Digital Forensics and Analysis of Deepfake Videos

Mousa Tayseer Jafar
mou20178003@std.psut.edu.jo

Mohammad Ababneh
m.ababneh@psut.edu.jo

Mohammad Al-Zoube
mzoube@psut.edu.jo

Ammar Elhassan
a.elhassan@psut.edu.jo

Princess Sumaya University for Technology
Amman, Jordan

2020 11th International Conference on Information and Communication Systems (ICICS) 978-1-7281-6227-0/20/$31.00 ©2020 IEEE 10.1109/ICICS49469.2020.239493

***ABSTRACT* – The spread of smartphones with high quality digital cameras in combination with easy access to a myriad of software apps for recording, editing and sharing videos and digital images in combination with deep learning AI platforms has spawned a new phenomenon of faking videos known as Deepfake. We design and implement a deep-fake detection model with mouth features (DFT-MF), using deep learning approach to detect Deepfake videos by isolating, analyzing and verifying lip/mouth movement. Experiments conducted against datasets that contain both fake and real videos showed favorable classification performance for DFT-MF model especially when compared with other work in this area.**

Keywords – Deep Learning, Python, Deepfake, Videos; Digital Forensics; Manipulation; Detection; Classification; Segmentation.

I. INTRODUCTION

Technological innovation especially in handheld devices with high fidelity cameras in combination with the spread of artificial intelligence tools, models and apps have resulted in a huge number of videos of world famous celebrities and leaders that have been doctored to convey fake news for either political gains or in order to ridicule certain people.

With billions of digital images and videos exchanged daily across several of the mainstream social media platforms together with the advent of deep learning apps allow us to create a fake video in minutes. Videos can be sourced from huge repositories that are available on the internet, and with a few steps in deep learning apps, it is very easy to create genuine-looking *deepfake* videos. The potential impact of a fake video of a world leader can be a catastrophic and can have far-reaching implications for the world's economy and political stability.

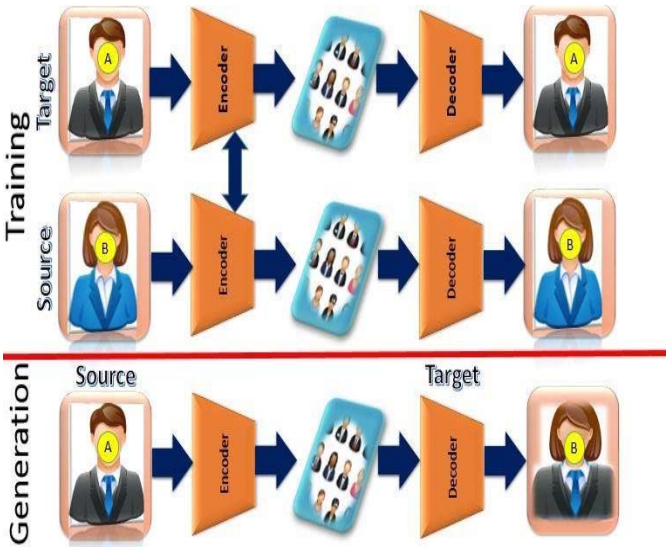

Figure 1 : Creation Deepfake videos

This paper introduces a mechanism for detecting deepfake videos; it is organized as follows: Section 1 introduces definition and methods for the creation deepfake videos. In section 2, the related, background work on deepfake detection methods will be reviewed. In section 3, we present a selection of two datasets: Celeb-DF [1] and Deepfake Video-TIMIT [2][3] that contain fake and genuine videos. In section 4, the proposed methodology for detecting deepfake videos will be discussed in more detail. Experimental results and their evaluation will be presented in section 5.

II. CREATION OF DEEPFAKE VIDEOS

Deepfake is a video that has been constructed to make a person appear to say or do something that they never said or did. The first appearance of Deepfake was in early 2018 [4], through the utilization of generative adversarial networks (GANs), which have led to the development of tools like Open Face, Swap Face2Face and Fake App that can generate videos from a large volume of images with minimum manual editing. Table 1 contains some of Common tools that Used to Create Deepfake videos. The main target of deepfake algorithm allows users to /transpose/replace the face of one person in a video with the face of another in a seemingly, realistic manner. To build deepfake videos, the fakers need to assign two important GAN algorithm components, namely: **the encoder network** that will help to achieve a dimensional reduction by encoding the data starting from the input layer until it reduces the number of variables. The second one is the **decoder network** which reduces variables to create a new output very similar to the original [5] as illustrated in figure 1.

In this work, the novel method will be introduced to detect deepfake videos and to get high detection accuracy more effectively and efficiently than other, common methods. The DFT-MF model will be tested on two new datasets: The Deepfake Forensics (Celeb-DF) dataset and the Deepfake Vid-TIMIT dataset. The proposed method utilizes Convolutional Neural Networks (CNN) to export videos into frames (images) and subsequently convert these images into gray-scale images to be processed and classified within one of the various deep learning type. Note that deep learning can be of 4 (architecture) types: Deep Boltzmann Machin (DBM), Deep Belief Networks (DBN), Convolutional Neural Networks (CNN) and Stacked Auto-Encoders. In this study, Convolutional Neural Networks (CNN) will be chosen to construct our deep learning model [21].

There are several apps utilize AI algorithm to enable users with limited knowledge of programming and machine learning to create deepfake video in order to transpose the face of one person keeping the facial expression of the original (speaker) person. Users can make fake videos of state leaders addressing

press and similar serious scenarios. These videos could, potentially, be presented as court evidence exploiting our inclination to trust the reliability of evidence that we see with our own eyes or to create political distress to destabilize international relations, blackmail someone or fake terrorism events.

Deepfake can be used in ways that are highly fraudulent with extreme repercussions; government officials in security, health, transport and defense or candidates in a political campaign can be targeted by manipulated videos in which they appear to say things that could harm their chances for election. Deepfake can affect people and may also lead to the devastation of own lives. This poses a new set of forensic challenges regarding the reliability of digital video evidence.

Table 1: Tools Commonly Used to Create Deepfake videos

App	Link
DeepFaceLab	https://github.com/iperov/DeepFaceLab
DFaker	https://github.com/dfaker/df
DeepFake-tf	https://github.com/StromWine/DeepFake-tf
Faceswap-GAN	https://github.com/shaoanlu/faceswap-GAN
FaceCrop	https://www.luxand.com/facecrop
Face Swap Live	http://faceswaplive.com
Face Swap Online	https://faceswaponline.com
NaturalFront	https://naturalfront.com
Reflect	https://reflect.tech
Voicery	https://www.voicery.com
Face2Face	https://web.stanford.edu/~zollhoef/papers/CVPR2016_Face2Face/page.html

III. RELATED WORK

Much research work has been conducted on different methods in this area. In the work presented by Xin Yang, Yuezun Li, and Siwei Lyu [6], they used a new Support Vector Machines (SVM) based method to detect deepfake videos by comparing the face landmarks between the real images and fake images; the authors found landmark locations were changed in both real and fake images. That means the face landmark locations in the altering face are huge but in the real face, it is small. On the other hand, the authors used the head pose as a benchmark to detect deepfake videos. Because the difference between central and whole face boundary in the fake video tends to be large, in contrast, it is a lot smaller in real video. The authors applied this approach on two datasets: the UADFV dataset which comprises 49 real videos and using it to generate 49 deepfake videos, the other dataset is DARPA MediFor which contains 241 real images and 252 Deepfake images. These datasets were fed to (SVM) classifier to detect deepfake videos. They used a small database and after evaluating their method using the Area Under ROC curve (AUROC) [22] for analysis the performance. With an attained ratio of 0.89 against the UADFV dataset and 0.843 against the DARPA MediFor dataset, a performance figure of 54.8 was obtained when applying this approach on the Celeb-DF Dataset [1].

Other methods presented by Darius Afchar, Vincent Nozick, Junichi Yamagishi and Isao Echizen [7], they used deep learning algorithms to detect deepfake videos and focusing on the compression artifacts for images in the videos. Their approach is based on two Convolutional Neural Networks using Meso-4 and MesoInception-4 models to analyze intrinsic characteristics of images. No dataset was available for deepfake videos and hence, they created bespoke test videos from the internet and subsequently evaluated their approach using the ROC for performance analysis and found the Meso-4 scored approximately 0.891 and the MesoInception-4 approximately 0.917 with an average deepfake detection rate of 98%. In contrast, when applying this method against the Celeb-DF dataset the Meso-4 model scores 53.6 and the MesoInception-4 49.6. The performance discrepancy between results is because the dataset was collected by authors, which exhibit promising detection performance. Furthermore, the database was verified manually to eliminate misalignment and mis-classification. This is time consuming in databases containing thousands of images.

David Guera and Edward J. Delp in [4] integrated two deep learning algorithms to detect fake videos. Each video frame is analyzed as it passes through two stages of scrutiny and analysis. The first algorithm is Convolutional Neural Network (CNN) that is used to retrieve features from frames of fake videos, the second algorithm is Recurrent Neural Network (RNN) and Long Short Term Memory (LSTM) which is used to detect fake videos after the training phase. Relying on the inconsistencies between the frames and the time discrepancies that appear after the creation of deepfake video they classify fake and real videos. They categorized videos into fake and real videos based on the number of fake frames in each video, and these frames should be sequential without jumping into other frames in each video.

Their approach analyzes fake video of the face which is part of the video that appears in a small portion of the total video time such that the duration of the fake does not exceed a certain number of frames. This algorithm, Conv-LSTM, was applied to several frames, including 20, 40 and 80 frames such that it does not exceed two seconds per video. The accuracy of each classification appeared as follows: ConvLSTM with 20 frames sequential 96.7, Conv-LSTM with 40 frames sequential 97.1 and Conv-LSTM with 80 frames sequential 97.1. The authors used videos collected from different websites as a dataset in combination with two deep learning algorithms which is very time-consuming. In DFT-MF model (below), 50 frames will be used in the whole video for more powerful and accurate analysis that does not exceed two seconds.

Yuezun Li, Siwei Lyu in [8] presented a model based on comparing the face and the sides of the face in both real and fabricated image frames from the video. They used four models to check their method VGG16 [9], ResNet50, ResNet101 and ResNet152 [10]. They collected real images from different resources on the internet and the fake images were created by reducing the resolution and changing some properties like brightness, contrast, distortion and sharpness and ran epoch times of 100 for VGG16, and 20 For ResNet50, ResNet101 and ResNet152 models. By evaluating their model against two datasets: UADFV and Deepfake TIMIT datasets, they managed to obtain high accuracy using the Area Under Curve (AUC) metric for ResNet50 of 97.4, 99.9 and 93.2 for

UADFV, Deepfake TIMIT LQ and Deepfake TIMIT HQ respectively.

The model above was built from images that were not released from deepfake video programs, but from real images that were manipulated via opacity settings to deliberately reduce their fidelity to become similar to fake images produced from deepfake videos. The accuracy of this method will increase due to the visible differences between low and high-resolution images, especially as most deepfake videos tend to have high resolution, making it difficult to distinguish original frames from fabricated ones. The accuracy for this model was 53.8 on the Celeb-DF dataset.

IV. DEEPFAKE DATASETS

In this section, we introduce two of the most common datasets used to test deepfake detection algorithms. The first one is The Deepfake Forensics (Celeb-DF) dataset [1], A new dataset for deepfake that was created for development and evaluation of deepfake detection methods/models. This dataset was distinguished by a notable reduction in the visible artifacts that were recognized in existing datasets, such as color variation with the surrounding area of the synthesized face combined with grained or reduced video quality, visible boundaries. Although existing datasets provide plenty of synthesized videos, the low quality and easy to note defects make it easier for deepfake detection methods to attain high accuracy on these datasets.

The Celeb-DF dataset contains synthesized videos of excellent visual quality. Specifically, it comprises 408 real videos and 795 synthesized videos generated with deepfake tools. The average length of videos in this dataset is 13 seconds, and they all run at a standard 30 frames per second.

The synthesized videos in the Celeb-DF dataset have been enhanced from the original auto-encoder structure by including extra convolutional layers to the decoder and generating synthesized faces of 256×256 pixels to counter the low resolution of the synthesized face.

Color contrast between synthesized faces and the original ones is another popular visual artifact in existing dataset. This is frequently caused by the mismatch between color and brightness between the target and original faces used in training. To mitigate this issue, the Celeb-DF dataset applies an enhancement process/noise reduction to the training faces: by randomly changing the illumination, variation, color distortion and sharpness of the input image with a random scale during training step. This improves the diversity of the training data and effectively reduces color mismatch in the Celeb-DF dataset.

Existing datasets normally use a narrow mask to determine the regions of the detected face landmarks. Combined with incorrect choices for the mask regions increases the visible facial parts of the original face and many artifacts will appear in the *doctored* face. The Celeb-DF dataset magnified the first mask to obtain a smoother and non-rectangular mask that models the fake face much more accurately.

The second dataset is the Deepfake Vid-TIMIT dataset [2] [3] which was built for development of systems to detect deepfake videos. Its designers found that the existing methods do not meet the requirements and consistently achieved low detection rates when applying different models against existing datasets which called for the development of this bespoke dataset.

The total videos generated using generative adversarial networks (GANs) is 320 low quality and 320 high quality videos. Deepfake Video-TIMIT dataset was evaluated by implementing VGG [17] and Facenet [18] neural network algorithms in order to classify deepfake videos and also to detect discrepancies between lip movement and voice synchronization in both videos.

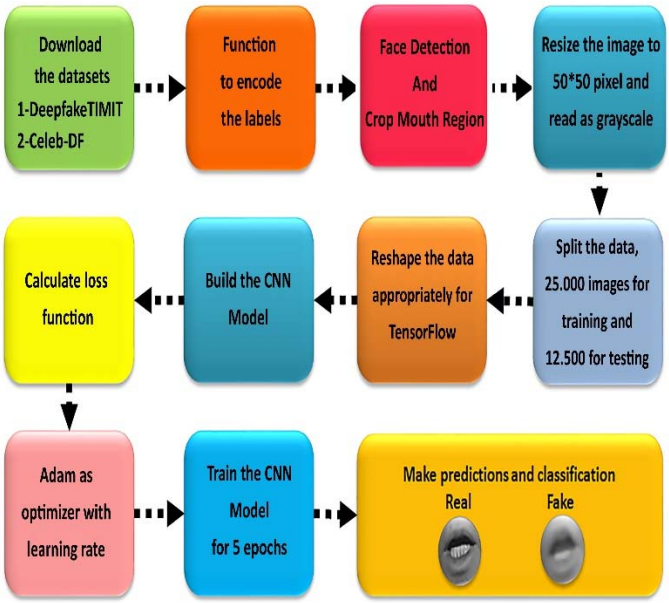

Figure 2: phases of the DFT-MF model

V. METHODOLOGY

In this section, The DFT-MF model will be described to detect Deepfake videos which uses Convolutional Neural Networks (CNN) to classify fake videos and real videos as in figure 2. The main component in our work relies on images extracted as frames from videos; this image extraction process is resource hungry and requires a great amount of time and computing power. To mitigate this issue, we have used MoviePy, which is an open-source software that was written in Python for editing and cutting videos, to cut the video based on certain word occurrences in which the mouth appears open and the teeth are visible.

Using the above approach with MoviePy, we can exclude all irrelevant images thus saving time and resources, which distinguishes DFT-MF model from other algorithms that process all images from the video and then work to discover the facial region of this image, thus reducing their efficiency.

Step 1: Data Collection

In this stage, the information will be collected from relevant sources to create a new dataset that contains a combination of fake and real videos. In our work, we will use the Deepfake Forensics (Celeb-DF) dataset and the Deepfake Vid-TIMIT dataset that have been explained above and we will extract all frames from the video before filtering irrelevant frames.

Step 2: Pre-processing

Prior to performing analysis on the image frames, some preprocessing is required. Face detection is one of the most essential steps of this work to enable us to filter out image frames (or parts thereof) that do not contain faces [19]. TO this end, the Dlib classifier [20] will be used to detect face landmarks and eliminate all unnecessary frames.

Step3: Mouth Cropping

The DFT-MF model focuses on area surrounding the mouth, especially the teeth; therefore, the mouth area will be cropped from a face in the frame. Working on a typical image frame of a face, the facial landmark detector inside the Dlib library is used to estimate the location of 68 (X, Y)-coordinates that graph to specific facial structures, these coordinates can be visualized as follows:

• The mouth can be located through points (49, 68).

• The right eyebrow through points (18, 22).

• The left eyebrow through points (23, 27).

• The right eye using points (37, 42).

• The left eye using points (43, 48).

• The nose is defined with points (28, 36).

• The jaw via points (1, 17).

the Dlib face landmark detector will return a shape object containing the face bounding box at 68 (x, y)-coordinates of the facial landmark regions in the image figure 3 illustrates that and at the points (49,68) the mouth can be located. This area will be used by DFT-MF model to crop the mouth region based on the ratio between each two-point upper lips and the lower lips.

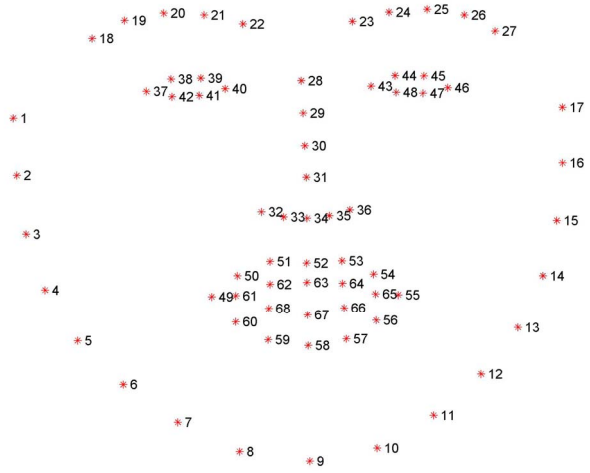

Figure 3: Visualizing the 68 Face landmarks

The next step is to exclude all frames that contain a closed mouth by calculating distances between lips. This is because an image with a closed mouth has no fake value as nothing is being uttered in that frame. We will be tracking the open mouth, which the teeth with reasonable clarity so as to obtain high accuracy and increase efficiency of the model, figure 4 illustrates the idea.

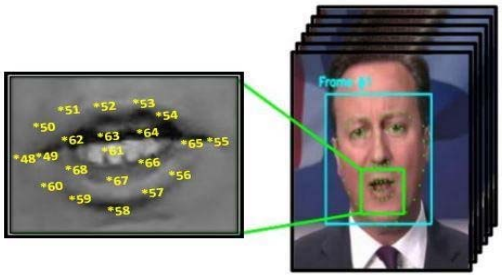

Figure 4: Mouth Cropped

Step 4: Deepfake Video Classification

The test data is split into fake and real videos for training 25000 Frames: 12500 frames were labeled as Real and 12500 frames were labeled as Fake videos. As seen in figure 5 and figure 6.

Figure 5: Samples Real cropped mouth

Figure 6: Samples Fake cropped mouth

The DFT-MF model will use the deep learning supervised, Convolutional Neural Network (CNN), to classify videos into fake or real based on a (threshold) number of the fake frames that are identified in the entire video Based on calculate three variables word per sentence, speech rate and frame rate.

The word per sentence, according to a study by the American Press Institute (API), readers understood more when sentences were shorter and 100% percent of information when sentences averaged 8 words or fewer. As such, The DFT-MF model will be calculated the number of words in the sentence and the perfect sentence should contain only five words and thus will be clearer, for that assume a threshold of 5 words as a clear sentence indicator.

Speech rate describes the speed at which people speak, it is calculated in the number of words spoken per minute (wpm). Studies show that speech rate differs depending on several factors including culture, location, gender, emotional state, fluency, profession, setting or audience. Speech rate for conversational and official speeches ranges from 120 to 150 wpm, for this work, The DFT-MF model will be considered 120 words per minute as the standard, which means 2 words per second, especially for the presidents and world leaders to speak more comfortable and clear in front of people.

Frame rate or frames per second (FPS) refers to the number of individual frames or consecutive images that are displayed per second for a video, all videos in our datasets run at 30 frames per second.

To alter facts, a fake video needs to contain two seconds of doctored frames, in these two seconds, the *faker* can change at least one sentence to make a fake video, as shown in table 2. For our work, to build a stronger model than the other models, the DFT-MF model will use less than two seconds to detect fake videos, which is equivalent to 50 frames from each video. This implies that 50 frames in a typical video determine the video as fake or real.

Table 2: Number of fake frame

Time in second	Number of Frames	Number of Words	If (number of fake frames in videos >50) Video is Fake; Else Video is Real;
1	30	2	
2	60	4	
3	90	6	

VI. EVALUATION OF EXPERIMENTAL RESULTS

In this section, the DFT-MF is evaluated and the most important evaluation metrics for checking classification model performance, namely: AUROC (Area Under the Receiver Operating Characteristics) curve is used. The DFT-MF model is evaluated against two of the latest datasets: The Deepfake Forensics (Celeb-DF) dataset and the Deepfake Vid-TIMIT dataset. The DFT-MF model has achieved 71.25 accuracy rate when run against the Deepfake Forensics (Celeb-DF) dataset, Accuracy of 98.7 was obtained against Deepfake Vid-TIMIT-LQ dataset and, 73.1 with the Deepfake Vid-TIMIT-HQ dataset. The results also indicate the following findings: Deepfake forensics (Celeb-DF) dataset that contains 795 fake videos and 408 real videos: True Positive (TP) = 761, False Positive (FP) =34, False Negative (FN)=307, and True Negative (TN) =101. As for the Deepfake Vid-TIMIT dataset which contains 640 fake videos of mixed low quality (LQ) and high quality (HQ): For high quality (HQ): True Positive (TP) = 234, False Positive (FP) =86, False Negative (FN) = 0, and True Negative (TN)=0. These results are summarized in Table 3 which contrasts the DFT-MF model with various other approaches. Figure 7 shows the comparative results on two datasets. Figures 8, 9 and 10 show the Comparison of the DFT-MF model with other methods.

Table 3: The result of DFT-MF model vs. other approaches

Methods	Approaches	Celeb-DF Deepfake Dataset	Deepfake-TIMIT Dataset	
			LQ	HQ
DFT-MF	CNN	71.25	98.7	73.1
Two-stram	DNN+SVM	55.7	83.5	73.5
HeadPose	SVM	54.8	55.1	53.2
FWA	ResNet50	53.8	99.9	93.2
Meso-4	CNN	53.6	87.8	68.4
Mesolnception-4	CNN	49.6	80.4	62.7
VA-MLP	Multilayer Perceptron (MLP)	48.8	61.4	62.1
VA-LogReg	Logistic regression (LogReg)	46.9	77.0	77.3
XceptionNet	CNN	38.7	56.7	54.0
Multi-task	CNN	36.5	62.2	55.3

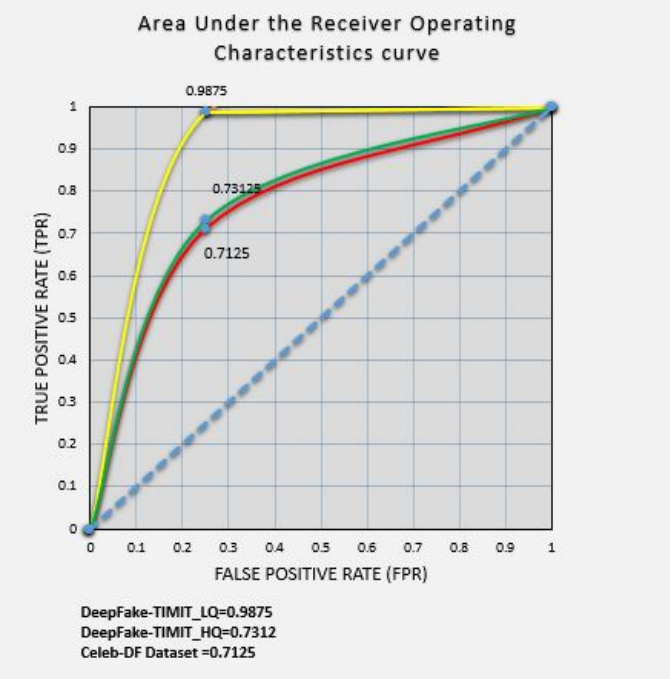

Figure 7 :Our comparative result on two datasets

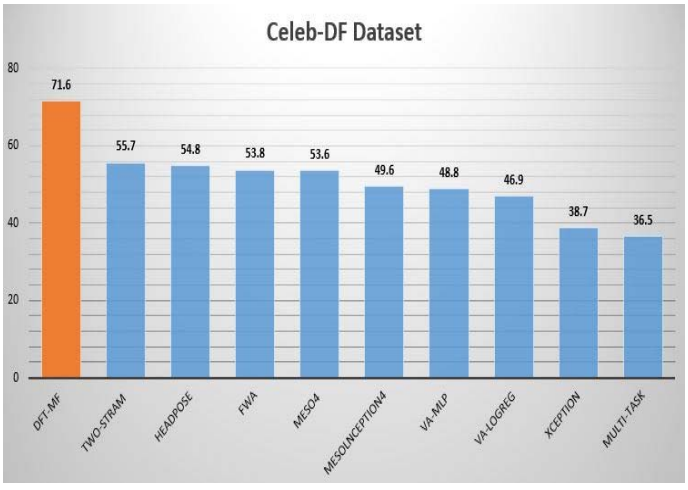

Figure 8: Comparison of the DFT-MF model with other methods on the Celeb-DF dataset

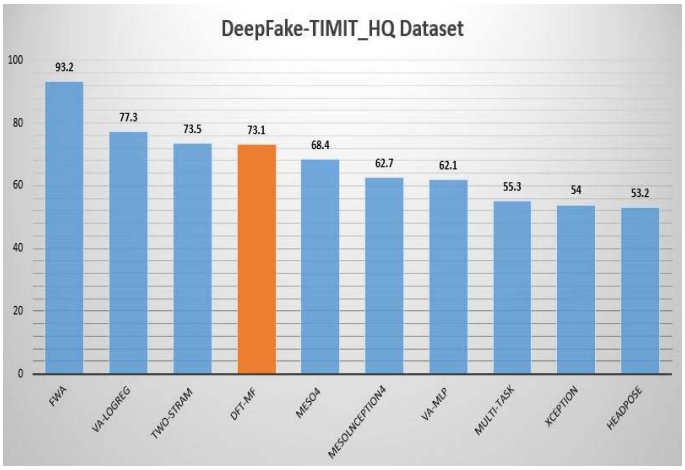

Figure 9: Comparison of the DFT-MF model with other methods on the Deepfake TIMIT-HQ dataset

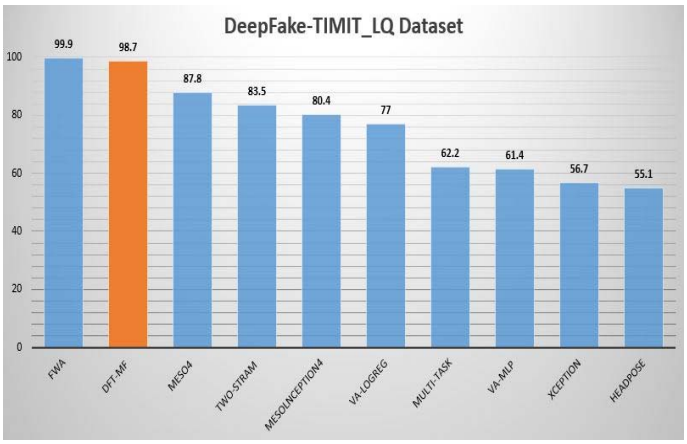

Figure 10: Comparison of the DFT-MF model with other methods on the Deepfake TIMIT-LQ dataset

I. CONCLUSIONS

In this paper, the CNN deep learning algorithm is used to classify deepfake videos. We have used MoviePy, which is an open-source application for editing and cutting videos, to cut the video based on certain words in which the mouth appears open and the teeth are visible. We eliminated all images that are irrelevant to our thus saving time and resources. This is in contrast to other algorithms, that extract all images from the video and then attempt to identify the facial region within the extracted images.

The DFT-MF model was built to detect deepfake videos by using the mouth as a biological signal. First, the datasets were used that contain both fake and real videos Celeb-DF and Deepfake-TIMIT. Secondly, deep learning (CNN) was applied to classify fake videos, depending on the features that will be taken from the mouth as a biological signal. Our work demonstrably improves on other methodologies; the results were compared to show these performance gains.

REFERENCES

[1] Y. Li1, X. Yang, P. Sun, H. Qi and S. Lyu, "CelebDF: A New Dataset for DeepFake Forensics," in Oct 2019.

[2] P. Korshunov and S. Marcel, "Deepfakes: a new threat to face recognition? assessment and detection," Idiap-RR Idiap-RR-18-2018, Idiap, 2018. 1

[3] P. Korshunov and S. Marcel, "Vulnerability of Face Recognition to Deep Morphing," in Oct 2019.

[4] T. Nguyen, C. Nguyen, D.Nguyen ,D. Nguyen and S. Nahavandi, " Deep Learning for Deepfakes Creation and Detection," in Sep 2019.

[5] 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), 2018.

[6] X. Yang, Y. Li, and S. Lyu, "Exposing Deep Fakes Using Inconsistent Head Poses," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019.

[7] 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.

[8] Y. Li and S. Lyu, "Exposing DeepFake Videos By Detecting Face Warping Artifacts," arXiv preprint arXiv:1811.00656,2018.

[9] K. Simonyan and A. Zisserman., "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556 , 2014.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," In CVPR. 2016.

[11] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Two-Stream Neural Networks for Tampered Face Detection," 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017.

[12] 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), 2019.

[13] H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen, "Multi-task learning for detecting and segmenting manipulated facial images and videos," arXiv preprintarXiv:1906.06876, 2019.

[14] A. R¨ossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, " Faceforensics++: Learning to detect manipulated facial images," arXiv preprint arXiv:1901.08971,2019.

[15] Y. Li, M.-C. 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), 2018.

[16] Y. Li, M.-C. Chang, and S. Lyu, "In ictu oculi: Exposing ai generated fake face videos by detecting eye blinking," CVPR, vol. abs/1806.02877, 2018. [Online]. Available: http://arxiv.org/abs/1806.02877

[17] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in British Machine Vision Conference, 2015.

[18] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015, pp.815–823.

[19] T. Mita, T .Kaneko, and O .Hori , "Joint Haar-like Features for Face Detection," 2005.

[20] D. King, "Dlib-ml: A machine learning toolkit," Journal of Machinen Learning Research, vol. 10, pp. 1755–1758, 2009.

[21] L. Deng, "A tutorial survey of architectures, algorithms, and applications for deep learning," APSIPA Transactions on Signal and Information Processing, vol. 3, 2014.

[22] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," Pattern Recognition, vol. 30, no. 7, pp. 1145–1159, 199