Review of Techgaurd on Deepfake detection

Ishu Sharma
Department of CSE,
Chandigarh University,
Punjab, India
sharmaishu8191@gmail.com

Arzit Mahajan
Department of CSE,
Chandigarh University,
Punjab, India
nidhimahajan251@gmail.com

Adil Husain Rather
Department of CSE,
Chandigarh University,
Punjab, India
adilhusain5057@gmail.com

***Abstract*—Deep fake videos are AI-generated videos that look real but are fake. Deep fake videos are generally created by face-swapping techniques. It started as fun but like any technology, it is being misused. In the beginning, these videos could be identified by human eyes. But due to the development of machine learning, it became easier to create deep fake videos. It has almost become indistinguishable from real videos. Deep fake videos are usually created by using GANs (Generative Adversarial Network) and other deep learning technologies. The danger of this is that technology can be used to make people believe something is real when it is not. Smartphone desktop applications like Face App and Fake App are built on this process. These videos can affect a person's integrity. So identifying and categorizing these videos has become a necessity. This paper evaluate methods of deepfake detection and discuss how they can be combined or modified to get more accurate results. Hopefully, we will be able to make the internet a safer place. Finally, the paper highlights the promising directions and future research opportunities in the field of deepfake detection. As the arms race between deepfake generators and detectors continues, collaborative efforts from academia, industry, and policymakers are crucial to developing robust defenses against the misuse of deepfake technology.**

Index Terms:-Detection, Classification, Deepfake Video, Gen- erative Adversarial Network, Artificial Neural Netwok, Machine Learning.

Keywords- Deepfake, AI, ML, GAN's, Deep learning, human motion transfer.

I. INTRODUCTION

A growing disquiet as settled around the emerging deepfake that make it possible to create evidence of scenes that has never ever happened. Celebrieties and politcians are the ones who are considerably affected by this.

Deepfake can optimally stitch anyone into a video or photo that they never have actually knowledge with. Nowadays since technologies are elevating widely the systems can synthesize images and videos more quickly. A creator would first train a neural network on many hours of real video footage to give it a realistic understanding of what he or she looks like on many angles or lighting inorder to create a deepfake video of someone.Then they would combine the trained network into graphics techniques to superimpose a copy of person into different one.

AI-Generated synthetic media, which is also known as deepfakes, ofcourse have many positive sides.Deepfakes en-ables clear benefits in areas such as education, accessibility, film production, criminal forensics, and artistic expression. It can ac-celerate the artistic quest into equity. Creative use of synthetic voice and video can enhance overall success and learning outcomes with scale and limited expenditure. Deepfakes can democratize VFX technology as a strong tool for independent story tellers. It could give individuals new tools for self-expression and amalgamation in online world. Deepfakes also has disadvantages which affects different groups of our society. It is being used to revenge porn to defame certain celebrities, creating fake news and propaganda etc...

II. DISSCUSSION

Different types of deepfake detection methods are available today and each method has its own advantages and disadvan-tages. This paper tries to evaluate such methods from different papers and points out how these methods can be combined and modified in a new project in order to get more accurate results.

In the paper [1] ”Deepfake Video Detection Using Recurrent Neural Network”, David Guera and Edward J Delp propose a temporalaware pipeline to automatically detect deepfake videos. In order to detect deepfake videos, firstly we need to have a clear knowledge of how it is created, which helps us to understand the weak points of deepfake generation so that by exploiting those weak points, deepfake detection can be done. In the approach discussed in this paper, framelevel scene inconsistency is the first feature that is exploited. If the encoder is not aware of the skin or other scene information, there will be boundary effects due to a seamed fusion between the new face and the rest of the frame which is another weak point. The third major weakness that is exploited here is the source of multiple anomalies and leads to a flickering phenomenon in the face region.

This flickering is common to most of the fake videos. Even though this is hard to find with our naked eye, it can be easily captured by a pixellevel CNN feature extraction.

Dataset used here contains 300 videos from the HOHA dataset. Preprocessing steps are clearly described in this paper. Here the proposed system is composed of a convolution LSTM structure for processing frame sequences.

CNN for frame feature extraction and LSTM for temporal sequence analysis are the 2 essential components in a convolutional LSTM. For an unseen test sequence,set of features for each frame are generated by CNN. After that features of multiple consecutive frames are concatenated and pass them to the LSTM for analysis which finally produces an estimated likelihood of the sequence being either a deepfake or nonmanipulated video.

With less than 2 seconds this system could accurately predict if the fragment being analyzed comes from a deepfake video or not with an accuracy greater than 97 percentage.

In the paper [2] ”Effective and Fast Deepfake detection method based on Haarwavelet Transform” by Mohammed Akram Younus and Taha Mohammed Hasan describes another method to detect deepfake videos by haar wavelet transform. The method described here take the advantage of the fact that during deepfake video generation, deepfake algorithm could only generate fake faces with specific size and resolution. In order to match and fit the arrangement of the source’s face on original videos, a further blur function must be added to the synthesized faces.

This transformation causes exclusive blur inconsistency between the generated face and its background outcome deepfake videos. The method detects such inconsis-tency by comparing the blurred synthesized areas ROI and the surrounding context with a dedicated Haar Wavelet transform function. The two main advantage of this Haar Wavelet transform function is that it first distinguishes different kinds of edges and the retrieves sharpness from the blurred image.

It is very effective and fast since the uniform background of the faces in the images will have no effect and it does not need to reconstruct the blur matrix function. To estimate the blur extend, two methods such as direct and indirect can be used. Direct method can measure the blur function extent by testing some distinctive features in an image. Eg: edge feature. The indirect method depends on the blur reconstruction function when the H matrix is unknown (H matrix is blur’s estimation and blur identification). Dirac structure, Step structure, and Roof structure are the different types of edges present in an image.

A blur extends is identified by taking the sharpness of roof structure and G step structure into account. The sharpness of the edge is indicated by the parameter (0¡¡ /2), if is larger means the edge is sharper. By comparing the blur extent of the ROI with the blur extend of the rest of the image, we can determine if the images(frames of video) have tampered or not. UADFV dataset which contains 49 unmanipulated and 49 manipulated videos is used here.

Videos are divided into frames and from each frame, the face region is extracted and deepfake detection algorithm using haar wavelet transform is applied. This algorithm is clearly described in this paper. This proposed model contains an accuracy of 90.5 percentage.

In the paper [3],” OC Fake Dect: Classifying Deepfakes using OneClass Variational Autoencoder” by Hasam Khalid and Simon S. Woo, the proposed model needs only real images for training. As new methods for deepfake video creation are increasing today due to technology advancement, for a model to detect such videos, datasets containing fake videos are very scarce for training. It affects the model's accuracy. But in the model proposed in this paper needs only real videos for training so that it can overcome data scarcity limitation.
FaceForensic ++ is the dataset used here. It contains real images and 5 sets of fake images: FaceSwap dataset, Face2Face dataset (F2F), Deepfake dataset(DF), Neural Tex-tures dataset (NT), Deepfake detection Dataset(DFD), After collecting the video datasets, they are converted into frames and face detection and alignment is done using MTCNN.
One class variational encoder is used here. It consists of an encoder and a decoder. At the encoder side, image is given as input, and scaling is done using convolutional layer and mean and variance is calculated and the result is given as input into decoder and the RMSE value is calculated which is low for real image and high for fake images. Two methods are discussed in this paper: OCFakeDect1 and OCFakeDect2.
In OCFakeDect1 from input and output image itself, reconstruction score is computed directly and in OCFakeDect2 contains additional encoder structure which computes reconstruction score from input and output latent information.
Eventhough it has 97.5 percentage accuracy, better performance is only on NT and DFD datasets.
Deepfake technology has a huge range of applications which could use both positively or negatively. Although most of the time it is used for malicious purposes. The unethical uses of Deepfake technology has harmful consequences in our society either in short term or long term. People regularly using social media are in a huge risk of Deepfake. However , proper use of this technology could bring many positive results. Below both negative and positive applications of Deepfake technology described in details.

In the paper [4] ”Deep Fake Source Detection via Interpreting Residuals with Biological Signals”, Umur Aybars Ciftci, Ilke Demir and Lijun Yin presented a deep fake source detection technique via interpreting residuals with biological signals. To their knowledge it is the first method to apply biological signals for the task of deep fake source detection. In addition to this they had experimentally validated this method through various ablation studies their experiments had achieved 93.39accuracy on FaceForensics++ dataset on source detection from four deep fake generators and real videos. Other than this they had demonstrated the adaptability of the approach to new generative models, keeping the accu- racy unchanged.

In the paper [5] “ Digital Forensics and Analysis of Deep-fake Videos” by Mousa Tayseer Jafar, Muhammed Ababneh, Muhammad Al-Zoube, Ammar Elhassan proposed a method detect deepfakes using mouth features.
Nowadays deepfake videos can have an adverse effect on a society and these videos can challenge a person's integrity. Deepfake is a video that has been constructed to make a person appear to say or do some- thing that they never said or did. Therefore there shows the increase in demand to detect methods to identify deepfakes. In this proposed model mouth features is used to detect deepfake video.
A deepfake detection model with mouth features(DFT-MF),using deep learning approach to detect deepfake videos by isolating analysing and verifying lip/mouth movement is designed and implemented here. Here, dataset contains the combination of fake and real videos. Some preprocessing is done prior to performing analysis. Then the mouth area is been cropped from a face. There will be fixed coordinates for face. Working on a typical image frame facial landmark detector is used to estimate the location of 68 (X,Y)coordinates. In next step all face containing closed mouth is excluded and face with only open mouth is been tracked having teeth with reasonable clarity.
CNN is used to classify videos into fake or real based on a threshold number of fake frames based on calculating three variable word per sentence, speech rate and frame rate. If the number of fake frames is greater than 50 the video is been classified as fake or else as real.
After studying biological signal analysis on deepfake videos, it is found that ground truth PPG data along side original and manipulated videos enabled new direction in research on deepfake analysis and detection. In the next stage of their work, . With ground truth PPG, they planned to create a new dataset with certain distribution variation as well as source variations.

GANS can be used in various field to give realistic experiences such as in retail sector, it might be possible to see the real product what we see in shop going physically [36]. Recently Reuters collaborated with AI startup Synthesis has made first ever synthesized news presenter by using Artificial intelligence techniques and it was done using same techniques that used in Deepfakes and it would be helpful for personalized news for individuals [37].

III. DEEPFAKE DETECTION

It is often difficult sometimes impossible to detect Deepfake contents by a human being with a untrained eye. A good level of expertise is needed to detect irregularities in Deep fake videos. Till now there are several approaches have been proposed including machine detection, forensics, authentication as well as regulation to combat Deepfake.

Fig1: Visualization of accuracy measure.

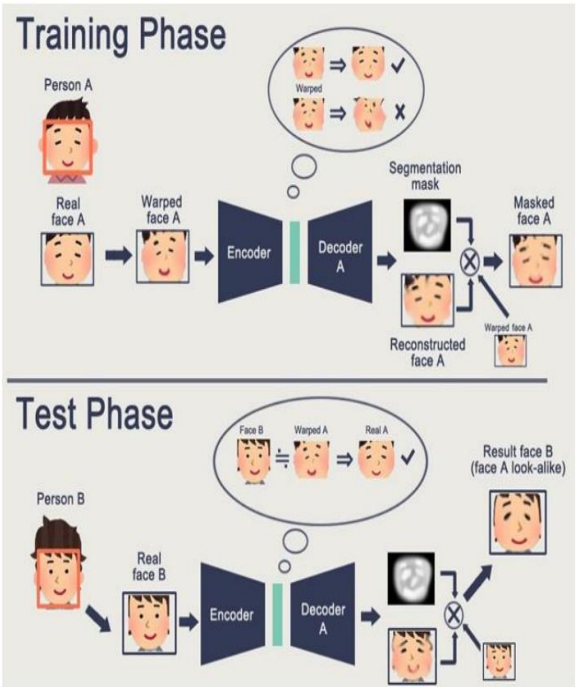

Fig2: Working procedures of autoencoder.

Experts say as Deepfake video is created by algorithm, while real video is made by a actual camera, it is possible to detect Deepfake from existing clues and artifacts. There are also some anomalies like lighting inconsistencies, image warping, smoothness in areas and unusual pixel formations which could help to detect Deepfake. Detecting Deepfakes Korshunov et al. [65] described a process that use to find inconsistency in the middle of visible mouth motion and voice in recording. In this article they also applied several approaches including simple principal component analysis (PCA) , linear discriminant analysis (LDA), image quality metrics (IQM) and support vector machine (SVM) [66]. VidTIMIT database [67], is a publicly available database of videos, used for generating several Deepfakes videos with combination of different features of Deepfake creation technique including face swapping, mouth movements, eye blinking etc. After using these videos to verify several methods of Deepfake detection, it is found that several face swap identification technique fail to detect fake contents.

As example deep learning based face recognition technique VGG [68] and Facenet [69] are unable to detect Deepfakes properly. Although earlier Deepfake detection method was not able to measure the blink of eye, recent methods showed promising results to detect eye blink in source video and target video. The authors in [70] aimed to detect facial forgeries automatically and properly by using recent deep learning techniques convolutional neural networks (CNNs) with the help of neural network.

Forensics model face extreme difficulties to detect forgeries when the source data are made by CNN and GAN deep learning method [71]. To avoid the problems of adaptability, Huy H. Nguyen et al. proposed a method that supports generalization and could locate manipulated location easily. Ekraam Sabir et al. [72] tried to detect face swapping that generate by several available software such as Deepfake, face2face, faceswap by using conventional networks and recurrent unit.

TABLE I

SUMMARY OF DEEPFAKE ARTICLES

Serial	Paper Name	Authors	Date and Publication	Key findings	Limitations
1	Reconstructing detailed dy-namic face geometry from monocular video[44]	Pablo Garrido,Levi Valgaert ,Chenglei Wu ,Christian Theobalt	ACM Trans. Graph. 32, 6, Article 158 (November 2013), 10 pages	Monocular tracking of fa-cial expressions, In some cases, less susceptible to occlusions and drift, Well suited for video augmentation	Fail in extreme head angle, Not free from artifacts like tracking inaccuracies of teeth and lips
2	Displaced dynamic expression regression for real-time facial tracking and animation[46]	C. Cao, Q. Hou, and K. Zhou	ACM TOG, 33(4):43, 2014	Robustness against fast motions, Large head rotations, lighting . Use of single web camera.	Not ideal for low quality images.
3	Real-Time High-Fidelity Facial Performance Capture[47]	Cao, C., Bradley, D., Zhou, K., Beeler	ACM Trans. Graph. 34, 4, Article 46 (August 2015)	Able to reconstruct person specific ankle details in real time from a single camera, Mesh tracking is improved and enhanced the details through novel local wrinkle regression methods.	Changing illumination can affect both optical flow and the local regresser, Problem of occlusions.
4	Real-time Expression Transfer for Facial Reenactment[49]	Thies, J., Zollhöfer, M., Nießner, M., Valgaerts, L., Stamminger, M., Theobalt, C. 2015	ACM Trans. Graph. 34, 6, Article 183 (November 2015), 14 pages	Real time facial reenactment, This model tracked from real time rgb-d input, GPU based tracking	Tracking could fail if hands occludes the face, if occlusion is extreme the tracker will fail.
5	Face2Face: Real-time Face Capture and Reenactment of RGB Videos[51]	Justus Thies, Michael Zollhöfer, Christian Theobalt, Marc Stamminger, and Matthias Nießner	Proc. Computer Vision and Pattern Recognition (CVPR), IEEE 1 (2016)	RGB data input, Real time video manipulation, Using the temporal and photo- metric similarity	Hard shadows or Spectacular highlights, Capturing subjects with long hair and beard is very challenging, Mouth behavior can not learn if subject is too static or insufficient expressions
6	Synthesizing Obama: Learning Lip Sync from Audio[54]	Supasorn Suwajanakorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman.	ACM Trans. Graph. 36, 4, Article 95 (July 2017), 13 pages	Audio input is converted to a time-varying sparse mouth shape,Can synthesize visual speech only from audio.	In final composition, mouth shape and chin location can vary depending on target frame, Target person's non frontal face can cause mouth texture composited outside face, Face texture synthesis depends on complete mouth expressions.

IV. DISCUSSIONS AND CONCLUSIONS

In this paper, we have presented a brief review of some papers which describes different methods to detect deepfake videos and images. Also how those methods can be modified or combined in our new project inorder to get more accurate results than prevailing methods. Hope we will succeed in our project.

Improvement of Deepfake generation technique make the detection work difficult day by day. Improved detection tech- niques and accurate dataset are important issues for detecting Deepfake properly [94] [95]. From our study it is shown that Because of Deepfake technology, people are loosing trust on online content day by day. As Deepfake contents creation technique improving day by day, any person having a high configuration computer instead of having less technological knowledge could create Deepfakes content of any individuals for malicious purposes. Moreover, the advancement ofinternet and networking made it possible to spread Deepfake videos in a moment. This technique could influence the decisions of world leaders, important public figures which could be harmful for world peace [96]. From our study, we found that the battle between Deepfake creator and detector is growing rapidly. Although having lots of negativity, we have also dis- cussed several positive use of Deepfake technology such as in film industry, fundraising etc. It is possible to return back the voice of a individual who have already lost it by using the Deepfake generation methods. There are lots of debate against and in favor of Deepfake technology. Our study tried to analyze the both sides of Deepfake technology. Deepfake technology have a detrimental effects on film industry, commercial media platform, social media. Deepfake technology used to gain trust of people via social media in any political context or any other social media context. To make profit by generating traffic of fake news via web platform is increasing rapidly [97]. Zannettou et al. [98] showed a number of people behind Deepfake including politicians, public figures, celebrity ,creating Deepfakes and spreading it via social media for various beneficial purposes of their own. We have also found that many organizations are involved in making Deepfake contents for useful purposes. According to our study Deepfake has lots of threats towards individuals, society, politics as well as business. Because of increasing fake contents , the situation become worse for media person to detect real one specially journalists.

V. REFERENCES

[1] Choi, N. and Kim, H., 2023. DDS: Deepfake Detection System through Collective Intelligence and Deep-Learning Model in Blockchain Environment. *Applied Sciences*, *13*(4), p.2122.

[2] Rani, R., Kumar, T., Sah, M.P. (2022). A Review on Deepfake Media Detection. In: Sharma, H., Shrivastava, V., Kumari Bharti, K., Wang, L. (eds) Communication and Intelligent Systems . Lecture Notes in Networks and Systems, vol 461. Springer, Singapore.

[3] P. Korshunov and S. Marcel, “Deepfakes: a New Threat to Face Recogni- tion? Assessment and Detection,” arXiv preprint arXiv:1812.08685,2018.

[4] Mittal, T., Bhattacharya, U., Chandra, R., Bera, A. and Manocha, D., 2020, October. Emotions don't lie: An audio-visual deepfake detection method using affective cues. In Proceedings of the 28th ACM international conference on multimedia (pp. 2823- 2832).

[5] Li, Y. and Lyu, S., 2018. Exposing deepfake videos by detecting face warping artifacts. arXiv preprint arXiv:1811.00656.

[6] Baltrušaitis, T., Ahuja, C. and Morency, L.P., 2018. Multimodal machine learning: A survey and taxonomy. IEEE transactions on pattern analysis and machine intelligence, 41(2), pp.423-443.

[7] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A. and Schulman, J., 2022. Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, pp.27730-27744.

[8] GitHub. 2019. Shaoanlu/Faceswap-GAN. [online] Available at: <https://github.com/shaoanlu/faceswap-GAN> [Accessed 25 March 2020].

[9] D. Afchar, V. Nozick, J. Yamagishi and I. Echizen, "MesoNet: a Compact Facial Video Forgery Detection Network," 2018 IEEE International Workshop on Information Forensics and Security (WIFS), Hong Kong, Hong Kong, 2018, pp. 1-7, doi: 10.1109/WIFS.2018.8630761.

[10] F. Matern, C. Riess and M. Stamminger, "Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations," 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW), Waikoloa Village, HI, USA, 2019, pp. 83-92, doi: 10.1109/WACVW.2019.00020.

[11] Zannettou, S., Sirivianos, M., Blackburn, J., Kourtellis, N. 2019. The Web of False Information: Rumors, Fake News, Hoaxes, Clickbait, and Various Other Shenanigans. Journal of Data and Information Quality, 1(3): Article No. 10. https://doi.org/10.1145/3309699

[12] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 770-778). [86] Sabour, S., Frosst, N., and Hinton, G. E. (2017). Dynamic

[13] Andreas R o¨ ssler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2019. FaceForensics++: Learning to Detect Manipulated Facial Images. arXiv (2019)