‘DEEPFAKE DETECTION’

Project Report submitted to Shri Ramdeobaba College of Engineering & Management, Nagpurin partial fulfillment of the requirement for the award of the degree of

Bachelor of Technology

in

COMPUTER SCIENCE AND ENGINEERING (DATA SCIENCE)

by

Mr. Aryan Kashikar

Ms. Disha Vasani

Mr. Pranav Shivhare

Guide

Prof. Priya Khobragade

RCOEM

Shri Ramdeobaba College of Engineering and Management, Nagpur

Computer Science and Engineering (Data Science)

Shri Ramdeobaba College of Engineering & Management, Nagpur
(An Autonomous Institute affiliated to Rashtrasant Tukdoji Maharaj Nagpur University, Nagpur)

December 2023

SHRI RAMDEOBABA COLLEGE OF ENGINEERING & MANAGEMENT, NAGPUR

(An Autonomous Institute affiliated to Rashtrasant Tukdoji Maharaj Nagpur University, Nagpur)

Department of Computer Science and Engineering (Data Science)

CERTIFICATE

This is to certify that the project **"DEEPFAKEDETECTION"** is a bonafide work of

1.Mr. Aryan Kashikar

2.Ms. Disha Vasani

3.Mr. Pranav Shivhare

submitted to the Rashtrasant Tukdoji Maharaj Nagpur University, Nagpur in partial fulfillmentof the award of a Degree of Bachelor of Engineering, in Computer Science and Engineering (Data Science). It has been carried out at the Department of Computer Science and Engineering (Data Science), Shri Ramdeobaba College of Engineering and Management, Nagpur during theacademic year 2023-24.

Date: 30.12.2023

Place: Nagpur

Prof. A.M. Karandikar
H.O.D
Department of Computer Science and Engineering (Data Science)

Prof. Priya Khobragade
Project Guide

Dr. R. S. Pande
Principal (RCOEM)

DECLARATION

I, hereby declare that the project titled **"DEEPFAKE DETECTION"** submitted herein has been carried out in the Department of Computer Science and Engineering (Data Science) of Shri Ramdeobaba College of Engineering & Management, Nagpur. The work is original and has not been submitted earlier as a whole or part for theaward of any degree/diploma at this or any other institution / University

Date: 30.12.2023
Place: Nagpur

Mr. Aryan Kashikar
(Roll no.: 32)

Ms. Disha Vasani
(Roll no.: 04)

Mr. Pranav Shivhare
(Roll no.: 46)

ACKNOWLEDGEMENT

We would like to express our deep and sincere gratitude to our guide **Prof. Priya Khobragade,** Assistant Professor in the Computer Science and Engineering Department (Data Science), RCOEM, for allowing us to work on this project and providing valuable guidance throughout the project. It was a great privilege and honor to work under his guidance. We are extremely grateful for the experience we had in this project with him.

We express our sincere gratitude to **Prof. A. M. Karandikar**, Head of the Department of Computer Science Department (Data Science), RCOEM for his guidance. Talent wins games, but teamwork and intelligence win championships. We would like to take this opportunity to express our deep gratitude to all those who extended their support and guided us to complete this project.

Mr. Aryan Kashikar **Ms. Disha Vasani**

Mr. Pranav Shivhare

Approval Sheet

This report entitled **"DEEPFAKE DETECTION"** by Aryan Kashikar, Disha Vasani, Pranav Shivhare is approved for the degree of Bachelor of Technology, in Computer Science andEngineering (Data Science)

Name & signature of Supervisor(s) Name & Signature of External Examiner(s)

--- ---

Prof. A. M. Karandikar

H.O. D

Department of Computer Science and Engineering

(Data Science)

Date: 30.12.2023

Place: Nagpur

ABSTRACT

The growing computation power has made the deep learning algorithms so powerful that creating an indistinguishable human-synthesized video popularly called deep fakes has become very simple. Scenarios where these realistic face-swapped deep fakes are used to create political distress, fake terrorism events, revenge porn, and blackmail people 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 reenactment of 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 a 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 a large amount of balanced and mixed data.

Keywords:

Res-Next Convolution neural network,

Recurrent Neural Network (RNN),

Long Short-Term Memory (LSTM),

Computer vision.

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

CHAPTER 1: INTRODUCTION

1.1 Project Idea

In the world of ever-growing Social media platforms, Deepfakes are considered the major threat of AI. There are many scenarios where these realistic face-swapped deepfakes are used to create political distress, fake terrorism events, revenge porn, and blackmail people are easily envisioned.

Some examples are Brad Pitt and Angelina Jolie's 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 Snapchat and Face Swap, which use pre-trained neural networks like GAN or Auto encoders for these deepfakes creation.

Our method uses an 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.

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.

1.2 Aim and Objectives

1. The primary goal of our project is to investigate and unveil the distorted truth propagated through deepfakes. Deepfakes, driven by artificial intelligence, have emerged as a formidable tool for manipulating audiovisual content. These manipulations often lead to the creation of synthetic media that can deceive and mislead individuals, blurring the lines between reality and fiction.

2. Our project targets a reduction in online abuses and misinformation, striving to create a safer digital space for common internet users. Through advanced tools, education, and collaborations, we aim to significantly diminish instances of online harassment and misleading content. By empowering users with media literacy and fostering partnerships, the project seeks to establish a more secure and trustworthy online environment, promoting responsible digital engagement globally.

3. Our project aims to create a system that accurately distinguishes between deepfake and authentic videos, utilizing advanced computer vision and machine learning techniques. This tool will contribute to the identification of manipulated content, promoting media authenticity and trust in video content across platforms.

4. Our project aims to provide a user-friendly system for effortless video upload and instant verification, using advanced technology to distinguish between real and fake content. This streamlined tool empowers users to assess video authenticity quickly, fostering a safer online environment.

1.3 Motivation of the Project

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 transformative technology, this has created new challenges. So-called "deep fakes" are 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.

Until recently, the number of fake videos and their degrees of realism have been increasing due to availability of the editing tools, and the high demand for domain expertise. Spreading of the Deep fakes over the social media platforms has become very common leading to spamming and peculating wrong information over the platform.

Just imagine a deep fake of our prime minister declaring war against neighboring countries, or a Deep fake of a reputed celebrity abusing the fans. These types of deep fakes will be terrible, and lead to threatening and misleading 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 to develop technology that can spot fakes so that the deep fakes can be identified and prevented from spreading over the internet.

1.4 Proposed system Flow Graph

DFD Level-0

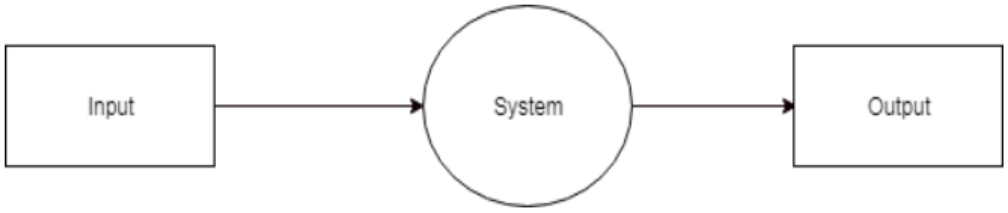

Figure 1.1

DFD level – 0 indicates the basic flow of data in the system. In this System Input is given equal importance as that for Output.

- Input: Here input to the system is uploading video.
- System: In system it shows all the details of the Video.
- Output: Output of this system is it shows the fake video or not.

  Hence, the data flow diagram indicates the visualization of system with its input and output flow.

DFD Level-1

[1] DFD Level – 1 gives more in and out information about the system.

[2] Where the system gives detailed information of the procedure taking place.

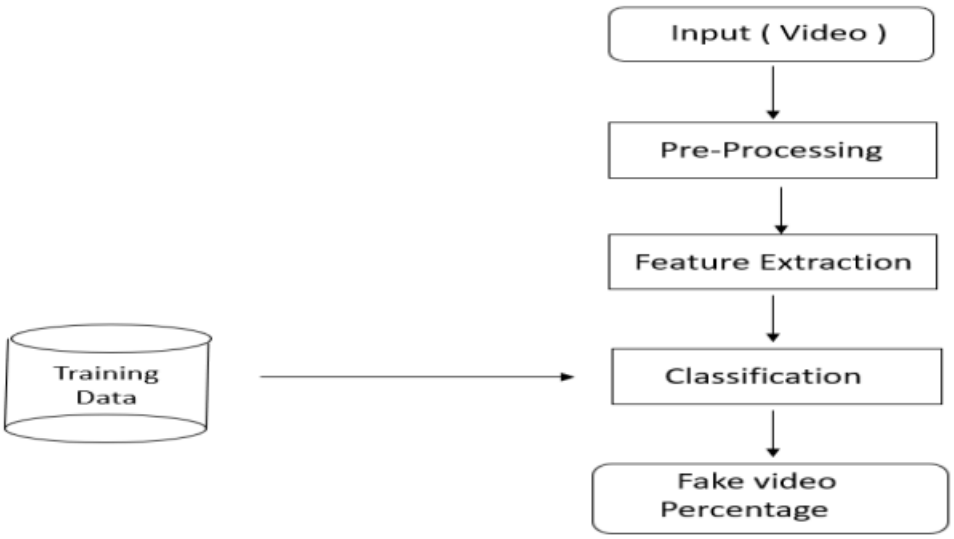

Figure 1.2

DFD Level-2

[1] DFD level-2 enhances the functionality used by user etc.

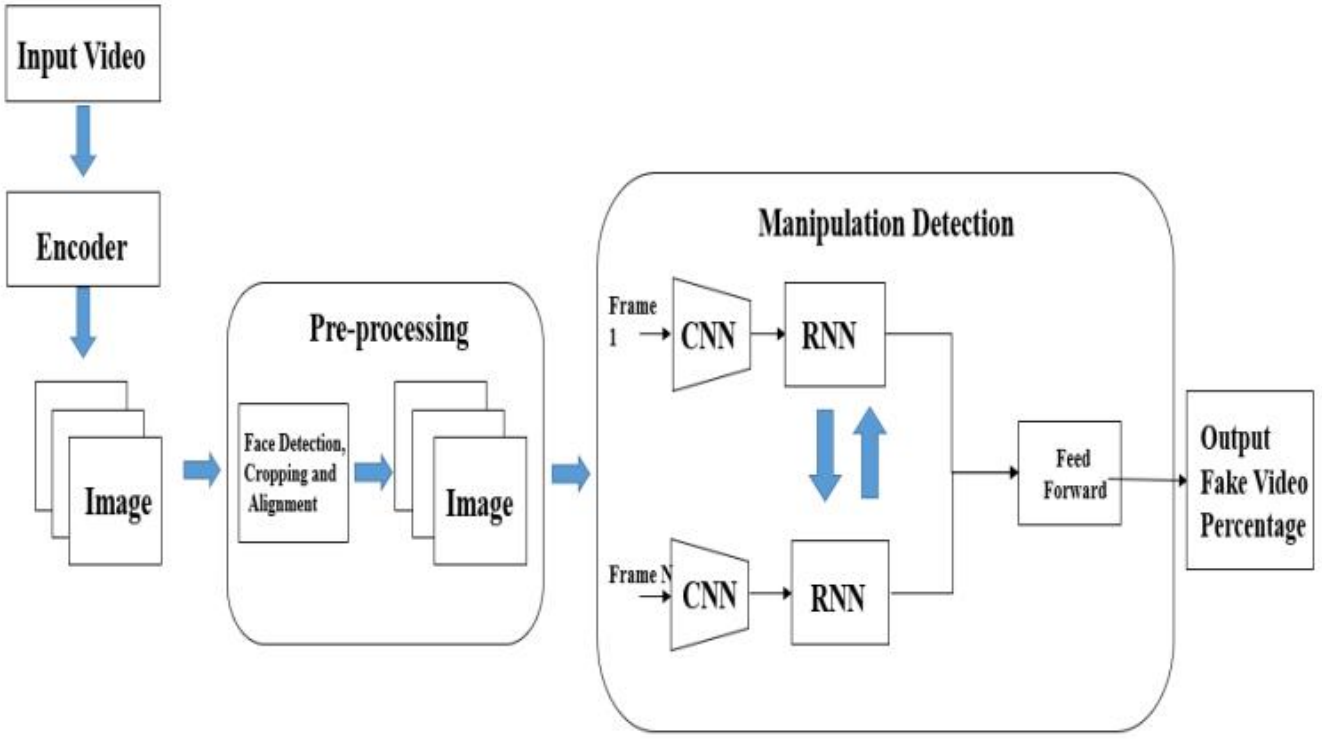

Figure 1.3

CHAPTER 2: LITERATURE SURVEY

2.1 Overview and Survey

Face Warping Artifacts used the 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 the current deepfake algorithm can only generate images of limited resolutions, which then need to be further 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 describes a new method for detecting deepfakes by eye blinking as a crucial parameter leading to the 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. Today the deepfake generation algorithms have become so powerful that lack of eye blinking can not be the only clue for the detection of the deepfakes. There must be certain other parameters must be considered for the detection of deepfakes like teeth enchantment, wrinkles on faces, wrong placement of eyebrows, etc.

Capsule networks to detect forged images and videos use 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.

[1] Nicolo Bonettini et.al, "**Video Face Manipulation Detection Through Ensemble of CNNs**". 16 Apr 2020-A project which able to detect whether a video contains manipulated content is nowadays of paramount importance, given the significant impact of videos in everyday life and in mass communications. In this vein, we tackle the detection of facial manipulation in video sequences, targeting classical computer graphics as well as deep learning-generated fake videos.

[2] MD Shohel Ranaet.al, "**Deepfake Detection: A Systematic Literature**" Review24 February 2022-This systematic literature review (SLR) covers 112 studies from 2018 to 2020, exploring state-of-the-art methods for detecting Deepfake. Key findings include the widespread use of deep learning, particularly convolutional neural networks (CNN), and a focus on the FF++ dataset. Detection accuracy is the primary performance metric. The results indicate that deep learning techniques are effective, with deep learning models generally outperforming non-deep learning models. Despite progress in multimedia technology, Deepfake detection still presents challenges, and this SLR aims to guide the development of effective detection methods and countermeasures for the research community.

[3] Andreas Rossler Davide Cozzolinoet.al, "**FaceForensics++: Learning to Detect Manipulated Facial Images**". 26 Aug 2019-The paper's primary focus is on the impact of compression on the detectability of advanced manipulation methods. It also proposes a standardized benchmark for future research, with all data, trained models, and the benchmark being publicly accessible. This dataset aids in detecting fakes with minimal training data, advancing digital media forensics research, especially in the realm of facial forgeries.

CHAPTER 3: TOOLS AND TECHNOLOGIES

3.1 Programming Languages

- Python3
- JavaScript

3.2 Programming Frameworks

- PyTorch
- Django

3.3 IDE

- Google Colab
- Jupyter Notebook
- Visual Studio Code

3.4Application and Web Servers

- Google Cloud Engine

3.5Libraries

- pandas
- numpy
- sklearn
- os
- random

CHAPTER 4: METHODOLOGY

4.1 Analysis

We analyzed the problem statement and found the feasibility of the solution to the problem. We read different research papers. After checking the feasibility of the problem statement.

The next step is the dataset gathering and analysis. We analyzed the data set in different approaches of training like negatively or positively trained i.e. training the model with only fake or real video but found that it may lead to the addition of extra bias in the model leading to inaccurate predictions.

So after doing a lot of research, we found that the balanced training of the algorithm is the best way to avoid the bias and variance in the algorithm and get a good accuracy.

We analyzed the solution in terms of cost, speed of processing, requirements, level of expertise, and availability of equipment.

Parameter Identified

1. Blinking of eyes
2. Bigger distance for eyes
3. Double edges, eyes, ears, nose
4. Iris segmentation
5. Wrinkles on face
6. Inconsistent head pose
7. Face angle
8. Facial Expressions

4.2 Design

After research and analysis, we developed the system architecture of the solution as mentioned in Chapter 5. We decided on the baseline architecture of the Model which includes the different layers and their numbers.

4.3 Development

Following an exhaustive analysis, our selection for programming entails harnessing the Python3 language synergistically with the PyTorch framework. The rationale behind this preference lies in PyTorch's adeptness, offering robust support for CUDA, empowering seamless utilization of Graphic Processing Unit (GPU) acceleration.
Additionally, PyTorch's inherent customizability aligns with our specific requirements. For the ultimate model training on an extensive dataset, we strategically opt for the Google Cloud Platform, leveraging its scalable infrastructure to efficiently handle the computational demands inherent in processing large volumes of data..

4.3 Training and Testing Workflow

Training Workflow

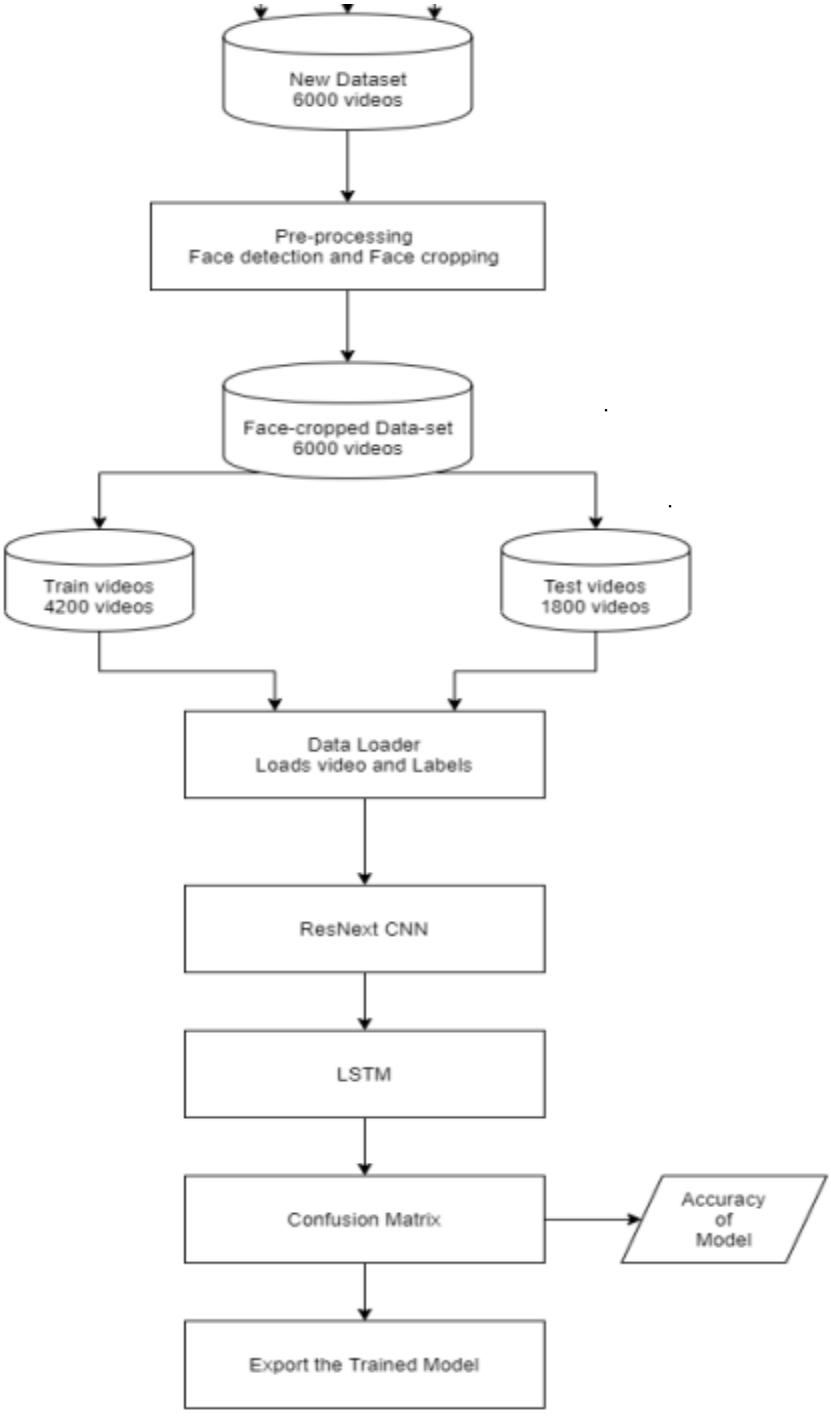

Figure 4.1

Testing Workflow

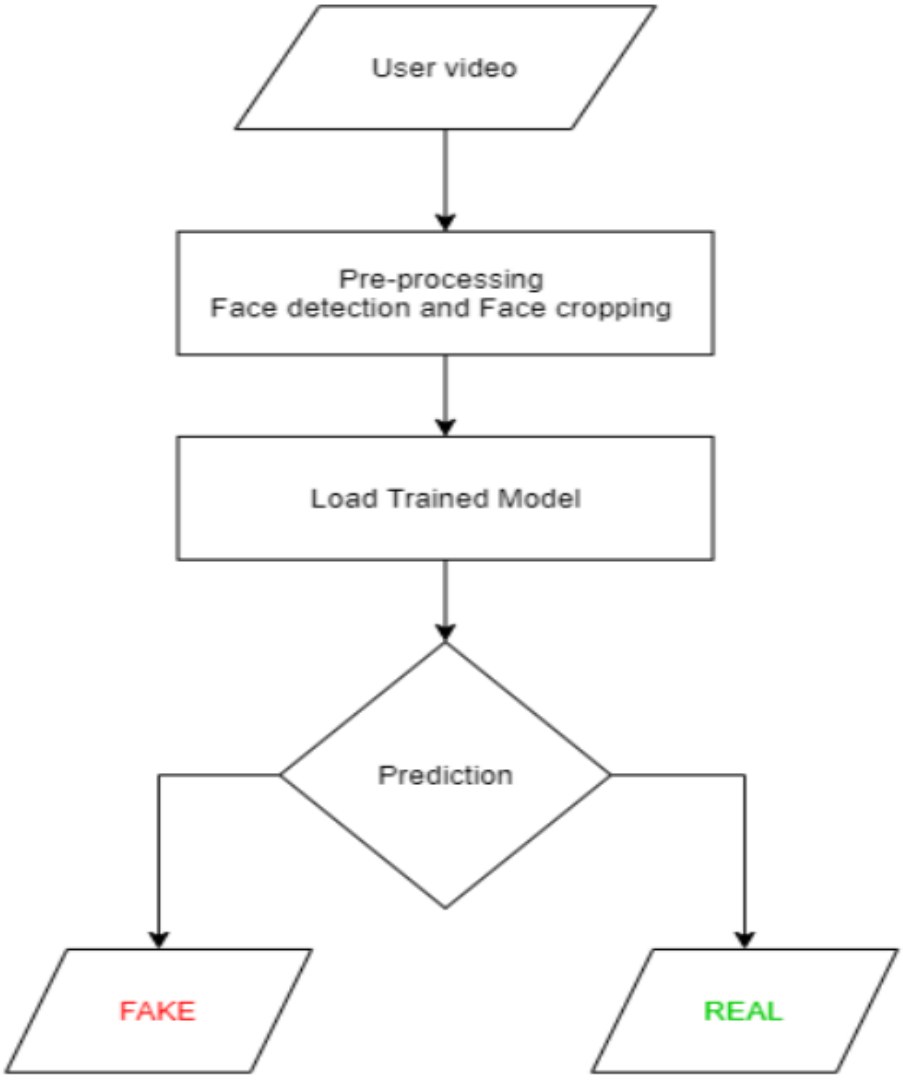

Figure 4.2

CHAPTER 5: IMPLEMENTATION OF METHODOLOGY

5.1 Architectural Design

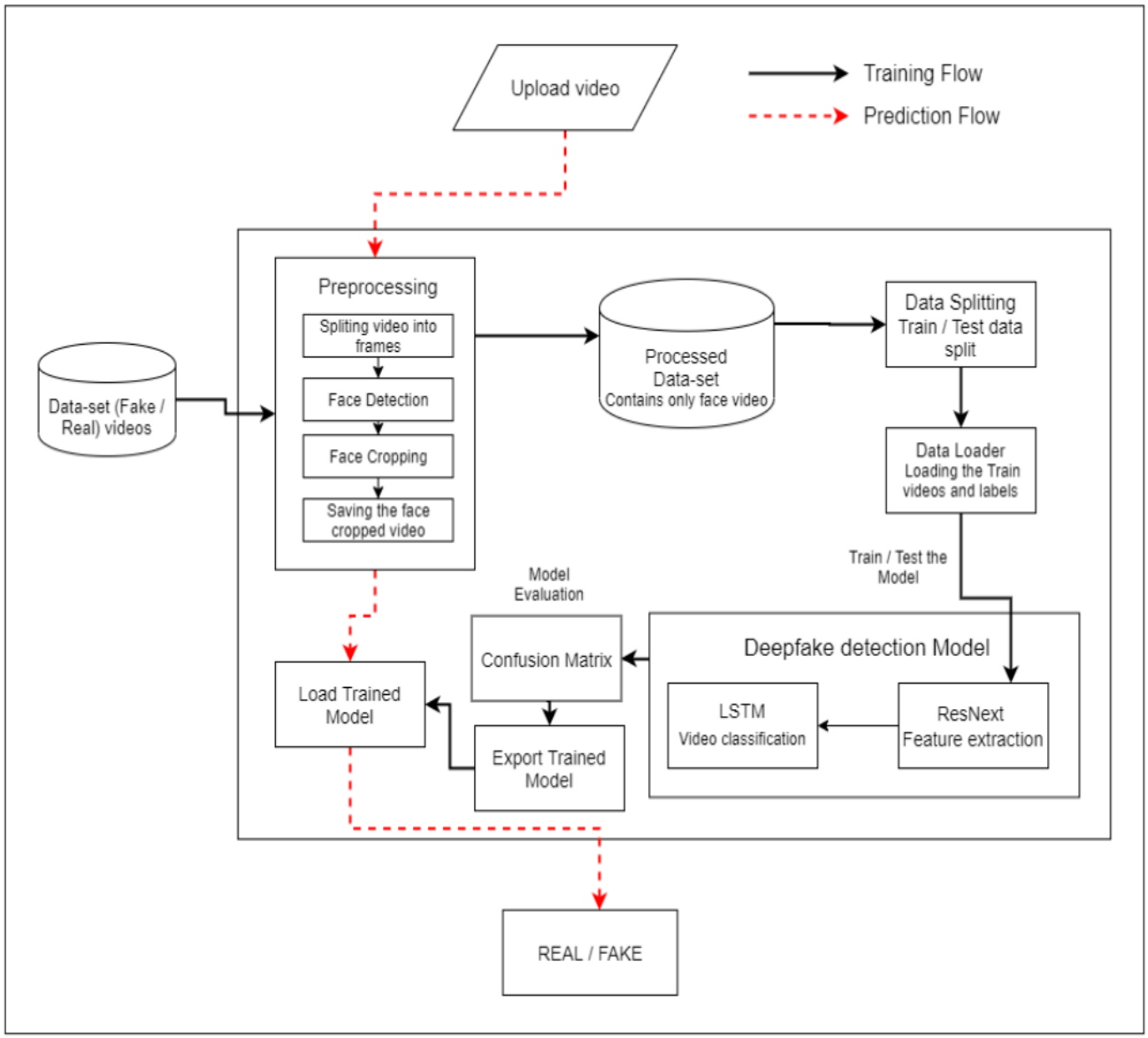

Figure 5.1

Module 1: Data-set Gathering

Developed a meticulously crafted dataset, characterized by a balanced 50% representation of authentic (Real videos) and manipulated videos (Fake videos). This deliberate distribution serves as the cornerstone for a strategically designed approach, aiming to encapsulate a diverse array of content within the dataset. The intention behind this diversity is to facilitate comprehensive analysis and ensure the efficacy of model training. By exposing models to a wide spectrum of video types, the dataset empowers them to discern patterns, intricacies, and variations, fostering a more robust understanding that enhances their performance in handling real-world scenarios with accuracy and adaptability.

Module 2: Pre-processing

In this step, the videos are preprocessed and all the unrequired noise is removed from the videos. Only the required portion of the video i.e. face is detected and cropped. The first step in the preprocessing of the video is to split the video into frames. After splitting the video into frames the face is detected in each of the frames 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 the creation of a processed dataset containing face-only videos. The frame that does not contain the face is ignored while preprocessing.

A video of 10 seconds at 30 frames per second(fps) will have a total of 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 the 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 sequentially considered the frames i.e. first 150 frames and not randomly.

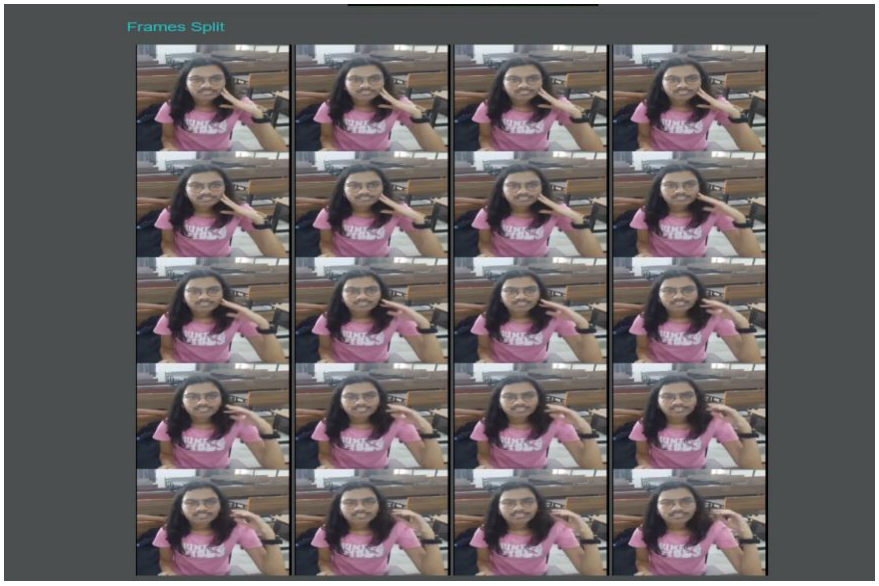

Figure 5.2

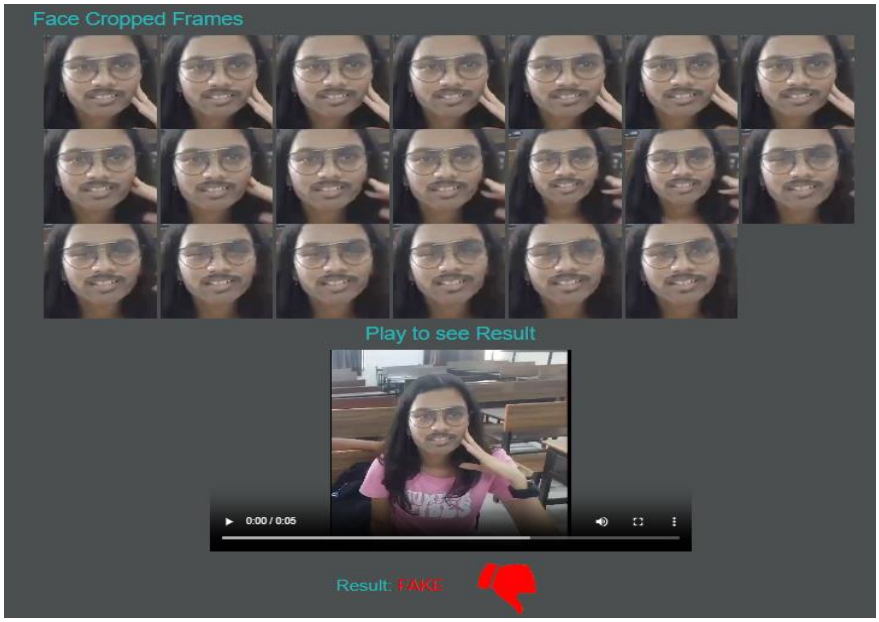

Figure 5.3

Module 3: Data-set split

The dataset is split into train and test datasets with a ratio of 70% train videos and 30% test videos. The train and test split are balanced split i.e. 50% of the real and 50% of fake videos in each split.

5.2 Model Details

The model consists of the following layers:

- **ResNext CNN:** ResNext is a convolutional neural network architecture that enhances traditional residual networks by employing a cardinality parameter, allowing for increased model capacity and improved performance by aggregating diverse information through parallel pathways, fostering richer feature representations.

- **Sequential Layer:** Sequential is a container of Modules that can be stacked together and run at the same time. The sequential layer is used to store the feature vector returned by the ResNext model in an ordered way. So that it can be passed to the LSTM sequentially.

- **LSTM Layer:** LSTM is used for sequence processing and spotting the temporal change between the frames.2048-dimensional feature vectors are fitted as the input to the LSTM. We are using 1 LSTM layer with 2048 latent dimensions and 2048 hidden layers along with a 0.4 chance of dropout, which is capable of achieving our objective. LSTM is used to process 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.

- **Confusion Matrix:** A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions is summarized with count values and broken down by each class. This is the key to the confusion matrix. The confusion matrix shows how your classification model is confused when it makes predictions. It gives us insight not only into the errors being made by a classifier but more importantly the types of errors that are being made. Confusion matrix is used to evaluate our model and calculate the accuracy.

5.3Model Training Details

• **Train Test Split:** The dataset is split into train and test datasets with a ratio of 70% train videos and 30% 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.

• **Data Loader:** It is used to load the videos and their labels with a batch size of 4.

• **Training:** The training is done for 20 epochs with a learning rate of 1e-5 (0.00001), and weight decay of 1e-3 (0.001) using the Adam optimizer.

• **Adam optimizer:** To enable the adaptive learning rate Adam optimizer with the model parameters is used.

• **Cross Entropy:** To calculate the loss function Cross Entropy approach is used because we are training a classification problem.

• **Confusion Matrix:** A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions is summarized with count values and broken down by each class. This is the key to the confusion matrix.

5.4 Model Prediction Details:

• The model is loaded into the application

• The new video for prediction is preprocessed and passed to the loaded model for prediction

• The trained model performs the prediction and returns if the video is real or fake along with the confidence of the prediction.

• Our project outperforms existing benchmarks, notably surpassing the FaceForensics++:Learning to Detect Manipulated Facial Images research paper, which reports a 70% accuracy in detecting manipulated facial images. Our achievement stands at an impressive 90% accuracy.

CHAPTER 6: RESULTS, FUTURE SCOPE AND CONCLUSION

6.1 RESULTS

Screenshots:

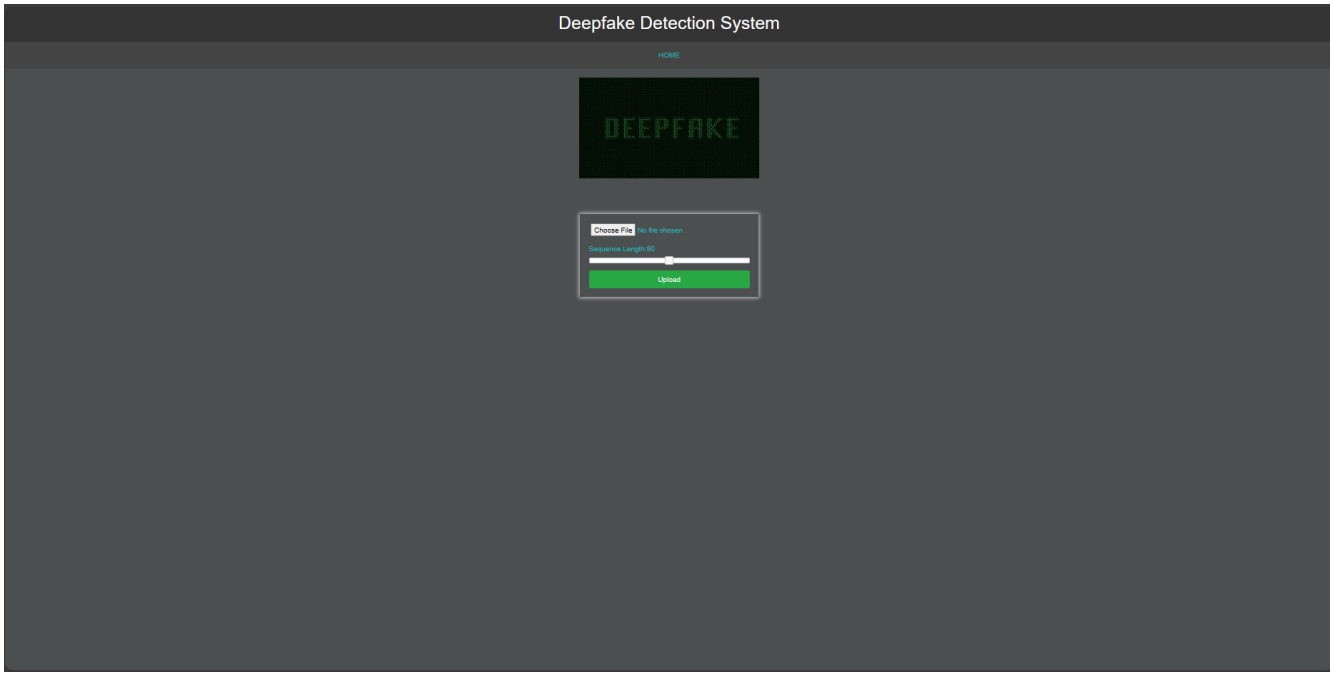

Figure 6.1

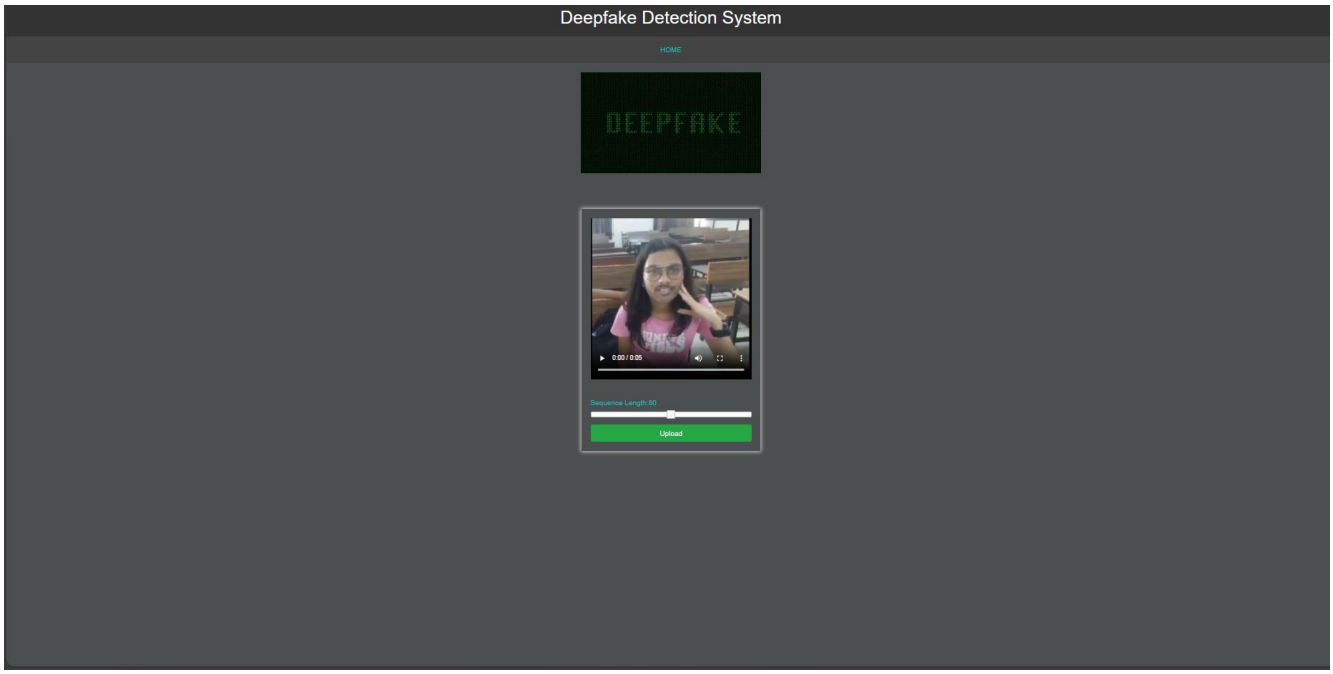

Figure 6.2

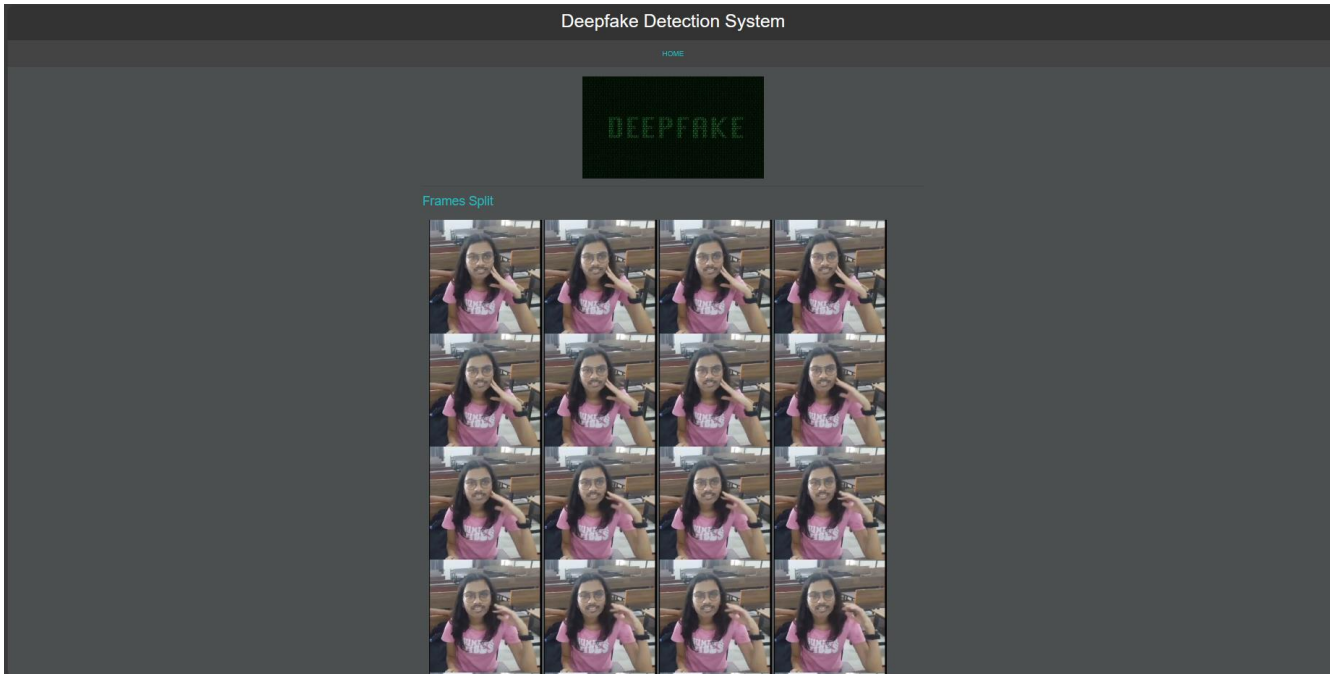

Figure 6.3

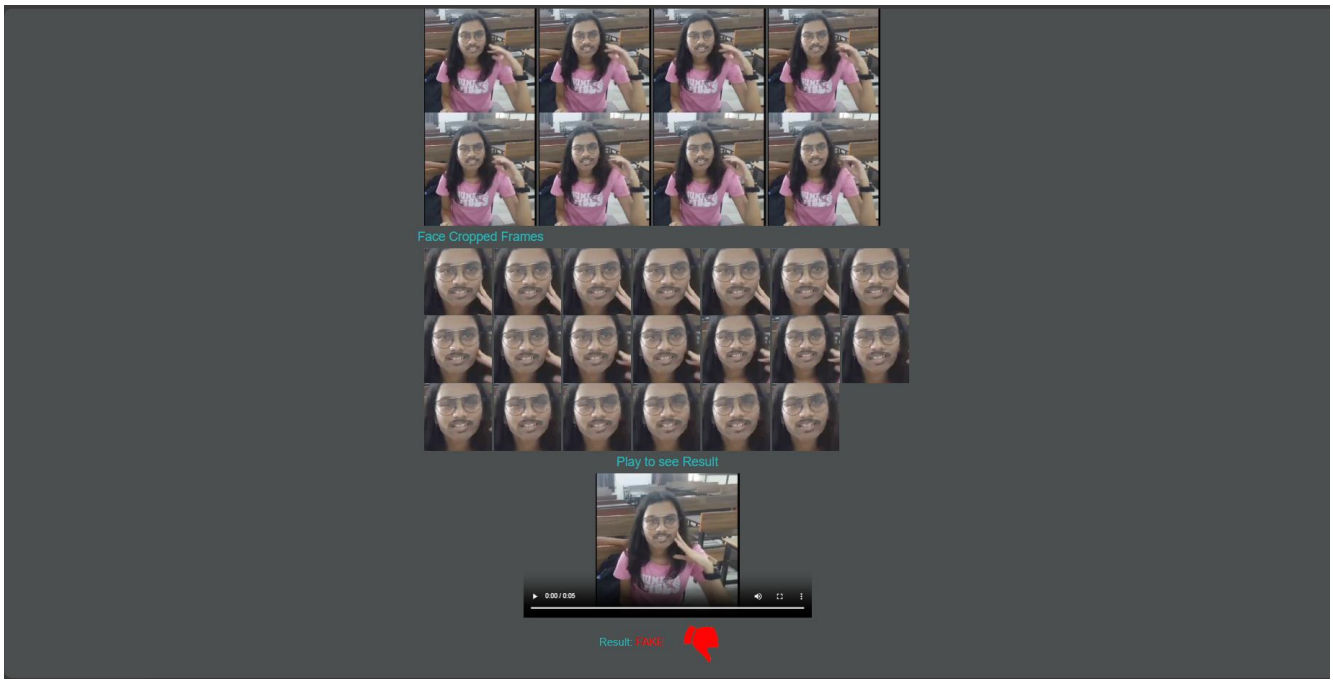

Figure 6.4

6.2 Future Scope:

There is always a scope for enhancements in any developed system, especially when the project is built using the latest trending technology and has a good scope in the future.

Web-based platforms can be upscaled to a browser plugin for ease of access to the user.

Currently only Face Deep Fakes are being detected by the algorithm, but the algorithm can be enhanced in detecting full-body deep fakes.

6.3 Conclusion:

We presented a neural network-based approach to classify the video as deep fake or real, along with the confidence of the proposed model. Our method is capable of predicting the output by processing 1 second of video (10 frames per second) with good accuracy. We implemented the model by using a 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] Nicolo Bonettini et.al,'' Video Face Manipulation Detection Through Ensemble of CNNs.

[2] MD Shohel Rana et. al," Deepfake Detection: A Systematic Literature Review24

[3] Andreas Rossler Davide Cozzolinoet.al," FaceForensics++: Learning to Detect Manipulated Facial Images.

[4] Deepfake Video Detection Using Long Short-Term Memory

[5] 10 Deepfake examples that terrified and amused the internet: https://www.creativebloq.com/features/deepfake-examples

[6] TensorFlow: https://www.tensorflow.org/

[7] Keras: https://keras.io/

[8] PyTorch: https://pytorch.org/