ISSN: 2366-1313

DEEPFAKE DETECTION

[1]Mrs. Roja Ramani Adapa, [2]M A Akram,[3]M Manish,[4]Mohd Farhan Akhter

[1]Assistant Professor, Dept.of CSE, Teegala Krishna Reddy Engineering College, Meerpet, Hyderabad,

roja.adapa@tkrec.ac.in

[2]BTech student, Dept.of CSE, Teegala Krishna Reddy Engineering College, Meerpet, Hyderabad,

muhammadakram12017@gmail.com

[3]BTech student, Dept.of CSE, Teegala Krishna Reddy Engineering College, Meerpet, Hyderabad,itsurmanishmehra@gmail.com

[4]BTech student, Dept.of CSE, Teegala Krishna Reddy Engineering College, Meerpet, Hyderabad,farhanakhterr123@gmail.com

Abstract:*The growing computation power has made the deep learning algorithms so powerful that creating a indistinguishable human synthesized video popularly called as deep fakes have become very simple. Scenarios where this realistic face swapped deep fakes are used to create political distress, fake terrorism events, revenge porn, blackmail peoples are easily envisioned. In this work, we describe a new deep learning-based method that can effectively distinguish AI-generated fake videos from real videos. Our method is capable of automatically detecting the replacement and re-enactment deep fakes. We are trying to use Artificial Intelligence(AI) to fight Artificial Intelligence (AI). Our system uses a Res-Next Convolution neural network to extract the frame-level features and these features and further used to train the Long Short Term Memory(LSTM) based Recurrent Neural Network(RNN) to classify whether the video is subject to any kind of manipulation or not, i.e., whether the video is deep fake or real video. To emulate the real time scenarios and make the model perform better on real time data, we evaluate our method on large amount of balanced and mixed data-set prepared by mixing the various available data-set like Face-Forensic++[1] , Deepfake detection challenge, and Celeb-DF. We also show how our system can achieve competitive result using very simple and robust approach.*

Keywords*: Res-Next Convolution neural network, Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), Computer vision.*

I. INTRODUCTION

In the world of ever-growingsocial media platforms, Deepfakes are considered as the major threat of the AI. There are many

ISSN: 2366-1313

Scenarios where these realistic face swapped deepfakes are used to create political distress, fake terrorism events, revenge porn, blackmail peoples are easily envisioned.Some of the examples are Brad Pitt, Angelina Jolie nude videos.

It becomes very important to spot the difference between the deepfake and pristine video. We are using AI to fight AI.Deepfakes are created using tools like FaceApp[1] and Face Swap , which using pre-trained neural networks like GAN or Auto encoders for these deepfakes creation. Our method uses a LSTM based artificial neural network to process the sequential temporal analysis of the video frames and pre-trained Res-Next CNN to extract the frame level features. ResNext Convolution neural network extracts the frame-level features and these features are further used to train the Long Short-Term Memory based artificial Recurrent Neural Network to classify the video as Deepfake or real. To emulate the real time scenarios and make the model perform better on real time data, we trained our method with large amount of balanced and combination of various available dataset like FaceForensic++[1], Deepfake detection challenge, and Celeb-DF.

Further to make the ready to use for the customers, we have developed a front-end application where the user the user will upload the video. The video will be processed by the model and the output will be rendered back to the user with the classification of the video as deepfake or real and confidence of the model.

While some deepfakes can be created by traditional visual effects or computer-graphics approaches, the recent common underlying mechanism for deepfake creation is deep learning models such as autoencoders and generative adversarial networks (GANs), which have been applied widely in the computer vision domain [2]. These models are used to examine facial expressions and movements of a person and synthesize facial images of another person making analogous expressions and movements. Deepfake methods normally require a large amount of image and video data to train models to create photo-realistic images and videos. As public figures such as celebrities and politicians may have a large number of videos and images available online, they are initial targets of deepfakes. Deepfakes were used to swap faces of celebrities or politicians to bodies in porn images and videos. The first deepfake video emerged in 2017 where face of a celebrity was swapped to the face of a porn actor. It is threatening to world security when deepfake methods can be employed to create videos of world leaders with fake

ISSN: 2366-1313

speeches for falsification purposes [3]. Deepfakes therefore can be abused to cause political or religion tensions between countries, to fool public and affect results in electioncampaigns, or create chaos in financial markets by creating fake news. It can be even used to generate fake satellite images of the Earth to contain objects that do not really exist to confuse military analysts, e.g., creating a fake bridge across a river although there is no such a bridge in reality. This can mislead a troop who have been guided to cross the bridge in a battle.

MOTIVATION

The increasing sophistication of mobile camera technology and the ever-growing reach of social media and media sharing portals have made the creation and propagation of digital videos more convenient than ever before. Deep learning has given rise to technologies that would have been thought impossible only a handful of years ago. Modern generative models are one example of these, capable of synthesizing hyper realistic images, speech, music, and even video. These models have found use in a wide variety of applications, including making the world more accessible through text-to-speech, and helping generate training data for medical imaging.

Like any trans-formative technology, this has created new challenges. So-called "deep fakes" produced by deep generative models that can manipulate video and audio clips. Since their first appearance in late 2017, many open-source deep fake generation methods and tools have emerged now, leading to a growing number of synthesized media clips. While many are likely intended to be humorous, others could be harmful to individuals and society. Until recently, the number of fake videos and their degrees of realism has been increasing due to availability of the editing tools, the high demand on domain expertise.

Spreading of the Deep fakes over the social media platforms have become very common leading to spamming and peculating wrong information over the plat-form. Just imagine a deep fake of our prime minister declaring war against neighbouring countries, or a Deep fake of reputed celebrity abusing the fans. These types of the deep fakes will be terrible, and lead to threatening, misleading of common people.

To overcome such a situation, Deep fake detection is very important. So, we describe a new deep learning-based method that can effectively distinguish AI-generated fake videos (Deep Fake Videos) from real videos. It’s incredibly important

ISSN: 2366-1313

to develop technology that can spot fakes, so that the deep fakes can be identified and prevented from spreading over the internet.

II. LITERATURE SURVEY

Face Warping Artifacts [4] used the approach to detect artifacts by com-paring the generated face areas and their surrounding regions with a dedicated Convolutional Neural Network model. In this work there were two-fold of Face Artifacts.

Their method is based on the observations that current deepfake algorithm can only generate images of limited resolutions, which are then needed to be fur-ther transformed to match the faces to be replaced in the source video. Their method has not considered the temporal analysis of the frames.

Detection by Eye Blinking [5] describes a new method for detecting the deep-fakes by the eye blinking as a crucial parameter leading to classification of the videos as deepfake or pristine. The Long-term Recurrent Convolution Network (LRCN) was used for temporal analysis of the cropped frames of eye blinking. As today the deepfake generation algorithms have become so powerful that lack of eye blinking cannot be the only clue for detection of the deepfakes. There must be certain other parameters must be considered for the detection of deep-fakes like teeth enchantment, wrinkles on faces, wrong placement of eyebrows etc.

Capsule networks to detect forged images and videos [6] uses a method that uses a capsule network to detect forged, manipulated images and videos in different scenarios, like replay attack detection and computer-generated video detection.

In their method, they have used random noise in the training phase which is not a good option. Still the model performed beneficial in their dataset but may fail on real time data due to noise in training. Our method is proposed to be trained on noiseless and real time datasets.

Recurrent Neural Network [7] (RNN) for deepfake detection used the approach of using RNN for sequential processing of the frames along with Image-Net pre-trained model. Their process used the HOHO dataset consisting of just 600 videos.

Their dataset consists small number of videos and same type of videos, which may not perform very well on the real time data. We will be training out model on large number of Real-time data.

Synthetic Portrait Videos using Biological Signals approach extract bio-logical signals from facial regions on pristine and deepfake portrait video pairs. Applied

ISSN: 2366-1313

transformations to compute the spatial coherence and temporal consistency, capture the signal characteristics in feature vector and photoplethysmography (PPG) maps, and further train a probabilistic Support Vector Machine (SVM) and a Convolutional Neural Network (CNN). Then, the average of authenticity probabilities is used to classify whether the video is a deepfake or a pristine.

Fake Catcher detects fake content with high accuracy, independent of the generator, content, resolution, and quality of the video. Due to lack of discriminator leading to the loss in their findings to preserve biological signals, formulating a differentiable loss function that follows the proposed signal processing steps is not straight forward process.

There have been existing survey papers about creating and detecting deepfakes, presented in[8]. For example, Mirsky and Lee focused on reenactment approaches (i.e., to change a target's expression, mouth, pose, gaze or body), and replacement approaches (i.e., to replace a target's face by swap or transfer methods). Verdoliva separated detection approaches into conventional methods (e.g., blind methods without using any external data for training, one-class sensor-based and model-based methods, and supervised methods with handcrafted features) and deep learning-based approaches (e.g., CNN models). Tolosana et al. [9 categorized both creation and detection methods based on the way deepfakes are created, including entire face synthesis, identity swap, attribute manipulation, and expression swap. On the other hand, we carry out the survey with a different perspective and taxonomy. We categorize the deepfake detection methods based on the data type, i.e., images or videos, as presented in Fig. 1. With fake image detection methods, we focus on the features that are used, i.e., whether they are handcrafted features or deep features. With fake video detection methods, two main subcategories are identified based on whether the method uses temporal features across frames or visual artifacts within a video frame. We also discuss extensively the challenges, research trends and directions on deepfake detection and multimedia forensics problems

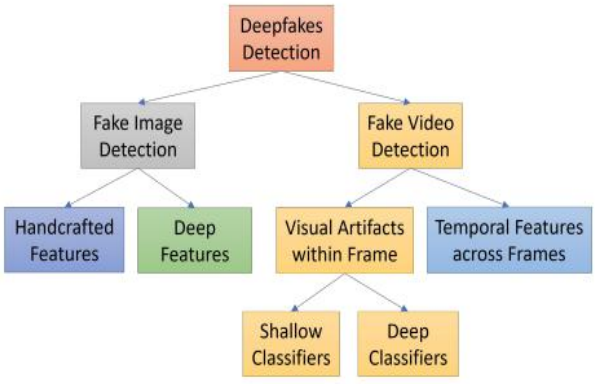

Fig.1 Categories of reviewed papers relevant to deepfake detection methods

ISSN: 2366-1313

where we divide papers into two major groups, i.e., fake image detection and face video detection.

III. PROPOSED WORK

This document lays out a project plan for the development of Deepfake video detection using neural network.The intended readers of this document are current and future developers working on Deepfake video detection using neural network and the sponsors of the project. The plan will include, but is not restricted to, a summary of the system functionality, the scope of the project from the perspective of the "Deepfake video detection" team (me and my mentors), use case diagram, Data flow diagram,activity diagram, functional and non- functional requirements, project risks and how those risks will be mitigated, the process by which we will develop the project, and metrics and measurements that will be recorded throughout the project.

Deepfake Creation

Deepfakes have become popular due to the quality of tampered videos and also the easy-to-use ability of their applications to a wide range of users with various computer skills from professional to novice. These applications are mostly developed based on deep learning techniques. Deep learning is well known for its capability of representing complex and high-dimensional data. One variant of the deep networks with that capability is deep autoencoders, which have been widely applied for dimensionality reduction and image compression [33–35]. The first attempt of deepfake creation was FakeApp, developed by a Reddit user using autoencoder-decoder pairing structure. In that method, the autoencoder extracts latent features of face images and the decoder is used to reconstruct the face images. To swap faces between source images and target images, there is a need of two encoder-decoder pairs where each pair is used to train on an image set, and the encoder's parameters are shared between two network pairs. In other words, two pairs have the same encoder network. This strategy enables the common encoder to find and learn the similarity between two sets of face images, which are relatively unchallenging because faces normally have similar features such as eyes, nose, mouth positions. Fig. 3 shows a deepfake creation process where the feature set of face A is connected with the decoder B to reconstruct face B from the original face A. This approach is applied in several works such as DeepFaceLab, DFaker, DeepFaketf (tensorflow-based deepfakes)

ISSN: 2366-1313

By adding adversarial loss and perceptual loss implemented in VGGFace to the encoder-decoder architecture, an improved version of deepfakes based on the generative adversarial network, i.e., faceswap-GAN, was proposed in [10]. The VGGFace perceptual loss is added to make eye movements to be more realistic and consistent with input faces and help to smooth out artifacts in segmentation mask, leading to higher quality output videos. This model facilitates the creation of outputs with 64x64, 128x128, and 256x256 resolutions. In addition, the multi-task convolutional neural network (CNN) from the FaceNet implementation is used to make face detection more stable and face alignment more reliable. The CycleGAN [60] is utilized for generative network implementation in this model

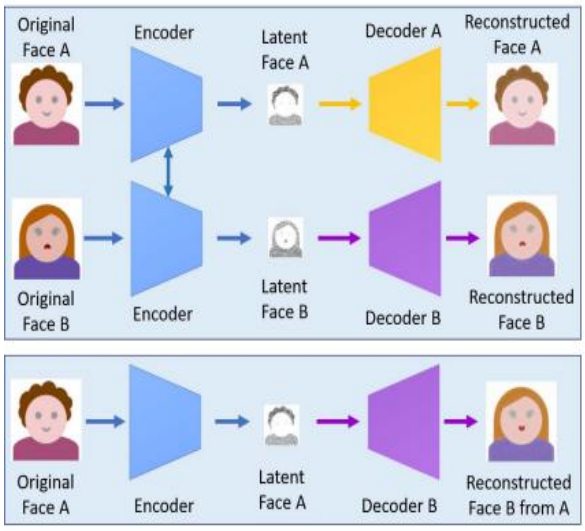

Fig.2 A deepfake creation model using two encoder-decoder pairs. Two networks use the same encoder but different decoders for training process (top). An image of face A is encoded with the common encoder and decoded with decoder B to create a deepfake (bottom). The reconstructed image (in the bottom) is the face B with the mouth shape of face A. Face B originally has the mouth of an upside-down heart while the reconstructed face B has the mouth of a conventional heart.

SYSTEM ARCHITECTURE

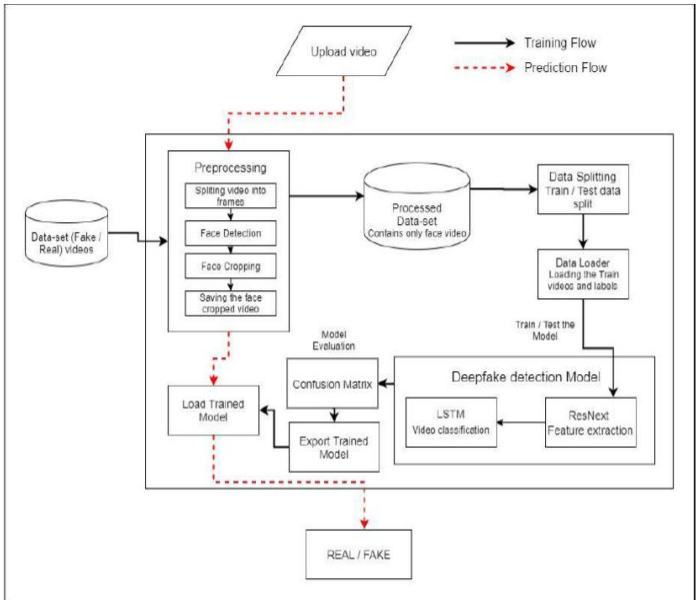

Fig.3 System Architecture

In this system, we have trained our PyTorch deepfake detection model on equal number of real and fake videos in order to avoid the bias in the model. The system architecture of the model is showed in the figure. In the development phase, we havetaken a dataset, preprocessed the dataset and created a new processed

ISSN: 2366-1313

dataset which only includes the face cropped videos.

Creating deepfake videos

To detect the deepfake videos it is very important to understand the creation process of the deepfake. Majority of the tools including the GAN and autoencoders takes a source image and target video as input. These tools split the video into frames, detect the face in the video and replace the source face with target face on each frame. Then the replaced frames are then combined using different pre-trained models. These models also enhance the quality of video my removing the left-over traces by the deepfake creation model. Which result in creation of a deepfake looks realistic in nature. We have also used the same approach to detect the deepfakes. Deepfakes created using the pretrained neural networks models are very realistic that it is almost impossible to spot the difference by the naked eyes. But in reality, the deepfakes creation tools leaves some of the traces or artifacts in the video which may not be noticeable by the naked eyes. The motive of this paper to identify these unnoticeable traces and distinguishable artifacts of these videos and classified it as deepfake or real video.

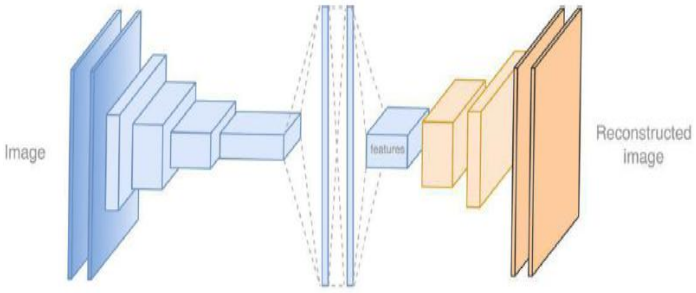

Fig.4 Deepfake generation

Fig.5 Face Swapped deepfake generation

Data-set Gathering

For making the model efficient for real time prediction. We have gathered the data from different available data-sets like FaceForensic++(FF)[1], Deepfake detection challenge(DFDC)[2], and Celeb-DF[3]. Futher we have mixed the dataset the col-lected datasets and created our own new dataset, to accurate and real time detection on different kind of videos. To avoid the training bias of the model we have consid-ered 50% Real and 50% fake videos.

Deep fake detection challenge (DFDC) dataset [3] consist of certain audio alerted video, as audio deepfake are out of scope for this paper. We preprocessed the DFDC

ISSN: 2366-1313

dataset and removed the audio altered videos from the dataset by running a python script.

After preprocessing of the DFDC dataset, we have taken 1500 Real and 1500 Fake videos from the DFDC dataset. 1000 Real and 1000 Fake videos from the FaceForensic++(FF)[1] dataset and 500 Real and 500 Fake videos from the Celeb-DF[3] dataset. Which makes our total dataset consisting 3000 Real, 3000 fake videos and 6000 videos in total. Figure 2 depicts the distribution of the data-sets.

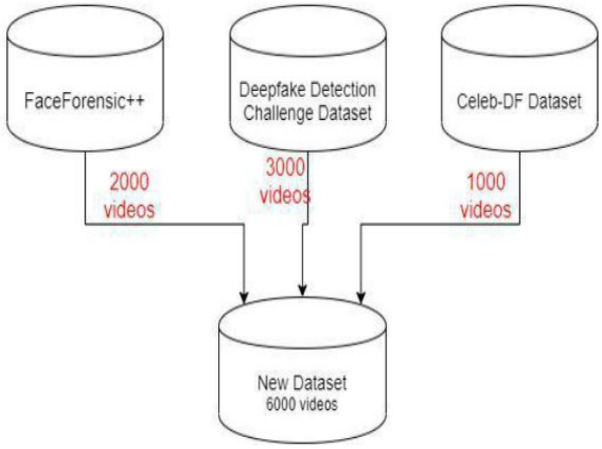

Fig.6 Dataset

Pre-processing

In this step, the videos are pre-processed and all the unrequired and noise is removed from videos. Only the required portion of the video i.e. face is detected and cropped.

The first steps in the pre-processing of the video are to split the video into frames. After splitting the video into frames, the face is detected in each of the frame and the frame is cropped along the face. Later the cropped frame is again converted to a new video by combining each frame of the video. The process is followed for each video which leads to creation of processed dataset containing face only videos. The frame that does not contain the face is ignored while pre-processing.

To maintain the uniformity of number of frames, we have selected a threshold value based on the mean of total frames count of each video. Another reason for selecting a threshold value is limited computation power. As a video of 10 second at 30 frames per second(fps) will have total 300 frames and it is computationally very difficult to process the 300 frames at a single time in the experimental environment. So, based on our Graphic Processing Unit (GPU) computational power in experimental environment we have selected 150 frames as the threshold value. While saving the frames to the new dataset we have only saved the first 150 frames of the video to the new video. To demonstrate the proper use of Long Short-Term Memory (LSTM) we have considered the frames in the sequential manner i.e., first 150 frames and not randomly. The newly created video is

ISSN: 2366-1313

saved at frame rate of 30 fps and resolution of 112 x 112.

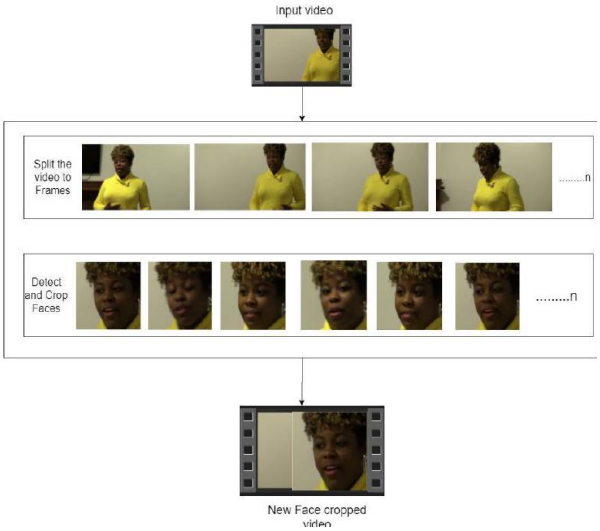

Fig.7 Pre-processing of video

Data-set split

The dataset is split into train and test dataset with a ratio of 70% train videos (4,200) and 30% (1,800) test videos. The train and test split is a balanced split i.e 50% of the real and 50% of fake videos in each split

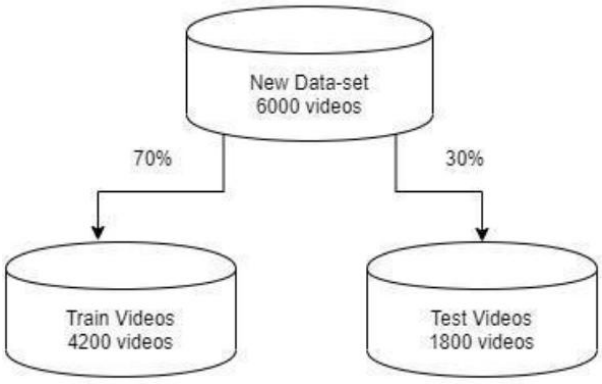

Fig.8Train test split

Model Architecture

Our model is a combination of CNN and RNN. We have used the Pre- trained ResNext CNN model to extract the features at frame level and based on the extracted features a LSTM network is trained to classify the video as deepfake or pristine. Us-ing the Data Loader on training split of videos the labels of the videos are loaded and fitted into the model for training.

ResNext :

Instead of writing the code from scratch, we used the pre-trained model of ResNext for feature extraction. ResNext is Residual CNN network optimized for high performance on deeper neural networks. For the experimental purpose we have used resnext50_32x4d model. We have used a ResNext of 50 layers and 32 x 4 dimensions.

Following, we will be fine-tuning the network by adding extra required layers and selecting a proper learning rate to properly converge the gradient descent of the model. The 2048-dimensional feature vectors after the last pooling layers of ResNext is used as the sequential LSTM input.

LSTM for Sequence Processing:

2048-dimensional feature vectors are fitted as the input to the LSTM. We are using 1 LSTM layer with 2048 latent dimensions

ISSN: 2366-1313

and 2048 hidden layers along with 0.4 chance of dropout, which is capable to do achieve our objective. LSTM is used toprocess the frames in a sequential manner so that the temporal analysis of the video can be made, by comparing the frame at ‘t’ second with the frame of ‘t-n’ seconds. Where n can be any number of frames before t.

The model also consists of Leaky Relu activation function. A linear layer of 2048 input features and 2 output features are used to make the model capable of learning the average rate of correlation between eh input and output. An adaptive average polling layer with the output parameter 1 is used in the model. Which gives the the target output size of the image of the form H x W. For sequential processing of the frames a Sequential Layer is used. The batch size of 4 is used to perform the batch training. A SoftMax layer is used to get the confidence of the model during predication.

IV. RESULTS

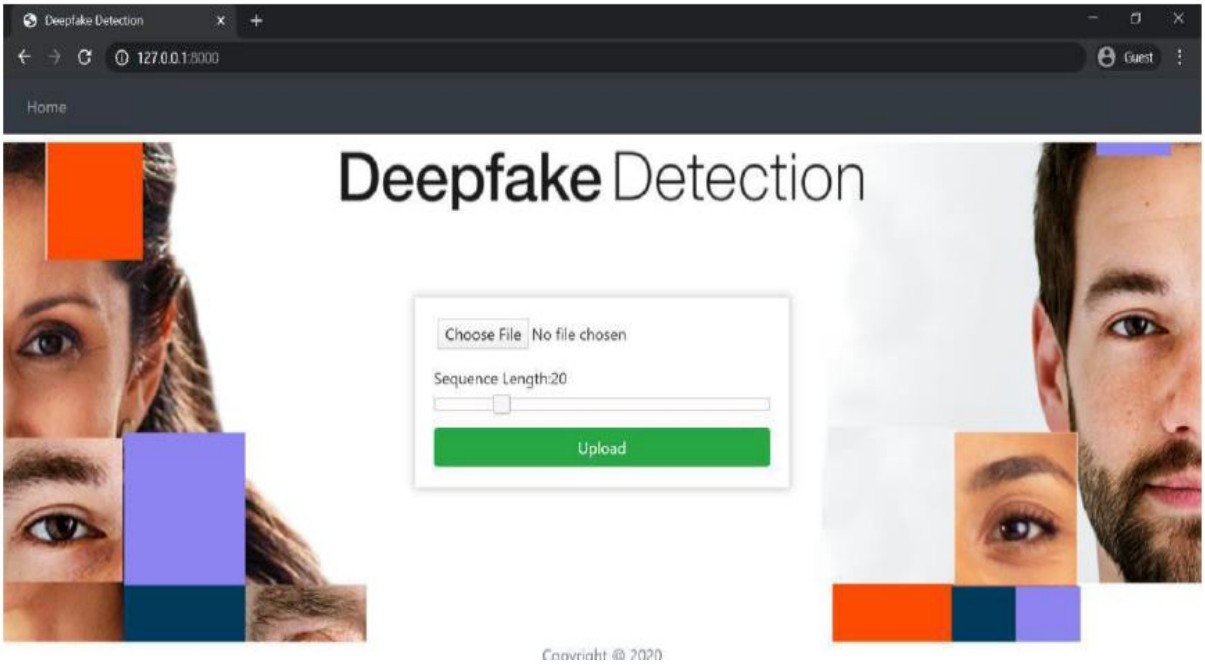

Fig.9 Home page

ISSN: 2366-1313

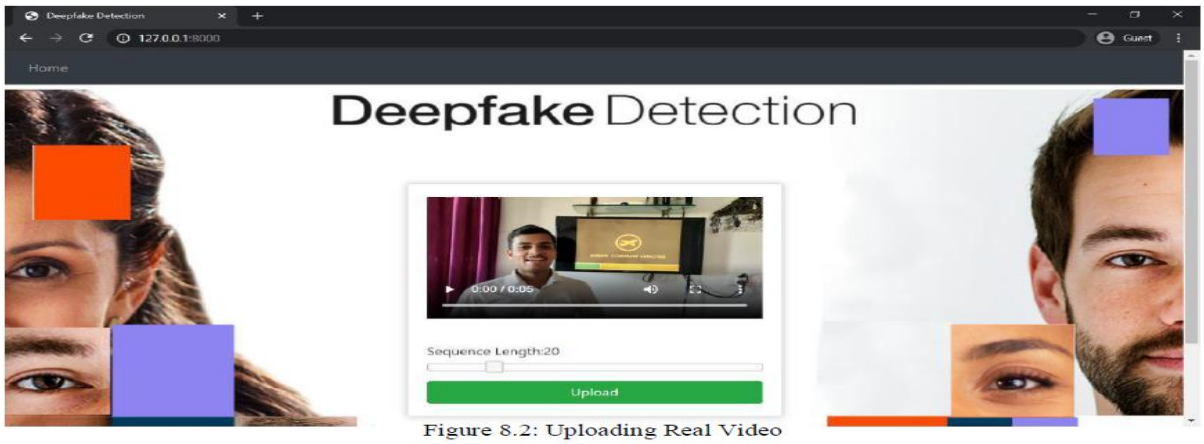

Figure 8.2: Uploading Real Video

Fig.10 Uploading real video

Fig.11 Real video output

Figure 8.4: Uploading Fake Video

Fig.12 Deepfake detection

ISSN: 2366-1313

Fig 13 Fake result

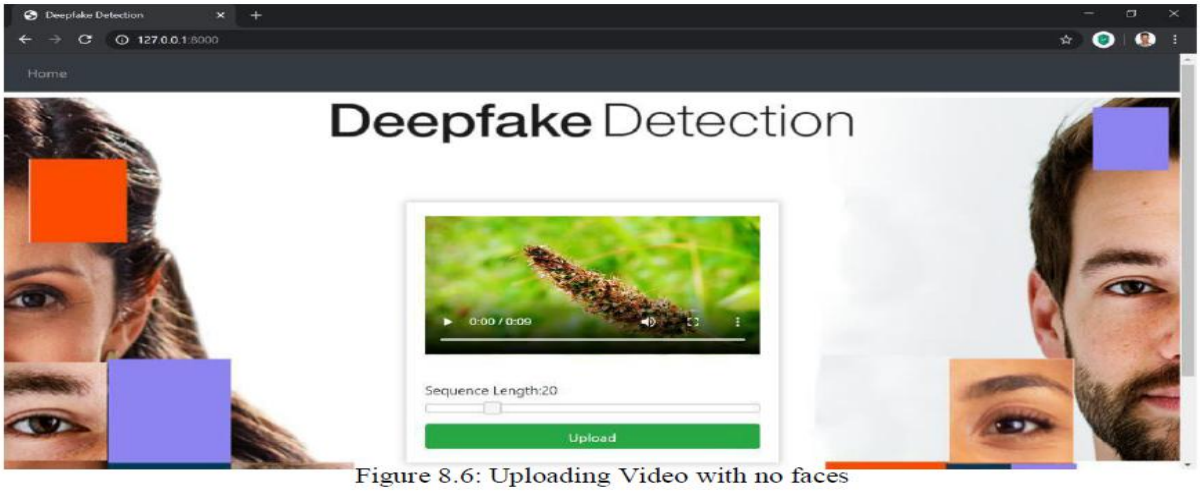

Figure 8.6: Uploading Video with no faces

Fig.14 uploading video without faces

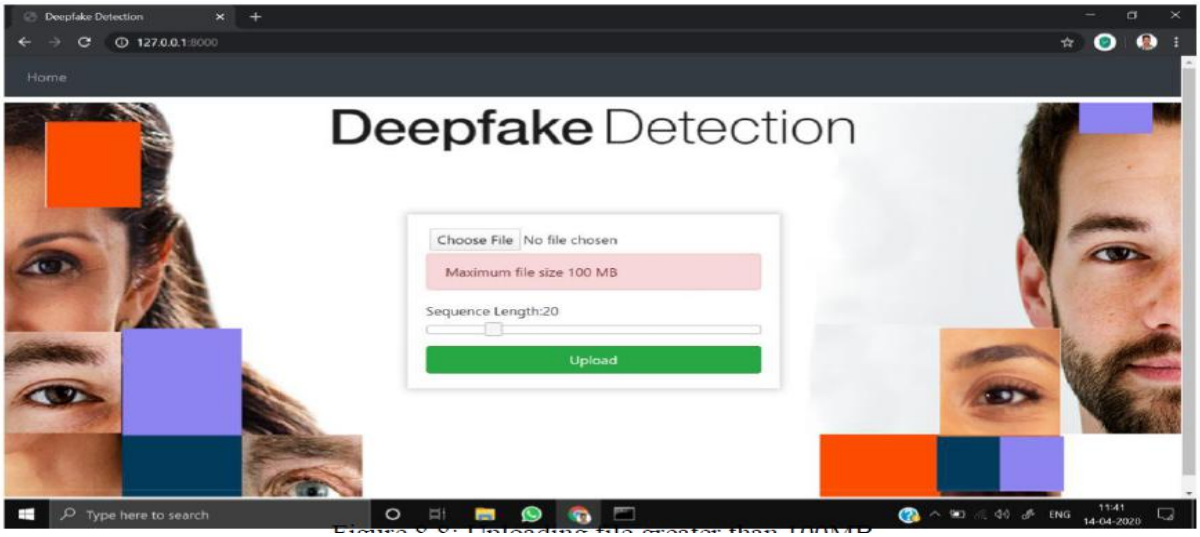

Fig.15 Uploading file greater than 100MB

ISSN: 2366-1313

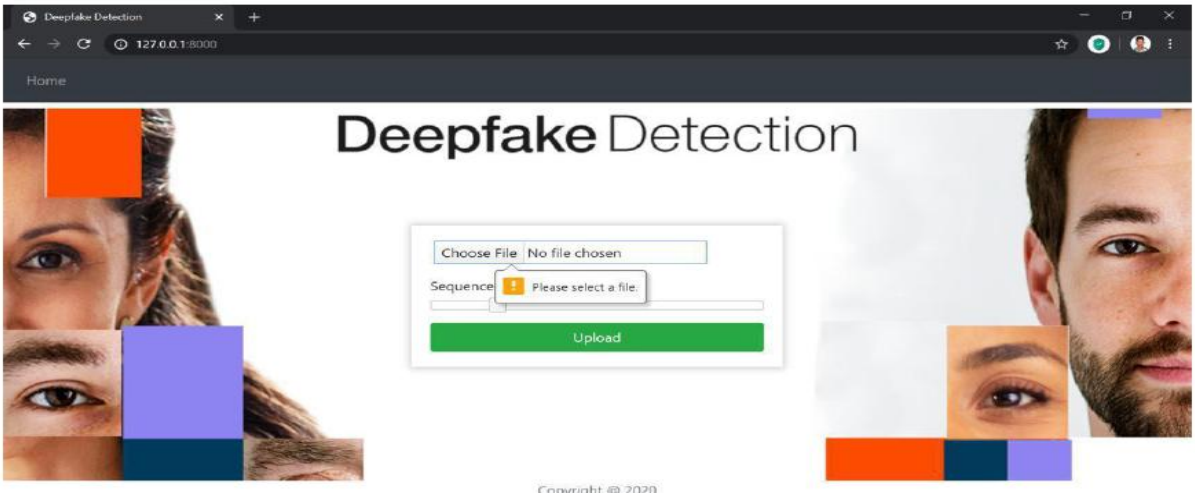

Fig.16 Pressing Upload button without selecting video

V. CONCLUSION

We presented a neural network-based approach to classify the video as deep fake or real, along with the confidence of proposed model. Our method is capable of predicting the output by processing 1 second of video (10 frames per second) with a good accuracy. We implemented the model by using pre-trained ResNext CNN model to extract the frame level features and LSTM for temporal sequence processing to spot the changes between the t and t-1 frame. Our model can process the video in the frame sequence of 10,20,40,60,80,100.

REFERENCES

[1] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies,MatthiasNießner, “FaceForensics++: Learning to Detect Manipulated Facial Images” in arXiv:1901.08971.

[2] Deepfake detection challenge dataset:https://www.kaggle.com/c/deepfake-detection-challenge/data Accessed on 26 March, 2020

[3] Yuezun Li , Xin Yang , Pu Sun , Honggang Qi and SiweiLyu “Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics” in arXiv:1909.12962

[4] Deepfake Video of Mark Zuckerberg Goes Viral on Eve of House A.I. Hearing :https://fortune.com/2019/06/12/deepfake-mark-zuckerberg/ Accessed on 26 March, 2020

[5] 10 deepfake examples that terrified and amusedtheinternet:https://www.creativebloq.com/features/deepfake-examples Accessed on 26 March, 2020

[6]TensorFlow:https://www.tensorflow.org/ (Accessed on 26 March, 2020)

[7] Keras: https://keras.io/ (Accessed on 26 March, 2020)

ISSN: 2366-1313

[8] PyTorch:https://pytorch.org/ (Accessed on 26 March, 2020)

[9] N Srivani, Prasadu Peddi (2021), Face Assessment Learned From Existing Images In Order To Classify The Gender Of The Images Based On Improved Face Recognition, (TURCOMAT), Vol 12, issue 6, pp: 5724-5735

[10] J. Thies et al. Face2Face: Real-time face capture and reenactment of rgb videos. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2387–2395, June 2016. Las Vegas, NV.

[11] Prasadu Peddi (2019), "Data Pull out and facts unearthing in biological Databases", International Journal of Techno-Engineering, Vol. 11, issue 1, pp: 25-32.

[12]FaceSwap:https://faceswaponline.com/ (Accessed on 26 March, 2020)

[13] Deepfakes, Revenge Porn, And The Impact On Women :

https://www.forbes.com/sites/chenxiwang/ 2019/11/01/deepfakes-revenge-porn-an d-the-impact-on-women/

[14] Prasadu Peddi (2019), "Data Pull out and facts unearthing in biological Databases", International Journal of Techno-Engineering, Vol. 11, issue 1, pp: 25-32.