A Multimodal Approach Harnessing AI to Expose Digital Deceptions

Sharves. P [1], Shobana V [2], Sri Dharrshan S[3], Senthil Kumar S [4]
[1,2,3] Student [4] Professor
[1,2,3,4] Department of Artificial Intelligence and Data Science
Sri Krishna College of engineering and technology

Abstract - **The rapid advancement in computational power has greatly empowered deep learning algorithms, enabling the creation of indistinguishable human-synthesized videos known as deep fakes. This trend has raised concerns about the potential misuse of these realistic face-swapped deep fakes for various nefarious purposes such as political manipulation, fabricating terrorism events, spreading revenge porn, or blackmailing individuals. In this study, we propose a novel deep learning-based method designed to effectively differentiate between AI-generated fake videos and real ones. Our approach aims to combat the proliferation of AI-generated content with AI itself. Our system leverages a Res-Next Convolutional Neural Network (CNN) to extract frame-level features. These features are then used to train a Long Short-Term Memory (LSTM) based Recurrent Neural Network (RNN) to classify videos as manipulated (i.e., deep fakes) or genuine.To ensure the effectiveness of our model in real-world scenarios, we evaluate it using a large, diverse dataset that combines samples from various sources such as FaceForensic++, the Deepfake Detection Challenge, and Celeb-DF. Furthermore, we demonstrate how our approach achieves competitive performance using a straightforward and robust methodology.**

Keywords: Convolutional Neural Network (CNN), Long Short – Term Memory (LSTM), Recurrent Neural Network (RNN)

I. INTRODUCTION

In the rapidly expanding domain of social media platforms, the widespread emergence of deepfakes presents a formidable challenge within the realm of artificial intelligence. These convincingly altered videos, often featuring face-swapped content, are increasingly deployed in various malicious contexts, such as instigating political upheaval, fabricating terrorist incidents, and circulating illicit materials, as evidenced by incidents involving notable figures like Brad Pitt and Angelina Jolie. It is imperative to devise methodologies capable of distinguishing genuine videos from deepfakes. Employing AI to counter AI, our approach integrates a Long Short-Term Memory (LSTM) based artificial neural network to scrutinize the sequential temporal patterns of video frames. Furthermore, we harness a pre-trained ResNext Convolutional Neural Network (CNN) to extract essential frame-level features crucial for accurate classification. Our system undergoes rigorous training on a diverse range of datasets, including FaceForensic++, the Deepfake Detection Challenge, and Celeb-DF, to simulate real-world scenarios and bolster model performance on real-time data. For user convenience, we've developed a user-friendly frontend application. Users simply upload videos, which are subsequently analyzed by the model. The application provides a classification verdict, labeling videos as either authentic or deepfakes, accompanied by the model's confidence score. The evolution of mobile camera technology alongside the widespread adoption of social media and media-sharing platforms has streamlined the creation and dissemination of digital videos. Deep learning has played a pivotal role in unlocking previously unattainable technologies. Contemporary generative models exemplify this advancement, capable of generating hyper-realistic images, speech, music, and videos. These models find application across diverse domains, from enhancing accessibility through text-to-speech systems to facilitating the generation of medical imaging training data. However, as with any transformative technology, new challenges have emerged. Deep generative models have birthed "deep fakes," manipulated video and audio clips that have garnered considerable attention since their emergence in late 2017. An array of open-source methods and tools for creating deep fakes has since proliferated, resulting in a surge of synthesized media. While some are crafted for entertainment, others pose significant threats to individuals and society. The accessibility of editing tools and the demand for domain expertise have fueled the proliferation of fake videos, amplifying their realism. The rampant dissemination of deep fakes across social media platforms has become commonplace, contributing to spamming and the propagation of false information. Consider the impact of a deep fake depicting a national leader declaring war against neighboring nations or a revered celebrity disparaging their fan base. Such instances of deep fakes can wield considerable damage, inducing confusion and instilling fear among the populace. To confront this challenge, the detection of deep fakes assumes paramount importance. Hence, we introduce a novel deep learning-based method engineered to effectively discern AI-generated counterfeit videos (deep fake videos) from authentic ones. It is imperative to develop technology adept at identifying and curtailing the proliferation

of deep fakes on the internet, encompassing the detection of fake news through web scraping and sentiment analysis.

II. LITREATURE SURVEY

The exponential growth of deep fake video and its illicit usage represents a significant threat to democracy, justice, and public trust, driving an increased demand for fake video analysis, detection, and intervention. Several techniques for deep fake detection have emerged, including: ExposingDF Videos by Detecting Face Warping Artifacts [1]: This method detects artifacts by comparing the generated face areas and their surrounding regions with a dedicated Convolutional Neural Network (CNN) model. The approach focuses on identifying two types of face artifacts, leveraging the observation that current deep fake algorithms typically generate images of limited resolutions, necessitating further transformation to match the faces in the source video. Exposing AI Created Fake Videos by Detecting Eye Blinking [2]: This novel method uncovers fake face videos generated using deep neural network models by detecting eye blinking in the videos, a physiological signal typically absent or misrepresented in synthesized fake videos. The method shows promising performance in detecting videos generated with deep neural network-based software. Using capsule networks to detect forged images and videos [3]: This approach employs a capsule network to identify forged and manipulated images and videos across various scenarios, including replay attack detection and computer-generated video detection. Although their method involves using random noise during training, which may limit its effectiveness in real-time data, our proposed approach advocates for training on noiseless and real-time datasets.

Detection of Synthetic Portrait Videos using Biological Signals [5]: This method involves extracting biological signals from facial regions in authentic and fake portrait video pairs. By applying transformations to compute spatial coherence and temporal consistency, capturing signal characteristics in feature sets and PPG maps, and training probabilistic Support Vector Machines (SVM) and Convolutional Neural Networks (CNN), this approach aggregates authenticity probabilities to determine whether a video is fake or authentic.

Fake Catcher: This technique detects fake content with high accuracy, regardless of the generator, content, resolution, or quality of the video. However, the lack of a discriminator may result in the loss of biological signals, complicating the formulation of a differentiable loss function that adheres to the proposed signal processing steps.

III. PROPOSED SYSTEM

There is a plethora of tools available for creating deep fakes (DF), but there is a scarcity of tools for DF detection. Our approach to detecting DF will make a significant contribution to preventing the spread of DF across the internet. We will provide a web-based platform for users to upload videos and classify them as fake or genuine. This project can be expanded from developing a web-based platform to a browser plugin for automatic DF detection. Even major applications like WhatsApp and Facebook could integrate this project into their applications for easy pre-detection of DF before sending to another user. One of the primary goals is to evaluate its performance and effectiveness in terms of security, user-friendliness, accuracy, and reliability.

Our method focuses on detecting all types of DF, including replacement DF, conservation DF, and interpersonal DF. Figure 1 illustrates the basic framework architecture of the proposed system.

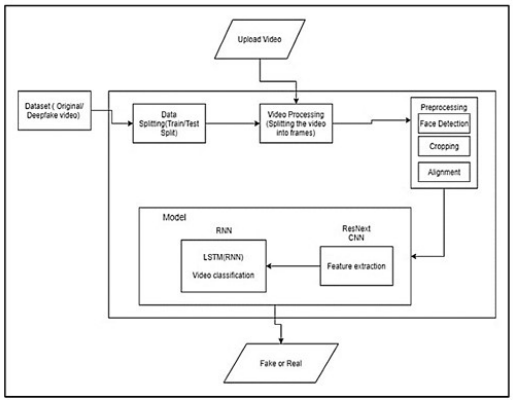

Fig. 1: System Architecture

A. Dataset:

We employ a blended dataset comprising an equal number of videos sourced from various datasets such as YouTube, FaceForensics++[14], and the Deepfake Detection Challenge dataset[13]. Our newly curated dataset consists of 50% original videos and 50% manipulated deepfake recordings. The dataset is divided into 70% for training and 30% for testing.

B. Preprocessing:

Dataset preprocessing involves segmenting the video into frames, followed by face detection and cropping of the frames with detected faces. To maintain consistency in the number of frames, the duration of the dataset videos is calculated, and a new processed face-cropped dataset is created containing frames equal to the calculated duration. Frames without detected faces are omitted during preprocessing. Since processing a 10-minute video at 30 frames per second, totaling 300 frames, requires significant computational power, we propose using only the first 100 frames for training the model for experimental purposes.

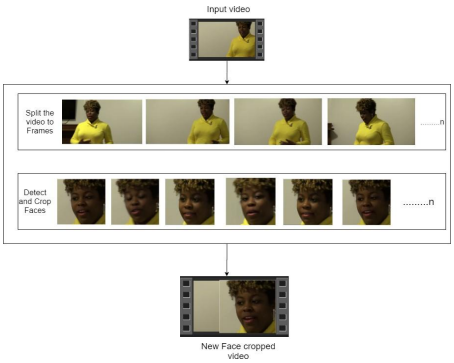

Fig. 2: Preprocessing of Video

C. Model:

The model consists of ResNext50_32x4d followed by one LSTM layer. The Data Loader ingests the preprocessed face-cropped videos and partitions them into training and testing sets. Subsequently, frames from the training videos are passed to the model for training and testing in mini-batches.

D. ResNext CNN for Feature Extraction:

Instead of redesigning the classifier, we propose using the ResNext CNN classifier to extract features and accurately identify the frame-level features. Next, we fine-tune the network by adding additional required layers and selecting an appropriate learning rate to properly converge the gradient descent of the model.

E. LSTM for Sequence Training:

Assuming a sequence of ResNext CNN feature vectors of input frames, we employ a 2-node neural network with the probabilities of the sequence being part of a deepfake video or an untampered video. The key challenge we need to address is designing a model to recursively process a sequence effectively. For this purpose, we propose the use of a 2048 LSTM unit with a 0.4 dropout probability, capable of achieving our goal. LSTM is used to process the frames sequentially, enabling temporal analysis of the video by comparing the frame at time 't' with the frame of 't-n' seconds, where 'n' can be any number of frames before 't'.

F. Prediction:

A new video is passed to the trained model for prediction. The new video is also preprocessed to align with the structure of the trained model. The video is segmented into frames followed by face cropping, and instead of storing the video in local storage, the cropped frames are directly passed to the trained model for detection.

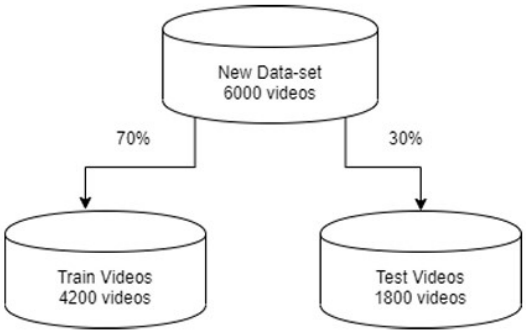

Fig. 3. Train Test split

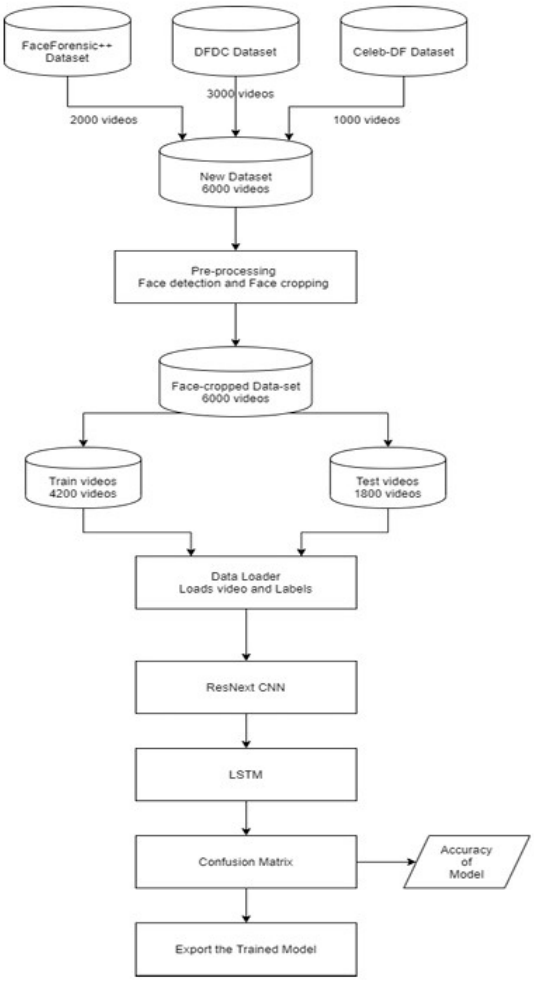

Fig. 4: Training Flow

IV. RESULT

The output depicted in Figure 3 exemplifies the outcome of a specialized model tasked with discerning between deepfake videos and genuine recordings, complemented by a confidence rating. In this context, the determination serves as a clear declaration regarding the nature of the video, indicating whether it is classified as a deepfake or a genuine recording. This determination is pivotal for users seeking to evaluate the authenticity of multimedia content. Alongside the determination, the confidence level furnishes users with valuable insight into the model's certainty regarding its classification. Expressed as a percentage or a score, this metric quantifies the model's level of confidence, aiding users in gauging the reliability of the classification. By presenting both the determination and the confidence level, this output equips users with essential information to make informed assessments of video content, thereby fostering greater discernment in an era marked by the proliferation of manipulated media.

Fig. 5: Expected Results

V. CONCLUSION

Our approach to video categorization employs a neural network-based methodology designed to differentiate between deep fake and genuine videos, while also providing a confidence level indicative of the model's certainty. Motivated by the mechanisms behind deep fake generation, particularly the utilization of Generative Adversarial Networks (GANs) coupled with Auto encoders, our method adopts a multi-stage process. Initially, we conduct frame-level detection utilizing a Resent Convolutional Neural Network (CNN). This stage involves analyzing individual frames of the video to identify potential manipulations or anomalies characteristic of deep fakes. Subsequently, the video undergoes classification using Recurrent Neural Networks (RNN), specifically Long Short-Term Memory (LSTM) networks. The RNN processes temporal dependencies within the video sequence, capturing patterns and nuances that may indicate the presence of deep fakery. Through this dual-stage approach, our proposed method leverages both spatial and temporal information inherent in videos, enhancing its ability to discern between authentic and manipulated content.

Furthermore, our method is supported by a comprehensive set of parameters outlined in the accompanying paper, meticulously selected and optimized to facilitate effective video categorization. These parameters encompass various aspects of the neural network architecture, including but not limited to convolutional filter sizes, LSTM hidden layer dimensions, learning rates, and dropout probabilities. Their careful calibration ensures that the model can effectively extract discriminative features indicative of deep fakes while maintaining robustness against noise and variations in input data.

REFERENCES

[1] Yuezun Li, Siwei Lyu, “ExposingDF Videos By Detecting Face Warping Artifacts,” in arXiv:1811.00656v3.
[2] Yuezun Li, Ming-Ching Chang and Siwei Lyu “Exposing AI Created Fake Videos by Detecting Eye Blinking” in arxiv.
[3] Huy H. Nguyen , Junichi Yamagishi, and Isao Echizen “ Using capsule networks to detect forged images and videos ”.
[4] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari and Weipeng Xu “Deep Video Portraits” in arXiv:1901.02212v2.
[5] Umur Aybars Ciftci, ˙Ilke Demir, Lijun Yin “Detection of Synthetic Portrait Videos using Biological Signals” in arXiv:1901.02212v2.

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In NIPS, 2014. David G¨uera and Edward J Delp. Deepfake video detection using recurrent neural networks. In AVSS, 2018.
[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, 2016.
[8]]An Overview of ResNet and its Variants : Vision and Pattern Recognition, pages 5967–5976, July 2017. Honolulu, HI.
[9] Long Short-Term Memory: From Zero to Hero with Pytorch: https://blog.floydhub.com/long-short-term memory-from-zero-to-hero-with-pytorch/
[10] Sequence Models And LSTM Networks https://pytorch.org/tutorials/beginner/nlp/sequence_mod els_tutorial.html
[11] C.Nagarajan and M.Madheswaran - 'Experimental verification and stability state space analysis of CLL-T Series Parallel Resonant Converter' - Journal of ELECTRICAL ENGINEERING, Vol.63 (6), pp.365-372, Dec.2012.
[12] C.Nagarajan and M.Madheswaran - 'Performance Analysis of LCL-T Resonant Converter with Fuzzy/PID Using State Space Analysis'- Springer, Electrical Engineering, Vol.93 (3), pp.167-178, September 2011.
[13] C.Nagarajan and M.Madheswaran - 'Stability Analysis of Series Parallel Resonant Converter with Fuzzy Logic Controller Using State Space Techniques'- Taylor & Francis, Electric Power Components and Systems, Vol.39 (8), pp.780-793, May 2011.
[14] C.Nagarajan and M.Madheswaran - 'Experimental Study and steady state stability analysis of CLL-T Series Parallel Resonant Converter with Fuzzy controller using State Space Analysis'- Iranian Journal of Electrical &Electronic Engineering, Vol.8 (3), pp.259-267, September 2012.
[15] Nagarajan C., Neelakrishnan G., Akila P., Fathima U., Sneha S. "Performance Analysis and Implementation of 89C51 Controller Based Solar Tracking System with Boost Converter" Journal of VLSI Design Tools & Technology. 2022; 12(2): 34–41p.
[16] C. Nagarajan, G.Neelakrishnan, R. Janani, S.Maithili, G. Ramya "Investigation on Fault Analysis for Power Transformers Using Adaptive Differential Relay" Asian Journal of Electrical Science, Vol.11 No.1, pp: 1-8, 2022.
[17] G.Neelakrishnan, K.Anandhakumar, A.Prathap, S.Prakash "Performance Estimation of cascaded h-bridge MLI for HEV using SVPWM" Suraj Punj Journal for Multidisciplinary Research, 2021, Volume 11, Issue 4, pp:750-756
[18] G.Neelakrishnan, S.N.Pruthika, P.T.Shalini, S.Soniya, "Perfromance Investigation of T-Source Inverter fed with Solar Cell" Suraj Punj Journal for Multidisciplinary Research, 2021, Volume 11, Issue 4, pp:744-749
[19] C.Nagarajan and M.Madheswaran, "Analysis and Simulation of LCL Series Resonant Full Bridge Converter Using PWM Technique with Load Independent Operation" has been presented in ICTES'08, a IEEE / IET International Conference organized by M.G.R.University, Chennai.Vol.no.1, pp.190-195, Dec.2007
[20] M Suganthi, N Ramesh, "Treatment of water using natural zeolite as membrane filter", Journal of Environmental Protection and Ecology, Volume 23, Issue 2, pp: 520-530,2022
[21] M Suganthi, N Ramesh, CT Sivakumar, K Vidhya, "Physiochemical Analysis of Ground Water used for Domestic needs in the Area of Perundurai in Erode District", International Research Journal of Multidisciplinary Technovation, pp: 630-635, 2019 https://github.com/ondyari/FaceForensics
[22] Y. Qian et al. Recurrent color constancy. Proceedings of the IEEE International Conference on Computer Vision, pages 5459–5467, Oct. 2017. Venice, Italy.