DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

Major Project Pre Final Report- [VIII Sem B.E]

on

"A reliable solution to detect deepfakes using Deep Learning"

Submitted by

Ashutosh Mishra 1SI18CS018

M N M Varun 1SI18CS052

under the guidance of

Mr Vedamurthy H K

Assistant Professor

SIDDAGANGA INSTITUTE OF TECHNOLOGY, TUMAKURU

(An Autonomous Institute under Visvesvaraya Technological University, Belagavi

Approved by AICTE, New Delhi, Accredited by NAAC and ISO 9001:2015 certified)

B.H. road, Tumkur 572103, Karnataka, India

AY 2021-22

Contents

Abstract **1**

1 Introduction **3**

1.1 Background Study 3

1.2 Related Works 4

1.3 Project Problem Statement and Objectives 5

1.3.1 Problem Statement 5

1.3.2 Objectives 5

1.4 Organization of the Report 6

2 Literature Survey **7**

3 High-level Design **10**

3.1 Software development methodology 10

3.1.1 Why Agile for this project? 11

3.2 Architecture 13

3.3 Functional Requirements 14

3.4 Non-Functional Requirements 14

3.4.1 Data Preparation 14

3.4.2 Feature Selection 15

3.4.3 Prediction 15

Abstract

In recent months, free deep learning-based software tools has facilitated the creation of credible face exchanges in videos that leave few traces of manipulation, in what they are known as "DeepFake"(DF) videos. Manipulations of digital videos have been demonstrated for several decades through the good use of visual effects, recent advances in deep learning have led to a drastic increase in the realism of fake content and the accessibility in which it can be created. These so-called AI-synthesized media (popularly referred to as DF). Creating the DF using the Artificially intelligent tools are simple task. But, when it comes to detection of these DF, it is major challenge. Because training the algorithm to spot the DF is not simple. At times it may also happen that the video is genuine but a deep fake voice is wrapped over it to convince the the victim. Today, high-quality fraudulent deepfake voices are used to convince a human listener. The problem of recognizing fake or computer-generated audio requires identifying where it differs from real audio. Audio files contain subtle differences which are imperceptible to the human ear. Also, advancements in the field make of speech synthesis make it impractical to try picking up on differences between these audio classes without first transforming the audio file. In this project our focus is to develop a system that can reliably detect these deepfakes based on visual and audio analysis using deep learning.

List of Figures

3.1 The structure of Software Development Life Cycle 10
3.2 An Agile Architecture 11
3.3 System Architecture 13

Chapter 1

Introduction

This part of the report fundamentally manages the concise foundation depiction of DFs identification which is one of the most slanting themes in the present period. This segment additionally clarifies what are the issues that individuals are confronting at present and why this model come into picture out of nowhere. This segment additionally clarify upon the issue articulation and the target of the venture.

1.1 Background Study

It is a fact that DeepFakes or DFs are a problem of the future and it needs to dealt using advanced tools. We propose to develop a Deep Learning method to classify the video as deepfake or pristine. The increasing sophistication of smartphone cameras and the availability of good internet connection all over the world has increased the ever-growing reach of social media and media sharing portals have made the creation and transmission of digital videos more easy than ever before. The growing computational power has made deep learning so powerful that would have been thought impossible only a handful of years ago. Like any transformative technology, this has created new challenges. So-called "DeepFake" produced by deep generative adversarial models that can manipulate video and audio clips. Spreading of the DF over the social media platforms have become very common leading to spamming and peculating wrong information over the platform. These types of the DF will be terrible, and lead to threating, misleading of common people.
Voice cloning technology has advanced rapidly over the years, finding application in many industries such as security and biometric verification, medicine, entertainment and gaming. With this increased use, there is growing concern about the security challenges which come with the vulnerabilities exposed by such technology. Along with voice cloning, computer-generated audio has advanced, and it has become increasingly difficult to tell real speech from synthesized or computer-generated speech. This opens the world to a new plethora of attacks by fraudsters aiming to manipulate unsuspecting individuals for their objectives. Equally speedy response in the security field is required combatting this

problem.

1.2 Related Works

The explosive growth in deep fake video and its illegal use is a major threat to democracy, justice, and public trust. Due to this there is a increased the demand for fake video analysis, detection and intervention. Some of the related word in deep fake detection are listed below: ExposingDF Videos by Detecting Face Warping Artifacts [1] used an approach to detects artifacts by comparing 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 DF algorithm can only generate images of limited resolutions, which are then needed to be further transformed to match the faces to be replaced in the source video. Exposing AI Created Fake Videos by Detecting Eye Blinking [2] describes a new method to expose fake face videos generated with deep neural network models. The method is based on detection of eye blinking in the videos, which is a physiological signal that is not well presented in the synthesized fake videos. The method is evaluated over benchmarks of eye-blinking detection datasets and shows promising performance on detecting videos generated with Deep Neural Network based software DF. Their method only uses the lack of blinking as a clue for detection. However certain other parameters must be considered for detection of the deep fake like teeth enchantment, wrinkles on faces etc. Our method is proposed to consider all these parameters. Using capsule networks to detect forged images and videos [3] 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. This paper [4] analyzes the spoof audio created by recording the speech and replay it which dodge an automatic speaker verification mechanism. There is not much research conducted on building spoofing detection system which can handle the vulnerability of ASV systems. The detection system is based on the convolutional neural network. The training dataset consists of a speaker, genuine and replays audio clips. Replay attacks [5] present a significant threat to Automatic Speaker

Verification systems as they can be easily mounted using everyday smart devices by any non-professional imposter. The ASVspoof 2017 challenge was an initiative to develop solutions to counteract such replay attacks. The proposed solution builds on the fact that all the distinguishing features between genuine and spoofed audio are not effectively captured by conventional feature extraction techniques.

1.3 Project Problem Statement and Objectives

To overcome such a situation, DF detection is very important. So, we describe a deep learning-based method that can effectively distinguish AI-generated fake videos (DF Videos) from real videos. It's incredibly important to develop technology that can spot fakes, so that the DF can be identified and prevented from spreading over the internet. Our response to this problem is the use of machine learning techniques to identify real and computer-generated audio, training a model to tell the difference even if the human ear cannot. To achieve this, we make use of Convolutional Neural Network (CNN) classifier model. The CNN is a Deep neural learning technique which mimics the learning process of the human brain to train a model.

1.3.1 Problem Statement

We propose to develop a Deep Learning method to classify the video as deepfake or pristine.

1.3.2 Objectives

The chosen project tries to fulfill the following objectives:

- To describe a deep learning-based method that can effectively distinguish DF videos from the real ones
- To achieve the detection of the audio deep fakes based on Convolutional Neural Networks.
- To develop a web application to integrate both models and detect whether the uploaded video is pristine or deepfake.

1.4 Organization of the Report

The report comprises of various chapters which are explained in brief in this section:

1. Introduction : This part of the report mainly list out the background study of fake news detection and why is it needed currently.

2. Literature Survey : This part of the report is an overview of the previously published works on a specific topic.

3. High Level Design : This part explains the software model,its structure and the architecture and explains the functional requirements.

4. Implementation : It explains the detailed methodology and technology that is being used to make the proposed model.

5. Testing : It list out all the testing that is being done while developing the project.

6. Conclusion and Future Scope : This section contains the conclusion part and expected results of the project and the future scope of the project.

7. Published Paper : This part of the report contains original research results or reviews existing results.

The report then contains Bibliography.

Chapter 2

Literature Survey

Title of the work: Exposing DeepFake Videos By Detecting Face Warping Artifacts

Citation : Yuezun Li, Siwei Lyu, "ExposingDF Videos By Detecting Face Warping Artifacts," in arXiv:1811.00656v3. https://arxiv.org/abs/1811.00656

Description: This paper [?] uses an approach to detect artifacts by comparing 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 DF algorithm can only generate images of limited resolutions, which are then needed to be further transformed to match the faces to be replaced in the source video.

Title of the work: Exposing AI Created Fake Videos by Detecting Eye Blinking

Citation : Yuezun Li, Ming-Ching Chang and Siwei Lyu "Exposing AI Created Fake Videos by Detecting Eye Blinking" in arxiv.

https://www.cs.albany.edu/ lsw/papers/wifs18.pdf

Description: This paper [?] describes a new method to expose fake face videos generated with deep neural network models. The method is based on detection of eye blinking in the videos, which is a physiological signal that is not well presented in the synthesized fake videos. The method is evaluated over benchmarks of eye-blinking detection datasets and shows promising performance on detecting videos generated with Deep Neural Network based software DF. Their method only uses the lack of blinking as a clue for detection. However certain other parameters must be considered for detection of the deep fake like teeth enchantment, wrinkles on faces etc. Our method is proposed to consider all these parameters.

Title of the work: Using capsule networks to detect forged images and videos
Citation : Huy H. Nguyen , Junichi Yamagishi, and Isao Echizen " Using capsule networks to detect forged images and videos ".
https://ieeexplore.ieee.org/document/8682602

Description: This paper [?] 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.

Title of the work: Analysing The Predictions Of a CNN-Based Replay Spoofing Detection System
Citation : B. Chettri, S. Mishra, B. Sturm and E. Benetos, "Analysing The Predictions Of a CNN-Based Replay Spoofing Detection System", 2018 IEEE Spoken Language Technology Workshop (SLT), 2018.
Available: 10.1109/slt.2018.8639666 [Accessed 26 April 2020].
https://arxiv.org/pdf/1904.04589

Description: This paper [?] analyzes the spoof audio created by recording the speech and replay it which dodge an automatic speaker verification mechanism. There is not much research conducted on building spoofing detection system which can handle the vulnerability of ASV systems. The detection system is based on the convolutional neural network. The training dataset consists of a speaker, genuine and replays audio clips. State of the art LCNN machine learning method used which consist of five convolutional layers, four network layers, five max pool layers and two fully connected layers. RELU is used as an activation function as it was a state of the art as compared to Max- feature-map (MFM) activations. For data preprocessing mean-variance normalized spectrogram generated. To keep the input consistent audio clips are truncated to 4 seconds before putting into the convolutional neural network. The output classes are spoofing, and genuine so binary entropy used, and it needs to be optimized by training the network. Maximum 100 epochs used while training the network and equal error rate calculated for the overall dataset. Further, they used SLIME model to gain insights about the model and then for

the model prediction. SLIME is works based on interpretable sequence for each input to get the class explanation. It was concluded that the performance of spoofing detection system majorly depends on the first few audio samples and this model got the satisfactory results.

Title of the work: Replay attack detection with raw audio waves and deep learning framework

Citation : S. Shukla, J. Prakash and R. Guntur, "Replay attack detection with raw audio waves and deep learning framework", 2019 International Conference on Data Science and Engineering (ICDSE), 2019.

Available: 10.1109/icdse47409.2019.8971793 [Accessed 26 April 2020].

https://ieeexplore.ieee.org/document/8971793

Description: Replay attacks [**?**] present a significant threat to Automatic Speaker Verification systems as they can be easily mounted using everyday smart devices by any non-professional imposter. The ASVspoof 2017 challenge was an initiative to develop solutions to counteract such replay attacks. The proposed solution builds on the fact that all the distinguishing features between genuine and spoofed audio are not effectively captured by conventional feature extraction techniques. Hence we propose a 1D ConvNet system with raw audio waves as features to it. This approach is able to achieve an EER of 0.41% on development set and 5.29% on evaluation set and hence outperforming best submission to ASVspoof 2017 challenge which had EER of 3.95% and 6.73% on development and evaluation sets respectively.

Chapter 3

High-level Design

3.1 Software development methodology

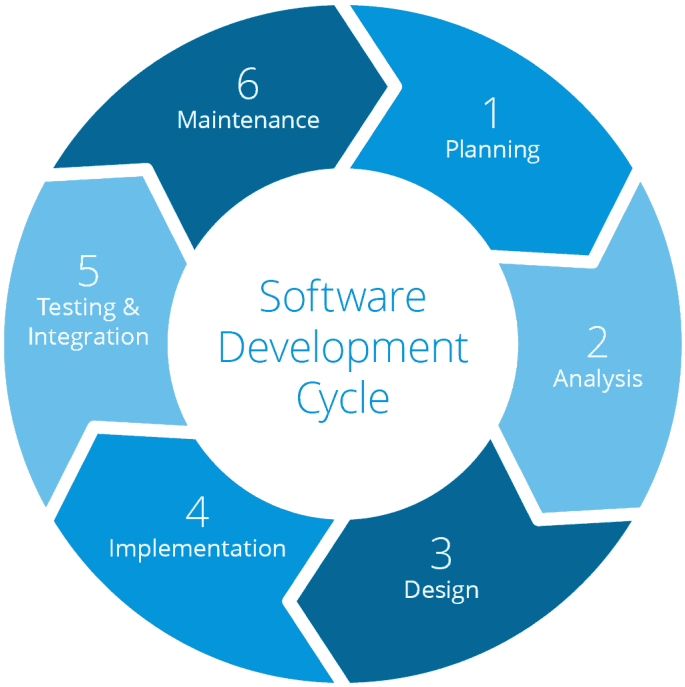

Figure 3.1: The structure of Software Development Life Cycle

It is a framework that is utilised to plan and structure the procedure of collecting the requirements in the form of data and information that is required in prior to start the project. Software Development is one of the part of system development.It incorporates the pre-meaning of expectations, ancient rarities made and finished by a task group for improvement or upkeep of the application.The software development methodology consists of a number of stages which are Requirements,Design,Implementation,Verification and Maintenance. Requirements phase includes the collection of all the requirement that is needed in order to develop the project and is required while developing it. It also includes the requirement analysis which is done in order to check whether the requirement collect is

of use or not. Design phase involves the completion of the product design. The team that is under design team works for the completion of all the elements of the design, specifications of the product and specifies a design for the manufacturing process as well. It is an iterative process. The third phase is the implementation phase which describes the part where the developers start coding which starts when the first two phases are completed which is the requirement and the design phase.This part includes all the coding section that is done developing a project. Here, the developers start converting the requirement and design into code fragments which is one of the most important part of the project development. The fourth phase is the verification phase where the products are tested and verified whether it performs the tasks as specified or not and whether it fulfils the needs of the customer or end user. This phase is important as it tests whether the model is doing for what it was developed. The last and the one of the important phase is Maintenance which occurs when the project is full working and needs any updates,repairs or any fixes specified by the user. The project is maintained and checked at regular interval to keep a track of what the user is demanding and to how much extent the project is fulfilling those requirements.

3.1.1 Why Agile for this project?

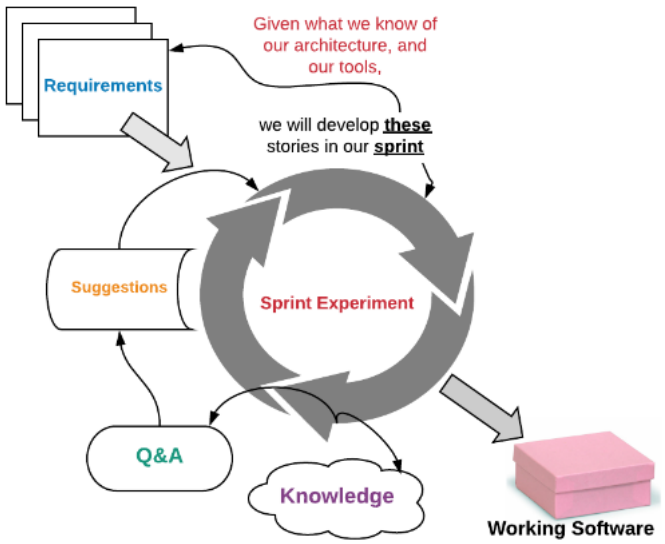

Figure 3.2: An Agile Architecture

Agile Model is more preferred than the traditional or the conventional models because the agile model is more efficient, reliable and easy to use. The model is chosen for its

speed because it is really quick. Nowadays a model is accepted which is ready to adapt the upcoming changes that occurs according to the business requirements and agile is the model which is adaptable to the changes occuring. So, this adds on one more point of choosing this model. A regular audit is done in agile model which improves the efficiency of the work done and the cooperation of the teammates involved. The agile model embraces iterative turn of events, and every cycle is intended to be little and reasonable that can be conveyed in a particular brief timeframe. i.e., a week or two or three weeks. An agile model is a gathering of advancement procedures, and its fundamental thought process is to evacuate/maintain a strategic distance from exercises that may not be required for the task and to expel anything which is an exercise in futility and exertion. Agile gives us a thought regarding the amount of our conveyed item is important and whether we've passed up a great opportunity onto something that matters. We have chosen agile so they can respond to changes in the marketplace or feedback from customers quickly without derailing a year's worth of plans. "Just enough" planning and shipping in small, frequent increments lets your team gather feedback on each change and integrate it into future plans at minimal cost. But it's not just a numbers game—first and foremost, it's about people. As described by the Agile Manifesto, authentic human interactions are more important than rigid processes. Collaborating with customers and teammates is more important than predefined arrangements. And delivering a working solution to the customer's problem is more important than hyper-detailed documentation. An agile team unites under a shared vision, then brings it to life the way they know is best. Each team sets their own standards for quality, usability, and completeness. Their "definition of done" then informs how fast they'll churn the work out. Although it can be scary at first, company leaders find that when they put their trust in an agile team, that team feels a greater sense of ownership and rises to meet (or exceed) management's expectations. Agile is a framework that characterizes how programming improvement should be finished. It is anything but a solitary or explicit strategy, and it is the assortment of different systems and best practices that follow the worth proclamation marked with the client. The model is adjusted to the Scrum model of agile improvement with certain changes as appeared above in the figure 3.2. The gatherings were directed week by week to guarantee that the task advancement is on target with the objective. The main motive of the meetings that are held regularly is to check how much work has been completed after the last meeting.

and what has to be discussed in the next meeting and what works should be completed by the next meeting.

3.2 Architecture

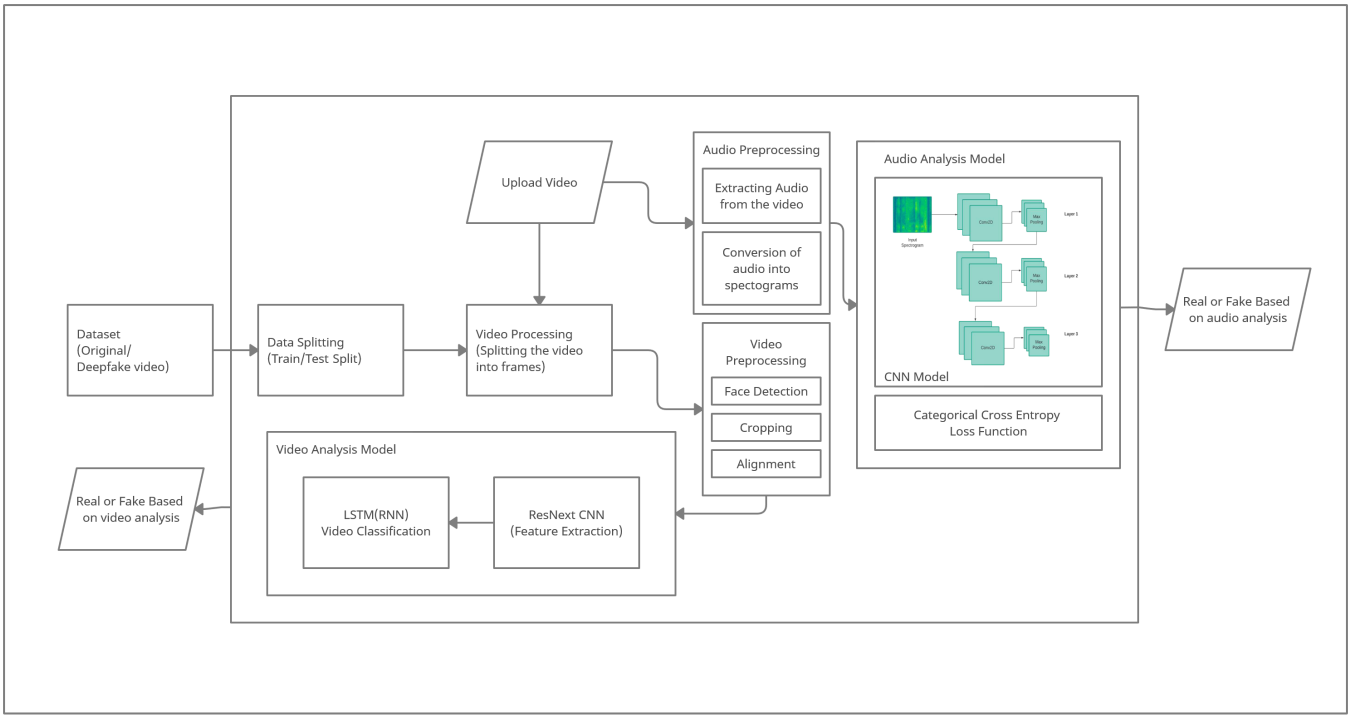

Figure 3.3: System Architecture

The above figure shown explains the architecture of the proposed model. The method is based on a properties of the DF videos, due to limitation of computation resources and production time, the DF algorithm can only synthesize face images of a fixed size, and they must undergo an affinal warping to match the configuration of the source's face. This warping leaves some distinguishable artifacts in the output deepfake video due to the resolution inconsistency between warped face area and surrounding context. Our method detects such artifacts by comparing the generated face areas and their surrounding regions by splitting the video into frames and extracting the features with a ResNext Convolutional Neural Network (CNN) and using the Recurrent Neural Network (RNN) with Long Short Term Memory(LSTM) capture the temporal inconsistencies between frames introduced by GAN during the reconstruction of the DF. To train the ResNext CNN model, we simplify the process by simulating the resolution inconsistency in affine face wrappings directly. In this architecture a sequential model created with three Conv2D

consisting of 32,64 and 128 layers. Batch normalization layers also added which convert the input values to standardized values automatically in Deep neural network. After normalization, max pooling layer added in the network which reduces the dimensionality of the input images features while keeping the maximum spatial data. The layer models look like the above figure. ReLU is used as an activation function in conv2d layers and it gave the nonlinear output and model compiled with categorical crossentropy because output classes are categorical and have more than 2 labels that are why binary cross-entropy is not used here. The goal here is to keep the loss minimum to get the optimized results, the loss is a cost which explained the values of variables associated with real integers lost. Adam is such loss function used here to gain the optimization.

3.3 Functional Requirements

1. A simple web application that can be easily accessed from anywhere.

2. User can easily upload a video which he/she wants to verify whether the video is pristine or deepfake.

3. The web app will give the result of the uploaded video whether it is genuine or fake depending upon the prediction it performs on the basis of the trained model.

3.4 Non-Functional Requirements

3.4.1 Data Preparation

Data preparation is the process of cleaning and changing crude data preceding handling and examination. It is a significant advance preceding handling and frequently includes reformatting data, making amendments to data and the consolidating of data sets to enhance data. This file contains all the pre processing techniques that are used to process the data. Dataset preprocessing includes the splitting the video into frames. Followed by the face detection and cropping the frame with detected face. To maintain the uniformity in the number of frames the mean of the dataset video is calculated and the new processed face cropped dataset is created containing the frames equal to the mean. The frames that doesn't have faces in it are ignored during preprocessing. As processing the 10 second video at 30 frames per second i.e total 300 frames will require a lot of computational power. So for experimental purpose we are proposing to used only first 100 frames for

training the model. In this study Automatic speaker verification ASVspoof 2019 dataset used which is publicly available. It consists of several audio files which contains real, text to speech and voice conversion audio data [21]. The data is downloaded and stored in the hard disk directory and referred from there. Due to computational constraints of the system, limited audio files (25380 audio files) have been used in this research along with the given text file containing description of each audio files. We utilized the text files to segregate the audio files according to their system id (A01- A19) and created three folders, one for each class (Real, Spoof_TTS and Spoof_VC). Audio files have not been standardized in the processing part to maintain the essential feature of the data as clips have only few seconds audio. For each audio class a respective image folder is created which contains the spectrograms of audio files belonging to that class. And finally, all the spectrograms are moved to a single directory to create dataframe of image data.

3.4.2 Feature Selection

In this file feature extraction and selection methods have been performed. Instead of writing the rewriting the classifier, we are proposing to use the ResNext CNN classifier for extracting the features and accurately detecting the frame level features. 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 are then used as the sequential LSTM input.

3.4.3 Prediction

A new video is passed to the trained model for prediction. A new video is also preprocessed to bring in the format of the trained model. The video is split into frames followed by face cropping and instead of storing the video into local storage the cropped frames are directly passed to the trained model for detection.