Deep Fake Detection using Neural Networks

Anuj Badale
Information Technology
St.Francis Institute of Technology
Mumbai, India

Chaitanya Darekar
Information Technology
St.Francis Institute of Technology
Mumbai, India

Lionel Castelino
Information Technology
St.Francis Institute of Technology
Mumbai, India

Joanne Gomes
Information Technology
St.Francis Institute of Technology
Mumbai, India

***Abstract*— Deepfake is a technique for human image synthesis based on artificial intelligence. Deepfake is used to merge and superimpose existing images and videos onto source images or videos using machine learning techniques. They are realistic looking fake videos that cannot be distinguished by naked eyes. They can be used to spread hate speeches, create political distress, blackmail someone, etc. Currently, Cryptographic signing of videos from its source is done to check the authenticity of videos. Hashing of a video file into "fingerprints" (small string of text) is done and reconfirmed with the sample video and thus verified whether the video is the one originally recorded or not. However, the problem with this technique is that the fingerprints and hashing algorithms are not available with common people. In this paper the proposed system follows a detection approach of Deepfake videos using Neural Networks. Binary classification of deepfakes was done using combination of Dense and Convolutional neural network layers. It was observed that 91% accuracy was obtained in Adam and 88% was obtained in sgd(stochastic gradient descent) for categorial cross entropy. In binary cross entropy, 90% accuracy was seen in Adam and 86% accuracy was noticed in sgd whereas, 86% accuracy in Adam and 80% accuracy in sgd was obtained in mean square.**

***Keywords*— Deepfake, Binary Classification, Neural Networks, Convolutions , Pooling.**

I. INTRODUCTION

Most DeepFakes on the Internet include pornographic images of men, usually by female celebrities such as their often used without their consent. Extraordinary pornography is being released surfing the Internet in 2017, especially Reddit. DeepFake is also used to misrepresent famous politicians. In separate videos, Argentine President Mauricio Macri's face has been replaced by Adolf Hitler's face, and Angela Merkel's face has been replaced by Donald Trump's. The first known attempt to make a face-to-face exchange was seen [1] in the photograph of Abraham Lincoln. The lithography superimposes his head with the body of John Calhoun. The engravings of his head on other bodies appeared quite often after his assassination.

Authors David Guera Edward J. Delp put forth a paper based on Artificial Intelligence. The topic discussed in the paper is Convolution Neural Network(CNN), Recurrent Neural Network(RNN). The author tried to evaluate method against a large set of DeepFake videos collected from multiple video websites. Scenarios where these realistic fake videos are used to create political distress, black-mail someone or fake terrorism events are easily envisioned. This paper proposes a temporal recognition pipeline to automatically detect deep videos. It presents end-to-end trainable recurrent deepfake video detection system. The author claimed that it is not unusual to find DeepFake videos where the manipulation is only present in a small portion of the video (i.e.the target face only appears briefly on the video, hence the DeepFake manipulation is short in time). To account for this, for every video in the training, validation and test splits, the system extracts continuous subsequences of fixed frame length that serve as the input of the system. This system works only with large dataset. The authors proposed an analysis composed of CNN to extract features followed by RNN network to capture erratic frames in the face swapping process. For the proposed system, a set of 600 videos were analysed that were collected from multiple hosting websites. [1]

Authors Yuezun Li et al. in [2] put forth a paper based on Artificial Intelligence. The topic discussed in the paper was Convolution Neural Networks (CNN) and Recursive Neural Network (RNN). The author tried to create a new system that exposes fake faces based on eye blinking, that have been generated using Neural Networks. New developments in deep reproduction networks have greatly improved the quality and efficiency of producing authentic face videos. Therefore, in his paper, the author aims at analysing the eye blin king in the videos, which is a psychological signal that is not well presented in the synthesized fake videos. The authors have performed Preprocessing in the first step inorder to locate the face areas in

each frame of the video. Then they have used Long Term Recurrent Convolution Network (LRCN) to capture temporal dependencies, as human eye blinking shows strong temporal dependencies. The Model training is then done using 2 steps: In the first step, they trained the VGG based CNN model based on a set of labeled training data consisting of eye regions corresponding to open and closed eyes. The model is trained using back-propagation implemented with stochastic gradient descent and dropout by probability 0.5 in fully connected layers In step 2, the LSTM-RNN and the fully connected part of the network are co-trained using a back-propagation-by-time (BPTT) algorithm. The authors claimed that they have evaluated the LRCN methods with comparison to Eye Aspect Ratio (EAR) and CNN. They also used VGG16 as their CNN model to distinguish eye state. The author was able to make such claims because EAR method replies on eye landmarks to analyze eye state, in terms of the ratio between the distance of upper and lower lid, and the distance between left and right corner point. This method runs fast as merely cost in ratio computation. Also, CNN shows an exceptional well performance to distinguish the eye state on image domain. The author has displayed the authenticity of his claim via eye blinking detection on an original video and DeepFake generated fake video. Some gaps were presented by the author as the author implemented a full system with the intention of inventing a new method to detect fake videos such as; only eye blinking detection was done which is a relatively easy cue in detecting fake face videos. The author can try to find a more efficient system by adding hardware to the proposed system. Authors Gustavo B. Souza et al. in [3] put forth a paper based on Artificial Intelligence. The authors have discussed about the use of Width Extended Convolution Neural Networks. The authors tried to solve the problem of inefficiency of CNNs by implementing Width Extended Convolution Neural Networks (wCNN). The face is considered to be one of the most promising biometric features of human identification, including mobile devices. However, facial recognition systems can be easily fooled, for example, by providing a printed image, a 3D mask, or video recorded on the face of an official user. Recently, although some of the CNNs used (Convolutional Neural Networks) have obtained technical results in the detection of face loops, in many cases the proposed structures are very deep, because they are not suitable for limited hardware devices. In this work, we propose a functional architecture for face recognition based on the expanded CNN, which we call wCNN. Each part of the wCNN is trained, separately, in a given face area, where their output is computed to decide whether the faces presented on the sensor are real or fake.In order to evaluate the efficiency of the proposed wCNN in terms of processing required for face spoofing detection, they compared its performance with two state-of-the-art CNNs: Fine-Tuned VGG-Face a newly updated CNN based on random cassettes Instead of presenting the processing times, they present the amount of multiplication operations required by the adopted CNNs in the forward pass of each face image (or patches) for classification, since this measure is independent of the hardware used. Since the pass of the images through the neural networks is the core of the back propagation algorithm, the training of the CNNs is also usually much more complex for the architectures with more expensive forward passes. Its complexity tends to increase substantially since the backpropagation algorithm calculates partial derivatives for all the weights of the network. The author claimed that, besides presenting results compatible with state-of-the-art very deep CNNs, which they could not even train with their limited GPU, it saves lots of processing and time in training and test, being very suitable for environments with significant hardware restrictions, including mobile ones. The author has made such claims because of efficiency provided by the wCNN technology. The author has performed Face Spoofing Detection and Patch Net analysis as evidence for the results. As future work, they plan to evaluate the wCNN in other image domains, such as texture-based representations of the faces, and investigate the learning of local features for face spoofing detection in other color spaces. No new research problems can be thought of, based on the work done by the author. Authors Haya R. Hasan and Khaled Salah in [4]" Combating deepfake Videos Using Blockchain and Smart Contracts" put forth a paper on Blockchain Technology and Artificial Intelligence. The author proposes a blockchain based system for deepFake videos. The author tried to solve the scenarios where Fake footage, images, audios, and videos (known as deepfakes) can be a scary and dangerous phenomenon, and can have the potential of altering the truth and eroding trust by giving false reality. The recent rise of AI, deep learning, and image processing have led the way to the production of deepfake videos. Deep videos are dangerous, and can have the power to distort the truth, confuse viewers and misleading facts. With the onset of social media networks, the proliferation of such content may remain unchanged and may add to the problems associated with the fabrication and ideas of corporate strategies. The owner (original artist) of a video first creates a smart contract where other artists can request a permission to edit, alter or distribute according to the terms and conditions of an agreement form. The agreement form is saved on the IPFS server and its hash is available as an attribute in the smart contract. The secondary artist requests first permission to edit, alter or share. A request sent by the secondary artist is also a confirmation to the terms and conditions of the agreement form. This request is assessed by the original artist and the result is then announced. The contract can handle multiple requests at the same time and can handle multiple different requests by the same artist. Once an artist gets an approval to their request, they create a child contract which is similar to the original contract and they update the parent's information. The second artist then asks for proof of his new contract from the first artist for the first video contract. The original artist then approves and grants the attestation after checking the newly created smart contract. A successfully attested smart contract would then be added as a child in the original smart contract. Hence, both the contracts point to each other as each one has the Ethereum address of the other as part of their attributes. The author claims that he made use of a decentralized storage system IPFS, Ethereum name service, and decentralized reputation system. The proposed solution framework, system design, algorithms, sequence diagrams, and implementation and testing details are generic enough and can be applied to other types of digital content such as audios, photos, images, and manuscripts. The author claims it as they can help combat deepfake videos and audios by helping users to determine if a video or digital content is

traceable to a trusted and reputable source. The system also provides a trusted way for secondary artists to request permission from the original artist to copy and edit videos. The author has cited an example to assist a user in tracing back a video with multiple versions to its origin. If a video cannot be traced to its original publisher, then it cannot be trusted. The authors are in the process of developing front- end DApps for users to automate the establishment of proof of authenticity of published videos. Also, they plan to develop a pluggable DApp component to provide traceability and establish authenticity when playing or displaying videos within a web browser. Also, work is underway for designing and implementing a fully functional and operational decentralized reputation system. No new research problems can be thought of based on the work done by the author. Authors Shuvendu Rana, Sibaji Gaj, Arijit Sur and Prabin Kumar Bora in [5] put forth a paper based on Neural Network. The topic discussed in the paper are Convolution Neural Network (CNN), Dual tree complex wavelet transform (DT DCT), Depth image-based rendering (DIBR), Multiview video plus depth (MVD),3D highly efficient-video-coding(3D-HEVC). In this, the author tried to detect method to differentiate fake 3D video and real 3D video using CNN. The author tries to identify the real and fake 3D and pre- filtration is done using the dual tree complex wavelet transform to emerge the edge and vertical and horizontal parallax characteristics of real and fake 3Dvideos. The efficiency of the fake 3D video is examined over the training and testing dataset. Using the CNN, each video sequences in the training dataset is used to train the CNN. The author claimed that due to this the time complexity and huge computing resources is required to achieve desired accuracy. High-resolution video sequences are used for training. The author implemented CNN architecture for proposed scheme. The author can try to find a more efficient a powerful mechanism for detecting real and fake videos. Authors Neelesh Bhakt, Pankaj Joshi, Piyush Dhyani in [6] put forth a paper based on Artificial Intelligence. The topic discussed in the paper are Support Vector Machine, GIST, Video Processing. The author tried to detect and differentiate between fake and real smiles. An indicator of emotions, a smile can be categorized into two types. Some are real, originating from an exhilarated atmosphere, while some are fake. Hence, it becomes utterly difficult to differentiate between the two smiles. This research work is based on capturing the movement of zygomatic major and obicularis oculli , which plays a vital role in detecting whether a smile is fake or real. The training result on the two rounds conducted using GIST feature on the facial parts. The author has made the distinction in fake and real smile frames were done by the appearance of wrinkles on the cheeks. The author states that this technology is useful only for smiles on face and can include feature of eye blinking in addition to face. The authors in [9], propose a method based on deepfake video property that relates to limitation of production time and computation resources. According to them, the face images should undergo affine warping in order to match the face of the source.

In our proposed system the dataset comprises of 900 deepfake videos out of which frames were gathered and then frame level feature extraction was done using combination of Dense and Convolutional Neural Networks to detect the pixel manipulations in the sample video. It was observed that 91% accuracy was obtained in Adam and 88% was obtained in sgd(stochastic gradient descent) for categorical cross entropy. In binary cross entropy, 90% accuracy was seen in Adam and 86% accuracy was noticed in sgd whereas, 86% accuracy in Adam and 80% accuracy in sgd was obtained in mean square. Section II talks about the algorithms used for proposed system.

II. ALGORITHMS USED:

Neural Networks consist of network of neurons which are the computational units. Neurons consist of a number (initial input or the output of the previous layer) and an activation function. Activation functions are the non-linear functions used which determine the output of the neuron. The commonly used activation functions are ReLU (Rectified Linear Unit), tanh, sigmoid, etc. The connection between layers have weights present and every layer has a bias. Backpropogation in neural networks adjusts these weights and biases according to the label of the training data. Thus, the values of weights and bias of the real and deepfake manipulated frames are different. Similar properties in images cause similar neurons to fire and thus they have similar values of weights and biases.

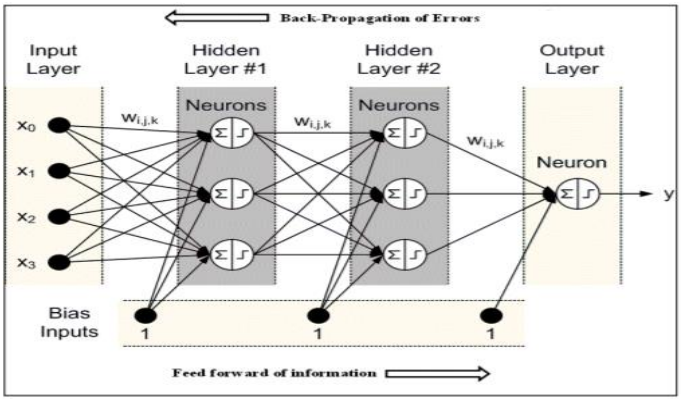

Fig. 1. Structure of Neural Network[11]

The convolutional neural network is composed of neurons that are not fully connected to the next layer. The neurons are connected on the basis of filters used i.e. if the filter used is a 3x3 filter nine neurons in the nth layer determine the output of one neuron in the (n+1)th layer. The convolution operation is done on normalized pixels of the image. The convolution operation multiplies filter values and pixels and then adds the values and thus features are extracted. If fxf filter is applied on nxn image, the resulting output is of dimension:

$$n_{out} = [\frac{n_{in}+2p-k}{s}] + 1 \quad (1)$$

n_{in}: number of input features

n_{out}: number of output features

k: convolution kernel size

p: convolution padding size

s: convolution stride size

The convolution operation thus reduces the dimension of image extracting features. Further Max pooling operation is done on pixels where the maximum value in the surrounding pixel is selected to reduce the spatial representation of the image and thus decreases the number of parameters in the network.

Optimizers in Neural Network:

Optimizers are the functions that change value of weights and bias in the network and thus adjust them appropriately for every output class. And it increases accuracy and decreases loss. Every problem given to the neural network is converted to an optimization problem. In most of the cases, the problem is solved using minimisation of error or the loss function. The loss function is given by

$$\frac{\partial Loss(y,\hat{y})}{\partial W} = \frac{\partial Loss(y,\hat{y})}{\partial y} * \frac{\partial \hat{y}}{\partial z} * \frac{\partial z}{\partial W} \quad where\ z = Wx + b$$

$$= 2(y - \hat{y}) *$$

$$derivative\ of\ sigmoid\ function * x$$

$$= 2(y - \hat{y}) * z(1 - z) * x \qquad (2)$$

Where y is the actual label and y' is the one calculated in each epoch. The optimization function adjusts every weights and biases after each epoch and thus steadily reduces error and increases accuracy. The commonly used optimization functions are sgd (stochastic gradient descent), categorical cross entropy, etc. Section III explains the Proposed methodology for the proposed system.

III. PROPOSED METHODOLOGY:

The pipeline gives all the detailed information about the implementation of our system. A dataset of real and manipulated images based on the videos was created in our system. After creating dataset, the images in our dataset were preprocessed. The CNN model was created after the preprocessing and then training was done on the training set. After training, the model was run on the testing and validation sets. Further, tuning the parameters and hyperparameters was done to increase accuracy and decrease loss of the system. Lastly, running the model on sample video and detecting classes for each frame of the video was done.

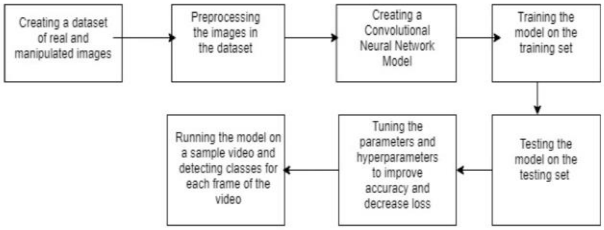

Fig. 2. Pipeline of Implementation for Deepfake Detection

Deepfakes and original Videos were used to extract frames of videos. Each frame was given a class i.e. "real" or "fake", depending on the video from which it is extracted. The images were pre-processed using Image Data Generator in Tensorflow and converted into an n dimensional array of size (256,256,3) where 256 is the height and width whereas 3 represents the number of channels. Further the inputs are flattened to make it of the size (256*256*3,1) to give it as an input to the first layer of the neural network. Thus, the data is made ready to be given as input. Further the model is made comprising sequential layers of Convolutional, Pooling and dense layers. The model comprises of 1 input layer,13 Convolutional layers,5 pooling and 3 dense layers. The hyperparameter tuning is done and the number of layers, activation function, optimizers and learning rate. Feature extraction is done using the convolution operation using 3x3 filters. The corresponding Fig. 3 explains 2D convolution operation as follows:

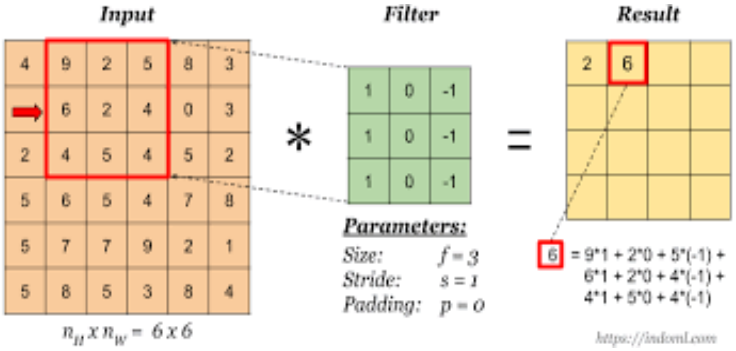

Fig. 3. 2D Convolution operation

Thus, the features of the corresponding 9 pixels are extracted and thus this convolution operation extracts features of parts of images. These features of parts of images are then used for detection which is done on the basis of result of the computation of the convolution operator. The pooling layers then select the maximum computed value among the values

selected in the n dimensional array. Pooling layers reduce the number of computations. These layers are then connected to the final layer which contains 2 neurons fully connected to the previous layer. Tanh activation function is used in the layers which squashes the output of the previous layer into the range of (-1,1) as shown in Fig. 4.

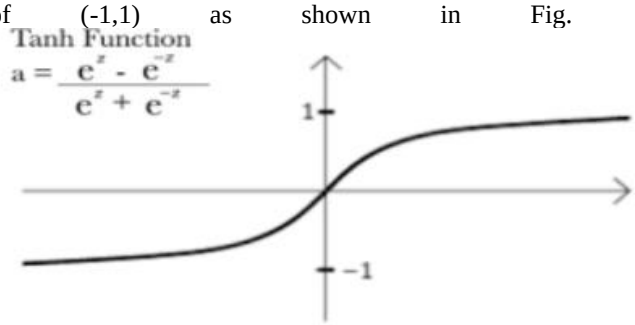

Fig. 4. Activation Function

Section IV displays the results and observations that were seen for the proposed methodology.

IV. RESULTS AND OBSERVATIONS:

The model was trained on images, labelled as 'Fake'(Deepfakes) and 'Original' from the dataset and accuracy and loss calculations were obtained.

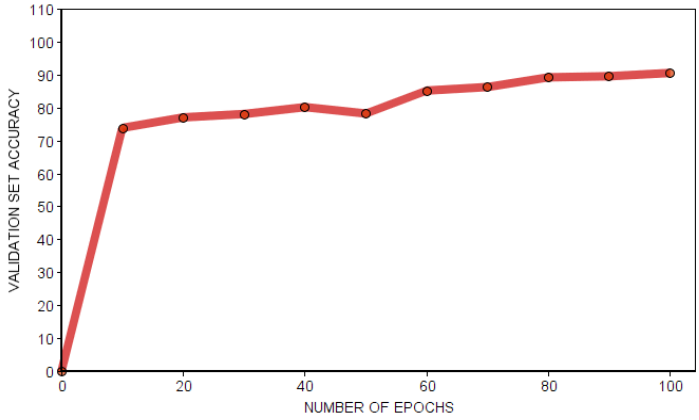

Fig. 5. Accuracy of validation set (Adam)

The predictions were done for the images in the dataset of the model. The function predict_ classes() predicts the class that the images belong to. The accuracies were observed and seen to be the most for Categorial Cross Entropy. The above Fig. 5 shows the increasing accuracy of validation sets as number of epochs increase for Adam optimizer of Categorial Cross Entropy. The below Fig. 6 shows the accuracy for SGD optimizer for Categorial Cross Entropy.

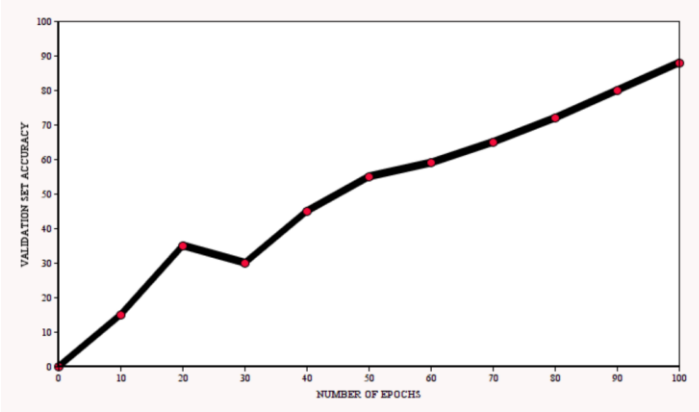

Fig. 6: Accuracy of validation set (SGD)

Following is the classification obtained for the dataset:

Fig. 6. Prediction for image dataset of Deepfake Detection

Combinations of different optimization functions and loss functions were tried and accuracy was recorded as shown in Table 1.

TABLE 1: Accuracies for various optimizers and loss functions.

	Adam	SGD
Categorial Cross Entropy	91%	88%
Binary Cross Entropy	90%	86%
Mean Square	86%	80%

Fake images are predicted as 0.1 and real images are predicted as 1.0.
Further the saved model was applied on each frame of the video and the manipulated part in the video is highlighted.

Fig. 7. Frame classified as REAL

When the model runs in real time on a sample video, the faces of the video are detected and the frames in the video that are real are highlighted as REAL as shown in Fig 7, whereas the frames that are fake are highlighted as FAKE as seen in Fig 8.

Fig. 8. Frame classified as FAKE

Section V summarizes and explains the future scope for the proposed work.

V. CONCLUSION:

Thus, Deepfake videos were studied and analyzed using this methodology and it also produces a good level of accuracy. The frames of the video were extracted and preprocessing was done. Subsequently, Image Classification was done and the images were labelled. With the help of Machine Learning algorithms, predictions were made on the dataset. Thus, any video can be analyzed using this methodology. It helps in detecting fake faces in a video which may have been manipulated, hence can prevent individuals from being defamed unknowingly.

Further, different combinations of hyperparameters with respect to Neural Networks can be used and hyperparameter tuning can be done for the purpose of studying Deepfakes and the outputs of those algorithm models can be analyzed and compared, so that Deepfakes can be combated in the most efficient way, as it is one of the major threats looming large over the authenticity of videos. Modern technologies like Blockchain can be used for immutable storage inorder to preserve the originality of videos.

REFERENCES

[1] D. Guera and E. J. Delp, Deepfake Video Detection Using Recurrent Neural Networks, 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Auckland, New Zealand, 2018, pp.1-6.

[2] Y. Li, M. Chang and S. Lyu, In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking, 2018 IEEE International Workshop on Information Forensics and Security (WIFS), Hong Kong, Hong Kong, 2018, pp. 1-7.

[3] G. Botelho de Souza, D. F. da Silva Santos, R. Gonsalves Pires, J. P. Papa and A. N. Marana, Efficient Width-Extended Convolutional Neural Network for Robust Face Spoofing Detection, 2018 7th Brazilian Conference on Intelligent Systems (BRACIS), Sao Paulo, 2018, pp. 230-235.

4] H. R. Hasan and K. Salah, Combating Deepfake Videos Using Blockchain and Smart Contracts, in IEEE Access, vol. 7,2019, pp. 41596-41606.

[5] S. Rana, S. Gaj, A. Sur and P. K. Bora, Detection of fake 3D video using CNN, 2016 IEEE 18th International Workshop on Multimedia Signal Processing (MMSP), Montreal, QC, 2016, pp. 1-5.

[6] N. Bhakt, P. Joshi and P. Dhyani, A Novel Framework for Real and Fake Smile Detection from Videos, 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, 2018, pp. 1327-1330.

[7] de Souza, G. B., da Silva Santos, D. F., Pires, R. G.,Papa, J. P., & Marana, A. N. (2018, October). Efficient Width-Extended Convolutional Neural Network for Robust Face Spoofing Detection. In 2018 7th Brazilian Conference on Intelligent Systems (BRACIS) (pp. 230-235). IEEE.

[8] Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2018). Faceforensics: A large-scale video dataset for forgery detection in human faces. arXiv preprint arXiv:1803.09179.

[9] Li, Y., & Lyu, S. (2018). Exposing deepfake videos by detecting face warping artifacts. arXiv preprint arXiv:1811.00656, 2.

[10] Zhang, Z. (2019). Detect forgery video by performing transfer learning on Deep Neural Network (Doctoral dissertation).

[11] Almogdady, H., Manaseer, S., & Hiary, H. (2018). A Flower Recognition System Based On Image Processing And Neural Networks. International Journal Of Scientific & Technology Research, 7(11).