BRAUDE
College of Engineering, karmiel

CAPSTONE PROJECT
PHASE B

Manipulated Reality

Machine Learning Deepfake Detection System

Presented By

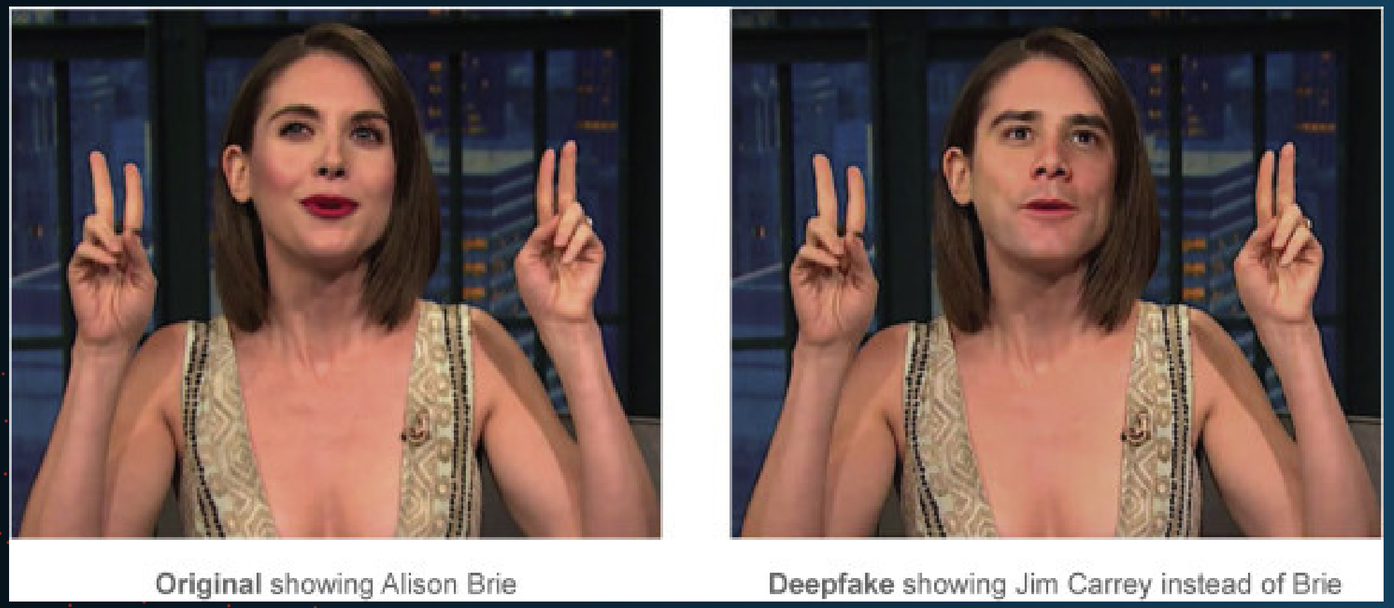
Maxim Lebedisnky 318171485 **Dmytri Kislitsyn** 323349738

supervised by Dr. Reuven Cohen

Table of Contents

Deepfakes
Introduction

Deepfakes are highly realistic, artificially generated content (images, videos, audio) that manipulate or replace existing media to depict people saying or doing things they never did. They leverage deep learning techniques to create convincing fakes that are often indistinguishable from real content.



Original showing Alison Brie

Deepfake showing Jim Carrey instead of Brie

The necessity of Deepfake Detection

Deepfakes pose a significant threat to our society. They can be used to:

- **Spread misinformation and disinformation:** Manipulating public opinion and influencing elections.
- **Harm individuals:** Disgrace reputations, blackmailing, and causing emotional distress.
- **Undermine trust:** Crumbling faith in institutions and the media.


South Korean Teenagers Detained Over Deepfake Sexual Images

The country has been hit by a wave of sexually explicit videos and pictures that have spread online, prompting the authorities to launch a sweeping investigation.

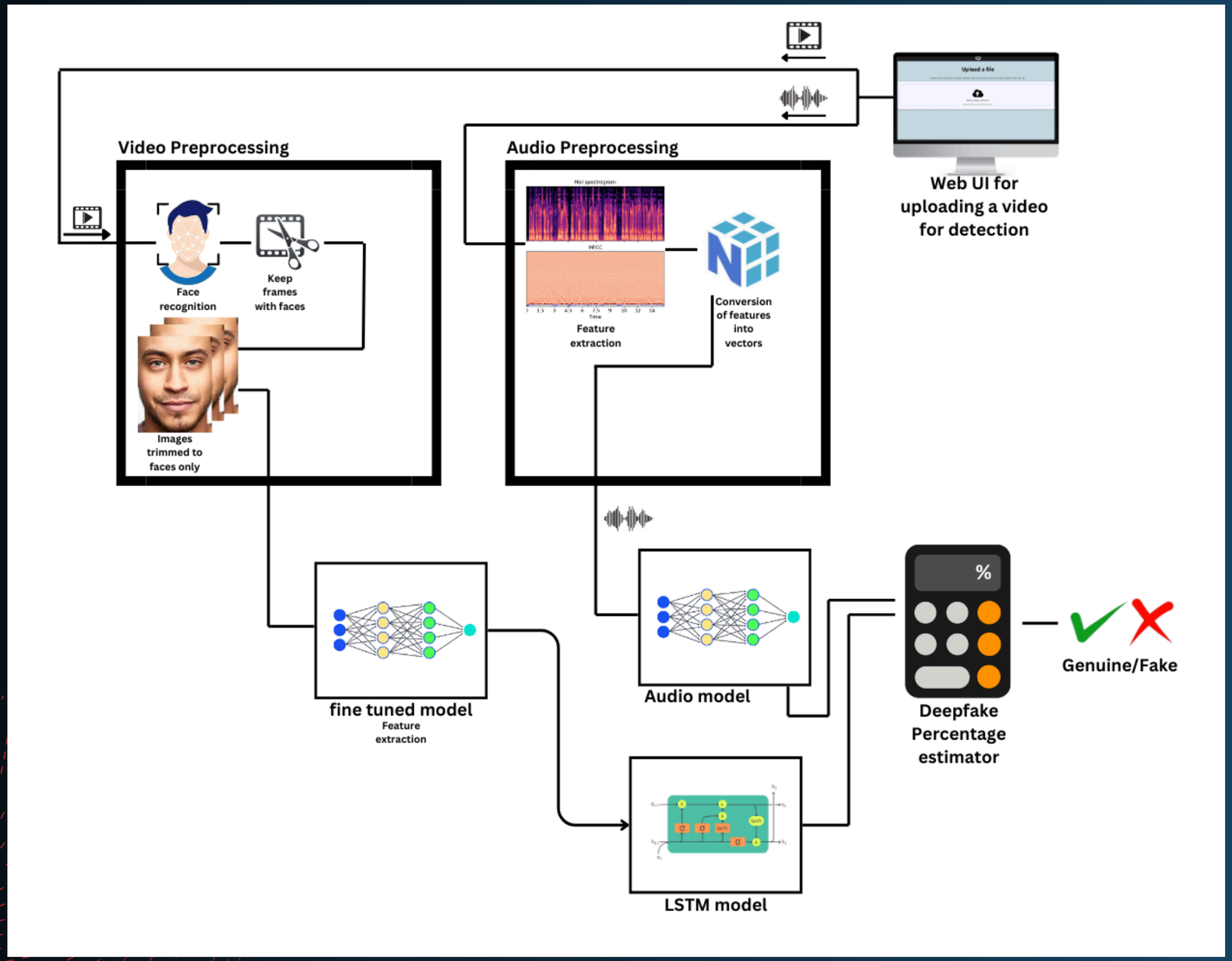

TECHNOLOGY

Fake babies, real horror: Deepfakes from the Gaza war increase fears about AI's power to mislead

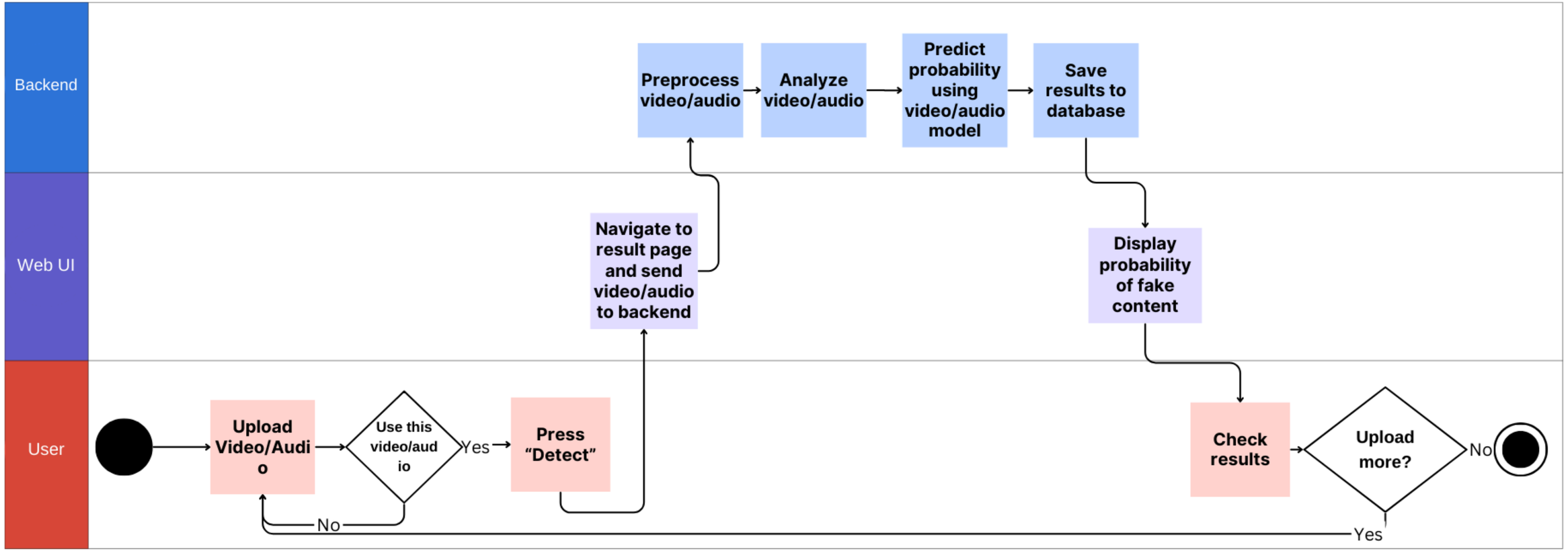
Project Requirements

Our system prioritizes **high accuracy** for both video and audio models while ensuring **minimal computational** demands on the user. We aim for a **simple and intuitive** user experience, providing straightforward results without complex interactions.

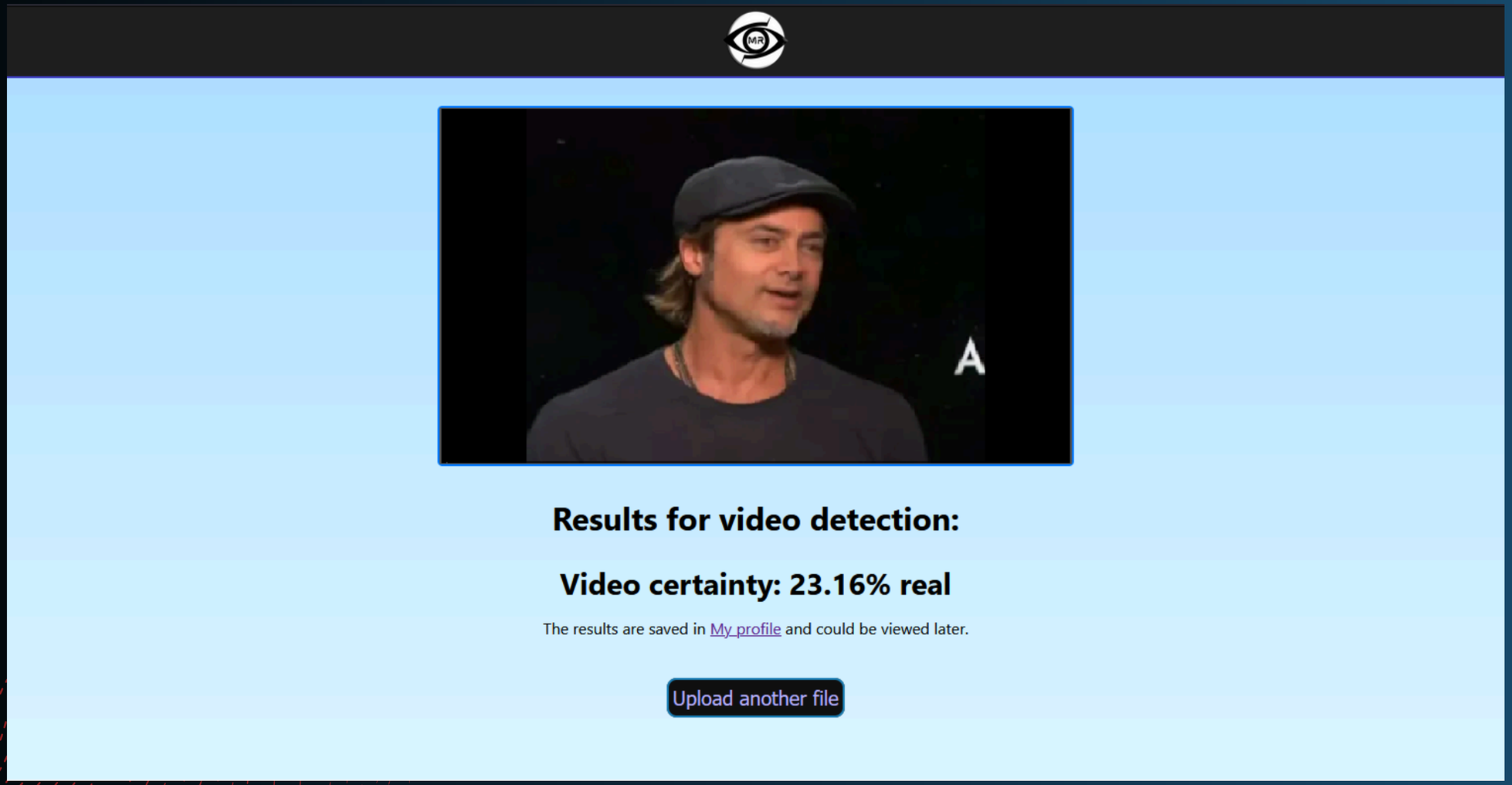
System Architecture



Activity Diagram



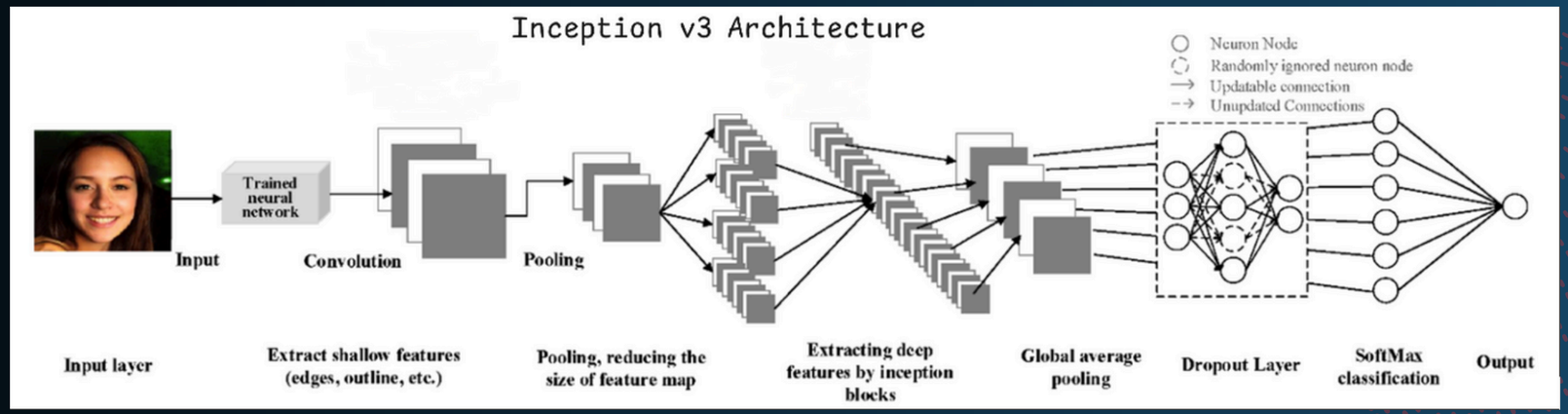
Website Page Example



Deepfake Detector

Video Model

A video model was trained on the **Celeb-DF-V2** dataset using a pre-trained **InceptionV3** model and a **Long Short-Term Memory** (LSTM) network. Facial extraction was performed using the face_recognition library, followed by frame rate reduction and total frame selection. The model was developed in Google Colab Pro+ using Python and Keras.



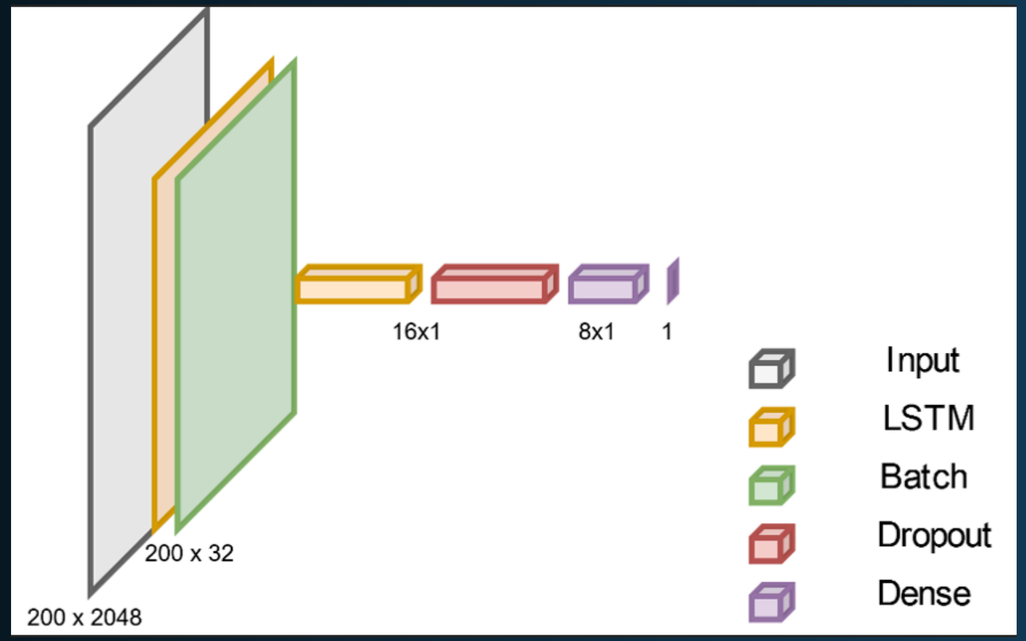
Video Model

Dataset

The **Celeb-DF (v2)** dataset, consist of 5639 deepfake videos and 590 real videos. To mitigate class imbalance, an additional 4900 real videos were sourced from YouTube, C23,and DFMNIST+.

Video Model Development Process

1. Preprocess the data
2. Fine tune InceptionV3
3. Feature extraction using the fine tuned model
4. Construct LSTM network
5. Input the feature extracted data into the LSTM network



Deepfake Detector

Audio Model

The audio model was trained on three audio datasets.

- Fast: Up to 10 seconds for a single audio file.
- Accurate: Delivers promising results on audio data.
- Simple structure: 12-layer machine learning model.
- Accessible: Provides machine learning analysis access to everyone.

Audio Model

Preparing the datasets

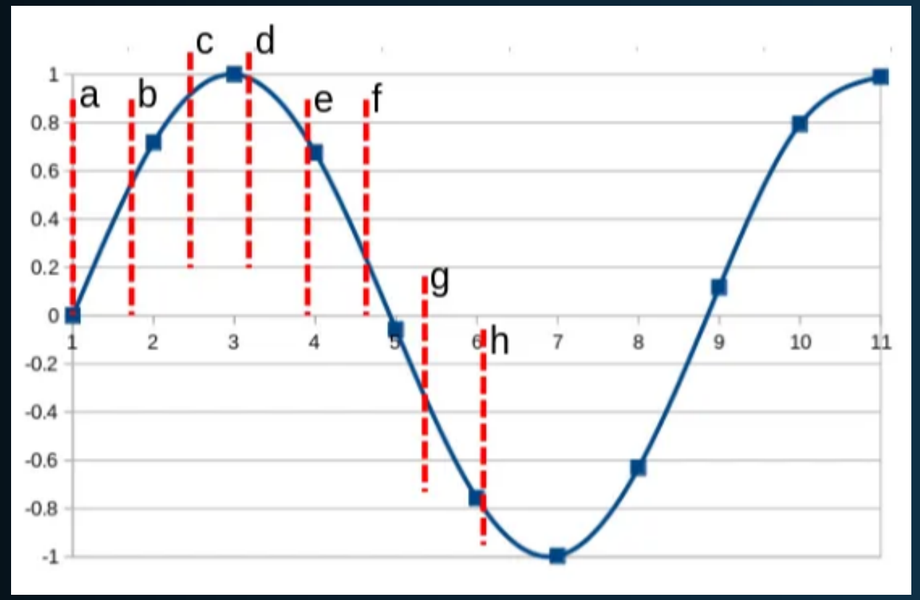
Steps taken before training:

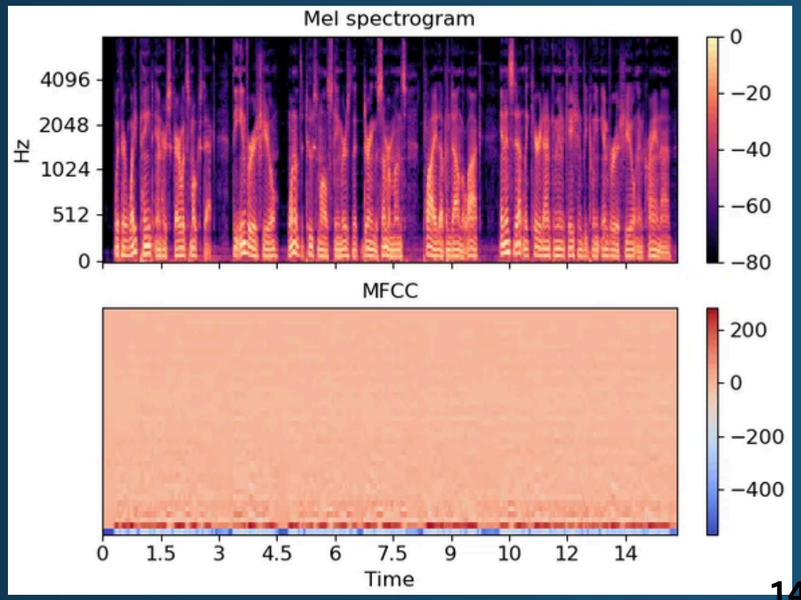
1. Preprocess data
2. Extract special frequency features
3. Create vectors from each preprocessed sample
4. Scale data
5. Construct a machine learning model

Audio Model

Feature Extraction

- Split each recording into chunks of length 1000ms.
- Resample each chunk into 22.05khz sample frequency.
- Extract MFCC, delta, delta² features.
- Create vectors for each chunk.
- Save labels [Real,Fake] for each chunk.



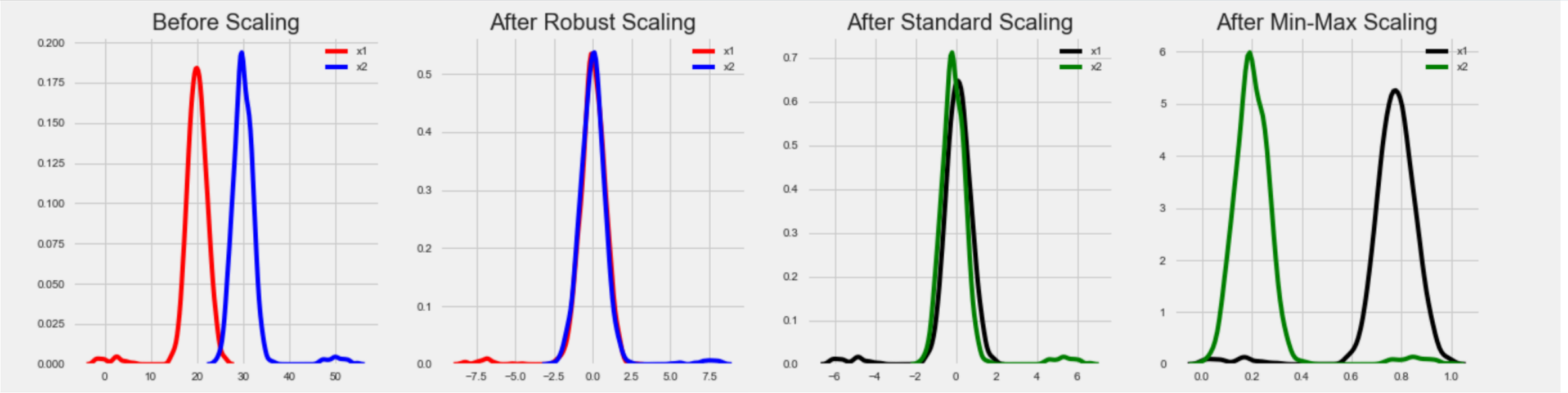


Audio Model

Rescaling the data

Ultimately, the robust scaler was chosen to normalize the data and deal with outliers.

```
# Scale the audio data using RobustScaler
scaler = RobustScaler()
train_data = scaler.fit_transform(train_data)
```



Audio Model

Datasets

In the wild dataset:

- 31,781 recordings of 58 celebrities.
- Average of 23 minutes of real recordings
- Average of 18 minutes of fake recordings

- Barack Obama - Real
- Barack Obama - Fake
- Donald trump - Real
- Donald trump - Fake

Deep Voice dataset:

- 8 real recordings of celebrities.
- 56 fake recordings of celebrities, voice swapped with a different celebrity.
- 10 minutes of speech per each recording

Fluent speech corpus dataset:

- 30,043 real recordings
- 97 speakers
- With and without background noise

- Fluent corpus - Speaker 1
- Fluent corpus - Speaker 2

Deepfake Detection Results

Video Model

Our model exceeded our initial goal of 80% accuracy. It achieved a remarkable **83% accuracy** on unseen testing data, demonstrating its effectiveness in detecting deepfakes.

Audio Model

As in the video model, the same goal of 80% was set. Our audio model has achieved an impressively high accuracy rate of approximately **99%** on the validation set.

Constraints & Challenges

Video Model

- Dataset Acquisition.
- Computationally Expensive.
- Model Architecture.

Audio Model

- Dataset availability.
- Feature extraction.

UI

- Learning new frameworks (Flask and React)

Tests

Case	Test Case	Expected Result	Results
1	Upload a video file (e.g., mp4)	The video is loaded and preprocessed, user taken into next page.	Passed
2	Upload an audio file (e.g., mp3)	The audio is loaded and preprocessed, user taken into next page.	Passed
3	Select other detection type (Audio, Video, or Audio + Video)	The system acknowledges the user's selection. Highlighting the chosen detection type.	Passed
4	Press “Detect”	The system displays a clear indication that analyzing has begun using a progress bar.	Passed
5	Press “choose other video”	Allows the user to upload a different video.	Passed
6	Upload a video/audio in an unsupported format	Error is shown: “File not supported! Please select a supported format: mp4, mkv, avi, mov, wav, mp3”	Passed
7	Upload a corrupted video/audio file	System displays an error message indicating bad upload.	Passed
8	Upload a large video/audio file (exceeding size limit)	System displays an error message indicating bad upload.	Passed
9	Upload an empty file	System displays an error message indicating bad upload.	Passed
10	Attempt to detect audio with video detection	Detection button disappears. The user cannot press detect.	Passed
11	Attempt to detect audio with video + audio detection	Detection button disappears. The user cannot press detect.	Passed
12	Attempt to detect audio from a video containing no sound	Server does not detect audio, responds with “null”. Client-side displays “Error: This video does not contain audio!”.	Passed

Manipulated Reality
Demonstration

Questions

THANK YOU