VIVA-Tech International Journal for Research and Innovation
ISSN(Online): 2581-7280

Volume 1, Issue 4 (2021)
Article No. X
PP XX-XX

VIVA Institute of Technology
9th National Conference on Role of Engineers in Nation Building – 2021 (NCRENB-2021)

VIVA-TECH INTERNATIONAL JOURNAL FOR RESEARCH AND INNOVATION
ANNUAL RESEARCH JOURNAL
ISSN(ONLINE): 2581-7280

DEEPFAKE DETECTION TECHNIQUES: A REVIEW

Neeraj Guhagarkar[1], Sanjana Desai[2] Swanand Vaishyampayan[3], Ashwini Save[4]

[1]*Department Computer Engineering, Mumbai University, MUMBAI*
[2]*Department Computer Engineering, Mumbai University, MUMBAI*
[3]*Department Computer Engineering, Mumbai University, MUMBAI*
[4]*Department Computer Engineering, Mumbai University, MUMBAI*

Abstract : *Noteworthy advancements in the field of deep learning have led to the rise of highly realistic AI generated fake videos, these videos are commonly known as Deepfakes. They refer to manipulated videos, that are generated by sophisticated AI, that yield formed videos and tones that seem to be original. Although this technology has numerous beneficial applications, there are also significant concerns about the disadvantages of the same. So there is a need to develop a system that would detect and mitigate the negative impact of these AI generated videos on society. The videos that get transferred through social media are of low quality, so the detection of such videos becomes difficult. Many researchers in the past have done analysis on Deepfake detection which were based on Machine Learning, Support Vector Machine and Deep Learning based techniques such as Convolution Neural Network with or without LSTM .This paper analyses various techniques that are used by several researchers to detect Deepfake videos.*

Keywords - *Convolutional Neural Networks, Deepfake Detection, Long Short Term Memory , Super Resolution, Facial Forgery.*

I. INTRODUCTION

Remarkable improvements in the field of Deep Learning have led to the growth of Deepfake videos. With the help of Deep Learning architectures such as Generative Adversarial Neural Networks (GANs) and autoencoders and a considerable amount of footage of a target subject, anyone can create such convincing fake videos [4]. Head Puppetry, Face swapping and Lip-syncing are the 3 major types of Deepfake videos [3]. This technique has provided potential harm to society. For instance, Indian journalist Rana Ayyub had become the victim of a sinister Deepfake plot. A fake pornographic video that showed her in it was shared on social media platforms such as Twitter and WhatsApp [25]. Therefore, there is an urgent need for researchers to develop a system that would detect such Deepfake videos. This paper focuses on various techniques like Machine Learning techniques based on Support Vector Machine(SVM) [1][16][19][22], Deep learning techniques like Convolution Neural Network(CNN) [10][14][2], CNN with SVM [3], CNN with Long Short Term Memory(LSTM) [4][6][7][11][12][14] and Recurrent Neural Network(RNN) [20]. Also, many different approaches such as considering the manipulation in the background color [19], exposing inconsistent headposes [1], to detect Deepfake videos.

II. LITERATURE SURVEY

Xin Yang, et. al. [1] have proposed a system to detect Deepfake using inconsistent headposes. Algorithms used in the previous model create the face of different persons without changing the original expressions hence creating mismatched facial landmarks. The landmark locations of few false faces often vary from those of the real faces, as a consequence of interchanging faces in the central face region in the DeepFake process.The difference in the distribution of the cosine distances of the two head orientation vectors for real and Deepfakes suggest that they can be differentiated based on this cue. It uses the DLib package for face detection and to

extract 68 facial landmarks. The standard facial 3D model is created with OpenFace2, and then difference is calculated.The proposed system uses UADFV dataset. Trained SVM classifier with Radial basis function (RBF) kernels on the training data is used. Area Under ROC (AUROC) of 0.89, is achieved by the SVM classifier on the UADFV dataset. The crucial point that can be inferred from this paper is the focus on how the Deepfakes are generated by splicing a synthesized face region into the original image, and how it can also use 3D pose estimation for detecting synthesized videos.

Rohita Jagdale, et. al. [2] have proposed a novel algorithm NA-VSR for Super resolution. The algorithm initially reads the low resolution video and converts it into frames. Then the median filter is used to remove unwanted noise from video. The pixel density of the image is increased by bicubic interpolation technique. Then Bicubic transformation and image enhancement is done for mainly resolution enhancement. After these steps the design metric is computed. It uses the output peak signal-to-noise ratio (PSNR) and structural similarity index method (SSIM) to determine the quality of image. Peak signal-to-noise ratio and structural similarity index method parameters are computed for NA-VSR and compared with previous methods. Peak signal to noise ratio (PSNR) of the proposed method is improved by 7.84 dB, 6.92 dB, and 7.42 dB as compared to bicubic, SRCNN, and ASDS respectively.

Siwei Lyu,[3] has surveyed various challenges and also discussed research opportunities in the field of Deepfakes. One critical disadvantage of the current DeepFake generation methods is that they cannot produce good details such as skin and facial hairs. This is due to the loss of information in the encoding step of generation. Head puppetry involves copying the source person's head and upper shoulder part and then pasting it on the target person's body , so that target appears to behave in a similar way as that of the source. The second method is face swapping which swaps only the face of the source person with that of the target. It also keeps the facial expressions unchanged. The third method is Lip syncing which is used to create a falsified video by only manipulating the lip region so that the target appears to speak something that she/he does not speak in reality. The detection methods are formulated as frame level binary classification problems. Out of the three widely used detection methods, the first category considers inconsistencies exhibited in the physical/physiological aspects in the DeepFake videos. The second algorithm makes use of the signal-level artifacts. Data driven is the last category of Detection in this, it directly employs multiple types of DNNs trained on genuine and Fake videos but captures only explicit artifacts. It also sheds some light on the limitations of these methods such as quality of deepfake datasets, social media laundering, etc.

Digvijay Yadav, et. al. [4] have elaborated the working of the deepfake techniques along with how it can swap faces with high precision. The Generative Adversarial Neural Networks (GANs) contain two neural networks, the first is generator and other is discriminator. Generator neural networks create the fake images from the given data set. On the other hand, discriminator neural networks evaluate the images which are synthesized by the generator and check its authenticity. Deepfake are harmful because of cases like individual character defamation and assassination, spreading fake news, threat to law enforcement agencies. For detection of Deepfakes blinking of eyes can be considered as a feature. The limitations for making Deepfakes are the requirement of large datasets, training and swapping is time consuming, similar faces and skin tones of people, etc. Deepfake video detection can be done using recurrent neural networks. CNN is best known for its visual recognition and if it is combined with LSTM it can easily detect changes in the frames and then this information is used for detecting the DeepFakes. The paper suggests that Meso-4 and Mesoinception-4 architectures are capable of detecting the Deepfake video with the accuracy of 95% to 98% on Face2Face dataset.

Irene Amerini, et. al. [5] have proposed a system to exploit possible inter-frame dissimilarities using the optical flow technique. CNN classifiers make use of this clue as a feature to learn. The optical flow fields calculated on two consecutive frames for an original video and the corresponding Deepfake one are pictured and it can be noticed that the motion vectors around the chin in the real sequence are more vociferous in comparison with those of the altered video that appear much smoother. This is used as a clue to help neural networks learn properly. FaceForensics++ dataset was used, in that 720 videos were used for training, 120 videos for validation, and 120 videos for testing. They used two neural networks VGG16 and ResNet 50. For Face2Face videos, VGG gives detection accuracy of 81.61 % and ResNet50 gives detection accuracy of 75.46 %. The uniqueness of this paper is the consideration of inter-frame dissimilarities, unlike other techniques which rely only on intra-frame inconsistencies and how to overcome them using the optical flow based CNN method.

Peng Chen, et. al. [6] have developed FSSPOTTER which is a unified framework, which can simultaneously explore spatial and temporal information in the videos. The Spatial Feature Extractor (SFE) first divides the videos into several consecutive clips, each of which contains a certain number of frames. SFE takes clips as input and generates frame-level features. It uses convolution layers of Visual Geometry Group VGG16 with batch normalization as the backbone network which extracts spatial features in the intra-frames. Also, the superpixel-wise binary classification unit (SPBCU) is exploited to encourage the backbone network to extract more features. The Temporal Feature Aggregator (TFG) deploys a Bidirectional LSTM to find the temporal inconsistencies in the frame. Then, a fully connected layer and a softmax layer are exploited to compute the probabilities of whether the clip is real or fake. FaceForensics++, DeepfakeTIMIT, UADFV and Celeb-DF are used for the evaluation. FSSpotter takes a simple VGG16 as the backbone network and it is superior to Xception by 2.2% for UADFV, 5.4% for Celeb-DF, and 4.4%, 2.0% for DeepfakeTIMIT HQ, LQ respectively.

Mohammed A. Younus, et. al. [7] have compared notable Deepfake detection methods. From various methods, few techniques include Background Comparison, Temporal Pattern Analysis, Eye blinking, Facial Artifacts, etc. The initial method uses Long Term Recurrent CNN (LRCN) to learn the temporal pattern of eye blinking and the dataset used consists of 49 interview and presentation videos and their corresponding generated Deepfakes. With the help RCN that unites convolutional network DenseNet and the gated recurrent unit cells few temporal discrepancies across frames are explored in the background comparison technique. In this FaceForensics++ dataset is used. The third approach is Temporal Pattern Analysis which uses Convolutional Neural Network (CNN) to extract frame-level features and Long Short Term Memory (LSTM) for classification. The dataset used for this method is a batch of 600 videos obtained from multiple websites. In the fourth method, Artifacts are discovered using ResNet50 CNN models based on resolution inconsistency between the warped face area and the surrounding context. The Mesoscopic Analysis uses two networks Meso-4 and MesoInception-4 for determining the Deefakes. This study paper helped to understand various deepfake detection approaches currently used and recommends how other features could be found that would help to detect Deepfakes more efficiently.

Shivangi Aneja, et. al. [8] to overcome the challenge of zero and few-shot transfer they have proposed a transfer learning-based approach, called Deep Distribution Transfer (DDT). The fundamental idea behind this method was a new distribution-based loss formulation that can be efficiently equipped to span the gap between domains of different facial forgery techniques or obscure datasets. The proposed method outperforms the baselines by a significant margin in both zero-shot and few-shot learning. The model uses an ImageNet Large Scale Visual Recognition Challenge ILSVRC 2012-pretrained ResNet-18 neural network. A 4.88% higher detection efficiency for zero-shot and 8.38% for the few-shot case transferred is achieved. The suggested method tries to generalize the forgery detection techniques by using zero and a few shot transfers. Hence, a similar approach can be followed to broaden the scope of the project for the detection of forgery on several datasets.

XTao, et. al. [9] have proposed a system that emphasizes the fact that to achieve better results, proper frame alignment and motion compensation needs to be done. The authors have introduced a sub-pixel motion compensation layer (SPMC) layer in a CNN framework. Along with FlowNet-S CNN, frame alignment and motion compensation is achieved using motion compensation transformer (MCT) module. Also, they have collected 975 sequences from high-quality 1080p HD video clips publically available on the internet and downsampled the original frames to 540 × 960 pixels. The proposed method has a Peak signal-to-noise ratio (PSNR) of 36.71 and a Structural Similarity Index (SSIM) value of 0.96 which is better than that of the previously proposed SRCNN. This paper provides an insight into how to organize multiple frame inputs for getting better results. Also, it gives a foreknowledge about how data is to be sampled before feeding it to the CNN model.

Jin Yamanaka, et. al. [10] have asserted that mostly single image Super resolution is used for medical systems, surveillance systems. But the computational power required for the system is very high so it cannot be used for smaller devices. But as the need for super resolution is increasing, they have proposed a different way which will reduce the deep CNN computational power by 10 to 100 times still maintaining higher accuracy. The no of layers in the neural network used so far has 30 layers but with the authors' system, the no of layers will reduce to 11 only hence decreasing the computational power on a large scale. The computational power of the 3*3 matrix will be 9 times the computational power of nine 1*1 matrix. Datasets used have a total of 1,164 training images and a total size of 435 MB. It focuses on a super resolution algorithm and proposes a different

way to perform super resolution with reduced space complexity and reduced computational power. The system used in the paper also has a reduced number of neural layers which is the significant takeaway from this paper. Even if they are reducing the neural layers down to 11 from 30, the proposed system does not lose its accuracy but the accuracy was only incremented as per the shown outcomes in the paper.

David guera, et. al. [11] have demonstrated how Deepfake videos are created and how they can be detected using CNN and LSTM. GAN's are used for better quality deepfake videos. For the generation process, the encoder of the original image is used but for swapping faces with the target image, the decoder of the target image is used. They tried various techniques over Deepfake videos for devising accurate detection systems and came to the final conclusion that the best accuracy was found when the video was split into 80 frames per second along with a combination of CNN and LSTM. The maximum accuracy which was acquired was around 97.1%. But the accuracy which was acquired was on a set of high-resolution images. The above paper illustrates in great detail how the Deepfake videos are generated.

Mousa Tayseer Jafar, et. al. [12] have developed a system that mainly focuses on detecting Deepfake videos by focusing on mouth movements. The datasets used are Deepfake forensics and VID-TIMID which contains both real and deepfake videos. The algorithm used for Deepfake detection is CNN. For detection of faces from video frames, in the pre-processing stage, a Dlib classifier is used which will be used to detect face landmarks. For e.g. the face according to Dlib has coordinates (49,68). In this way, the coordinates of eyebrows, nose, etc can be known. The succeeding step excludes all frames that contain a closed mouth by calculating distances between lips. The model proposed works on the number of words spoken per sentence. Another thing that they have focused on is the speech rate. Generally, the speech rate is 120-150 words per minute. The system proposed in the system uses facial expressions and speech rate to determine the Deepfake videos. It shows that most of the Deepfake detection systems have found success in the classification of deepfake videos using CNN and LSTM.

Pranjal Ranjan, et. al. [13] have proposed a system is based on CNN LSTM combination for classifying the videos as fake or original. The datasets used are Faceforensics++, Celeb-DF, DeepFakeDetectionChallenge. The best performer on the custom test set from the single dataset train models is the DFD model with an accuracy of 70.57%, while the combined train split model achieves 73.20%. The worst performing cross-test pair is of Celeb-DF model tested on DFD. This makes sense since both datasets are opposite of each other in terms of manipulation and visual quality. The highest cross-test accuracy is 66.23% for the model trained on DFD and tested on DFDC. The combined test set performance is highest for the DFD model (74.03%) among the single distribution-based model, and the combined train model records the highest classification accuracy of 86.49%. This suggests that the Xception Net has a vast learning capacity since it can learn the various manipulations present in all the three distributions, and is still not overfitting, conveyed by its effective, combined test accuracy. The authors have used transfer learning in their system to increase the accuracy of thesystem.

Mingzhu Luo, et. al. [14] have proposed a system that can detect the faces present in the frame. The problem with the CNN algorithm is that the image is converted into a fixed size and given as input, but this causes image information to be lost and hampers the overall accuracy of the system. The authors have introduced spatial pyramid pooling between the convolution layer and fully connected layer. By this, the problem of information loss is subdued. The datasets used in the proposed system are CelebA, AFW, and FDDB. The image is divided into 11*11 grid first and each grid has two prediction frames. Each prediction box has a respective probability, confidence, and frame coordinates. The discrete accuracy of the proposed system is 95% and for continuous, it is 74%. The above system has given perspicacity into how the accuracy of CNN algorithm can drop because of certain factors. The system used also provided the solution that can be used so that the accuracy of CNN phase does not drop. Also, this system has improved loss function which has enabled them to get maximum accuracy out of the system.

Jiangang Yu, et. al. [15] have proposed a system that focuses on one drawback of the super resolution algorithm, which is low accuracy for videos with facial changes. For super resolution, first the video is broken down into multiple frames and then CNN is applied over each frame differently. Authors found out that when the video has facial expression changes, it is very difficult to produce higher accuracies for super resolution systems. To overcome this problem, the system proposed how handling of a facial image in a non-rigid manner can be done. The system proposed in the paper works in three steps 1) global tracking 2) local alignment 3) Super Resolution algorithm. For performance measurement of the system, authors recorded 10 video sequences

of 10 people lasting for about 3 minutes. The PSNR value was found to be 26.6261 which is a betterment from previous PSNR value 20.6261 using a global only approach.The above system has given insights on the problems faced by super resolution algorithm when there are facial changes. If the facial expression changes hamper the accuracy of super resolution, it will affect the system very badly because the output of the super resolution phase is given as input to CNN stage in our system. To overcome this problem, the paper has given a solution of using the handling of facial image in a non-rigid way. The PSNR value is also increased using this approach.

Yuqian Zhou, et. al. [16] have presented a survey paper. The face detection systems do not yield satisfactory results when the subject has poor illumination, extreme poses and when the input is having low resolution. Most of the systems are trained only on high resolution images therefore they perform badly when they are used in surveillance systems because of low resolution. The authors have used HoG-SVM and R-CNN and S3FD algorithms on low resolution images. The dataset used is FDDB. The FDDB dataset has a total 5171 faces in total 2845 images. Algorithms were tasted on performance degradation of the above models while changing the blur, noise, or contrast level. The conclusion was both the algorithms HoG-SVM and R-CNN-S3FD perform very badly when they are tested on low resolution images. This paper provides insights that care must be taken as R-CNN and S3FD performs very badly for face detection of low-resolution images. Also care must be taken for noise and contrast level as well because these factors also affect the accuracy of the algorithms.

Andreas Rossler , et. al. [17] have stated that it is difficult to detect Deepfake, either automatically or by humans in this paper. Including a human baseline, this paper also provides the benchmark for facial manipulation detection under random compression. The paper uses the CNN model to detect all this Deepfakes. They cast the forgery detection as a per-frame binary classification problem of the manipulated videos. They used total of 7 methods to detect the deepfake of the various quality of videos. In the Steganalysis method it uses the handcrafted feature and the SVM classifier. They provided a 128 × 128 central crop-out of the face as input to the method. It was observed that detection of raw images was good but when it came to compression factor its accuracy decreased. A constrained convolutional layer followed by two convolutional, two max-pooling and three fully-connected layers is used. Also a different CNN architecture is adopted with a global pooling layer that computes four statistics (mean, variance, maximum and minimum). XceptionNet gave best output among the other methods for the low quality video detection. The results demonstrated that the highest accuracy was 81% for the low quality images that was for XceptionNet algorithm.

Falko Matern, et. al. [18] have proposed the simple logistic regression model for detection. This paper shows how exactly the manipulation takes place in the generated faces and the Deep fake. They proposed the algorithm to detect completely generated faces. They demonstrated this by several visual features that focus on the eyes, teeth, facial contours. Specular reflection in faces is most prominent in the eyes. According to them, samples generated by Deepfake techniques show unconvincing specular reflections. They state that reflections in the eyes are either missing or appear simplified as a white blob. For the experiment purpose, they used these 3 datasets CelebA, ProGAN, Glow . To extract color features of each eye, they use commonly available computer vision methods. For the deepfake detection, they exploit missing reflections, and missing details in the eye and teeth areas. Then again detect facial landmarks and crop the input image to the facial region. They have used the neural network which is fully connected, which contains three layers with 64 nodes and ReLU activation functions. Two classifiers such as MLP and Log Reg are used. It concludes that classification done using only the features generated from teeth performs relatively poorly, with an AUC of 0.625 for both classifiers. Much better performances of 0.820 and 0.784 were observed when the features were extracted from the eye region. With the help of the combined feature vector, AUC of 0.851 is achieved by the neural network. This further concludes that the eye and teeth features are more accurate.

Scott McCloskey, et. al. [19] have proposed the model that uses the saturation cues to detect the deep fakes. With the help of saturation cues images can be distinguished as GAN-generated imagery or camera imagery. Two types of GAN-generated imagery can be exposed with the potency of this que. It is seen that HDR, camera images generally have regions of saturation under-exposure. For this forensic, the hypothesis stated is that the frequency of saturated and under-exposed pixels will be suppressed by the generator's normalization steps. They suggested the use of a GAN image detector, where one can simply measure the frequency of saturated and under-exposed pixels in each image for this purpose. Trained using Matlab's fitcsvm function, these features are classified by a linear Support Vector Machine (SVM). They used this method on 2 different datasets GAN Crop and GAN Full. This method clearly does a far better job of detecting fully GAN-generated images, where it

produces a 0.7 AUC. This paper provides an alternative for the features that can be considered while detecting the Deepfake videos.

Atra Akandeh, et. al. [20] have proposed two variations of LSTM networks. The first variant is called LSTM 6 which has three fixed constant gates and the second variant known as LSTM C6 which is an improvement over LSTM 6 having reduced memory-cell input block. When hyperparameters like learning rate, gate constants, number of hidden units are set properly, then reduced parameter LSTM 6 and LSTM C6 variants can perform as good as the standard LSTM networks. Also slim architectures enable training speedup. LSTM needs to be constantly updated so this alternative. LSTM RNN incorporates a memory cell (vector) and includes three gates: (i) an input gate, it (ii) an output gate ot, and (iii) a forget gate, ft. Then the number of (adaptive) parameters in LSTM 6 is n(m + n + 1) and for LSTM C6 the total number of (adaptive) parameters is n(m + 2). The standard LSTM (denoted as lstm0 in the figure) displays smooth profiles with (testing) accuracy around 88%. However, LSTM 6 (denoted as LSTM6 in the figure) shows fluctuations and also does not catch up with standard LSTM. Based on the hyper parameters the accuracy of both the models fluctuates. This paper gives insight about the alternative of the LSTM model.

Chao Dong, et. al. [21] have proposed a deep learning method for single image super resolution. For upscaling a single low resolution image, bicubic interpolation technique is used. They have not performed any other techniques for pre-processing. Various datasets like Set5, Set14, and BSD200 are used which contain 5, 14 and 200 images respectively. Four evaluation matrices, namely IFC, noise quality measure (NQM), weighted peak signal-to-noise ratio (WPSNR), and multi-scale structure similarity index (MSSSIM), are used by them other than OSNR and SSIM to check the accuracy. From all the various methods the SRCNN shows the highest accuracy for all the various indices based on 2,3 and 4 upscaling factors. From this, it can be inferred that using the CNN based model for the super resolution of the low quality video will give better results than any other model.

Faten F Kharbat, et. al. [22] have proposed a method to detect Deepfake videos using Support Vector Machine (SVM) regression. Their method uses feature points extracted from the video to train Artificial Intelligence (AI) classifiers to detect false videos. HOG, ORB, BRISK, KAZE, SURF, and FAST are the different feature point extraction algorithms that they have determined. This paper proposes a system that exploits this inconsistency by extracting feature points using traditional edge feature detectors. The dataset contains 98 videos, half of which are fake videos and the other half are real videos. All the videos are in the format of mp4 and have approximately 30 seconds of duration. 95% accuracy has been achieved using the HOG feature point extraction algorithm. The above-proposed system helps to find an alternative for feature detection. Feature Detection being the most important part of the project this paper suggests that the above stated conventional algorithm can also be used for the process of feature detection.

Chuan-Chuan Low, et. al. [23] have proposed the two experimental setups. First about the face detection performance and detection rate, with different skin color regions and processing sequence approach, and second about location. It proposes a framework for detecting multiple faces by using depth and skin color for digital advertising. It uses the Viola-Jones algorithm for face detection and uses the skin color to verify the human skin face. They have used two types of processing approaches for face and skin detection. In pre-filtering after the skin color filtering process, the Viola-Jones algorithm is applied to the image frame to detect the presence of the human skin face. Whereas the Post-processing approach applies the skin color analysis on the detected face image. After that, two skin color space is applied with the Viola-Jones algorithm for the true detection rate comparison i.e YCbCr with HSV and RGB-H-CbCr.The skin face information is processed to the next phase if the depth information for the detected face is within the region of interest (ROI). The experiment was implemented in Visual Studio 2017 on Intel I5 processor with 3.30GHz and 16GB RAM. It uses the david dataset and one more unknown dataset. The results show that RGB-H-CbCr achieves 88% true detection rate and low false-positive rate compared with the YCbCr color space under the post-processing category for multiple persons in the shortest processing time. This paper helps to understand how multiple face detection is carried out using the above mentioned algorithm.

Badhrinarayan Malolan, et. al. [24] have proposed a framework to detect these Deepfake videos using a Deep Learning Approach. They have trained a Convolutional Neural Network architecture on a database of extracted faces from FaceForensics Dataset. Besides, they have also tested the model on various Explainable AI techniques such as Layer-Wise Relevance Propagation (LRP) and Local Interpretable Model-Agnostic

Explanations (LIME) to provide crisp visualizations of the salient regions of the image focused on by the model. They used the FaceForensics++ dataset which consists of 363 source actor videos as the real counterpart and 3068 manipulated videos as the fake counterpart. They have used the Xception network, which is a traditional CNN with Depth-wise Separable Convolutions and LIME to interpret the predictions of the classifier. They have trained their model on datasets of images with two different scales of background namely 1.3x and 2x with the faces occupying roughly 80 to 85% and 60 to 65% area respectively and with 90.17% accuracy. First, the extraction of the face was done using the DLIB face extractor from the frames. Then the CNN XceptionNet was applied to extract the features. As this paper uses the same dataset of FaceForensics++ it suggests that by using the XceptionNet algorithm of CNN one can get accurate results.

III. ANALYSIS TABLE

The table 1 summarizes the research papers on the Deepfake detection as well as on Super resolution and it states the different techniques used for the Deepfake detection.

Table 1:Analysis Table

Sr. No	Title of Paper	Techniques used	Dataset used	Accuracy
1.	Deepfake: A Survey on Facial Forgery Technique Using Generative Adversarial Network[4]	1. Convolutional Neural Network (CNN) 2. Long Short-Term Memory (LSTM)	Face2Face, Reddit user deepfakes	95%
2.	Deepfake Video Detection through Optical Flow based CNN[5]	1.Convolutional Neural Network (CNN)	Face2Face	VGG16 81.61%, ResNet50 75.46%
3.	FSSPOTTER: Spotting Face-Swapped video by Spatial and Temporal Clues [6]	1.Convolutional Neural Network (CNN) 2.Long Short-Term Memory (LSTM)	FaceForensics++, Deepfake TIMIT, UADFV, Celeb-DF	77.6%
4.	Generalized Zero and Few-Shot Transfer for Facial Forgery Detection[8]	ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012-pretrained ResNet-18	Faceforensics++, Dessa, Celeb DF, Google DFD	92.23%
5.	Detail-revealing Deep Video Super-resolution[9]	1.FlowNet-S CNN With a sub-pixel motion compensation layer (SPMC) layer	975 sequences from high-quality 1080p HD video clips	Method(F3) 36.71/0.96, Method (F5) 36.62/0.96

6.	Deepfake Video Detection Using Recurrent Neural Network[11]	1. Convolution Neural Networks (CNN) 2. Long Short-Term Memory (LSTM)	1.HOHA dataset	Conv-LSTM(20 frames) 96.7%,Conv-LSTM(40 frames) 97.1%
7.	Improved Generalizability of Deep-Fakes Detection Using Transfer Learning Based CNN Framework[13]	1. Convolution Neural Networks (CNN) 2. Long Short-Term Memory (LSTM)	1. FaceForensics++ 2. Celeb-DF 3.DeepFake Detection Challenge	With Transfer Learning 84%,Without Transfer Learning 75%
8.	Multi-scale face detection based on convolutional neural network.[14]	1.Convolution Neural Networks (CNN)	1. CelebA 2. AFW 3. FDDB	Discrete- 95% and for continuous, it is 74%
9.	FaceForensics++: Learning to Detect Manipulated Facial Images[17]	1.Xception net (CNN) 2.LSTM	FaceForensics++	81%
10.	Exploiting Visual Artifacts to Expose Deep Fakes and Face Manipulations[18]	The neural network classifier as MLP and the logistic regression model as LogReg	1.CelebA 2.ProGAN , 3.Glow	MLP 84%(Eyes), LogReg 83%(Eyes)
11.	Detecting gan-generated imagery using saturation cues[19]	SVM classifier	Image net dataset	92%
12.	Image Feature Detectors for Deep Fake Video Detection[22]	1.SVM classifier 2.Feature extractor algorithms	Unnamed with 98 videos	HOG 94.5%, SURF 90%, KAZE 76.5%
13.	Experimental Study on Multiple Face Detection with Depth and SkinColour[23]	1.Voila jones face detection algorithm	Unnamed	88%
14.	Explainable Deep-Fake Detection Using Visual Interpretability Methods[24]	1. Xception net(CNN) 2. LRP and LIME	FaceForensics++	90.17%

The various algorithms and features used for the Deepfake detection are analyzed in the above table. It includes the Machine Learning and Deep Learning based techniques. From the analysis table above it can be seen that CNN algon with the LSTM gives better results and accuracy which can be further increased by using the Concept of Super resolution.

IV. CONCLUSION

With the increase in the use of Deepfake videos around the world, it is very much necessary to detect such videos before they could cause some sort of harm. Various Machine Learning and Deep Learning-based techniques along with the different features are used to classify the videos as fake or real. Among all the different techniques used, the one that uses CNN and LSTM has proved to be more accurate in the classification of the videos. Here, various datasets that contain several real and fake videos have been used for the classification. From studying papers it is apparent that CNN along with LSTM yields better results and accuracy.

REFERENCES

[1] Xin Yang, Yeuzen Li and Siwei Lyu, “EXPOSING DEEP FAKES USING INCONSISTENT HEAD POSES”, ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

[2] Rohita Jagdale and Sanjeevani Shah, “A Novel Algorithm for Video Super-Resolution”, Proceedings of ICTIS 2018, Volume 1, Information and Communication Technology for Intelligent Systems (pp.533-544).

[3] Siwei Lyu,”DEEPFAKE DETECTION: CURRENT CHALLENGES AND NEXT STEPS”,2020 IEEE International Conference on Multimedia & Expo Workshops(ICMEW).

[4] Digvijay Yadav, Sakina Salmani, “Deepfake: A Survey on Facial Forgery Technique Using Generative Adversarial Network”, Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2019).IEEE Xplore Part Number: CFP19K34-ART; ISBN: 978-1-5386-8113-8.

[5] Irene Amerini, Leonardo Galteri, Roberto Caldelli, Alberto Del Bimbo, “Deepfake Video Detection through Optical Flow based CNN”, 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW).

[6] Peng Chen, Jin Liu, Tao Liang, Guangzhi Zhou, Hongchao Gao, Jiao Dai1, Jizhong Han, “FSSPOTTER: SPOTTING FACE-SWAPPED VIDEO BY SPATIAL AND TEMPORAL CLUES”, 2020 IEEE International Conference on Multimedia and Expo (ICME).

[7] Mohammed A. Younus, Taha M. Hasan, “Abbreviated View of Deepfake Videos Detection Techniques”, 2020 6th International Engineering Conference “Sustainable Technology and Development" (IEC).

[8] Shivangi Aneja, Matthias Nießner, “Generalized Zero and Few-Shot Transfer for Facial Forgery Detection”, arXiv:2006.11863v1 [cs.CV] 2020.

[9] Xin Tao, Hongyun Gao, Renjie Liao, Jue Wang, Jiaya Jia, “Detail-revealing Deep Video Super-resolution”, 2017 IEEE International Conference on Computer Vision (ICCV).

[10] Jin Yamanaka, Shigesumi Kuwashima , and Takio Kurita ,“Fast and Accurate Image Super Resolution by Deep CNN with Skip Connection and Network in Network”, (2017 Springer).

[11] David Guera, Edward J. Delp, “Deepfake Video Detection Using Recurrent Neural Networks”, 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS).

[12] Mousa Tayseer Jafar, Mohammad Ababneh, Mohammad Al-Zoube, Ammar Elhassan,"Digital Forensics and Analysis of Deepfake Videos",(IEEE 2020).

[13] Ranjan, Sarvesh Patil, Faruk Kazi,"Improved Generalizability of Deep-Fakes Detection Using Transfer Learning Based CNN Framework",(IEEE 2020).

[14] Mingzhu Luo, Yewei Xiao, Yan Zhou," Multi-scale face detection based on convolutional neural network",IEEE 2018.

[15] Jiangang Yu and Bir Bhanu,"Super-resolution of Facial Images in Video with Expression Changes",IEEE 5th Conference on Advanced Video and Signal based Surveillance,2018.

[16] Yuqian Zhou, Ding Liu, Thomas Huang ,"Survey of Face Detection on Low-quality Images",2018 13th IEEE International Conference on Automatic Face and Gesture Recognition.

[17] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies,Matthias Nießner, " FaceForensics++: Learning to Detect Manipulated Facial Images"Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV),2019, pp. 1-11.

[18] Falko Matern ,Christian Riess , Marc Stamminger, Friedrich-Alexander ,"Exploiting Visual Artifacts to Expose Deep Fakes and Face",2019 IEEE Winter Application of Computer Vision Workshop.

[19] Scott McCloskey and Michael Albright,"DETECTING GAN-GENERATED IMAGERY USING SATURATION CUES", 2019 IEEE.

[20] Atra Akandeh and Fathi M. Salem,"Slim LSTM NETWORKS: LSTM 6 and LSTM C6",2019 IEEE.

[21] Chao Dong, Chen Change Loy, Member, IEEE, Kaiming He, Member, IEEE, and Xia oou Tang, Fellow,"Image Super-Resolution Using Deep Convolutional Networks",IEEE Transaction on Pattern Analysis and Machine Intelligence(2016).

[22] Faten F Kharbat, Tarik Elamsy, Ahmed Mahmoud, Rami ,"Image Feature Detectors for Deepfake Video Detection",IEEE 2019.

[23] Chuan-Chuan Low, Lee-Yeng Ong, Voon-Chet Koo,"Experimental Study on Multiple Face Detection with Depth and Skin Colour",IEEE 2019.

[24] Badhrinarayan Malolan, Ankit Parekh, Faruk Kazi ,"Explainable Deep-Fake Detection Using Visual Interpretability Methods",2020 3rd International conference on Information and Computer Technologies(ICICT).

[25] https://www.indiatoday.in/india-today-insight/story/india-s-deepfake-problem-videos-are-neither-deep-n or-fake-1643883-2020-02-06 (last accessed on 21 October 2020)