Unmasking Deepfakes: A Deep Learning Approach for Accurate Detection and Classification of Synthetic Videos

SK. Abdul Sattar[1],T. Guru Preetham[2], V.Kalyan[3],P.Venu[4],B.Avinash[5]

[1,2,3,4]B.Tech. Students, Department of IT, Vasireddy Venkatadri Institute of Technology, Guntur
[5]Assistant Professor, Department of IT, Vasireddy Venkatadri Institute of Technology, Guntur

---***---

ABSTRACT:

The growth of deepfakes fuels the spread of misinformation, undermining trust in media and information sources. Additionally, they worsen societal divisions by sharing fake content, leading to confusion and polarization. Deepfakes are becoming increasingly common, making it harder to spot them because they look so real. This paper addresses this problem by introducing a method to detect differences in facial features during video creation. Detection of deepfakes can be tricky due to their high realism, but our approach helps identify these fake videos by spotting changes in facial structures. Our model employs a Res-Next Convolutional Neural Network to extract frame-level features, which are then utilized to train a Long Short-Term Memory (LSTM)-based Recurrent Neural Network (RNN). This RNN classifies videos whether they are subjected to any manipulation or not. We have used a dataset called "Celeb-DF" to train our model to detect differences created around the face during deepfake construction. Integrated with a user-friendly interface utilizing ReactJs on the front end and Flask on the backend, our solution ensures robust defense against potential threats posed by deepfakes while prioritizing accessibility and usability.

***Key Words*: DeepFake Video, Res-Next Convolution Neural Network, Long Short-Term Memory (LSTM), Recurrent Neural Network (RNN), Flask**

1. INTRODUCTION:

This paper tackles the growing threat posed by deepfake videos, which convincingly alter reality and can be used to deceive or manipulate viewers. These videos, capable of fabricating events or speeches, present serious risks in areas such as politics and personal privacy. To address this issue, we propose a straightforward method to detect deepfakes by examining changes in facial features during video creation.

Our approach focuses on spotting subtle differences in facial expressions that distinguish genuine footage from manipulated content. Using advanced technologies like neural networks, specifically a combination of Res-Next CNN and LSTM-based RNN, our system can effectively identify deepfakes with a high level of accuracy. Additionally, we train our model on a diverse dataset called Celeb-DF , containing various examples of manipulated videos. We also prioritize accessibility by creating a user-friendly interface with ReactJs and Flask, ensuring that individuals can easily utilize our deepfake detection system to protect against the dangers of deceptive media.

2. LITERATURE REVIEW:

The widespread emergence of deepfake videos and their misuse poses a serious threat to democracy, justice, and public trust. As a result, there is a growing need for improved methods to analyze, detect, and intervene in fake videos.

Employing 26 distinct deep convolutional models, they proposed a technique to enhance detection precision, despite potential computational hurdles in managing and fine-tuning multiple models [1].

Introducing an innovative method, they utilize Cascaded Deep Sparse Auto Encoder (CDSAE) with temporal CNN for feature extraction, aiming to boost accuracy, although facing limitations due to the intricate nature of deepfake techniques [2].

Relying on feature fusion with MesoInception, their model extracts deep characteristics from various iterations of target faces, showcasing effectiveness in identifying altered faces in videos, albeit with performance variations depending on the complexity of facial manipulation techniques [3].

Their approach suggests merging Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) to expedite training by leveraging pre-trained CNN models while enhancing detection precision through combined CNN and RNN architectures, though encountering challenges in adapting to new deepfake variations [4].

Presenting the YOLO-CNN-XGBoost strategy, they leverage YOLO face detection for precise face region extraction in video frames, backed by XGBoost for thorough deepfake detection. However, challenges arise due to reliance on pre-trained models, possibly limiting effectiveness against emerging deepfake types [5].

Our model revolutionizes deepfake detection with a user-friendly interface, leveraging deep learning techniques. By integrating a Res-Next CNN and LSTM-based RNN, it offers accurate classification of videos, empowering users to combat manipulated media effectively, fostering a more trustworthy online space.

3.CREATION OF DEEPFAKES:

To detect the deepfake videos it is very important to understand the creation process of the deepfake. Majority of the tools including the GAN and autoencoders takes a source image and target video as input. These tools split the video into frames, detect the face in the video and replace the source face with target face on each frame. Then the replaced frames are then combined using different pre-trained models. These models also enhance the quality of video my removing the left-over races by the deepfake creation model. Which result in creation of a deepfake looks realistic in nature. We have also used the same approach to detect the deepfakes. Because pretrained neural network models are used to construct deepfakes, the differences between them are nearly hard to see with naked eye. However, some traces or artifacts that might not be seen to the unaided eye are really left in the video by the deepfakes production tools. This research aims to detect these subtle traces and discernible anomalies of these films and categorize them as authentic or deepfake.

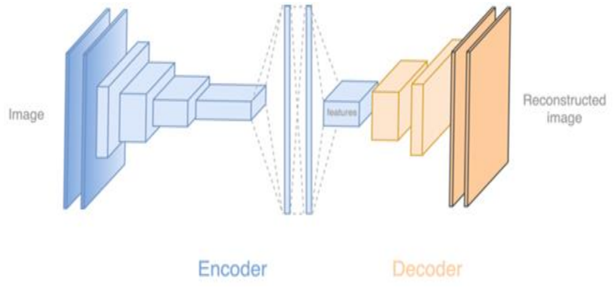


Figure 1: DEEPFAKE GENERATION

4.PROPOSED SYSTEM:

The proposed system leverages a hybrid approach combining ResNeXt, a powerful convolutional neural network (CNN) architecture, with Long Short-Term Memory (LSTM), a type of recurrent neural network (RNN). ResNeXt is employed for extracting high-level spatial features from video frames, while LSTM is utilized to analyze temporal dynamics crucial for discerning between real and fake videos. By integrating these two architectures, the system effectively captures both spatial and temporal aspects of deepfake manipulation. This approach enhances the model's ability to identify subtle anomalies in video content, making it a potent tool in combating digital misinformation.

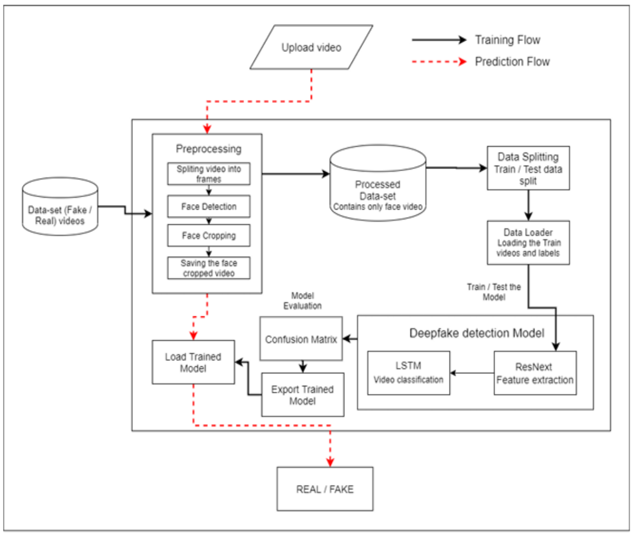


Figure 2: SYSTEM ARCHITECTURE

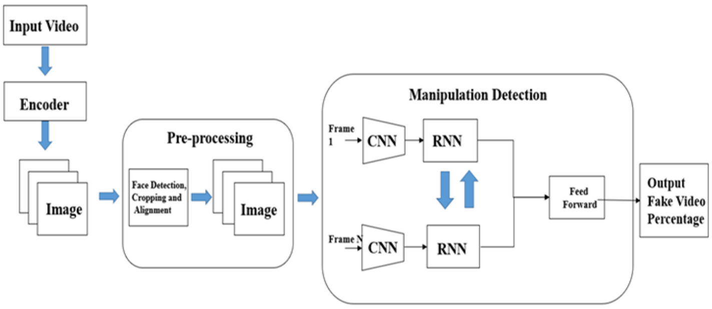


Figure 3: DATA FLOW DIAGRAM

4.1 DATA COLLECTION:

The Celeb-DF dataset comprises 408 authentic videos and 795 synthesized videos generated through a modified DeepFake algorithm. These videos have an average duration of 13 seconds, recorded at a frame rate of 30fps. Unlike previous datasets characterized by high resolution and numerous visual artifacts, the synthesized videos in Celeb-DF exhibit lower visual artifacts, resulting in higher quality. This lower quality of deepfakes presents a more challenging problem for detection purposes.

4.2 DATA PREPROCESSING:

The preprocessing pipeline extracts frames from input videos, detects faces using the face recognition library, and crops them for focus. Frames are transformed into PyTorch tensors with normalization. Sequential frames are stacked into sequences of fixed length (20 frames) for consistency. This processed data is then used to create a PyTorch dataset. These steps ensure standardized input preparation for robust deep learning model inference on video data, particularly for detecting fake videos.

4.3 MODEL ARCHITECTURE:

Our model utilizes a ResNeXt-50 CNN pretrained on ImageNet to extract important features from video frames. These features are then analyzed by an LSTM RNN to understand the time-based patterns crucial for distinguishing real from fake videos. To improve its ability to learn and handle complex data, the model incorporates leaky ReLU activation and dropout regularization techniques. These enhancements help the model generalize better and capture non-linear relationships in the data, ultimately improving its accuracy in identifying fake videos amidst real ones.

4.3.1 FEATURE EXTRACTION WITH RESNEXT CNN:

We employed the pre-trained ResNeXt model for extracting frame-level features, leveraging its design optimized for superior performance on deep neural networks. Specifically, we utilized the ResNeXt50_32x4d model, characterized by its 32 x 4 dimensions and 50 layers, within our project. To tailor the model to our task, additional layers were incorporated, and appropriate learning rates were selected to ensure smooth convergence of the gradient descent during fine-tuning. The input for our sequential LSTM comprises the 2048-dimensional feature vectors generated by the final pooling layers of the ResNeXt model.

4.3.2 SEQUENTIAL PROCESSING WITH LSTM:

2048-dimensional feature vectors were utilized as input to the LSTM. To accomplish the objectives, a single LSTM layer with 2048 latent dimensions and 2048 hidden units, incorporating a dropout probability of 0.4, was utilized. The temporal analysis of the video was made possible by sequentially processing frames using the LSTM, enabling comparison between the frame at time 't' seconds and the frame 't-n' seconds earlier, where 'n' represents the number of frames preceding time 't'.

4.4 HYPERPARAMETER TUNING:

Hyperparameter tuning is conducted iteratively to optimize model accuracy. We employ the Adam optimizer with an adaptive learning rate strategy, set to 1e-5, facilitating convergence to a superior global minimum during gradient descent. A Weight decay value of 0.001 is utilized for regularization. Cross-entropy loss is utilized for classification, while batch training with a size of 4 maximizes computation efficiency

4.5 PREDICTION:

The trained algorithm receives a fresh video to forecast. Additionally, a fresh video is preprocessed to import the format of the learned model. The video is divided into frames, then the faces are cropped. The cropped frames are sent straight to the trained model for detection rather than being stored locally.

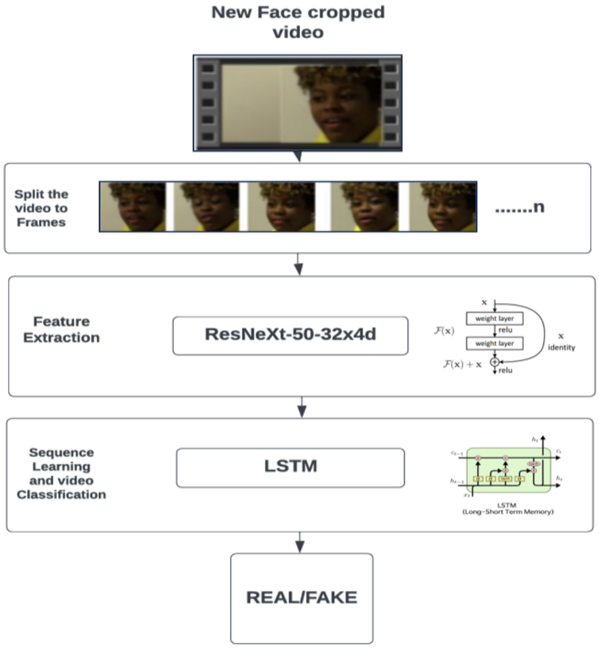


Figure 4: OVERVIEW OF OUR MODEL

5. RESULT ANALYSIS:

The thorough analysis shows the model's strong performance in detecting deepfakes, achieving an impressive 91% accuracy. By using both LSTM for understanding time-related patterns and ResNeXt for capturing detailed visual features, the model effectively identifies signs of deepfake manipulation. This combined approach allows for a thorough examination of video content, improving the model's accuracy and trustworthiness. With its reliable performance, the model offers a promising solution for countering digital misinformation and preserving the authenticity of video content across various platforms.

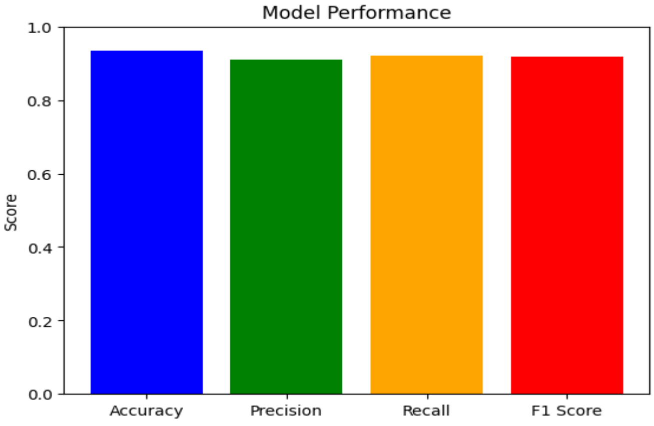


Figure 5: MODEL PERFORMANCE

The model underwent training using approximately 5000 authentic frames and 6000 fabricated frames. For testing purposes, 30% of the total data was used.

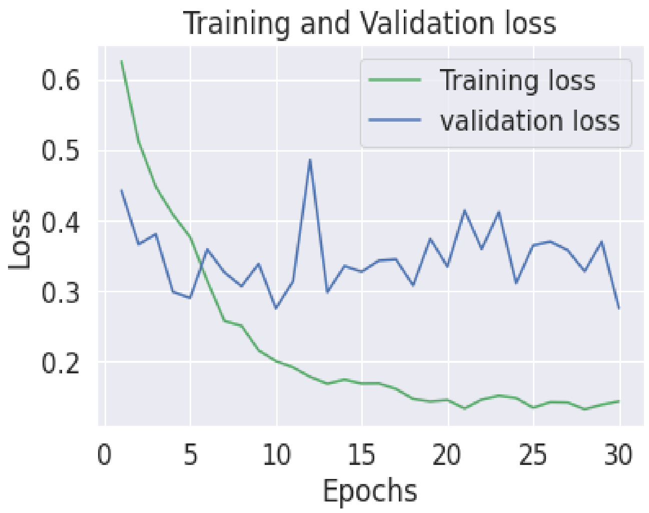


Figure 6: TRAINING AND VALIDATION LOSS CURVES

The loss curves reveal the model's successful training. Both training and validation loss steadily decrease, indicating efficient learning and good generalizability to unseen data. This suggests the model is effectively utilizing the data and has the potential for accurate predictions.

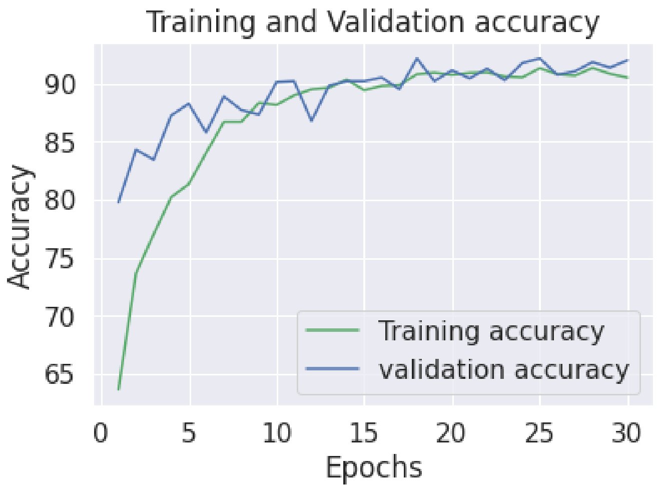
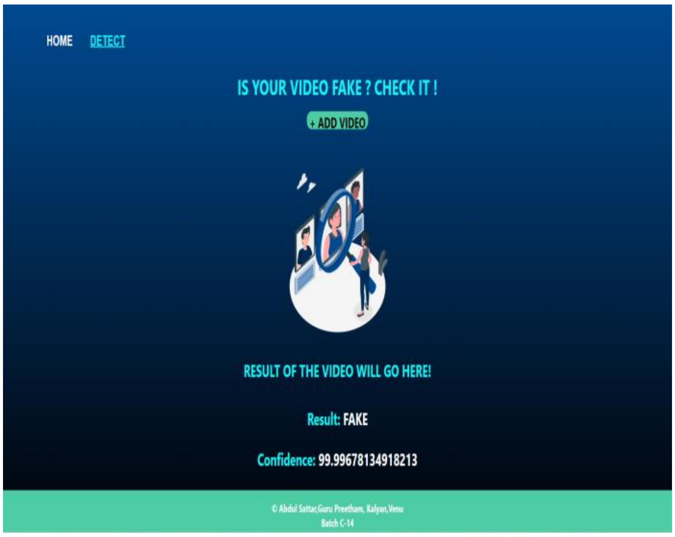


Figure 7: CONVERGENCE OF MODEL ACCURACY

The graph depicts a successful training process for the deep learning model. Both the training accuracy and validation accuracy curves increase steadily, indicating the model is effectively learning from the data. The validation accuracy remains close to training accuracy, suggesting the model is generalizing well to unseen data

6.OUTPUT SCREENS:



IS YOUR VIDEO FAKE ? CHECK IT !

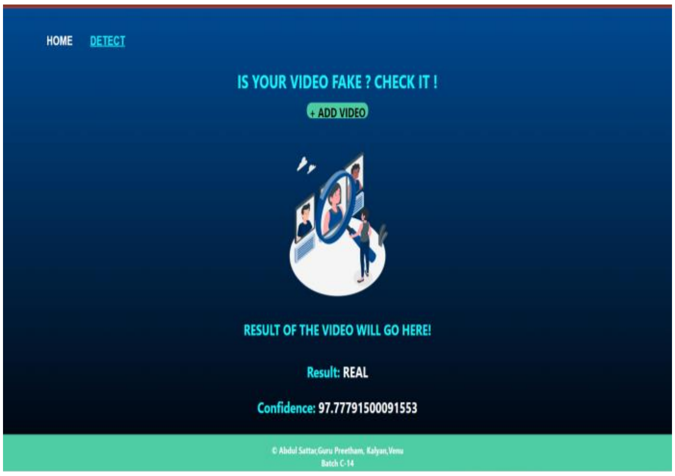
RESULT OF THE VIDEO WILL GO HERE!

Result: FAKE

Confidence: 99.99678134918213

© Abdul Sattar,Guru Preetham, Kalyan,Venu
Batch C-14

Figure 8: OUTPUT OF DEEPFAKE VIDEO



IS YOUR VIDEO FAKE ? CHECK IT !

RESULT OF THE VIDEO WILL GO HERE!

Result: REAL

Confidence: 97.77791500091553

© Abdul Sattar,Guru Preetham, Kalyan,Venu
Batch C-14

Figure 9: OUTPUT OF REAL VIDEO

7.CONCLUSION:

Our research delivers a groundbreaking approach to the critical challenge of deepfake detection, making it accessible to a wider audience. Leveraging deep learning techniques, we have developed a robust system for classifying videos as real or deepfake, complete with confidence scores. This system, built with a user-friendly interface powered by ReactJS on the frontend and a robust Flask backend, integrates a Res-Next Convolutional Neural Network (CNN) for feature extraction and a Long Short-Term Memory (LSTM) based Recurrent Neural Network (RNN) for video classification for detecting deepfakes.

This multifaceted strategy, combined with its user-friendly interface, significantly improves our ability to prevent the spread of manipulated media. By enabling in-depth video analysis through a readily accessible platform, our system empowers users to more effectively identify deepfakes. Furthermore, this research has the potential to contribute to a more trustworthy online environment.

8. FUTURE WORK:

Our current focus is on visual deepfake detection. However, recognizing, the growing threat posed by audio deepfakes, future work will explore techniques for audio analysis to enhance the system's comprehensiveness. This includes examining voice characteristics, speech patterns, and environmental sounds to identify signs of manipulation or synthesis. Given the potential risks associated with audio manipulation, integrating audio analysis with our existing visual detection framework is crucial for developing a robust multi-modal deepfake detection system. Utilizing both visual and auditory indicators, our goal is to enhance the system's ability to detect a broader spectrum of deepfake content, thereby reducing the spread of misinformation.

9.REFRENCES:

[1]. Kshirsagar, M., Suratkar, S., & Kazi, F. (2022, August). Deepfake Video Detection Methods using Deep Neural Networks. In 2022 Third International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICICT) (pp. 27-34). IEEE.

[2]. Balasubramanian, S. B., Prabu, P., Venkatachalam, K., & Trojovský, P. (2022). Deep fake detection using cascaded deep sparse auto-encoder for effective feature selection. PeerJ Computer Science, 8, e1040.

[3]. Ismail, A., Elpeltagy, M., S. Zaki, M., & Eldahshan, K. (2021). A new deep learning-based methodology for video deepfake detection using xgboost. Sensors, 21(16), 5413.

[4]. Al-Dhabi, Y., & Zhang, S. (2021, August). Deepfake video detection by combining convolutional neural network (CNN) and recurrent neural network (RNN). In 2021 IEEE International Conference on Computer Science, Artificial Intelligence and Electronic Engineering (CSAIEE) (pp. 236-241). IEEE.

[5]. Mohiuddin, S., Ganguly, S., Malakar, S., Kaplun, D., & Sarkar, R. (2021, December). A feature fusion-based deep learning model for deepfake video detection. In International Conference on Mathematics and its Applications in new Computer Systems (pp. 197-206). Cham: Springer International Publishing.

[6].Masood M, Nawaz M, Malik KM, Javed A, Irtaza A (2021) Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward. arXiv preprint arXiv:2103.00484

[7]. Perov I, Gao D, Chervoniy N, Liu K, Marangonda S, Umé C, Dpfks M, Facenheim SC, RP L, Jiang J, Zhang S, Wu P, Zhou B, Zhang W (2020) DeepFaceLab: a simple, flexible and extensible face swapping framework.

[8].Prajwal KR, Mukhopadhyay R, Namboodiri VP, Jawahar CV (2020, October) A lip sync expert is all you need for speech to lip generation in the wild. In Proceedings of the 28th ACM International Conference on Multimedia, pp 484-492.

[9].Kumar A, Bhavsar A, Verma R (2020) Detecting deepfakes with metric learning. In 2020 8th international workshop on biometrics and forensics (IWBF), IEEE, pp 1-6.

[10]. Halpern M, Kelly F, van Son R, Alexander A (2020). Residual networks for resisting noise: analysis of an embeddings-based spoofing countermeasure.

[11].Chen T, Kumar A, Nagarsheth P, Sivaraman G, Khoury E (2020, November) Generalization of audio deepfake detection. In Proc. Odyssey 2020 The Speaker and Language Recognition Workshop, pp 132-137.

[12].Rana MS, Sung AH (2020, August) Deepfakestack: a deep ensemblebased learning technique for deepfake detection. In 2020 7th IEEE International Conference on yber Security and loud omputing (S loud) 2020 6th IEEE International onference on Edge omputing and Scalable loud (Edge om), IEEE, pp 70-75.

[13].Antoniou A (2019) Zao's deepfake face-swapping app shows uploading your photos is riskier than ever. The Conversation.

[14]. Reimao RAM (2019) Synthetic speech detection using deep neural networks.

[15]. Güera D, Delp EJ (2018, November) Deepfake video detection using recurrent neural networks. In 2018 15th IEEE international conference on advanced video and signal-based surveillance (AVSS), IEEE, pp 1-6.