Proceedings of the Sixth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC-2022).
IEEE Xplore Part Number: CFP22OSV-ART; ISBN: 978-1-6654-6941-8

2022 Sixth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC) | 978-1-6654-6941-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/I-SMAC55078.2022.9987265

Comparative Analysis on Different DeepFake Detection Methods and Semi Supervised GAN Architecture for DeepFake Detection

Jerry John
M.Tech Scholar
Muthoot Institute Of Technology And Science
Kochi, India
jerryjohn1995@gmail.com

Ms. Bismin V. Sherif
Assistant Professor
Muthoot Institute Of Technology And Science
Kochi, India
mtech20ai07@mgits.ac.in

***Abstract*—The use of deep learning results in solving a wide range of real-world problems and applications, but there are some drawbacks along with this positive side. One of the most recent and advanced problems among them is the wide use of deepfakes. Deepfakes are digital tampered images or videos created using different deep learning methods. In a deepfake, the face of a targeted person is superimposed on a source image so that this digital tampered data can be used for digital frauds, blackmailing, pornography etc. With the developments in the deep learning field, it is becoming challenging to distinguish between real and fake manually. So it is essential to do research and development in the area of deepfake detection. In this paper, an extensive discussion and timely overview on different deepfake detection methods are done under the classification of feature-based, temporal-based, and deep feature-based deepfake detection. The comparison study is mainly done based on the key features used, face detection architecture, deep learning architecture, video-based or image-based, the dataset used, frames size, and dataset size used. Along with the comparison, a semi-supervised GAN architecture is also proposed and developed to detect the deepfake images.**

***Index Terms*—DeepFake, SGAN, DeepFake Detection**

I. Introduction

Artificial Intelligence along with deep learning techniques have brought considerable advancement in our day-to-day lives. One of the major contributions among them is in the field of computer vision, where deep learning models are able to identify and classify various parameters like human faces, human key points, medical images etc. Even though deep learning concepts has been introduced in the late nineteens but the effective use of all these technologies been started in the early twenties. Some of the main reasons for this sudden uptrend id due to the better computational capacity with low cost and availability of a large amount of data because of the internet boom. Data is the powerhouse of artificial intelligence and data science. As if the model needs to learn and work with good accuracy, it needs more accurate data to be learned, as the amount of data increases, the model performance also increases.

Even after considering all the positive sides of artificial intelligence, we also have to face the negative drawbacks of this modern technology. That is with modern technologies come modern problems. One of the most recent and highly threatening ones of such technological fraud is called deep fake. Deep fake as such is not a negative name. It can be simply defined as creating fake images or videos using deep neural network technologies. Fake images or videos means, placing some target person's face over a source video, so that when the final results come we will able to see the target person in the source video. These deepfakes can be created using different methods. One of the easy classifications is deepfake with voice and without voice. The deepfakes with a cloned voice make the situation more worst, as this will increase the genuineness of the video content. The deepfakes with voice consist of a combination of fake visual effects with

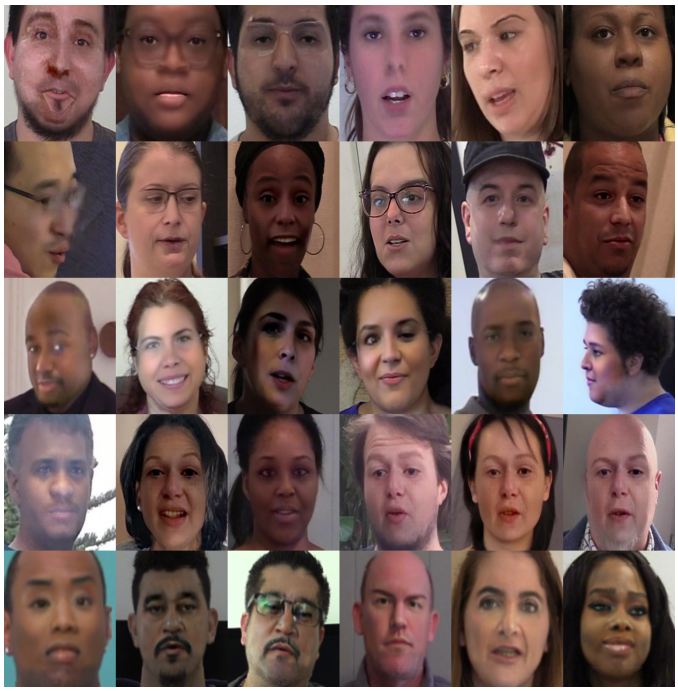

Fig. 1. DFDC Dataset

cloned voices. Cloned voice can be developed using a target person's voice, with source audio. Where this source audio will contains content that the target person never said. After imposing the target person's voice to the source voice using deep learning, it source audio will sound just like the target person.

The growth of such kinds of deepfake technological frauds has been increasing in recent years. Because of the increase in the availability of applications and open source codes to generate this kind of visual deepfakes very easily. One of the other reasons for this rapid increase is the increase in the availability of videos and images. Because to make an effective deepfake, the deepfake model has to be trained perfectly. To do that a large amount of data is needed to be trained in the model. In old days it was not that available. But now because of the internet revolution, it become possible. One of the other reasons is the easy availability of high capacity computer systems to do training.

This increase in the deepfake creation methods results in many social issues, it has been negatively used by school children to blackmail their friends to politicians to make fraud allegations against their opponents. Because of these kinds of widespread social issues, it has been high time that we find some methods to accurately detect and prevent these kinds of video and visual deepfakes. But one of the greatest problems developers face is that, once this deepfake is published on the internet then it is very difficult to detect each copy and remove them. So one of the main solutions we bring here is to use the deepfake detection models in social media and visual media platforms so that they can check the video even before publishing. If they find a video contains some deepfake elements, they will reject that video and in that way prevent the deepfake to be published.

II. Literature Survey

A. Deepfake creation methods

The deepfakes can be generated using different neural network architectures like autoencoders [17], GANs [16], Convolutional Neural Networks [20] etc. The deepfakes are generated by collecting a target person's videos or images and a source video or photo, to which the target person's face should be implanted. The video files are converted into frames for this process, and after that face, detection is done, as face detection is one of the most important steps in deepfake creation and detection. Many face-and-face landmark detection libraries are available like OpenCV [18], media pipe etc. The face detection model is trained using several face images, with different gender, age groups, colours etc. The face key points or landmark detection model has developed my training the model with face images with face key points marked. After that, this source person's image is passed through an encoder, where this encoder converts the images into their smallest form. After that, it is passed to a decoder, which generates the original image. This encoder and decoder are trained using these input images of the target person. Using the same way another encoder and decoder are trained using the source image. After that this encoder and decoders are replaced with each other. In the new encoder, if we give the target person's image, it will be placed with the source video.

Another way of creating deep learning is using affine transform [15]. Where initially the target persons and source persons' face are detected using face recognition libraries, these recognised face region is then cropped out and from that facial key points are detected. Different libraries have different numbers of keypoint detection methods, it is found that 68keypoints [14] are one of the ideal numbers for this method. After that, these key points are joined to the nearby key points to create a triangle-like structure. This process is done for both of the faces. After that, each triangular part is marked with a unique number. In the end, these unique numbers are used to transfer the source person's face. For that, the triangular part is taken from the source person's face and placed on the corresponding position of the target image. There will be a misfit, this problem is solved using the affine transform technique.

The voice deepfake is also created using encoder-decoder technology, where two separate encoder-decoder is trained using the source person's audio and the target person's script. After the proper training, the encoder and the decoders are interchanged. The final result will contain the target person's voice with the source person's content.

B. Deepfake detection methods

In [1] Li *et al.* introduce a method that mainly focuses on the lack of eye blinking to detect the digital tampered video. This method is based on the conclusion that this type of physiological signs are not well captured in the digital tampered video. Convolutional Neural Network and LSTM architectures [19] are combined. Where the CNN is used to detect whether the eye is closed or open, and the LSTM is used to identify the temporal information, as the eye blinking has a good correlation between the nearby frames. A cross-entropy loss is used as the loss function.

There are different methods to detect deepfakes, the detection methods can be classified based on the deepfake creation method. For each deepfake creation method, [11] [12] there will be a corresponding detection method. The deepfake detection works by taking the input video or image, if it is a video it is converted into frames (as the video is a collection of frames), and from each of these frames [13], the facial region is identified, using the face recognition libraries, this facial part is cropped out. This cropped-out face part is combined to create a new video. This video is considered a preprocessed video. This process is done because deepfake detection mainly concentrates on the face region rather than the other regions.

One of the research by [1] Li *et al.* are showing a method, where they use an eye blinking pattern to detect the deepfake. One of the main findings is that the deepfakes created will not have good eye-blinking patterns. Some of them have even no eye blinking because deepfakes are not that expert in correctly mimicking the eye blinking pattern. Which can be

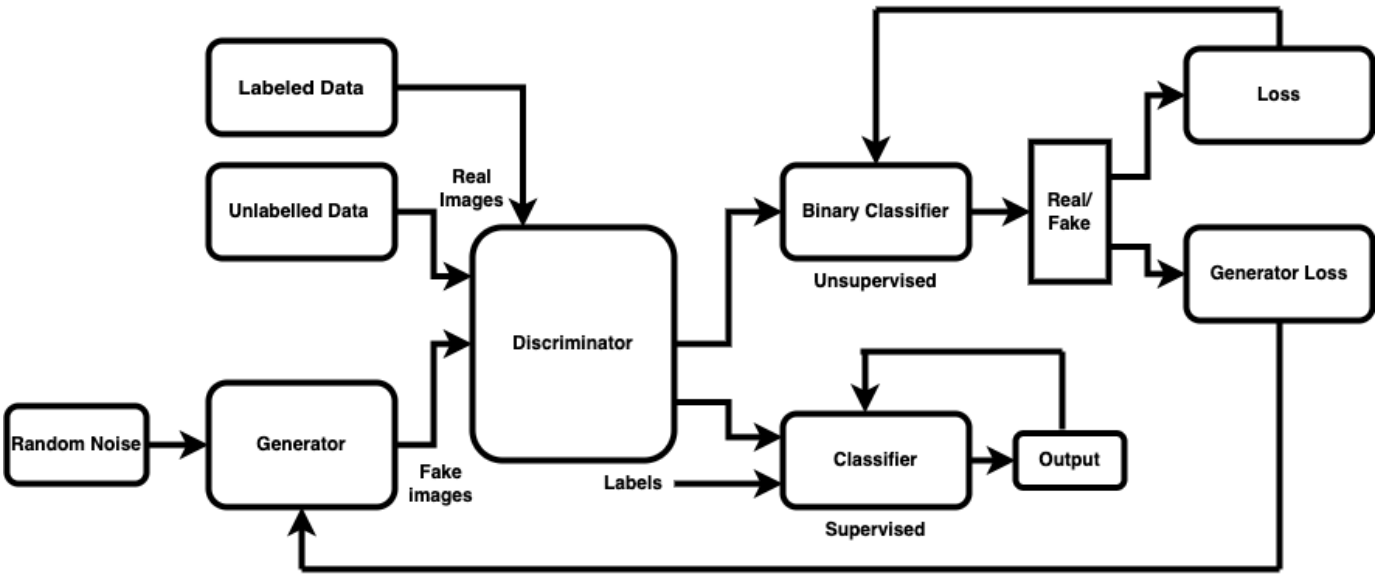

Fig. 2. SGAN Model Architecture

easily identified using a combination of convolutional neural networks and long short-term memory.

Some methods are only concentrating on deepfakes which are created by superimposing the target person's face on the source face, research on such methods is done by [2] X. Yang *et al.* The face position and superimposing are based on the face key points. Because of that reason on the deepfake detention also this face key point difference is calculated and given to the SVM for the final decision.

The face X-ray method [3] is another deepfake detection method, in which anyone can train the model without the fake images. That is because the model itself has the capability to create fake images. In this deepfake detection method also face key points are detected and based on that a face mask is created, which is then compared with the face x-ray to check the fakeness of the video.

One other major deepfake detection is deep feature-based detection [4] [5], in this method, deep neural networks like a convolutional neural network, or some advanced CNN architecture (transfer learning also) are used like Meso -4. The main working flow of such methods is like converting the video into frames, after that from each frame using face landmark and keypoint detection faces are cropped out. These cropped faces are given to the deep neural architecture for the training. At the time of classification, a new unknown image or video is taken, if it's a video it is converted into frames. This is then given to the trained model so that, the model can classify the input media as fake or not.

In some other areas of research on deepfake detection, temporal-based methods[8] [9] can be seen. Some researchers like Chintha *et al.* [6], Guera *et al.* [7] and Daniel *et al.* [9] purposes such kinds of works. Where not only is the deep pixel-wise information used, but they also consider the correlation between the nearby frames. This is one of the important factors to consider, and it can improve the performance [10] of deepfake detection by a great margin.

III. Semi- Supervised GAN based deepfake detection

A. Dataset Used

Two datasets are used; the first is the Flickr-Faces-HQ(FFHQ) dataset. It consists of more than 65,000 PNG images with a resolution of 1024x1024. This image dataset consists of variations like image background, age, eyeglasses, hats etc. After cropping and proper alignment, these final images are obtained using the image processing library called dlib. The occasional statues, paintings and photos are removed using Amazon Mechanical Turk.

The second is the Deepfake Detection Challenge Dataset, which consists of more than 100,000 videos created using different deepfake creation techniques and is sourced from more than 3000 paid actors of different ages, colours, and types.

B. Model Creation

In supervised learning techniques, the data with their corresponding labels are present. The semi-supervised method used in this paper comes in between the supervised and unsupervised learning techniques. Some labelled data and some unlabeled data are used for training. In semi-supervised models, only a small amount of labelled data is taken, which is one of the main advantages. In normal GAN, it has a discriminator and a generator. The discriminator is trained unsupervised for the final classification as real or fake. In this model, an SGAN is used. The discriminator is trained using both supervised and unsupervised methods. So, in the end, this discriminator acts as a multiclass classifier. The unsupervised mode learns the features, and the supervised model learns

TABLE I
OBSERVATIONS AND COMPARISON

Reference No	Technique used	Face detection	Method used	Key Features	Data Type	Datasets Used	Dataset size	Frame size
1	Eye blinking	-CNN based model (for detecting the eye) -CEW dataset is used to train this model.	-LRCN	-Use LRCN to learn the temporal patterns of eye blinking. -Based on blinking frequency	-Videos	-Consist of 50 interview and presentation videos.	-50 videos, each of approximate 30s.	-30 frames
2	Head poses	-Dlib -3D facial landmark model OpenFace2	-SVM	-Features are extracted using 68 landmarks. -SVM classifier	-Videos/ Images	-UADFV dataset -DARPA MediFor GAN Challenge.	-98 videos -493 images	-Single frame model.
3	Face X-ray	-Not Mentioned	-CNN	-Based on blending boundary -Can be trained without fake images.	-Images	FaceForensics++, DFDC and Celeb-DF.	-2000 videos	-Single frame model
4	Face warping artifacts	-Dlib package	-CNN	-For affine transformed deepfakes. -Can be trained without fake images.	Images/Video	DeepFake video dataset UADFV.	-98 videos, which have 49 real videos and 49 fake videos respectively.	-Single frame model
5	MesoNet	-Viola Jones detector	-CNN -Meso-4	-Examines mesoscopic analysis level.	Video	-Online videos - FaceForensics	-450 videos.	-50 faces per scene.
6	Spatio-temporal features with LSTM	-Dlib face extractor	Convolutional bidirectional recurrent LSTM network -DLib face extractor -XceptionNet	-XceptionNet CNN is used for facial feature extraction. -Kullback-Leibler divergence.	Video	FaceForensics++ and Celeb-DF	-5639 videos.	-Eight frames with a stride of eight.
7	Intra-frame and temporal inconsisten-cies	-Face recognition python library	-CNN and LSTM	-CNN extract frame-level features, LSTM extract temporal information.	Video	Multiple websites.	-600 videos	-Comparison of 20,40,80 frames
8	Automatic Face Weighting	-MTCNN	-MTCNN -AFW -GRU	-Automatic Face Weighting (AFW) -Gated Recurrent Unit (GRU).	Video	Deepfake Detection Challenge (DFDC) dataset	-3000 videos	-Extract face from 1 every 10 frames.

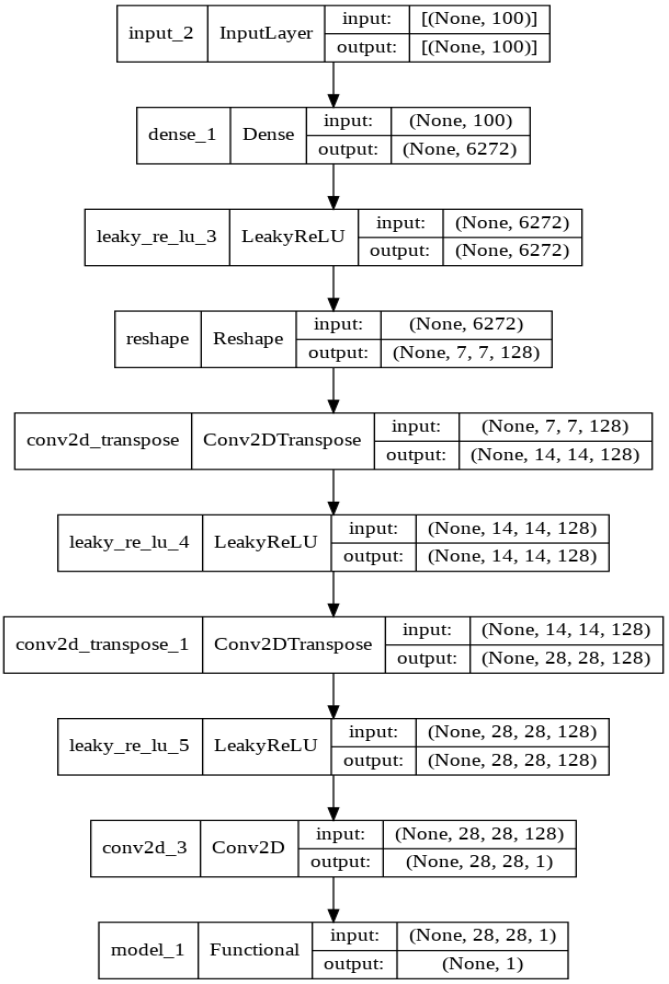

Fig. 3. GAN model

to do the classification. Here more concentration is done on the discriminator than the generator because of the semi-supervised learning technique.

In our training for the discriminator, three inputs are given, the first one is the actual images with their corresponding labels, the second is the real images without the labels, and the last one is the fake images from the generator. An efficient way to implement instead of creating two discriminator models is to reuse the output layers of one model as input to the next model. The final unsupervised method takes the output of the previous supervised model just before the softmax activation function calculates a normalized sum of the exponential output.

SGAN- Generator : In this model, the generator takes in a latent vector; initially, a size of one hundred is used and which is then reshaped(increased) later that can be sent through transpose convolutions and Leaky ReLU activation functions. The tanH activation function is used at the output to get the result between -1 and 1.

SGAN- Discriminator : A stacked discriminator approach is used. The image is given as the input. Then they are down-sampled using a few convolutions and Leaky Relu activation functions. Dropouts are added to reduce the over-fitting concerns. In the end, a softmax activation function is used. The final unsupervised model takes the output of the supervised model just before the softmax activation function and calculates a normalized sum of the exponential outputs. This can be implemented as a lambda layer using Keras.

SGAN Training: SGAN training process is similar to normal GAN training process, the only difference is that supervised model weights are updated using the labelled data. First, the supervised discriminator gets updated using the labelled data. Then the supervised discriminator gets updated using unlabeled and fake data, that is, the generated data. Finally, using the composite GAN model, the generator gets updated.

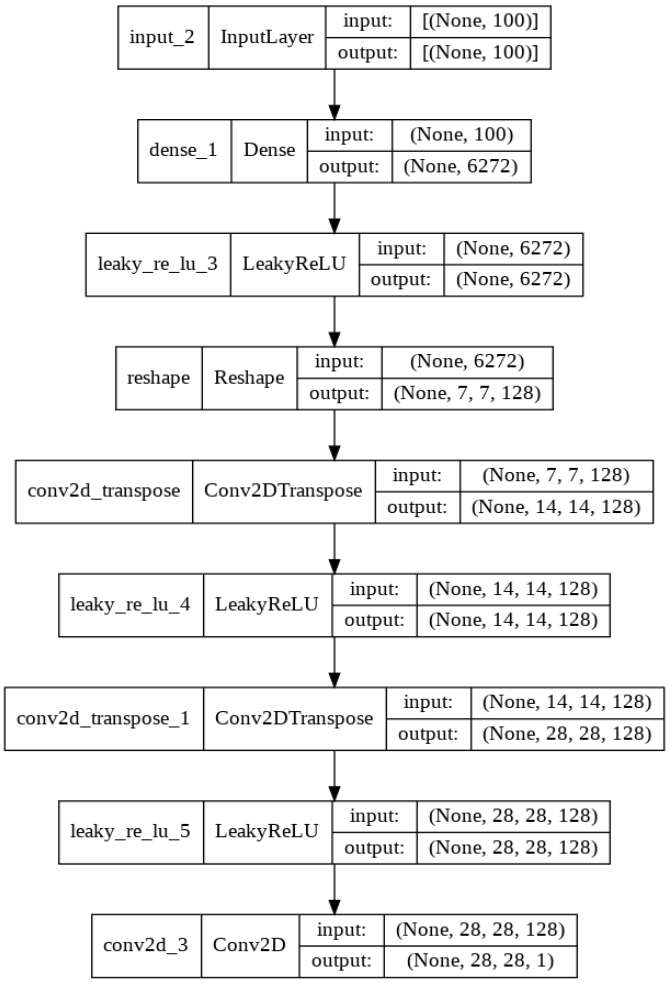

Fig. 4. SGAN Generator

IV. Result and Observation

A combination of four different sizes of datasets is used for testing the accuracy and for comparison purposes, and the highest accuracy of 92.3% is obtained for the dataset, which contains almost 40,000 images. It can also be observed that the dataset, which includes nearly 10,000, receives an accuracy of 86.7%, and the dataset containing 20,000 and 30,000 obtain

an accuracy of 88.9% and 91.6%, respectively. From these observations, a conclusion can be made that the accuracy is also increasing as the training data increases.

TABLE II
PERFORMANCE PARAMETERS

Models NO	Dataset Size	Accuracy
1	10,000	86.7%
2	20,000	88.9%
3	30,000	91.6%
4	40,000	92.3%

V. Conclusion

Even though deep learning is bringing a lot of advancement in our day-to-day life in all areas, like medical, technological, educational etc there are some negative sides also, one of such most relevant technological fraud is deepfake, in which deep learning technology is used to create digital frauds. To detect these kinds of deepfakes, so that we can prevent them even before publishing to the outer world, in this paper a neural network architecture is used to train a model to detect the deepfakes. With a large dataset of 40,000, this model has achieved the highest accuracy of 92.30 %. This is a promising result compared to the relative research works with this particular method and dataset. The accuracy score also shows that the model is neither overfitted nor under fitted also.

References

[1] Li, Y., Chang, M. C., and Lyu, S. (2018, December). “Exposing AI created fake videos by detecting eye blinking” 2018 IEEE International Workshop on Information Forensics and Security (WIFS) (pp. 1-7). IEEE.

[2] X. Yang, Y. Li, and S. Lyu, (May 2019)“Exposing Deep Fakes Using Inconsistent Head Poses,” in ICASSP 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom,, pp. 8261–8265.

[3] Li,L.,Bao,J.,Zhang,T.,Yang,H.,Chen,D.,Wen,F., Guo, B. (2020). “Face X-ray for more general face forgery detection”. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5001-5010).

[4] Li, Y., and Lyu, S. (2019). “Exposing deepfake videos by detecting face warping artifacts ”. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (pp. 46-52).

[5] Afchar, D., Nozick, V., Yamagishi, J., and Echizen, I. (2018, December). “MesoNet: a compact facial video forgery detection network”. In 2018 IEEE International Workshop on Information Forensics and Security (WIFS) (pp. 1-7). IEEE.

[6] Chintha, A., Thai, B., Sohrawardi, S. J., Bhatt, K. M., Hickerson, A., Wright, M., and Ptucha, R. (2020). “Recurrent convolutional structures for audio spoof and video deepfake detection”. IEEE Journal of Selected Topics in Signal Processing.

[7] Guera, D., and Delp, E. J. (2018, November). “Deepfake video detection using recurrent neural networks”. In 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS) (pp. 1-6). IEEE.

[8] Ekraam Sabir, Jiaxin Cheng, Ayush Jaiswal, Wael AbdAlmageed, Iacopo Masi, Prem Natarajan(2019). ”Recurrent Convolutional Strategies for Face Manipulation Detection in Videos” . Applications of Computer Vision and Pattern Recognition to Media Forensics at CVPR 2019.

[9] Daniel Mas Montserrat, Hanxiang Hao, S. K. Yarlagadda, Sriram Baireddy, Ruiting Shao (2020), ”Deepfakes Detection with Automatic Face Weighting”. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 2851–2859. (2020).

[10] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, Cristian Canton Ferrer(2020). ”The DeepFake Detection Challenge (DFDC) Dataset” . Computer Vision and Pattern Recognition 2020.

[11] Hinton, G. E., Krizhevsky, A., and Wang, S. D. (2011, June). Transforming auto-encoders. In International Conference on Artificial Neural Networks (pp. 44-51). Springer, Berlin, Hei- delberg.

[12] Agarwal, S., Farid, H., Gu, Y., He, M., Nagano, K., and Li, H. (2019, June). Protecting world leaders against deep fakes. In Computer Vision and Pattern Recognition Workshops (pp. 38-45).

[13] Lin, J., Li, Y., Yang, G. (2021). FPGAN: Face de- identification method with generative adversarial networks for social robots. Neural Networks, 133, 132-147.

[14] Zhang, H., Goodfellow, I., Metaxas, D., and Odena, A. (2018). Self-attention generative adversarial networks. arXiv preprint arXiv:1805.08318

[15] Hao, H. et al. (2022). Deepfake Detection Using Multiple Data Modalities. In: Rathgeb, C., Tolosana, R., Vera-Rodriguez, R., Busch, C. (eds) Handbook of Digital Face Manipulation and Detection. Advances in Computer Vision and Pattern Recognition. Springer, Cham.

[16] Jiang, L., Wu, W., Qian, C., Loy, C.C. (2022). DeepFakes Detection: the DeeperForensics Dataset and Challenge. In: Rathgeb, C., Tolosana, R., Vera-Rodriguez, R., Busch, C. (eds) Handbook of Digital Face Manipulation and Detection. Advances in Computer Vision and Pattern Recognition. Springer, Cham.

[17] Suratkar, S., Sharma, P. (2022). A Simple and Effective Way to Detect DeepFakes: Using 2D and 3D CNN. In: Rao, U.P., Patel, S.J., Raj, P., Visconti, A. (eds) Security, Privacy and Data Analytics. Lecture Notes in Electrical Engineering, vol 848. Springer, Singapore.

[18] Kaliyar, R.K., Goswami, A. Narang, P. DeepFakE: improving fake news detection using tensor decomposition-based deep neural network. J Supercomput 77, 1015–1037 (2021).

[19] Chen, Joy Iong Zong, and S. Smys. ”Social Multimedia Security and Suspicious Activity Detection in SDN using Hybrid Deep Learning Technique.” Journal of Information Technology 2, no. 02 (2020): 108-115.

[20] Kumar, T. Senthil. ”Construction of Hybrid Deep Learning Model for Predicting Children Behavior based on their Emotional Reaction.” Journal of Information Technology 3, no. 01 (2021): 29-43.