An Advanced Hybrid Approach to Detect Deepfake Videos

Raksha Pandey*[1],Dr. Alok kumar Singh Kushwaha[2]

[1,2]School of Studies (Engg & Tech), Guru Ghasidas Vishwavidyalaya,Bilaspur

ABSTRACT

Today, due to AI especially Deep – Learning, How Face manipulations in video seems quite realistic. They are known as Deep Fake which is the process of producing such a video through Artificial Intelligence. The term Deepfake employed primarily to clarify what kind of videos can be created with the use of such technology. Instead, it can prove to be somewhat complicated to distinguish one such Deep Fake (DF) pieces from the other. To identify this challanges, in this article, a methodology of Hybrid CNN and RNN is discussed. In particular, moderate CNN referred to as ResNet152 for feature extraction from the video and Long Short Term Memory as the RNN model for classifying the video manipulation as well as temporal artifacts in frames.

Keywords: Deep fake , LSTM, Video forgery, ResNet 152

1. INTRODUCTION

In recent advancements, deepfake videos arise from the application of sophisticated neural network methodologies, employing tools such as Generative Adversarial Networks (GANs) and Autoencoders [1]. These neural networks proficiently merge target images with source videos, yielding deceptively authentic deepfake content. Detecting the existence of deepfake videos poses a significant challenge, given their remarkably realistic visual appearance. This underscores the critical necessity for developing robust detection methods proficient in distinguishing between authentic and manipulated content. Presented strategy for deepfake detection revolves around capitalizing on the intrinsic limitations of the tools instrumental in creating deepfakes. These tools inadvertently introduce subtle artifacts within the frames of deepfake videos, which, though invisible to human observers, can be discerned by neural networks meticulously trained for this specific purpose. Implementing proposed detection methodology involves leveraging a Resnet152 Convolutional Neural Network for extracting intricate frame-level features from video sequences. These extracted features form the groundwork for training a specialized Recurrent Neural Network (RNN) based on Long Short-Term Memory (LSTM). The RNN is entrusted with the classification of videos, distinguishing between authentic content and deepfake creations. Proposed method amassed a substantial dataset of deepfake videos from the Deepfake Detection Challenge[2] and Celeb-DF[3] datasets. Proposed model's performance has been evaluated using real-time data, drawing from platforms such as YouTube. This exhaustive assessment ensures the practical viability and effectiveness of presented approach in real-world scenarios. To fortify the robustness of proposed neural network model, Training is done with a comprehensive amalgamation of available datasets. This strategic approach equips proposed model with the capacity to discern nuanced features across various image types. Following fig1. Shows the deep fake image.

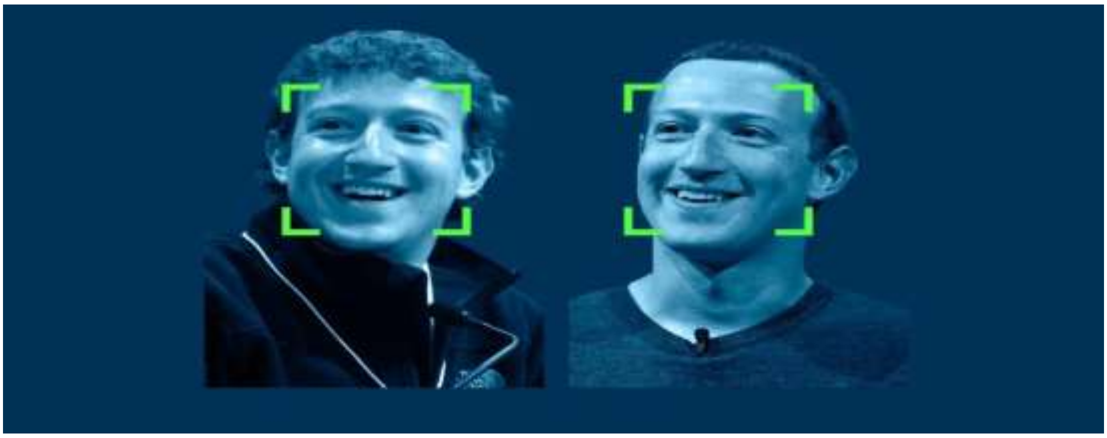


fig. 1: Deepfake image

2. RELATED WORK

Deep fake detection, acknowledging the success of current techniques in controlled settings. However, the real-world scenario, marked by dynamic social media dynamics and evolving manipulation tactics, necessitates a more adaptable approach.

Videos and photographs have become crucial legal evidence, scrutinized by digital forensics experts at the intersection of computer science and law enforcement. Intelligenceservices exploit this technology to influence national and international security decisions. Theresearch community emphasizes not only identifying manipulations but also understanding the intent behind digital deceptions, calling for a user-centric lens within the social context where deep fakes emerge.Machine learning and AI algorithms emerge as sentinels in discerning digital media authenticity. The symphony of detection and understanding involves harmonizing effective methodologies with social context discernment. Current methodologies, including

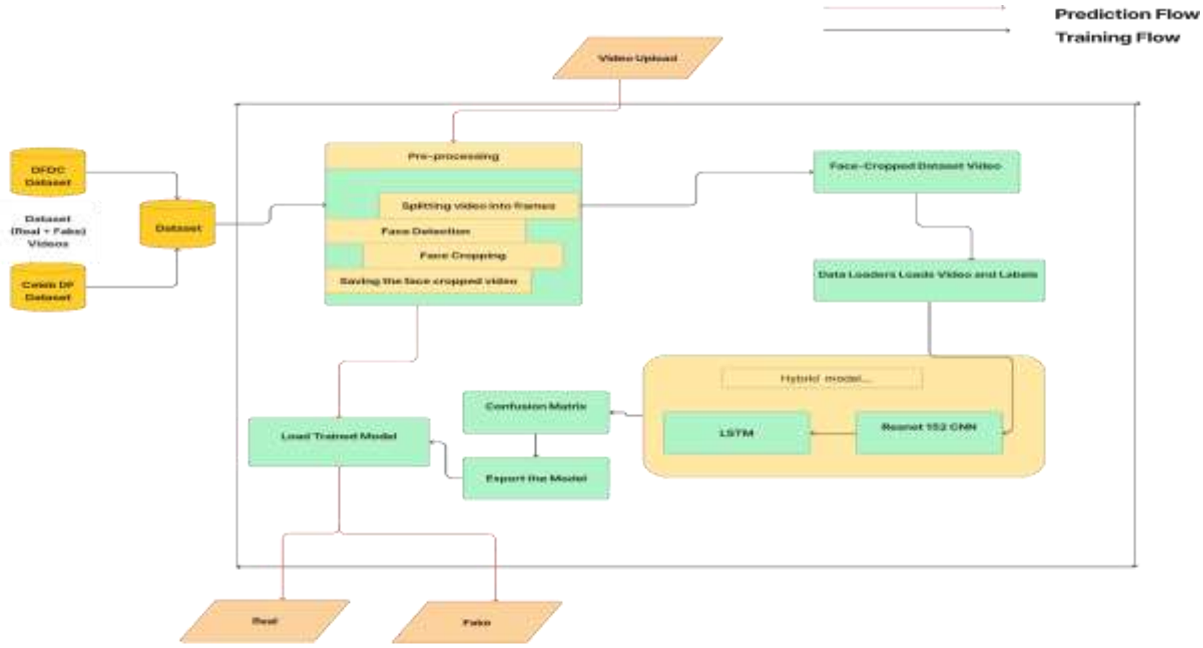


fig.2: flow chart of deepfake detection.

face-warping artifacts[4], eye-blinking analysis[5], Capsule Networks[6], Recurrent Neural Networks, and biological signal extraction[7], contribute to the field but face limitations. Fig.2 shows the deepfake detection flowchart.Ongoing efforts focus on enhancing accuracy and addressing real-world challenges in the intricate dance of perception and reality.

3. PROPOSED MODEL

The proposed deepfake detection system combines the strengths of LSTM networks and ResNet-152 architecture to offer a versatile and accurate solution. Leveraging LSTM's abilityto capture temporal dependencies and ResNet-152's robust spatial feature extraction, the model provides a holistic approach to identifying deepfake manipulations in videos. Itsadaptability to evolving manipulation tactics makes it well-suited for real-world scenarios, with applications spanning social media, legal proceedings, and national security. By enhancing detection accuracy and understanding the intent behind digital deceptions, this system contributes to a more user-centric and protective approach to combating the challenges posed by deepfake technology.

3.1 Preprocessing

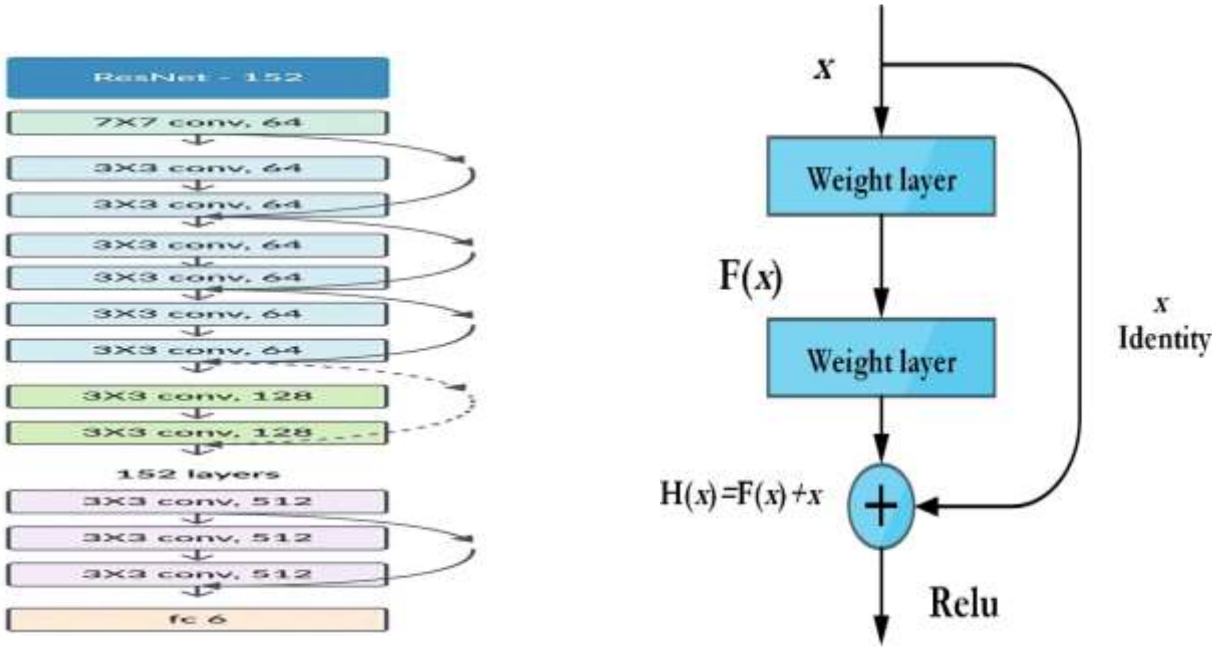
The dataset preprocessing ensures uniformity and efficiency. Videos are divided into frames, faces are detected and cropped, and a new dataset is generated with frames equal to the mean video length. Frames without detectable faces are excluded. Acknowledging the computational demands, proposed model using the first 150 frames of a 10-second video for training, balancing efficiency, and representative sample size for meaningful experimentation.

3.2 Model

The model uses a ResNet152 [8] followed by an LSTM layer. The Data Loader manages preprocessed face-cropped videos, dividing them for training and testing. Frames are processed in mini-batches, ensuring efficient data handling for optimization during bothtraining and testing.

3.3 Resnet152 CNN for feature extraction

Presented model utilized a pre-trained ResNet CNN model known for high performance [9-11]. Fine-tuning involved adding layers and optimizing the learning rate for effective convergence during gradient descent. The ResNet output, 2048-dimensional feature vectors, serves as input for proposed sequential LSTM. The formula $H(x) = F(x) + x$ captures the essence of residual blocks, combining the original input (x) with the transformed output (F(x)).Following fig. 3 (a) and (b) shows the process.

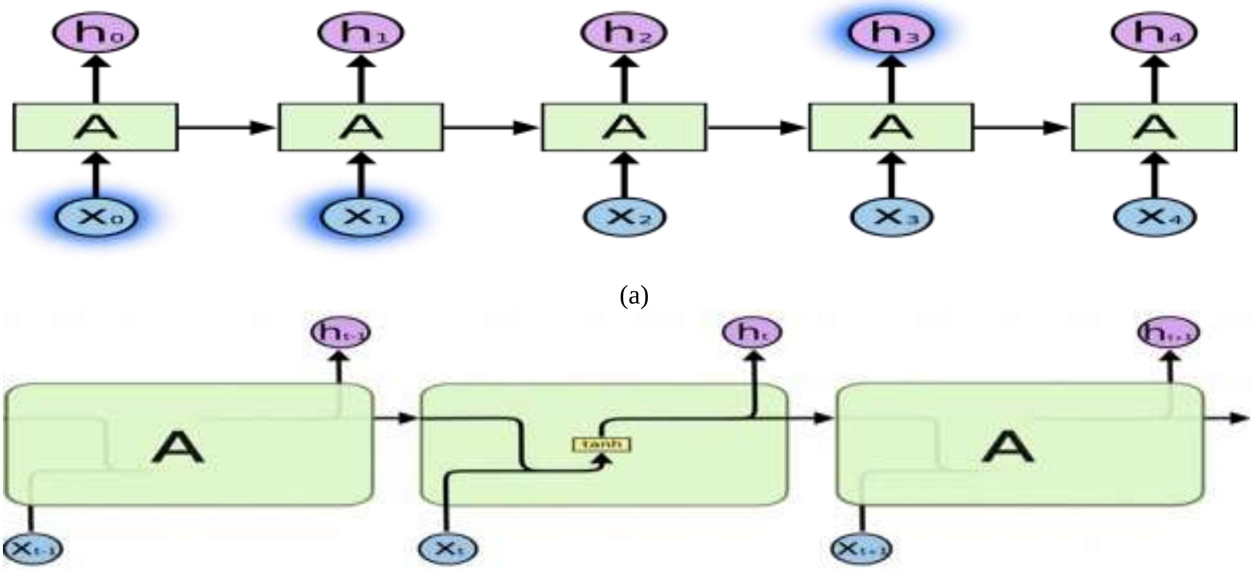


(a) (b)

fig. 3 (a) and (b) shows the process of feature extraction

3.4 LSTM for sequential

Proposed model use a sequence of ResNet152 CNN feature vectors as input for a model with a 2-node neural network, predicting the probabilities of the sequence being from a deepfake or unaltered video. To handle recursive processing effectively, a 2048-unit Long Short-Term Memory (LSTM)[12-13] with 0.5 dropout probability is employed. This configuration enables temporal analysis by sequentially processing frames, comparing the frame at a time 't'with the frame at 't-n'. This recursive approach captures temporal dependencies, facilitating a comprehensive examination of the video's sequential evolution. This is shown in fig 4(a) and (b).



(a)

(b)

fig.4 (a) and (b) shows the role of LSTM

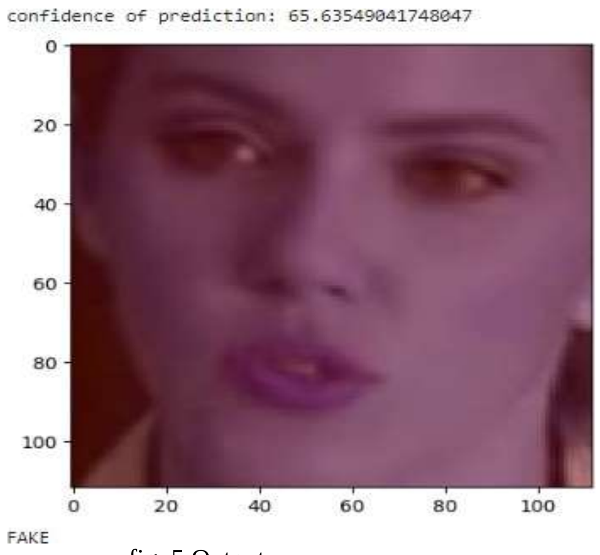
4. RESULT AND ANALYSIS

4.1 Dataset

Different dataset applied such as Deep Fake Detection Challenge and CelebDF. The dataset comprises 50% authentic and 50% manipulated deep fake videos, split into a 70% training and a 30% test set.

4.2 Prediction

In the operational phase of presented deepfake detection approach, a novel video undergoes prediction through the trained model. The new video is preprocessed to align with the format expected by the trained model. This preprocessing entails video segmentation into frames, subsequent face cropping and a distinctive feature of bypassing local storage by directly forwarding the cropped frames to the trained model for detection. In fig 5 output is displayed. Comparison with existing work shown in Table 1.



```
confidence of prediction: 65.63549041748047
```

```
FAKE
```

fig. 5.Output

Table 1: Comparison with state-of-the-art

Model	Dataset	Accuracy	Sample Size
ResNet-18	Celeb-DF+ DFDC	Approximately 61%	100 frames
ResNet-30	Celeb-DF+ DFDC	Approximately 79%	100 frames
ResNet-50	Celeb-DF+ DFDC	Approximately 85.8%[1]	100 frames
ResNet-152	Celeb-DF+ DFDC	Approximately 98%	100 frames
Hybrid Model (ResNet152 + LSTM)	Celeb-DF+ DFDC	Approximately 85%	100 frames

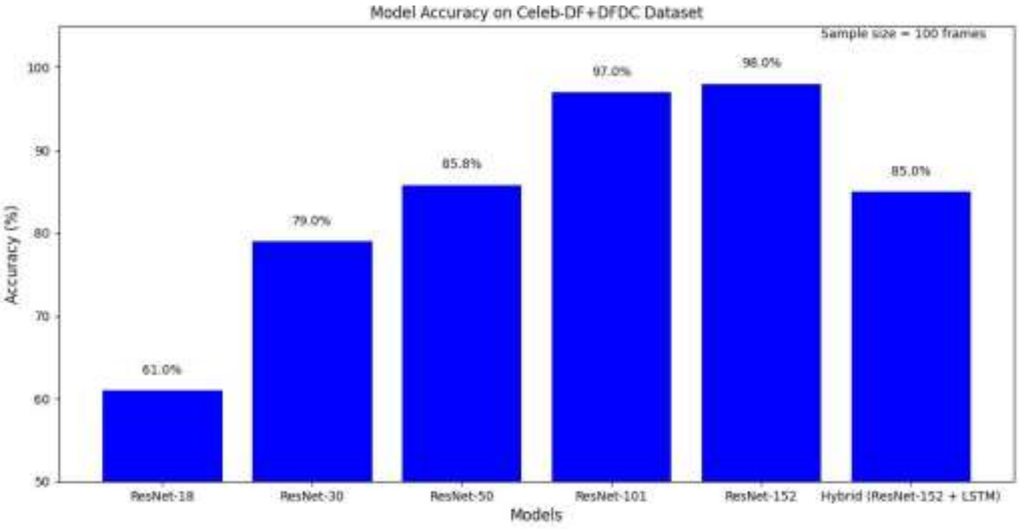


fig.6: Graph of proposed model with existing work

5. CONCLUSION

Presented deepfake detection model is a meticulously crafted solution designed to identify manipulated video content accurately. Through data preprocessing, an advanced neural network architecture, and careful video data handling, proposed approach is a notable contribution to ongoing efforts addressing the challenges posed by deepfake technology.

REFERENCES

[1] Dave Bergmann, Cole Stryker, What is an autoencoder, 23 November 2023, IBM.

[2] https://www.kaggle.com/c/deepfake-detectionchallenge/data

[3] https://github.com/danmohaha/celeb-deepfakeforensics

[4] A. Das, K. S. A. Viji and L. Sebastian, "A Survey on Deepfake Video Detection Techniques Using Deep Learning," 2022 Second International Conference on Next Generation Intelligent Systems (ICNGIS), Kottayam, India, 2022, pp. 1-4, doi: 10.1109/ICNGIS54955.2022.10079802.

[5] Yuezun Li, Ming-Ching Chang, and Siwei Lyu "Exposing AI Created Fake Videos by Detecting Eye Blinking" in arxiv

[6] Huy H. Nguyen , Junichi Yamagishi, and Isao Echizen " Using capsule networks to detect forged images and videos ".

[7] Umur Aybars Ciftci, ˙Ilke Demir, Lijun Yin "Detection of Synthetic Portrait Videos using Biological Signals" in arXiv:1901.02212v2.

[8] ResNet with TensorFlow (Transfer Learning) | by mrgrhn | The Startup | Medium

[9] He, Kaiming & Zhang, Xiangyu & Ren, Shaoqing & Sun, Jian. (2016). "Deep Residual Learning for Image Recognition". 770–778. 10.1109/CVPR.2016.90.

[10] Szegedy, Christian & Liu, Wei & Jia, Yangqing & Sermanet, Pierre & Reed, Scott & Anguelov, Dragomir & Erhan, Dumitru & Vanhoucke, Vincent & Rabinovich, Andrew. (2014). "Going Deeper with Convolutions".

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2017. ImageNet classification with deep convolutional neural networks. Commun. ACM 60, 6 (June 2017), 84–90. https://doi.org/10.1145/3065386.

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

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