2022 2nd Asian Conference on Innovation in Technology (ASIANCON)
Pune, India. Aug 26-28, 2022

CNN based Deep Learning model for Deepfake Detection

1st Vedant Jolly
Computer Engineering Department
Sardar Patel Institute of Technology
Mumbai, India
vedant.jolly@spit.ac.in

2nd Mayur Telrandhe
Computer Engineering Department
Sardar Patel Institute of Technology
Mumbai, India
mayur.telrandhe@spit.ac.in

3rd Aditya Kasat
Computer Engineering Department
Sardar Patel Institute of Technology
Mumbai, India
aditya.kasat@spit.ac.in

4th Atharva Shitole
Computer Engineering Department
Sardar Patel Institute of Technology
Mumbai, India
atharava.shitole@spit.ac.in

5th Kiran Gawande
Computer Engineering Department
Sardar Patel Institute of Technology
Mumbai, India
kiran_gawande@spit.ac.in

2022 2nd Asian Conference on Innovation in Technology (ASIANCON) | 978-1-6654-6851-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/ASIANCON55314.2022.9908862

***Abstract*—In the recent period there has been massive progress in synthetic image generation and manipulation which significantly raises concerns for its ill applications towards society. This would result in spreading false information, leading to loss of trust in digital content. This paper introduces an automated and effective approach to get facial expressions in videos, and especially focused on the latest method used to produce hyper realistic fake videos: Deepfake. Using faceforenc++ dataset for training our model, we achieved more that 99% successful detection rate in Deepfake, Face2Face, faceSwap and neural texture. Regular image forensics techniques are usually not very useful, because of the strong deterioration of data due to the compression. Thus, this paper follows a layered approach with first detecting the subject with the help of existing facial recognition networks followed by extracting facial features using CNN, then passing through the LSTM layer, where we make use of our temporal sequence for face manipulation between frames. Finally use of the Recycle-GAN which internally makes use of generative adversarial networks to merge spatial and temporal data.**

***Index Terms*—Face Detection, FaceForensics++, DeepFake, Face2Face, FaceSwap, Neural Texture, Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM)**

I. Introduction

Over the decades, the popularity of smartphones and the growth of social media have led to the adoption of digital photos and videos into the most popular digital assets. On Youtube alone, 300 hours of video are uploaded every minute. Every day, 5 billion videos are viewed and 1 billion hours are broadcast, with Facebook and Netflix streaming combined. This major use of digital photography is followed by an increase in photo editing techniques, using editing software such as Photoshop as an example. The proliferation of deepfakes in recent years raises serious concerns about the authenticity of digital content by the media and other online forums. For example, Deepfake (based on "deep reading" and "deception") is a method that can put more than a person's facial expression directed at a source video to create a video of a target person acting or referring to a source has demonstrated how computer graphics and visual effects can be used insult people by changing their faces to look different faces person. A basic way to create deepfake in-depth learning models such as autoencoders and competing production networks, widely used in the field of computer vision. These models are used to assess a person's facial expressions and movements and to combine images of another person's face making similar expressions and movements [1]. In-depth fraudulent methods often require large amounts of image and video data to train models to make real photos and videos. While public figures such as celebrities and politicians may have a large number of videos and photos available online, they are the first deep victims [1]. Many politicians and actors became victims of Deepfakes. For criminal purposes, forensic videos are converted using novel methods such as face swap and faceswap-GAN. Analysing this issue there have been several methods of obtaining deceptive images, most of them or analyze the inconsistencies compared to the conventional ones the camera pipe will be or rely on the release of something image changes in the resulting image. Opposition (seeing) as video fraud, a few algorithms using hand-made features, in-depth learning algorithms, and more recently GAN-based methods are being tested. For example, manual methods include steganalysis methods, detect 3D head position inconsistencies, etc. However, there is still a room for the development of the modern industry finding deepfakes, especially in challenging data such as a database of Face Forensics (FF ++). In this paper we have selected the base architecture of our model as ResNet18. The reason for choosing the ResNet18 model was because it takes care of a major problem like extinction and explosion of the gradient. Make this happen by using something known as skipping communication. The advantage of adding this type of link is that if any of the layers damage the performance of the structures it will be omitted normally. The biggest

challenge in getting a deep mock image is how well we process our image data, based on which we will highlight the most important similarities. Our model used the CNN model in feature releasing. This is followed by transferring them to the LSTM layer.

II. Literature Review

We cover the most important related research done in the deepfakes in the following paragraphs.

1) Deepfake Methods: In last couple of decades interest in virtual face manipulation has increased greatly. Deepfakes methods could be divided into different types of face manipulation methods. STYGAN model synthesizes entire non-existent face through GAN model. These approaches produce incredible outcomes, such as high-resolution facial images with a great degree of realism. The identity swap technique, also called the face-swap method, is very popular for replacing the face of one person in an image or video with that of another person. This could be achieved through 2 different approaches. One is graphics based approaches such as FaceSwap and another is deep learning technique-based approaches such as DeepFakes.
   Attribute manipulation, also known as face editing or face retouching, entails changing aspects of the face, such as hair or skin color, gender, age, and the addition of spectacles This manipulation process is usually carried out through a GAN, such as the StarGAN approach. Expression swap, also known as face reenactment, modifies the facial expression of a person. Face2face changes the facial expression of the person in the video based on the expression input given by another person. Some more techniques exist for face morphing, which creates biometric face samples that resemble the given biometric information.
2) Deepfake Detection: DeepFake detection dominates research on monitoring multimedia information and has the positive intention to improve the confidentiality and integrity of multimedia content. In recent years CNN based generated multimedia detection has become more popular.
   A novel photo-response nonuniformity (PRNU) analysis method has been tested for its effectiveness at detecting DeepFake video manipulation. This PRNU analysis reveals a statistically significant difference in mean normalized cross-correlation scores between real and DeepFake Videos [2]. Lugstein designed a novel pipeline to detect DeepFakes using photoresponse nonuniformity (PRNU). Basically, the PRNU technique is famous for detecting facial retouching and face morphing attacks. Two types of mesoscopic (a compact facial video forgery detection network) models (Meso-4 and MesoInception-4) have been proposed by Afchar to classify hyperrealistic forged videos based on DeepFake and Face2Face. It is obvious that uncompressed videos are severely degraded by image noise, wherein microscopic investigation-based image noise is not applicable. Moreover, the models are efficient in detecting hyper realistic forged videos at a low computational cost. The average detection efficiency rate was found to be 98% for DeepFake videos and 95% for Face2Face videos under real conditions of diffusion on the internet [3].
   Faceforensic++ designed a novel large-scale dataset of manipulated facial imagery composed of more than 1.8 million images from 1,000 videos with pristine (i.e., real) sources and target ground truth to enable supervised learning [4]. Research paper published by by faceforensic++ in 2019 used faceforensic++ dataset to train CNN model, which was tailored to detect face manipulations. In Lips Don't Lie, Haliassos suggested a generalizable and robust approach based on resenet-18 to detect face forgery in videos using the semantic irregularities of lips movement, which is also known as LipForensics. Jeon proposed a transferable GAN-image detection framework (T-GD) technique, which efficiently detects DeepFake images. The model works on teacher and student relations, which mutually improve the detection performance.

III. System Algorithm

The overall algorithm consists of three major components:

1) Dataset Used:
   The dataset which we have used is FaceForensics++. This dataset consists of about 1000+ manipulated youtube videos as well as around 1 million+ images for the same. This dataset was provided by Google and JigSaw. The novelty of this dataset is that along with the data which it is providing, it also provides an automated benchmark for facial manipulation detection. In particular, the benchmark is based on DeepFakes, Face2Face, FaceSwap and NeuralTextures as prominent representatives for facial manipulations at random compression level and size [4]. Another unique aspect of the dataset is that, in the videos which are manipulated FaceShifter has also been applied, so that there is no lag when the DeefFake is applied over the video, which makes detecting DeepFake even more difficult.

2) DeepFake Methodology:
   Over the years, the DeepFake model has been developing steadily, where it has come to a point, that it is visually impossible to tell the difference between a DeepFake video and a video that was original.
   Firstly, let's look at some of the techniques which DeepFake makes use of for manipulating videos and images.
   a) Face2Face: It is also known as a facial reenactment. The main purpose of this method is to transfer expression from the source to the target photo.

b) FaceSwap: It is used for facial identity manipulation. It is a graphics-based approach that makes use of each frame to create a model of the source and then project this new model onto the target by minimizing the distance between 2 frames.
c) NeutalTextures: This is an old technique that makes use of GANs for facial reenactment.

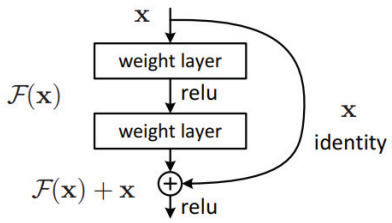

Fig. 1. Residual Block [6]

The base architecture of our model is based on ResNet18. The reason for choosing the ResNet model was because it takes care of the major problem like the vanishing and the exploding gradient. It makes this possible by using something known as skip connections. The basic idea behind the skip connection is that for some layers it skips training and directs connects to the output layer. This helps the model to learn the underlying mapping and thus allow the network to fit the residual mapping. The advantage of adding this type of skip connection is that if any layer hurt the performance of architecture then it will be skipped by regularization [5].

IV. Proposed System

The system consists of two major components:

1) Detecting a DeepFake Image:
The main challenge in detecting a deep fake image is how well are we pre-processing our image data, based on which we are going to highlight the most important aspects for the same. The primary task of our model is to detect the subject which is going to get analyzed further, for this we have made use of the existing facial recognition networks. After the face has been detected, the next step is to fine-tune the facial features of the current image so that most noises can be removed from the image.
We have made use of a CNN based deep learning model so that we can detect even the forensics model's deep fake images. The CNN model makes use of Gaussian Blur and Gaussian Noise so that we can ignore the noise as well as high-frequency sounds which are irrelevant in the detection of the face. The advantage we get by applying this model is that we can eventually recognize more meaningful characteristics, thus increasing the accuracy of our model.

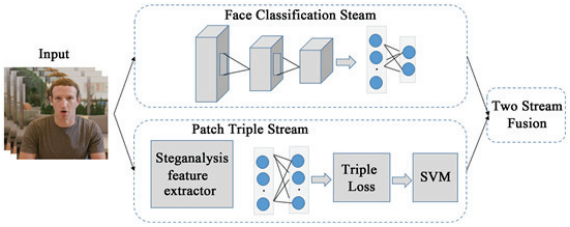

Fig. 2. Two Stream Neural Network Architecture [7]

2) Detecting the DeepFake Video:
We started our research based on analyzing a single frame at a given time, even though the accuracy achieved by this method was at-par but the time taken by this model was not justifiable towards the accuracy that was achieved. We then tried on a variety of techniques for parallelly running frames for detection. We made use of temporal sequence between frames which helped our model to detect the deep fake videos. Our model made use of a CNN model for the feature extraction. After the features are detected, we pass them on towards our LSTM layer, where we made use of our temporal sequence for face manipulation between frames. The last layer of our model consists of a softmax function which is used to classify the deep fake videos into the correct categories. We also made use of the Recycle-GAN which internally makes use of generative adversarial networks to merge spatial and temporal data [7]. The benefit of using the Recycle-GAN is that while it is evaluating, it passes the results back towards the start of the network so that at the same time the model can analyze its mistakes and manipulate the factors accordingly.

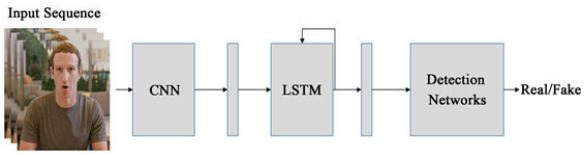

Fig. 3. ConvNet for spatial and temporal features analysis [7]

V. Result Analysis

1) FaceForensics++ dataset for trained models: For models trained only on paper databases, we realize that the model only learns to detect fraudulent strategies stated on paper and avoid all frauds in real-world data from data. This model has been zoomed in to the target image by minimizing the contrast between the proposed shape and local landmarks using the input image texture. At the end, the database model is combined with the image

with the specific correction related to color thesis applied accurately. We apply these steps to all the individual pairs and targeted pairs until one video ends. The frames detected for output are then used to make to surface populated with high concentration and density (refer Fig. 4a, 4b). Then these collected frames are used to connect properly with the dataset faces under various facial expressions and lighting conditions. To analyze the dataset videos precisely, we used the Face2Face method in order to duplicate the frames and achieve the required result. We process each video through a pre-processing world; here, we use the first frames to get a temporary face recognition (i.e., 3D model), and track additions over the remaining frames.

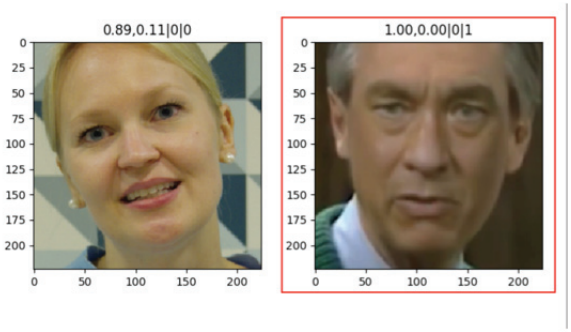

Fig. 4. Result Set 1

2) YouTube dataset for trained models: Models trained with one-to-one YouTube data learn to find real-world deepfakes, but also learn to find simple deepfakes on paper databases. These models however failed to detect any other type of deception (such as NeuralTextures). The large FaceForensics ++ database enables us to train a modern counterfeit image detector to detect surveillance (refer Fig. 5a, 5b). In this case, we use three default facial expressions, which are used in our database. To mimic real-life situations, we have chosen to collect videos anywhere online and on YouTube. The initial testing by inculcating the mentioned methods made us realize that the dataset face taken must be redirected with minimum delay in order for test to not fail and thereby produce accurate results. So, we did a personal review of the resulting clips to ensure the selection of high-quality video and to avoid videos with an explicit face. We have selected approxmiately 300,000 images for our dataset in order to be implemented by the above three mentioned algorithms.. All performed tests are done using the dataset videos from the set. The NeuralTextures method is based on the geometry that is used during the train and test times. The Face2Face module was used to produce the gathered information. It was used to identify and correct the expressions using the mouth regions only. The other parts like eye area were not modified, since if it were to modify then the network would additionally require extra input based on the movements of the eye.

Fig. 5. Result set 2

TABLE I
ACCURACY OF DIFFERENT ALGORITHMS TESTED

Method	Train	Validation	Tests	Raw	HQ	LQ
Bayar and Stamm	280374	52359	56382	98.74%	82.97%	66.84%
Rahmouni et al.	280342	52356	56371	97.03%	79.08%	61.18%
MesoNet	295164	55317	60540	95.23%	83.10%	70.47%
XceptionNet	295578	55384	60614	99.26%	95.73%	81.00%

VI. CONCLUSION

Deepfakes has led to people believe less in media and seeing them as less trustworthy and consistent with their contents. They may cause distress and ill effects to those who are targeted, nurture inadvertent knowledge and hate speech, and they may provoke political unrest, burning up society, violence, or war. This is especially important these days as the technology for creating deepfakes is very close and social media can spread that untrue content quickly. Sometimes deepfakes do not need to be distributed to a large audience to create harmful effects. People who build deepfakes with malicious intent only need to bring them to the target audience as part of their destructive strategy without using a social media platform. As new methods of deception emerge by the day, it is necessary to develop methods that can detect fakes with minimal training data. Our website is already being used for this legal transfer learning process, where the knowledge of one source of fraud is transferred to another targeted domain. We hope that the database and benchmark will be a stepping stone to future research in the field of digital media intelligence, and especially with a focus on face-to-face fraud. To summarize this, we were able to propose an automatic benchmark for face change acquisition under random pressure for standard comparison, including the human base. Comprehensive modern handmade experiments and counterfeiters learned in a variety of contexts are also shown in a modern way of finding counterfeit designed for facial modification.

REFERENCES

[1] Nguyen, Thanh Nguyen, Cuong M. Nguyen, Tien Duc, Thanh Nahavandi, Saeid. (2019). Deep Learning for Deepfakes Creation and Detection: A Survey.

[2] Korus, Pawel Huang, Jiwu. (2016). Multi-Scale Analysis Strategies in PRNU-Based Tampering Localization. IEEE Transactions on Information Forensics and Security. PP. 1-1. 10.1109/TIFS.2016.2636089.

[3] 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), 2018, pp. 1-7, doi: 10.1109/WIFS.2018.8630761.

[4] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies and M. Niessner, "FaceForensics++: Learning to Detect Manipulated Facial Images," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1-11, doi: 10.1109/ICCV.2019.00009.

[5] K. Zhang, M. Sun, T. X. Han, X. Yuan, L. Guo and T. Liu, "Residual Networks of Residual Networks: Multilevel Residual Networks," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 28, no. 6, pp. 1303-1314, June 2018, doi: 10.1109/TCSVT.2017.2654543.

[6] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

[7] Almars, Abdulqader. (2021). Deepfakes Detection Techniques Using Deep Learning: A Survey. Journal of Computer and Communications. 09. 20-35. 10.4236/jcc.2021.95003.

[8] B. Malolan, A. Parekh and F. Kazi, "Explainable Deep-Fake Detection Using Visual Interpretability Methods," 2020 3rd International Conference on Information and Computer Technologies (ICICT), 2020, pp. 289-293, doi: 10.1109/ICICT50521.2020.00051.

[9] S. Agarwal, N. Girdhar and H. Raghav, "A Novel Neural Model based Framework for Detection of GAN Generated Fake Images," 2021 11th International Conference on Cloud Computing, Data Science Engineering (Confluence), 2021, pp. 46-51, doi: 10.1109/Confluence51648.2021.9377150.

[10] N. S. Ivanov, A. V. Arzhskov and V. G. Ivanenko, "Combining Deep Learning and Super-Resolution Algorithms for Deep Fake Detection," 2020 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus), 2020, pp. 326-328, doi: 10.1109/EIConRus49466.2020.9039498.