Deepfake Face Morphing Video Detection using ResNext50 & LSTM

Vamsi Krishna[1] Ravi Kumar[2] Siri Vaishnavi[3] Aasritha[4] Pooja Ratnam[5]
[1,2,3,4]Student [5]Professor
[1,2,3,4,5]Department of Computer Engineering
[1,2,3,4,5]Gayatri Vidya Parishad COE(Autonomous) Visakhapatnam, AP, india

Abstract— The growing computation power has made the deep learning algorithms so powerful that creating a indistinguishable human synthesized video popularly called as deep fakes have became very simple. Scenarios where these realistic face swapped deep fakes are used to create political distress, fake terrorism events, revenge porn, blackmail peoples are easily envisioned. In this work, we describe a new deep learning-based method that can effectively distinguish AI-generated fake videos from real videos. Our method is capable of automatically detecting the replacement and reenactment deep fakes. We are trying to use Artificial Intelligence(AI) to fight Artificial Intelligence(AI). Our system uses a Res-Next Convolution neural network to extract the frame-level features and these features and further used to train the Long Short Term Memory(LSTM) based Recurrent Neural Network(RNN) to classify whether the video is subject to any kind of manipulation or not, i.e whether the video is deep fake or real video. To emulate the real time scenarios and make the model perform better on real time data, we evaluate our method on large amount of balanced and mixed data-set prepared by mixing the various available data-set like Face-Forensic++[14], Deepfake detection challenge[13], and Celeb-DF[22]. We also show how our system can achieve competitive result using very simple and robust approach.

Keywords: Res-Next Convolution neural network. Recurrent Neural Network (RNN). Long Short Term Memory(LSTM). Computer vision

I. INTRODUCTION

The improvement of smart phone cameras and easy access to the internet worldwide have made social media and video sharing websites more popular. It's now simpler than ever to create and share digital videos. Also, computers are much more powerful today, allowing advanced artificial intelligence (AI) techniques like deep learning to do things that seemed impossible not long ago.However, these advancements have brought new challenges. One big problem is the rise of "DeepFakes." These are fake videos and audios created using AI that can manipulate what people see and hear. They often spread quickly on social media, leading to a lot of false information and spam. This can deceive and harm people

To deal with this issue, it's crucial to be able to detect DeepFakes. In this study, we introduce a new method based on deep learning that can accurately tell apart fake AI-generated videos (DeepFake videos) from real ones. Developing technology that can spot DeepFakes is essential to stop them from spreading online and causing harm..

To detect DeepFakes, understanding how Generative Adversarial Networks (GANs) create them is crucial. GANs take a video and an image of a specific person (the 'target'), and then produce a new video where the target's face is replaced with someone else's (the 'source'). Deep adversarial neural networks, which form the basis of DeepFakes, are trained on face images and videos of the target to seamlessly swap faces and mimic facial expressions from the source. After processing, these videos can look very realistic.The GAN breaks down the video into frames and replaces the target's face with the source image in each frame. It then reconstructs the entire video. Autoencoders are often used in this process. Our new method for detecting DeepFakes is inspired by this GAN process. It focuses on specific characteristics of DeepFake videos: due to limits in computing power and time, the algorithms used for DeepFakes can only synthesize faces of a fixed size. They also apply a type of warping to match the source's face, which can leave noticeable artifacts because the resolution of the warped face area doesn't always match the surrounding context.Our method detects these artifacts by comparing the generated face areas with their surroundings. We achieve this by splitting the video into frames and using a ResNext Convolutional Neural Network (CNN) to extract features. Additionally, we employ a Recurrent Neural Network (RNN) with Long Short-Term Memory (LSTM) to capture any inconsistencies over time introduced during the video reconstruction by the GAN. To train our ResNext CNN model, we simulate the resolution inconsistencies caused by affine face wrappings directly.

II. LITERATURE SURVEY

The rapid proliferation of deep fake videos and their illicit use poses significant threats to democracy, justice, and public confidence. Consequently, there is a growing demand for methods to analyze, detect, and counteract fake videos. Several approaches in deep fake detection are discussed below:

Exposing DF Videos by Detecting Face Warping Artifacts [1]: This approach detects artifacts by comparing the generated face areas and their surroundings using a dedicated Convolutional Neural Network (CNN) model. It focuses on discrepancies caused by the limited resolution of current deep fake algorithms and subsequent transformations needed to align faces in source videos.

Exposing AI Created Fake Videos by Detecting Eye Blinking [2]: This method exposes fake face videos generated by deep neural networks by detecting eye blinking. Blinking is a physiological signal often absent or poorly represented in synthesized fake videos. The method shows promising performance in detecting videos created with deep neural network-based software, although it relies solely on detecting the absence of blinking.

Using capsule networks to detect forged images and videos [3]: This approach employs capsule networks to detect manipulated images and videos, including scenarios like replay attacks and computer-generated video detection.

It aims to improve detection accuracy by focusing on distinguishing features rather than using random noise during training, which may affect performance on real-time data.

Detection of Synthetic Portrait Videos using Biological Signals [5]: This approach extracts biological signals from facial regions in both authentic and fake portrait video pairs. It applies transformations to assess spatial coherence and temporal consistency, capturing signal characteristics in feature sets and photoplethysmogram (PPG) maps. The method trains probabilistic Support Vector Machines (SVM) and CNNs to aggregate authenticity probabilities and distinguish between fake and authentic videos.

Fake Catcher: This method claims to detect fake content with high accuracy, regardless of the generator, content, resolution, or quality of the video. It emphasizes the preservation of biological signals but acknowledges challenges in formulating a differentiable loss function aligned with proposed signal processing steps.

Each of these methods contributes uniquely to the ongoing effort to combat deep fake videos, addressing various technical challenges and considerations such as biological signals, image quality, and real-time applicability.

III. PROPOSED SYSTEM

It becomes very important to spot the difference between the deepfake and pristine video. We are using AI to fight AI. Deepfakes are created using tools like FaceApp[23] and Face Swap[24], which using pre-trained neural networks like GAN or Autoencoders for these deepfakes creation. Our method uses a LSTM based artificial neural network to process the sequential temporal analysis of the video frames and pre-trained Res-Next CNN to extract the frame level features. ResNext Convolution neural network extracts the frame-level features and these features are further used to train the Long Short Term Memory based artificial Recurrent Neural Network to classify the video as Deepfake or real. To emulate the real time scenarios and make the model perform better on real time data, we trained our method with large amount of balanced and combination of various available dataset like FaceForensic++[14], Deepfake detection challenge[13], and Celeb-DF[22].

Further to make the ready to use for the customers, we have developed a front end application where the user the user will upload the video. The video will be processed by the model and the output will be rendered back to the user with the classification of the video as deep fake or real and confidence of model.

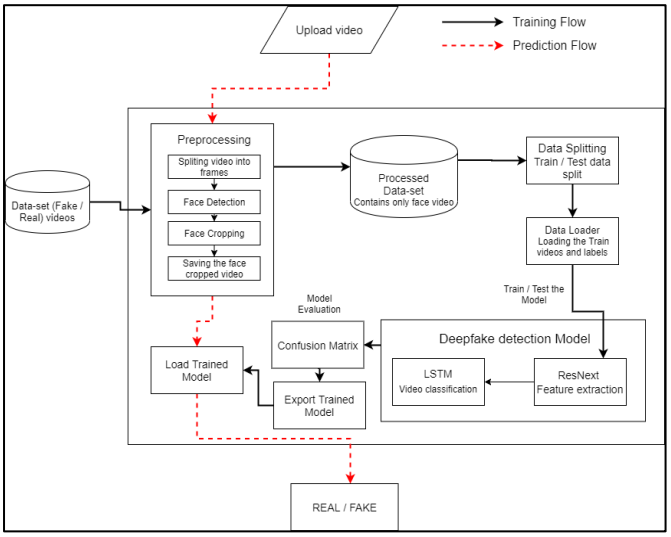

Fig. 1: System Architecture

A. *Dataset:*

We have collected data from various sources including FaceForensic++ (FF)[14], the Deepfake Detection Challenge (DFDC)[13], and Celeb-DF[22]. By combining these datasets, we created a new dataset aimed at achieving accurate and real-time detection across different video types. To prevent model bias, we ensured an equal split of 50% real and 50% fake videos. The dataset is divided into a 70% training set and a 30% test set.

B. *Preprocessing:*

The preprocessing steps involve splitting the video into frames, detecting faces within these frames, and cropping the frames around the detected faces. To standardize the number of frames, we calculate the mean number of frames across the dataset and create a processed face-cropped dataset with a frame count matching this mean. Frames without faces are excluded during preprocessing. Given the high computational demands of processing 300 frames from a 10-second video at 30 frames per second, we propose using only the first 100 frames for model training.

C. *Model:*

The model architecture comprises ResNext50_32x4d followed by an LSTM layer. The Data Loader handles the preprocessed face-cropped videos and divides them into training and testing sets. Frames from these videos are then fed into the model in mini-batches for both training and evaluation.

D. *ResNext CNN for Feature Extraction:*

Rather than developing a new classifier from scratch, we propose utilizing the ResNext CNN classifier for feature extraction to capture detailed frame-level features. We will fine-tune this network by adding necessary layers and optimizing the learning rate to ensure effective gradient descent convergence. The 2048-dimensional feature vectors obtained from the final pooling layers will then be used as input for the sequential LSTM.

E. *LSTM for Sequence Processing*

Assuming a sequence of ResNext CNN feature vectors as input and a 2-node neural network for classifying the sequence as either a deepfake or an untampered video, the challenge is designing a model to process sequences meaningfully.

We propose using a 2048-unit LSTM with a 0.4 dropout rate to achieve this goal. The LSTM will analyze frames sequentially, enabling temporal comparisons between frames at time 't' and those at 't-n' seconds, where 'n' can vary.

F. Predict:

For prediction, a new video is fed into the trained model after preprocessing it to match the format used during training. The video is divided into frames, face-cropped, and instead of saving these frames locally, they are directly sent to the trained model for detection.

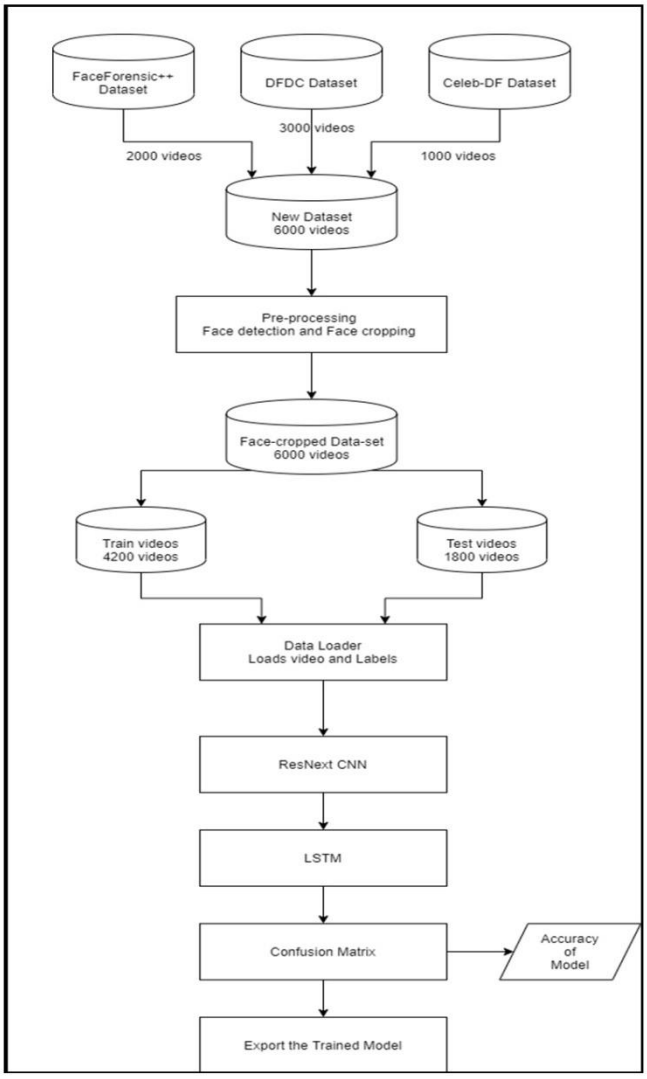

Fig. 2: Training Flow

IV. Result

The model outputs whether a video is deepfake or real, along with the confidence level of the prediction. An example of this output is illustrated in Figure 3.

Fig. 3: Expected Results

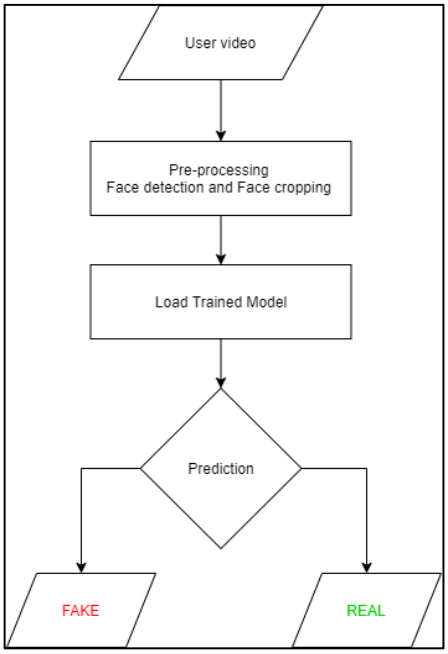

Fig. 4: Prediction flow

V. Conclusion

We introduced a neural network-based approach for classifying videos as either deepfake or real, providing confidence levels for each prediction. This method is inspired by the techniques used in GANs and autoencoders for deepfake creation.

Our approach utilizes ResNext CNN for frame-level detection and RNN with LSTM for video classification. The method is designed to achieve high accuracy in real-time data based on the parameters outlined.

VI. Limitations

Our approach does not account for audio, meaning it cannot detect audio deepfakes. We plan to develop methods to address audio deepfake detection in future work.

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.

[7] David G¨uera and Edward J Delp. Deepfake video detection using recurrent neural networks. In AVSS, 2018.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, 2016.

[9] An Overview of ResNet and its Variants : https://towardsdatascience.com/an-overview-of-resnet-and-its-variants-5281e2f56035

[10] Long Short-Term Memory: From Zero to Hero with Pytorch: https://blog.floydhub.com/long-short-term-memory-from-zero-to-hero-with-pytorch/

[11] Sequence Models And LSTM Networks https://pytorch.org/tutorials/beginner/nlp/sequence_models_tutorial.html

[12] https://discuss.pytorch.org/t/confused-about-the-image-preprocessing-in-classification/3965

[13] https://www.kaggle.com/c/deepfake-detection-challenge/data

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

[15] Y. Qian et al. Recurrent color constancy. Proceedings of the IEEE International Conference on Computer Vision, pages 5459–5467, Oct. 2017. Venice, Italy.

[16] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5967–5976, July 2017. Honolulu, HI.

[17] R. Raghavendra, Kiran B. Raja, Sushma Venkatesh, and Christoph Busch, "Transferable deep-CNN features for detecting digital and print-scanned morphed face images," in CVPRW. IEEE, 2017.

[18] Tiago de Freitas Pereira, Andr´e Anjos, Jos´e Mario De Martino, and S´ebastien Marcel, "Can face anti spoofing countermeasures work in a real world scenario?,"in ICB. IEEE, 2013.

[19] Nicolas Rahmouni, Vincent Nozick, Junichi Yamagishi, and Isao Echizen, "Distinguishing computer graphics from natural images using convolution neural networks," in WIFS. IEEE, 2017.

[20] F. Song, X. Tan, X. Liu, and S. Chen, "Eyes closeness detection from still images with multi-scale histograms of principal oriented gradients," Pattern Recognition, vol. 47, no. 9, pp. 2825–2838, 2014.

[21] D. E. King, "Dlib-ml: A machine learning toolkit," JMLR, vol. 10, pp. 1755–1758, 2009.

[22] Yuezun Li , Xin Yang , Pu Sun , Honggang Qi and Siwei Lyu "Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics" in arXiv:1909.12962

[23] Face app: https://www.faceapp.com/

[24] Face Swap: https://faceswaponline.com/