DeepFake Detection

Dataset creation, feature engineering and classification.

cheliotisnick@gmail.com

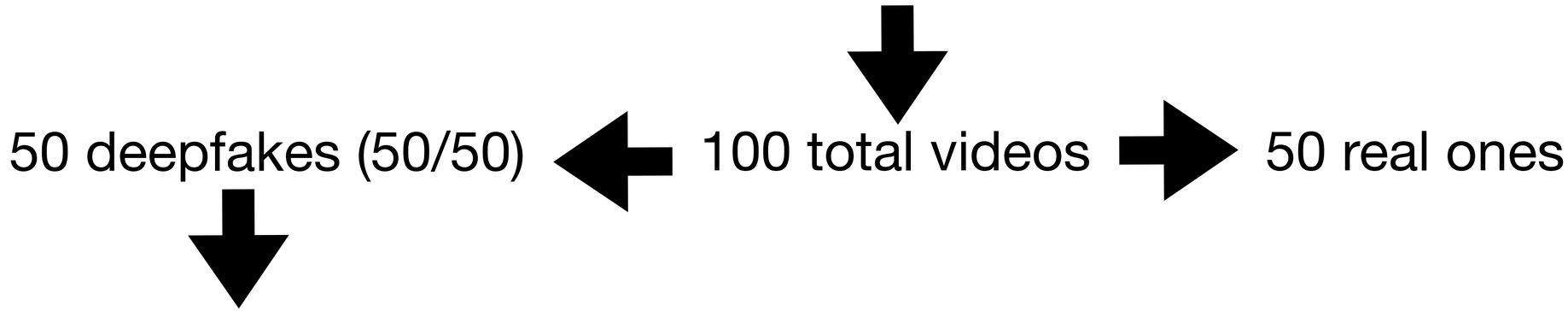
Dataset creation

Dataset creation

Video collection

We collected short, face-focused videos by searching YouTube Shorts.To ensure quality each video had to meet specific criteria:

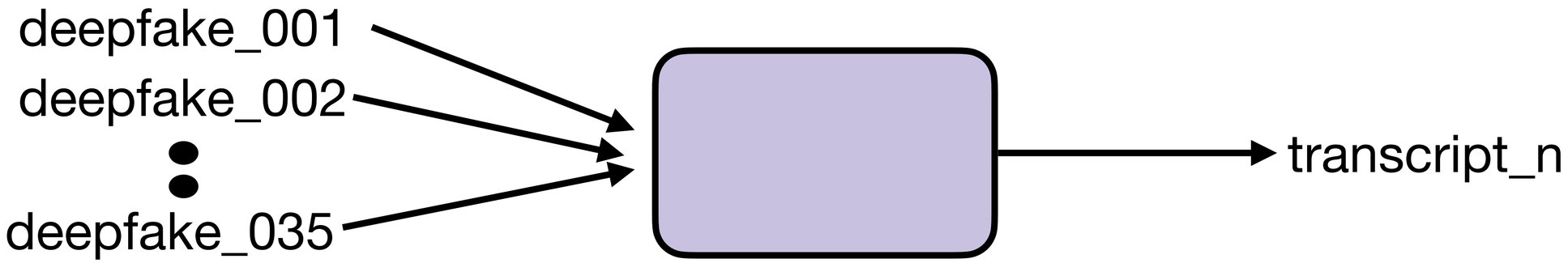
1. Only one person
2. No background music
3. No face occlusions
4. No sudden camera/angle changes
5. Exactly 10 seconds



Dataset creation

Deepfake Audio Generation

We used OpenAi's **Whisper*** model to convert audio into text.



We used ElevenLabs** to generate DeepFake voices.A variety of synthetic voice profiles were used to simulate different speech patterns.

*https://github.com/openai/whisper

**https://elevenlabs.io/

Dataset creation

Deepfake audio insertion

We used **Wav2Lip*** to sync the deepfake voices with the speaker's lip movements

*https://github.com/Rudrabha/Wav2Lip

Dataset creation

DeepFake video generation

We collected 35 unique female/ male faces.

We used **online face-swapping models** to generate deepfake videos by replacing the original face with synthetic ones.

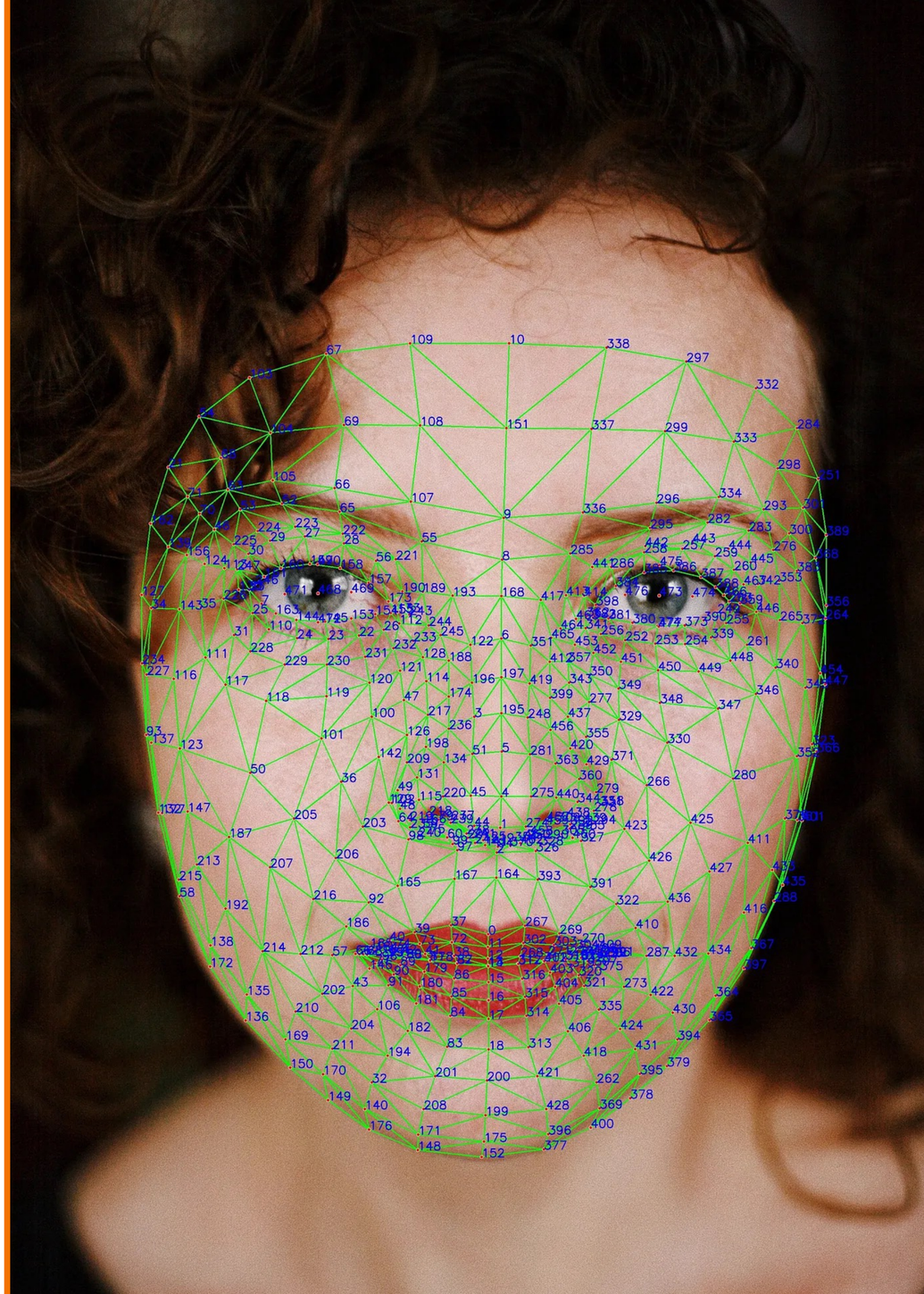
Feature engineering

Feature engineering

Visual features

To extract visual features we used MediaPipe, a library that detects facial landmarks.From these landmarks we applied custom logic to compute behavioural metrics.

We calculated **blink rate**, **mouth open ratio**, **head motion** (yaw and pitch) and **expression entropy** (using **DeepLab**)



Feature engineering

Visual features - Blink rate

1. Input frame
2. Calculate Eye Aspect Ratio (EAR)
3. If EAR<0.21 => potential blink
4. Get next frame
5. If EAR goes back up next frame => blink
6. Divide Blinks/seconds

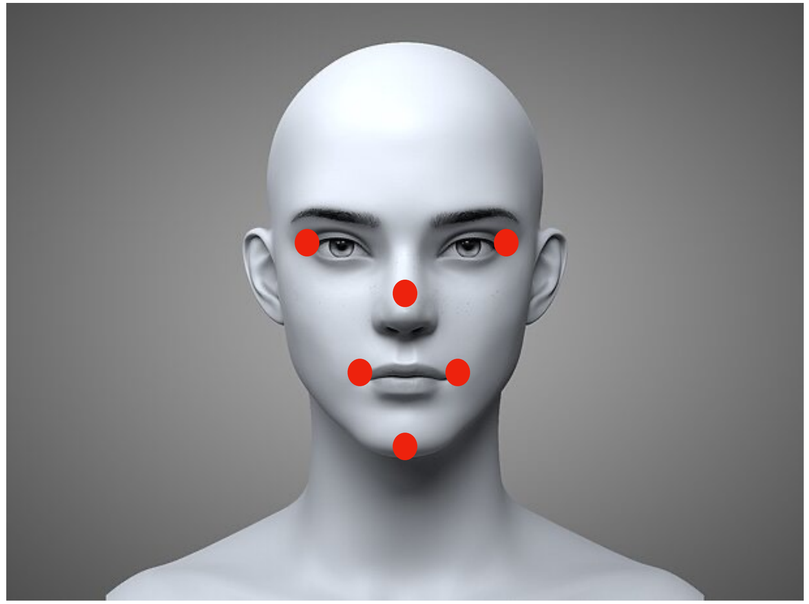
Feature engineering

Visual features - Mouth open ratio

1. Input frame
2. Calculate Mouth Open Ratio (MOR)
3. Store it
4. Calculate the mean (or variance).

Feature engineering

Visual features - Head motion (left-right and up-down)



3D Face

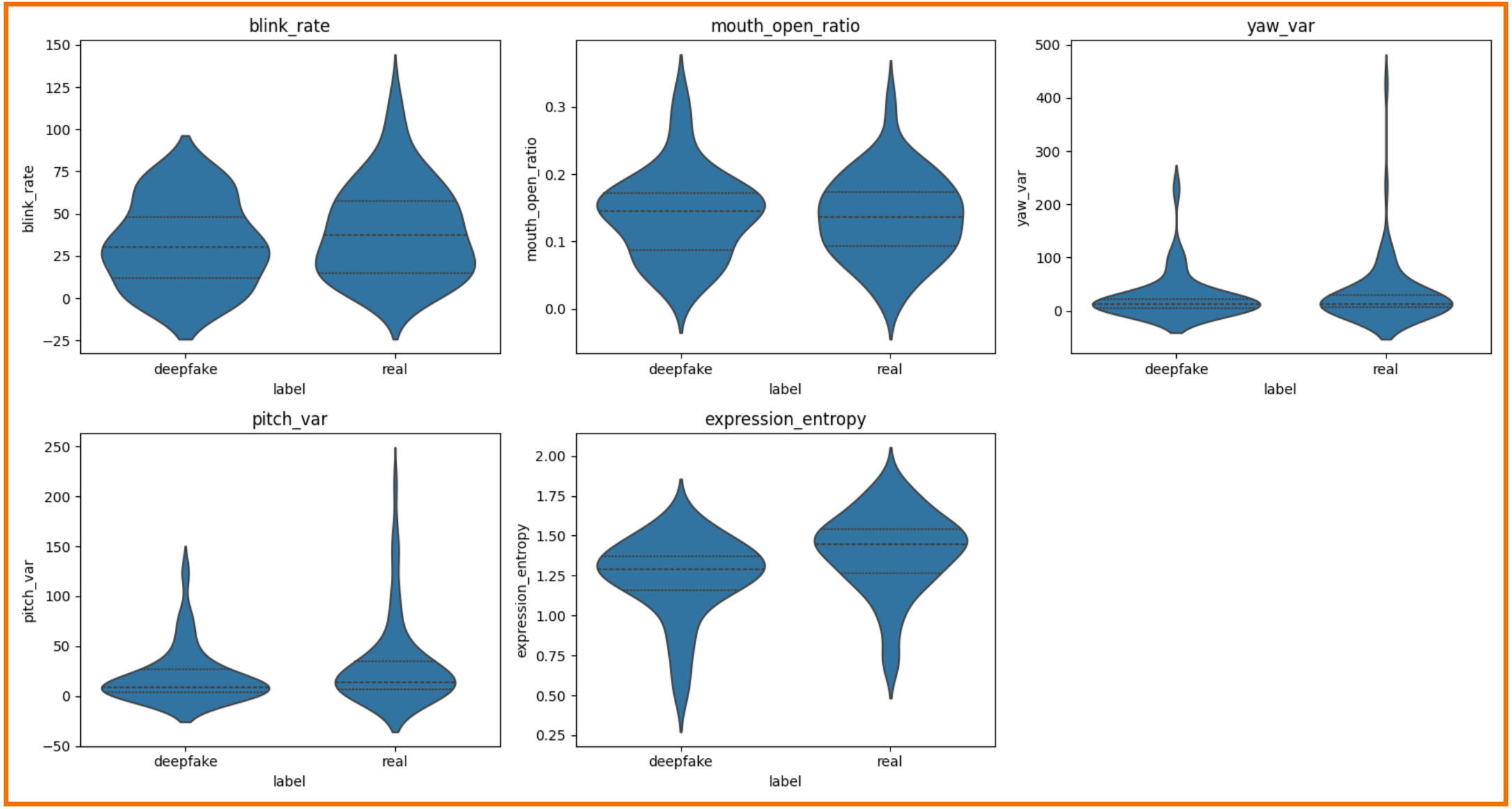
2D Frame

1.Rotate (yaw and pitch) the 3D face so it matches 2D frame.

2.Calculate the variance of yaw and pitch across all frames.

Feature engineering

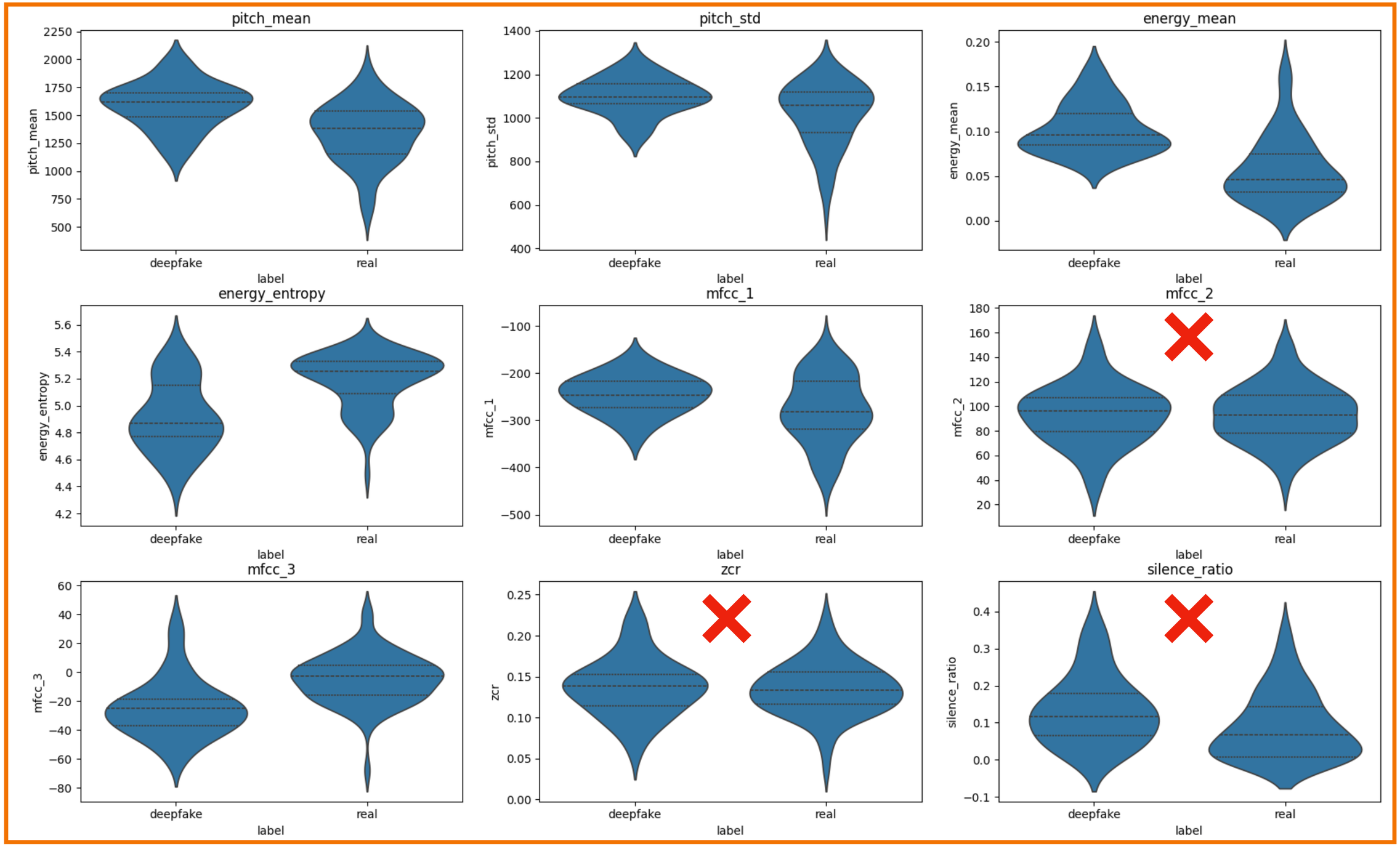
Visual features - Visual feature plots



Feature engineering

Audio features

We used Librosa to extract **pitch** (mean and std) ,**energy** (mean and entropy), **MFCC** (first 3), **zero cross rate** and **silence ratio**.


pitch_mean
pitch_mean
2250
2000
1750
1500
1250
1000
750
500
deepfake
real
label
pitch_std
pitch_std
1400
1200
1000
800
600
400
deepfake
real
label
energy_mean
energy_mean
0.20
0.15
0.10
0.05
0.00
deepfake
real
label
energy_entropy
energy_entropy
5.6
5.4
5.2
5.0
4.8
4.6
4.4
4.2
deepfake
real
label
mfcc_1
mfcc_1
−100
−200
−300
−400
−500
deepfake
real
label
mfcc_2
mfcc_2
180
160
140
120
100
80
60
40
20
deepfake
real
label
mfcc_3
mfcc_3
60
40
20
0
−20
−40
−60
−80
deepfake
real
label
zcr
zcr
0.25
0.20
0.15
0.10
0.05
0.00
deepfake
real
label
silence_ratio
silence_ratio
0.4
0.3
0.2
0.1
0.0
−0.1
deepfake
real
label

Model training

Model training

Models

We trained and evaluated two classification models on our Dataset.Svm and Logistic Regression.Each model was tasked with a binary classification problem, predict whether a video is real or deepfake.

SVM

	Precision	Recall	F1-score	Support
Real	**0.88**	0.64	0.74	11
Deepfake	0.67	**0.89**	0.76	9

Accuracy:75%

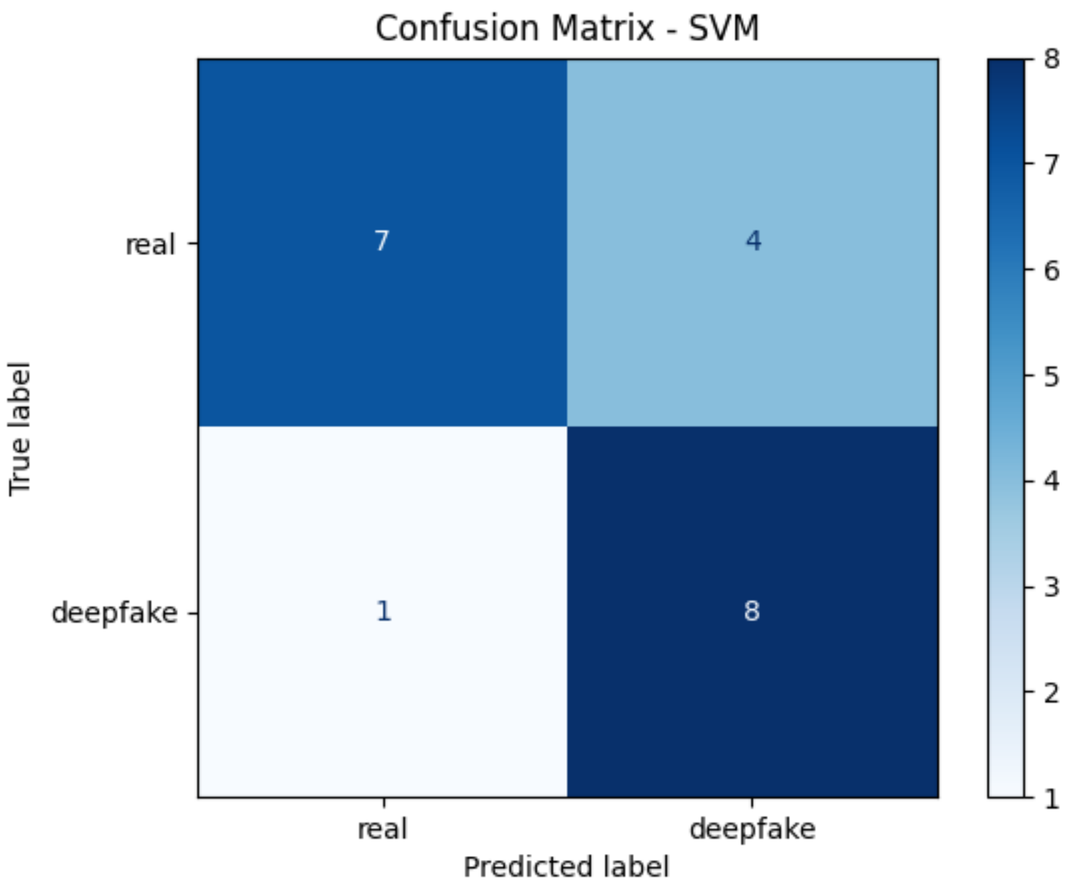
Logistic Regression

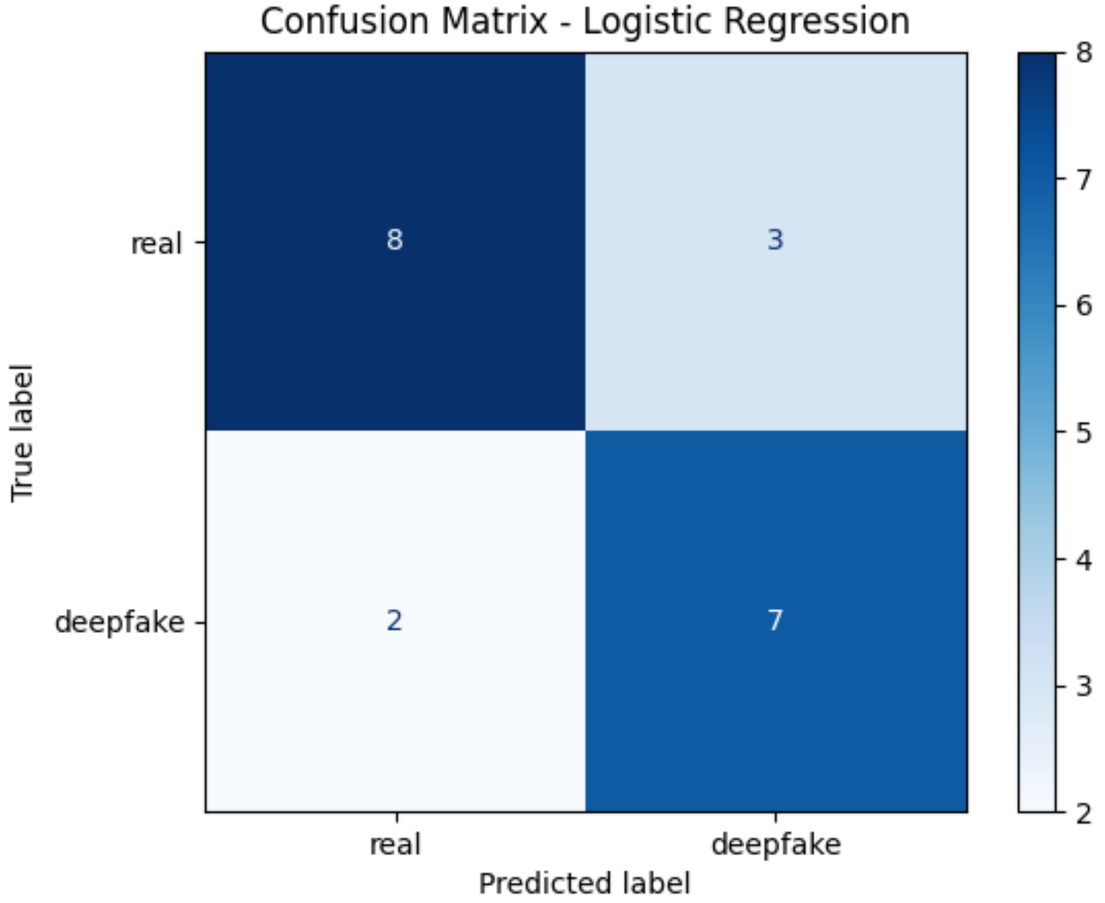
	Precision	Recall	F1-score	Support
Real	0.80	**0.73**	0.76	11
Deepfake	**0.70**	0.78	0.74	9

Accuracy:75%

Model training

Graphs

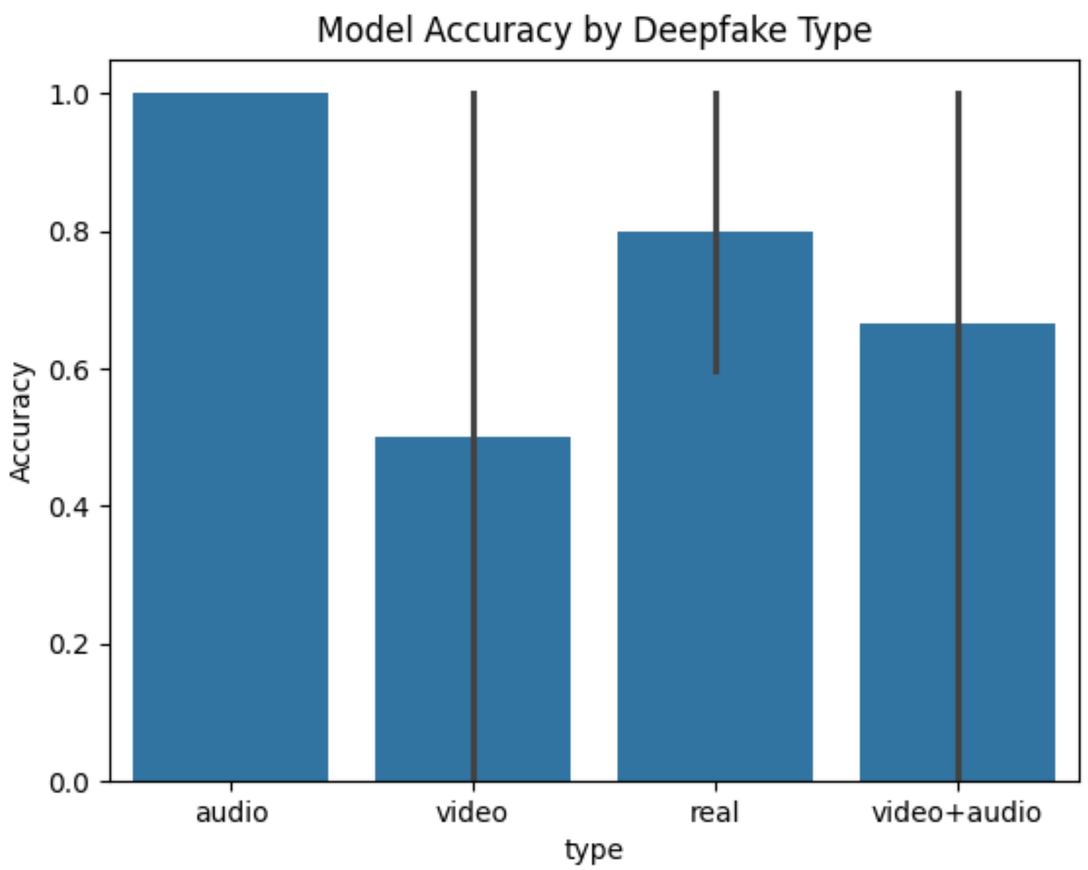




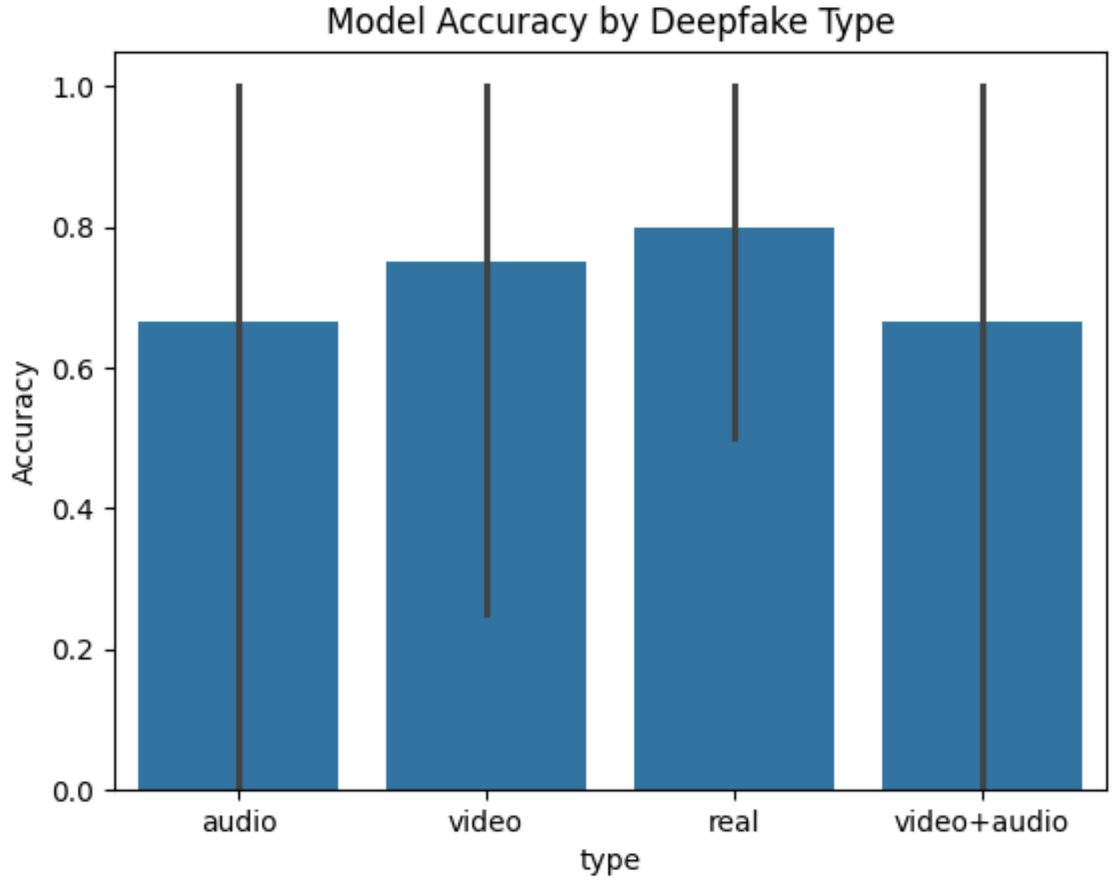
Model training

Graphs

SVM



Logistic Regression



How could we do better?

1. More data (quantity and diversity) to improve generalisation.
2. Better visual features.
3. Integrate cross modal consistency features, such as verifying whether mouth movements are synchronized with the spoken words.

Thank you!