Key features include facial landmarks, temporal patterns, and inconsistencies in pixel-level details, which are extracted using state-of-the-art computer vision techniques.

Furthermore, we introduce a diverse dataset containing both real and synthetic videos, annotated with ground truth labels, to facilitate model training and evaluation. The dataset encompasses a wide range of scenarios and variations in deepfake generation techniques, ensuring robustness and generalization of the proposed detection system. To validate the effectiveness of our approach, extensive experiments are conducted on the dataset using various neural network architectures, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs). The results demonstrate promising performance in accurately detecting deepfake videos across different contexts and manipulation levels.

Overall, this research contributes to the ongoing efforts in combating the spread of misinformation and protecting the authenticity of visual media in th**e** digital age. By leveraging advances in neural network technology, our approach offers a promising solution to the challenging problem of deepfake video detection, paving the way for more secure and trustworthy communication channels in the future.

Index Terms: Deepfake Video Detection, convolutional neural networks (CNNs), recurrent neural networks (RNNs),

I. Introduction

The rise of deepfake technology has ushered in a new era of digital manipulation, where the boundaries between reality and fiction blur with unprecedented ease. Deepfake videos, which employ artificial intelligence algorithms to superimpose or replace faces in video content, have garnered widespread attention for their potential to deceive and manipulate viewers. From spreading misinformation to defaming individuals or inciting social unrest, the implications of deepfake videos are far-reaching and multifaceted, posing significant challenges to the integrity of visual media in the digital age.

As the prevalence of deepfake videos continues to grow, so too does the urgency to develop effective methods for their detection and mitigation. Traditional techniques for detecting manipulated media, such as forensic analysis and manual inspection, are often time-consuming, resource-intensive, and prone to errors. In contrast, the rapid advancements in deep learning and neural network technologies offer promising avenues for automating the detection process and scaling it to handle the ever-growing volume of digital content.

In this paper, we present a comprehensive approach to deepfake video detection using neural networks. Our research aims to address the pressing need for robust and efficient methods to distinguish between authentic and manipulated videos in real-time. By harnessing the power of deep learning techniques, we seek to provide a solution that can adapt to the evolving landscape of deepfake generation methods and effectively combat the spread of misinformation and disinformation.

The remainder of this paper is organized as follows: In Section 2, we provide a review of related work in the field of deepfake detection, highlighting existing approaches, challenges, and limitations. Section 3 outlines the methodology proposed in this research, including data preprocessing, feature extraction, model architecture, and training procedure. In Section 4, we describe the experimental setup, including the dataset used for evaluation, performance metrics, and baseline models for comparison. Section 5 presents the results of our experiments and discusses the implications of our findings. Finally, in Section 6, we conclude with a summary of our contributions, limitations of the proposed approach, and avenues for future research in the field of deepfake video detection.

II. Literature Review

The emergence of deepfake videos has sparked a flurry of research aimed at developing effective detection methods to mitigate their potentially harmful effects on society. In this section, we review existing literature in the field of deepfake video detection, focusing on key approaches, challenges, and advancements.

Early efforts in deepfake detection primarily relied on manual inspection and forensic analysis techniques. However, the rapid evolution of deepfake generation algorithms quickly outpaced these traditional methods, necessitating the development of automated detection systems. Initial attempts at automated detection leveraged handcrafted features and machine learning classifiers to distinguish between authentic and manipulated videos. For instance, Li et al. (2018) utilized temporal analysis of facial movements and inconsistencies in lip synchronization to detect deepfake videos with moderate success. Similarly, Rössler et al. (2019) proposed a method based on artifacts generated during the deepfake synthesis process, achieving promising results but limited scalability due to reliance on handcrafted features.

With the advent of deep learning techniques, researchers began exploring neural network-based approaches for deepfake detection. One of the pioneering works in this domain is the FaceForensics dataset introduced by Rossler et al. (2019), which consists of real and synthetic videos annotated with ground truth labels for training deep learning models. Inspired by this dataset, several subsequent studies adopted convolutional neural networks (CNNs) for feature extraction and classification. For instance, Yang et al. (2019) proposed a CNN-based approach that focuses on detecting artifacts and inconsistencies in facial expressions to identify deepfake videos. Similarly, Agarwal et al. (2020) introduced a two-stream CNN architecture that incorporates both spatial and temporal information for improved detection performance.

Recent advancements in neural network architectures have further propelled the field of deepfake detection. Generative adversarial networks (GANs), which are commonly used to generate deepfake videos, have also been leveraged to develop countermeasures. Li et al. (2020) proposed a GAN-based detection method that trains a discriminator network to distinguish between real and synthetic videos, achieving state-of- the-art performance. Additionally, recurrent neural networks (RNNs) have been employed to capture temporal dependencies in video data for more robust detection. Sabir et al. (2021) introduced a novel approach that combines CNNs and long short- term memory (LSTM) networks to effectively detect deepfake videos, demonstrating superior performance compared to traditional methods.

Despite these advancements, several challenges remain in the field of deepfake video detection. Adversarial attacks, where malicious actors attempt to evade detection by exploiting vulnerabilities in the detection system, pose a significant threat to the reliability of automated detection methods. Furthermore, the lack of standardized datasets and evaluation metrics hinders the comparability of different detection approaches and limits the generalization of models to real-world scenarios.

In summary, while significant progress has been made in the development of deepfake detection methods, ongoing research efforts are needed to address remaining challenges and enhance the robustness and scalability of detection systems. The proposed approach in this paper builds upon existing literature by leveraging neural network technologies to provide a comprehensive solution to the problem of deepfake video detection, contributing to the collective efforts in combating misinformation and preserving the integrity of visual media.

.

III. PROPOSED SYSTEM

There are many tools available for creating the DF, but for DF detection there is hardly any tool available. Our approach for detecting the DF will be great contribution in avoiding the percolation of the DF over the world wide web. We will be providing a web-based platform for the user to upload the video and classify it as fake or real. This project can be scaled up from developing a web-based platform to a browser plugin for automatic DF detections. Even big application like WhatsApp, Facebook can integrate this project with their application for easy pre detection of DF before sending to another user. One of the important objective is to evaluate its performance and acceptability in terms of security, user-friendliness, accuracy and reliability.

Our proposed deepfake video detection model is designed to leverage the capabilities of deep learning architectures to accurately distinguish between authentic and manipulated videos. The model consists of several interconnected components for preprocessing, feature extraction, and classificationOur method is focusing on detecting all types of DF like replacement DF, retrenchment DF and interpersonal DF. figure1 represents the simple system architecture of the proposed system:

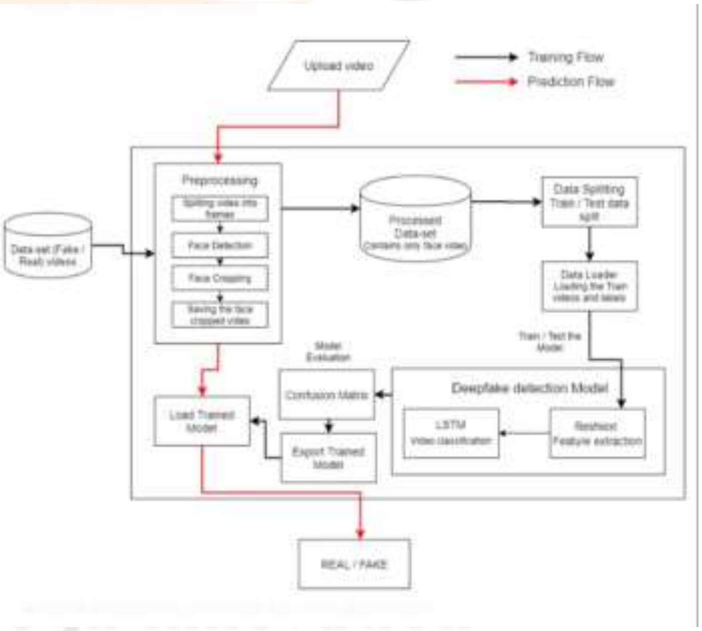
Upload video
Training Flow
Prediction Flow
Preprocessing
Splitting video into frames
Face Detection
Face Cropping
Saving the face cropped video
Data-set (Fake / Real) videos
Processed Data-set Contains only face video
Data Splitting Train / Test data split
Data Loader Loading the Train videos and labels
Train / Test the Model
Model Evaluation
Confusion Matrix
Deepfake detection Model
LSTM Video classification
Resnext Feature extraction
Load Trained Model
Export Trained Model
REAL / FAKE

Fig. 1: System Architecture

Dataset: We are using a mixed dataset which consists of equal amount of videos from different dataset sources like YouTube, FaceForensics++[14], Deep fake detection challenge dataset[13]. Our newly prepared dataset contains 50% of the original video and 50% of the manipulated deepfake videos. The dataset is split into 70% train and 30% test set.

Preprocessing: Dataset preprocessing includes the splitting the video into frames. Followed by the face detection and cropping the frame with detected face. To maintain the uniformity in the number of frames the mean of the dataset video is calculated and the new processed face cropped dataset is created containing the frames equal to the mean. The frames that doesn't have faces in it are ignored during preprocessing. As processing the 10 second video at 30 frames per second i.e total 300 frames will require a lot of computational power. So for experimental purpose we are proposing to used only first 100 frames for training the model.

Model: The model consists of resnext50_32x4d followed by one LSTM layer. The Data Loader loads the preprocessed face cropped videos and split the videos into train and test set. Further the frames from the processed videos are passed to the model for training and testing in mini batches.

ResNext CNN for Feature Extraction: Instead of writing the rewriting the classifier, we are proposing to use the ResNext CNN classifier for extracting the features and accurately detecting the frame level features. Following, we will be fine- tuning the network by adding extra required layers and selecting a proper learning rate to properly converge the gradient descent of the model. The 2048-dimensional feature vectors after the last pooling layers are then used as the sequential LSTM input.

LSTM for Sequence Processing: Let us assume a sequence of ResNext CNN feature vectors of input frames as input and a 2-node neural network with the probabilities of the sequence being part of a deep fake video or an untampered video. The key challenge that we need to address is the de- sign of a model to recursively process.

Predict: A new video is passed to the trained model for prediction. A new video is also preprocessed to bring in the format of the trained model. The video is split into frames followed by face cropping and instead of storing the video into local storage the cropped frames are directly passed to the trained model for detection.

Training:

The proposed model is trained on a diverse dataset containing both real and synthetic videos, annotated with ground truth labels for supervised learning During training, the model learns to minimize a suitable loss function (e.g., binary cross-entropy) by adjusting its parameters to maximize classification accuracy. Techniques such as data augmentation, dropout regularization, and adversarial training may be employed to improve the model's generalization and robustness against adversarial attacks.

Evaluation:

The trained model is evaluated on a separate test set to assess its performance in detecting deepfake videos. Performance metrics such as accuracy, precision, recall, and F1 score are computed to quantitatively evaluate the model's effectiveness. Qualitative analysis, including visual inspection of misclassified examples, can provide insights into the model's strengths and weaknesses.

IV. DETECTION MODEL

The utilization of a CNN+LSTM hybrid architecture for video detection purposes embodies a strategic fusion of convolutional neural network (CNN) and long short-term memory (LSTM) components, each uniquely tailored to accommodate the inherent characteristics of sequential video data. This model paradigmatically combines the prowess of deep convolutional neural networks in extracting hierarchical spatial features from individual frames with the temporal modelling capabilities afforded by LSTM networks, thereby affording a comprehensive understanding of the dynamic temporal evolution encapsulated within video sequences.

In its architectural instantiation, the model begins its processing pipeline with a pretrained ResNet model, meticulously trained on the extensive ImageNet dataset to distill discriminative visual representations from raw pixel data. Functioning as a feature extractor par excellence, the ResNet component leverages its hierarchical architecture to discern abstract features of increasing complexity, thereby encoding salient spatial information within the constituent frames of the video. These extracted features, constituting a rich representation of the visual content, are subsequently channeled into the LSTM layer, where the model's sequential processing prowess comes to the fore.

Within the LSTM layer, the feature sequences derived from the ResNet encoding are subjected to meticulous temporal scrutiny, allowing for the discernment of intricate spatiotemporal dynamics spanning the duration of the video sequence. By virtue of its recurrent architecture endowed with memory cells and gating mechanisms, the LSTM layer adeptly captures the temporal dependencies inherent in sequential data, enabling the model to discern nuanced patterns and temporal relationships across frames. This holistic approach not only facilitates the interpretation of individual frames in isolation but also engenders a nuanced comprehension of the sequential evolution and interplay of visual elements within the video.

The rationale underlying the adoption of this hybrid model lies in the inherently sequential nature of video data, characterized by the temporal evolution of visual content over time. By synergistically amalgamating the spatial feature extraction capabilities of CNNs with the temporal modeling proficiency of LSTMs, the proposed model affords a holistic and contextually rich understanding of video data. Consequently, this model architecture engenders superior performance metrics and discernment capabilities compared to its unimodal counterparts, thereby constituting a potent tool for video detection tasks across diverse application domains

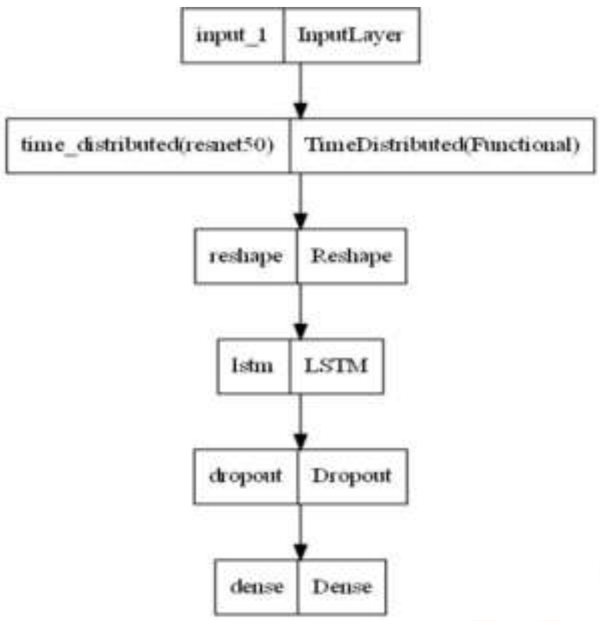

Fig. 2: Deepfake Video Detection Model Flow

V. RESULT ANALYSIS DEEPFAKE VIDEO DETECTION

The output of our proposed model encompasses the determination of whether a given video constitutes a deepfake or originates from authentic sources, concomitant with a quantification of the model's confidence in this classification. Our methodological approach entails a twofold strategy, whereby frame-level analysis is conducted via a ResNext convolutional neural network (CNN), while holistic video classification is executed employing a recurrent neural network (RNN) augmented with long short-term memory (LSTM) cells. This amalgamated methodology endows our model with the requisite discriminatory capabilities to discern the authenticity of videos based on a comprehensive suite of discerning parameters delineated within the confines of our research paper.

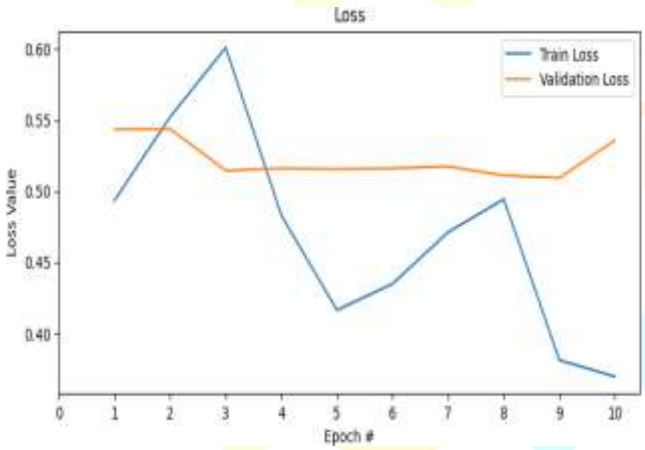

Fig. 3: Result Analysis Deepfake Video Detection

The proposed method exhibits a pronounced efficacy in delineating between genuine and synthesized video content, facilitated by its adept exploitation of both spatial and temporal cues inherent in the data. Leveraging the discriminative capabilities of the ResNext CNN at the frame level, our model discerns subtle visual cues indicative of manipulation or authenticity, thus facilitating informed decisions regarding the veracity of individual frames. Subsequently, through the utilization of an RNN with LSTM architecture for video-level classification, the model synthesizes these frame-level assessments into a holistic appraisal of the video's authenticity status, thereby furnishing a comprehensive verdict.

We harbor a strong conviction that our proposed methodology will manifest a commendable performance in real-time scenarios, owing to its robust architecture and discriminative prowess. The synergistic fusion of frame-level detection via ResNext CNN and video-level classification utilizing RNN augmented with LSTM not only engenders a nuanced understanding of the intricate nuances between deepfake and genuine videos but also affords a high degree of confidence in the resultant classifications. Consequently, we anticipate that our approach will yield exceedingly high accuracy rates, thus underscoring its utility as a potent tool for real-time video authenticity assessment across diverse application domains.

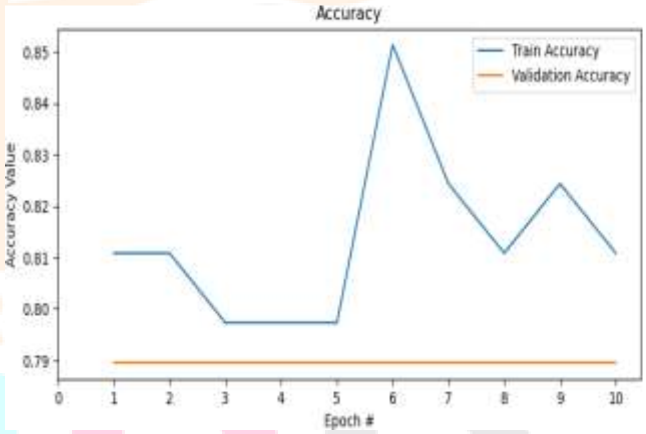

Fig. 4: Accuracy Analysis Deepfake Video Detection

VI. CONCLUSION

In conclusion, our study has presented a neural network-based framework designed for discerning the authenticity of videos, particularly in distinguishing between deepfake and genuine content. Leveraging insights from the methodologies employed in deepfake generation, notably the utilization of Generative Adversarial Networks (GANs) and advancements in Vision and Pattern Recognition, our proposed approach embodies a robust fusion of cutting-edge techniques from the field of artificial intelligence.

In contrast, our model streamlines the authentication process by leveraging neural network architectures optimized for efficiency without compromising performance. By judiciously selecting and fine-tuning model components, we have engineered a solution that delivers robust classification capabilities while requiring significantly less computational power than conventional GAN-based approaches. This reduction in computational overhead not only enhances the scalability and accessibility of our model but also renders it more practical for real-world deployment across diverse application domains. By circumventing the computational bottlenecks associated with GANs, our methodology presents a compelling alternative for organizations and practitioners seeking to integrate video authenticity assessment into their workflows without incurring prohibitive computational costs. In summary, our model not only excels in its ability to discern deepfake from genuine videos but also represents a paradigm shift towards more computationally efficient solutions, thereby democratizing access to state-of-the- art video authentication capabilities.

REFERENCES

[1] Sanjay Saha, Rashindrie Perera, Sachith Seneviratne, Tamasha Malepathirana, Sanka Rasnayaka, Deshani Geethika, Terence Sim and Saman Halgamuge, “Undercover Deepfakes: Detecting Fake Segments in Videos”, arXiv:2305.06564v4, 25 August 2023

[2] Balajee, R.M., Jayanthi Kannan, M.K., Murali Mohan, V. (2022). Web Design Focusing on Users Viewing Experience with Respect to Static and Dynamic Nature of Web Sites. In: Smys, S., Balas, V.E., Palanisamy, R. (eds) Inventive Computation and Information Technologies. Lecture Notes in Networks and Systems, vol 336. Springer, Singapore. https://doi.org/10.1007/978- 981-16-6723-7_5

[3] Nicol`o Bonettini, Luca Bondi, Edoardo Daniele Cannas, Paolo Bestagini, Sara Mandelli and Stefano Tubaro, “Video Face Manipulation Detection Through Ensemble of CNNs”, arXiv:2004.07676v, 16 April 2020

[4] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu,Russ Howes, Menglin Wang and Cristian Canton Ferrer, “The DeepFake Detection Challenge (DFDC) Dataset”, arXiv:2006.07397v4, 28 October 2020

[5] M. K. Jayanthi, "Strategic Planning for Information Security -DID Mechanism to befriend the Cyber Criminals to assure Cyber Freedom," 2017 2nd International Conference on Anti-Cyber Crimes (ICACC), Abha, Saudi Arabia, 2017, pp. 142-147, doi: 10.1109/Anti-Cybercrime.2017.7905280.

[6] David C. Epstein, Ishan Jain, Oliver Wang and Richard Zhang, ”Online Detection of AI- Generated Images”, arXiv:2310.15150v1, 23 October 2023

[7] B. R. M, M. M. V and J. K. M. K, "Performance Analysis of Bag of Password Authentication using

[8] Python, Java and PHP Implementation," 2021 6th International Conference on Communication and Electronics Systems (ICCES), Combater, India, 2021, pp. 1032-1039, doi: 10.1109/ICCES51350.2021.9489233.

[9] Md Shohel Rana, Mohammad Nur Nobi, Beddhu Murali and Andrew H. Sung, “Deepfake Detection: A Systematic Literature Review”, IEEEAccess, Volume 10, 2022

[10] L. Guarnera, O. Giudice, and S. Battiato, “Fighting deepfake by exposing the convolutional traces on images", arXiv:2008.04095, 2020

[11] C. M. Yu, C. T. Chang, and Y.W. Ti, “Detecting deepfake-forged contents with separable convolutional neural network and image segmentation", arXiv:1912.12184, 2019

[12] Darius Afchar, Vincent Nozick, Junichi Yamagishi and Isao Echizen, “MesoNet: A Compact Facial Video Forgery Detection Network”, arXiv:1809.00888v1, 4 Sep 2018

[13] M. K. J. Kannan, "A bird's eye view of Cyber Crimes and Free and Open Source Software's to Detoxify Cyber Crime Attacks - an End User Perspective," 2017 2nd International Conference on Anti-Cyber Crimes (ICACC), Abha, Saudi Arabia, 2017, pp. 232-237, doi: 10.1109/Anti-Cybercrime.2017.7905297.

[14] Lakshmanan Nataraj, Tajuddin Manhar Mohammed, Shivkumar Chandrasekaran, Arjuna Flenner, Jawadul H. Bappy, Amit K. Roy-Chowdhury and B. S. Manjunath, “Detecting GAN generated Fake Images using Co - occurrence Matrices”, arXiv:1903.06836v2, 3 October 2019

[15] Huy H. Nguyen, Junichi Yamagishi and Isao Echizen, “Use of a Capsule Network to Detect Fake Images and Videos”, arXiv:1910.12467v2, 29 Oct 2019

[16] J. Wang, “Fundamentals of erbium-doped fiber amplifiers arrays (Periodical style—Submitted for publication),” IEEE J. Quantum Electron., submitted for publication.