Deepfake Technology: Safeguarding Trust and Security in a Digital Age

Macklin Enico Salex.P
Department of Artificial Intelligence & Data Science
Chennai, India
macklinmacklin1@gmail.com

Lochan Abisheck.S
Department of Artificial Intelligence & Data Science
Chennai, India
lochanabisheck@gmail.com

R.Kethan deepak ram
Department of Artificial Intelligence & Data Science
Chennai, India
manga121943@gmail.com

Harshavardhan.V
Department of Artificial Intelligence & Data Science
Chennai, India
harshavardhan.v2324@gmail.com

***Abstract*—The advancements in Deep learning software have facilitated the creation of highly convincing facial swaps in videos, commonly referred to as Deepfake videos. Recent innovations in deep learning and AI tools have greatly enhanced the realism of synthetic content making it increasingly challenging to distinguish from authentic content.**

However, detecting these deep fake videos pose a significant barrier. Because it is difficult to teach the algorithm to detect the deep fake. Using CNN and LSTM, we are making progress in detecting fake videos. The system employs a convolutional neural network (CNN) at the frame level to extract features. These observations are noted, and this can train a recurrent neural network (RNN), which can learn and classify whether or not a video has been altered with and identify the physical inconsistencies in the frame introduced by Deep fake tools. Our approach implements the pre-detection methods, we can prevent the upload of deep fake videos from publishing on any digital platform. Thus, we can safeguard users from the spread of false information in deceptive media and ensure trust among them.

Keywords—component, formatting, style, styling, insert (*key words****)***

I. Introduction

The evolution of mobile and social media is happening at a rapid pace, driven by advancements in technology. With the use of these technologies, producing realistic-looking fake images, videos, and audio is getting easier. Artificial intelligence tools are used to create and disseminate deep fake content over the internet. The deep fake mainly affects high-profile people, such as politicians and actors. Deepfake spreads misinformation, which breaks security, privacy, and trust when it comes from a trustworthy source.

Software like face swap,adobe photoshop is used for the face replacement in videos, it can still be a somewhat tedious task, especially for users who are new to the software or working with complex footage .While it's true that processing a 20-second video with 25 frames per second would involve editing around 500 images, modern deepfake software is designed to handle tasks of this scale. These software tools often leverage powerful GPUs and parallel processing techniques to accelerate the editing process and reduce the overall processing time.

Deep neural networks enable the automated mapping and recognition of facial expressions from source to target, which is a critical component of deep fake video production. These networks utilize powerful machine learning algorithms to create incredibly realistic and convincing deep fake films, raising concerns about possible misuse and the need for effective detection and mitigation mechanisms.

As a result, detecting deep fake videos is really important. To do this, we have come up with a new method that uses deep learning for pre detection which is to figure out if a video is fake or real. This method looks at different aspects of the video, like the way people move and talk, to spot any signs that it might be fake.

Having technology like this is crucial because it helps us find fake videos before they upload or spread fake information online. It's like having a detective that can spot fake videos and stop them before they upload on any digital platforms. This way, we can better protect ourselves from being tricked by fake videos on the internet. An example of deep fake is shown in Figure 1.

Fig 1. A deep fake manipulate images examples [14]

It's essential to understand how Generative Adversarial Networks (GANs) produce deep fake videos in order to identify them. It takes an original video as input, along with a separate image of the person whose face will be manipulated (the "target"). The GAN then generates a modified version of the original video where the face of the target person is replaced with the face from another individual (the "source"). This process involves training the GAN on a dataset of original videos and images to learn how to generate realistic-looking deep fake videos.

The GAN breaks the videos down into frames and replaces each frame with an input image. It goes on to rebuild the video. Autoencoders are frequently utilized for this purpose. We introduce a fresh deep learning approach for distinguishing deepfake videos from genuine real-world videos. This solution operates on a similar principle as the generation of deepfakes using GANs.[7] The deep fake algorithm generates facial images of restricted size and requires undergoing affine warping to adapt and preserve the configuration of the source's face. However, because of the inconsistency in resolution between the surrounding context and the warped face area, this warping process leaves behind noticeable artifacts in the resulting deep fake video.[1]

Our method involves dividing the video into individual frames and analyzing the facial areas generated by the deep fake algorithm along with their surrounding regions. By comparing these areas, we can detect any noticeable artifacts resulting from the warping process used in deepfake creation. We employ Long Short-Term Memory (LSTM) in conjunction with Recurrent Neural Networks (RNN) to capture the temporal inconsistencies between frames produced by the Generative Adversarial Network (GAN) during the deepfake reconstruction process. Additionally, we utilize a ResNext Convolutional Neural Network (CNN) to extract relevant features from the analyzed frames, aiding in the detection of deep fake videos. This combined approach allows us to effectively identify anomalies indicative of deepfake manipulation within video content and prevent from uploading.

A simple GAN architecture diagram is shown in Figure 2, which gives a clear idea about how it works, and processes the image as fake or real.

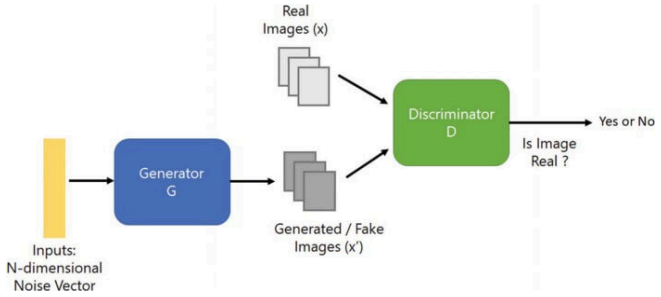

Fig. 2. GAN's General Architecture

I. LITERATURE SURVEY:

The recent development and use of deep fake videos pose a critical threat to democracy,justice and public trust. As a result the demand for fake video detection,analysis and intervention has been increasing rapidly.
Following are some of the relevant words in deep fake detection:

The approach outlined in the paper "ExposingDF Videos by Detecting Face Warping Artifacts" involves utilizing Convolutional Neural Networks (CNNs) to identify face warping artifacts by comparing the surrounding facial regions with the actual face area. This method focuses on detecting two specific types of face artifacts. It is grounded in the observation that low-resolution images often require further alterations to match the source face, prompting the need for advanced artifact detection techniques.

Similarly, the paper "Exposing AI-Created False Videos by Detecting Eye Blinking" introduces a novel strategy for uncovering deepfake videos generated using deep neural network models. This method relies on detecting the presence of blinking eyes in the video, leveraging blinking as a physiological signal that is challenging to fabricate convincingly in fraudulent videos. Figure 3 shows the total number of papers published from the year 2016 to 2024.

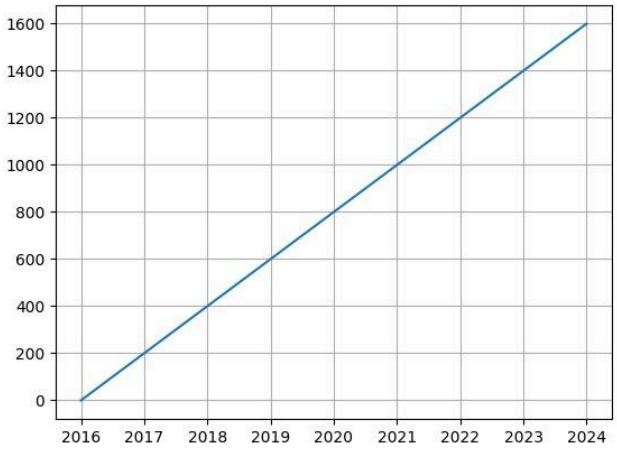

Fig 3. Graph showing the count of papers published from 2016-2024 based on deepfake project

The approach demonstrates efficacy in detecting videos generated by the Deep Neural Network-based program DF, particularly focusing on eye blinking datasets. The method yields promising outcomes by solely leveraging the absence of blinking as a detection indicator. However, it's crucial to acknowledge that comprehensive detection of deep fake videos requires consideration of additional factors, including teeth appearance, facial wrinkles, and other relevant parameters. Our project encompasses the evaluation of all these aspects to enhance detection accuracy.

Our approach is designed to be trained on datasets that are both noiseless and real-time. Table 1 shows a comparison of the several publications based on the key factors that distinguish each project.

Table 1. Comparison Between Different Projects

Parameters	Exposing DF Videos by Face Warping Artifacts [1]	DF Detecting using Eye Blinking [2]	DF detection by capsule networks [3]
Key Identification	Face Wrapping artifacts	Involuntary eye blinking movements	Inbuild capsule networks
Technology used	Neural Networks	Neural Networks	Neural Networks
Accuracy	This model is tested for moderate to low accuracy	This model is tested for medium accuracy	This has the highest accuracy among all
Consumption of resources	High consumption of computing resources	Moderate to High consumption of computing resources	Highest consumption of resources accounting for proprietary capsules.

Cele-DF is another project created to detect the deep fake video but it predicts based on the low resolution by improving the synthesized face to 256 × 256 pixels, checking for color mismatches and reducing the temporal blurring of fraudulent videos. [11]

The Effective and Fast Deep fake detection method, based on Haar Wavelet Transform, utilizes the Haar wavelet transform to identify deepfake content. This method focuses on extracting sharpness from blurred images, detecting edges within the images, and analyzing the synthesized surrounding areas using the Haar function. The project utilizes the UADFV dataset, comprising 49 fake and 49 real videos. The proposed accuracy of this method is reported to be 90.5%. Additionally, it operates on video frames, inspecting each frame individually and extracting the surroundings of the face.

II. Proposed System

There are numerous methods for creating DeepFake videos, but detecting them remains a challenge with only a limited number of viable approaches. The technique employed here for DeepFake detection offers a promising solution to safeguard the internet against the proliferation of such misleading content.

The project introduces a Django application, empowering users to upload videos and determine their authenticity as either real or fake. Additionally, the project offers a web-based platform accessible through browser plugins specifically designed for DeepFake pre-detections.

The project's utility extends to integration within popular applications like WhatsApp and Facebook, where it serves as a pre-detection tool for identifying DeepFake (DF) videos before users share them with connected groups or individuals. Its primary objective lies in enhancing performance and reliability through usability, accuracy, and security, thus ensuring that users can confidently verify the authenticity of videos prior to sharing them on these platforms.

This approach identifies different sorts of DeepFake, including retrenchment and replacement. Figure 4 illustrates the system architecture.

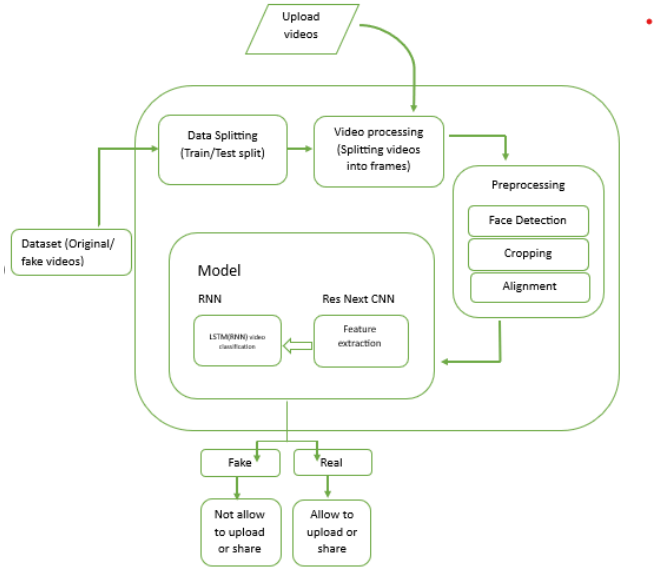

Fig. 4. System Architecture

A.Dataset Used:

The dataset utilized in this study comprises an amalgamation of videos sourced equally from various platforms, including Face Forensics + +, YouTube, and various challenge datasets. By amalgamating these videos, a new dataset is curated, consisting of an equal distribution of original videos and tampered DeepFake videos. Subsequently, the dataset is partitioned appropriately, allocating thirty percent for testing purposes and seventy percent for training the detection model.

Table 2 shows the total dataset created for deep fake till date.

Table 2. Basic details of the version dataset based on DeepFake [11]

Dataset	# Real		# DeepFake		Release Date
	Video	Frame	Video	Frame	
UADFV	49	17.3k	49	17.3k	2018.11
DF-TIMIT-LQ DF-TIMIT-HQ	320*	34.0k	320 320	34.0k 34.0k	2018.12
FF-DF	1,000	509.9k	1,000	509.9k	2019.01
DFD	363	315.4k	3,068	2,242.7k	2019.09
DFDC	1,131	488.4k	4,113	1,783.3k	2019.10
Celeb-DF	590	225.4k	**5,639**	2,116.8k	2019.11

B. Preprocessing Part

The pre-processing performed on the dataset involves several steps, including splitting each video into a series of frames, detecting faces within these frames, and subsequently cropping only the facial regions. To ensure consistency in the total frame count across all videos, the mean frame count of the dataset is calculated. Additionally, a new dataset is created containing only the cropped facial regions. This new dataset is standardized to have a frame count equal to the previously calculated mean, ensuring uniformity in the dataset's composition.

In the pre-processing phase, frames without detected faces are excluded from further analysis. Given the computational demands associated with processing a 15-second video at a frame rate of 20 frames per second, resulting in approximately 300 frames, significant computational resources would be required. To align with the hardware capabilities at hand, the model is trained using only the initial 100 frames of each video. This approach allows for more manageable computational requirements while still providing valuable data for training the model.

C. Model used:

The model used in this project involves several parameters, including a single layer of LSTM combined with resnext50 during the model creation. The Data Loader is responsible for loading preprocessed videos with cropped faces, splitting them into a test set and a train set. The frames obtained after preprocessing are then used for testing and training the model.

Table 3 shows the error rate.

Table 3. Error rate of various models [13]

model	top-1 err.	top-5 err.
VGG-16 [41]	28.07	9.33
GoogLeNet [44]	-	9.15
PReLU-net [13]	24.27	7.38
plain-34	28.54	10.02
ResNet-34 A	25.03	7.76
ResNet-34 B	24.52	7.46
ResNet-34 C	24.19	7.40
ResNet-50	22.85	6.71
ResNet-101	21.75	6.05
ResNet-152	**21.43**	**5.71**

D. Feature Extraction using ResNext CNN

We employ ResNext CNN to extract features from the video frames, enabling us to accurately capture frame-level features. Our network is carefully fine-tuned by adding extra layers and selecting optimal learning rates to ensure the model's gradient descent converges effectively.

After the final pooling layer, a feature vector of size 2048 dimensions is generated, which serves as the input for the LSTM layer. This process allows us to leverage the extracted features for further analysis and model training.

E.Sequence processing using LSTM

In our approach, we utilize LSTM for sequence processing. To illustrate, consider having 2 neural nodes and a sequence of feature vectors extracted by ResNext CNN from video frames as input. These feature vectors represent the likelihood of the sequence being part of an untampered or deep fake video.

Our primary objective is to design a model capable of consistently processing the sequence in the correct sequential order, thereby effectively analyzing the data for deepfake detection. Figure 5 represents the LSTM for sequence processing.

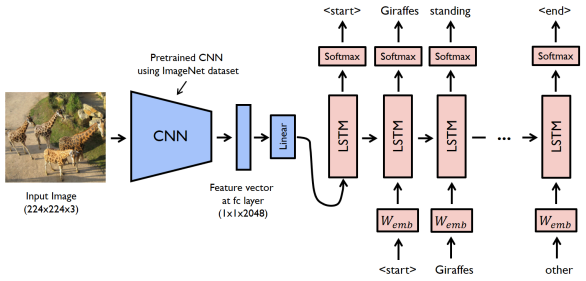

Fig.5. LSTM for Sequence Processing

To achieve this objective, we employ an LSTM unit consisting of 2048 units, incorporating a dropout probability of 0.4 to prevent overfitting. Utilizing LSTM allows us to process video frames sequentially, enabling temporal analysis of the video. Specifically, we compare frames at time 't' with those at 't-n' seconds, where 'n' denotes the total number of frames preceding time 't'. This approach facilitates the detection of temporal patterns and variations in the video, aiding in the identification of potential deep fake content.

F.Pre-detection of deep fake videos

The utilization of Long Short-Term Memory (LSTM) and Recurrent Neural Networks (RNNs) plays a crucial role in pre-detecting deep fake videos before they are uploaded onto the internet. By employing LSTM and RNN architectures, our system can effectively analyze the temporal sequences of frames within videos, enabling it to identify subtle patterns and discrepancies indicative of deepfake manipulation.

LSTMs, a type of RNN, are particularly adept at capturing long-range dependencies and temporal dynamics in sequential data. They excel at remembering information over extended periods, making them well-suited for discerning alterations or inconsistencies within video sequences. This capability enables our system to detect anomalies in facial expressions, movements, or other visual cues that are characteristic of deep fake videos.

Furthermore, RNNs provide the framework for learning and analyzing sequential data in real-time, enabling our system to process videos frame by frame as they are being uploaded. This real-time processing capability allows for swift detection of potential deep fakes before they are disseminated online, thereby mitigating the spread of misleading or fraudulent content.

By leveraging the power of LSTM and RNN architectures, our method offers a proactive approach to pre-detecting deep fake videos, helping to safeguard online platforms and users from the harmful effects of misinformation and deception.

G. Model Accuracy based on frame count

Table 4. Model Accuracy

No of videos	No of frames	Accuracy
6000	10	84.21461
6000	20	87.79160
6000	40	89.34681
6000	60	90.59097
6000	80	91.49818

H.Prediction:

To predict the authenticity of a video, the input video must undergo preprocessing to match the format accepted by the trained model. This typically involves splitting the video into frames and cropping out the faces.

Instead of storing the entire video locally and consuming memory, a more efficient approach is to directly pass the cropped frames to the trained model for analysis.

The model then outputs the confidence level regarding the presence of deepfake elements in the video, along with details indicating whether the video is deemed real or fake.

Figure 6 represents the Training flow.

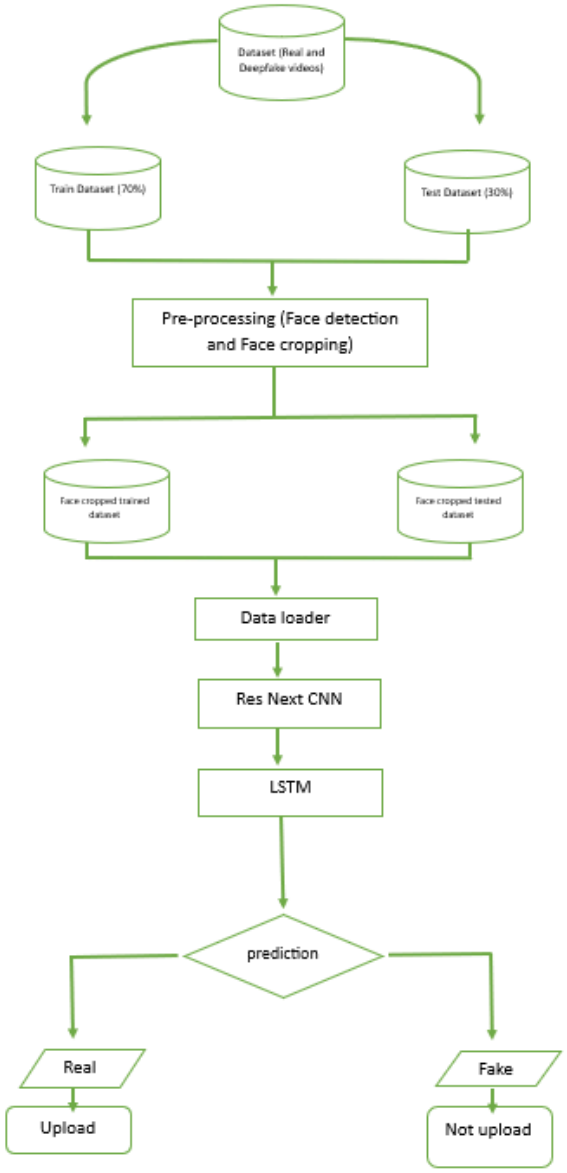

Fig. 6. Training Flow

IV. RESULT

A user-friendly GUI will be offered for users to upload their videos, and they will have the option to specify the frame size, such as 20, 40, 60, or 80. Uploading corrupted videos, excessively long videos, or images will result in an error message to ensure data integrity.

Upon processing, the final output will include the model's confidence level regarding the presence of deepfake elements in the video, along with a determination of whether the video is classified as real or fake. Figures illustrating these scenarios will be provided to enhance user understanding and interpretation of the results.

Fig. 7. GUI for video uploading

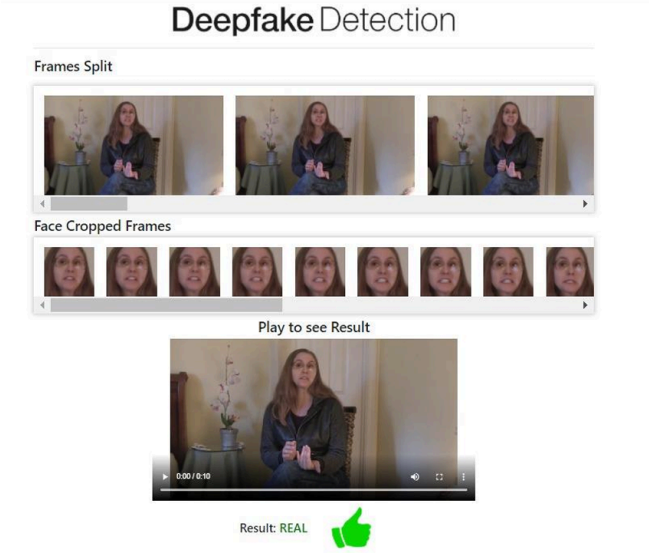

Fig. 8. Detection of Real Video

Fig. 9. Detection of Fake video

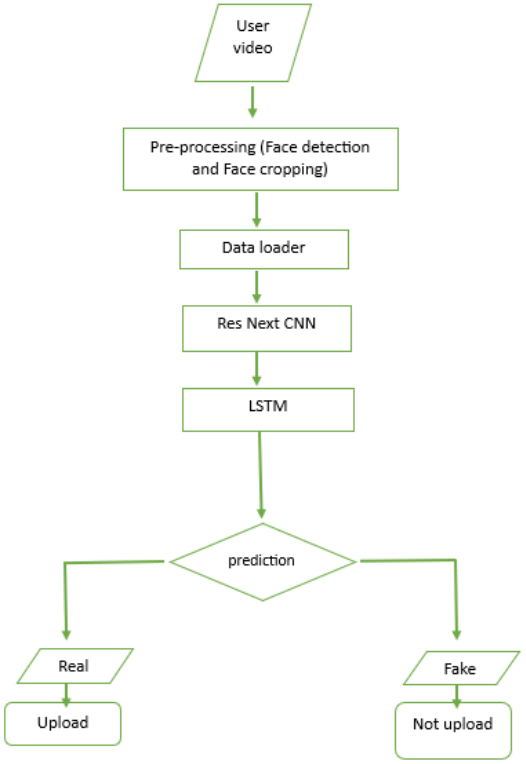

Fig. 10. Prediction Flow

V. CONCLUSION

The neural network-based approach is designed to classify videos as either real or deepfake while also indicating the confidence level of the model's classification. Deepfakes are videos that have been manipulated using techniques such as Generative Adversarial Networks (GANs) and autoencoders to alter their content, often by superimposing one person's face onto another person's body.

The proposed method utilizes a combination of convolutional neural networks (CNNs) and recurrent neural networks (RNNs) for the task. Specifically, it employs a ResNext CNN for frame-level detection, allowing it to analyze individual frames within the video. Then, an RNN, including Long Short-Term Memory (LSTM) cells, is used for video classification. RNNs are well-suited for sequential data analysis, making them suitable for processing video data over time.

The primary objective of this project is to accurately detect alterations in videos, indicative of deepfake manipulation, and then classify them as either real or fake based on predefined parameters outlined in the accompanying research paper. The passage suggests that the accuracy of the classification will improve when real-time data is incorporated into the model, likely through further training and refinement.

The proposed method aims to pre detect the deep fake videos before it is uploaded in any digital platforms or shared on the internet.It addresses the growing concern surrounding the proliferation of deepfake videos by providing a robust system for identifying and categorizing them with a high degree of accuracy.

VI. LIMITATIONS

Our approach has focused solely on analyzing the visual elements of videos, neglecting the audio component. Consequently, our method is not capable of detecting audio deepfakes. In the future, we plan to explore methods for identifying audio manipulations within videos to address this limitation.

REFERENCES

[1] Yuezun Li, Siwei Lyu, “ExposingDF Videos By Detecting Face Warping Artifacts,” in arXiv conference May, 2019

[2] Yuezun Li, Ming-Ching Chang and Siwei Lyu “Exposing AI Created Fake Videos by Detecting Eye Blinking” in IEEE conference 2018

[3] Huy H. Nguyen , Junichi Yamagishi, and Isao Echizen “ Using capsule networks to detect forged images and videos ”,IEEE conference, 2018.

[4] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner, “Face2Face: Real-time face capture and reenactment of RGB videos,” in CVPR. IEEE, 2019.

[5] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari and Weipeng Xu “Deep Video Portraits” in arXiv conference, May 2020

[6] MikaWesterlund, The Emergence of Deepfake Technology:, Technology Innovation Management, Version 9, November 2019.

[7] Thanh Thi Nguyena, Quoc Viet Hung Nguyenb, Dung Tien Nguyena, Duc Thanh Nguyena, Thien Huynh-ThecSaeid Nahavandid, Thanh Tam Nguyene, Quoc-Viet Phamf, Cuong M. Nguyen, arXiv conference, February 2022.

[8] Rushikesh Potdar, Ajay Gidd, Shreya Kulkarni, Rohit Chavan, Prof. Nikam, International Research Journal of Modernization in Engineering Technology and Science, Volume:03/Issue:07/July 2021.

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

[10] Umur Aybars Ciftci, ˙Ilke Demir, Lijun Yin “Detection of Synthetic Portrait Videos using Biological Signals”, IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. X, No. X, July 2020.
[11] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi and Siwei Lyu, arXiv conference, arXiv:1909.12962v4, March 2020.
[12] Karthik P C, Sanjana S, M P Adithya Vijayan, Thushara P, International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 10 Issue May-2021
[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Deep Residual Learning for Image Recognition, arXiv conference, arXiv:1512.03385v1, Dec 2018.
[14] Luisa Verdoliva, Media Forensics and DeepFakes: an overview , arXiv:2001.06564v1, IEEE, January 2020.
[15] 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, 2019.
[16] 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, 2020.
[17] Nicolas Rahmouni, Vincent Nozick, Junichi Yamagishi, and Isao Echizen, "Distinguishing computer graphics from natural images using convolution neural networks," in WIFS. IEEE, 2018
[18] Thanh Thi Nguyena, Quoc Viet Hung Nguyenb, Dung Tien Nguyena, Duc Thanh Nguyena, Thien Huynh-ThecSaeid Nahavandid, Thanh Tam Nguyene, Quoc-Viet Phamf, Cuong M. Nguyen, arXiv conference, February 2022.
[19] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, “Face Forensics ++: Learning to detect manipulated facial images,”in The IEEE International Conference on Computer Vision (ICCV), October 2019.October 2019.
[20] Ayush Tewari, Michael Zollhoefer, Florian Bernard, Pablo,Garrido, Hyeongwoo Kim, Patrick Perez, and Christian Theobalt. High-fidelity monocular face reconstruction based on an unsupervised model-based face autoencoder. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(2):357–370, 2018.