Department of Electrical & Electronic Engineering

Bangladesh University of Engineering and Technology

Dhaka, Bangladesh

EEE 6608: Machine Learning & Pattern Recognition

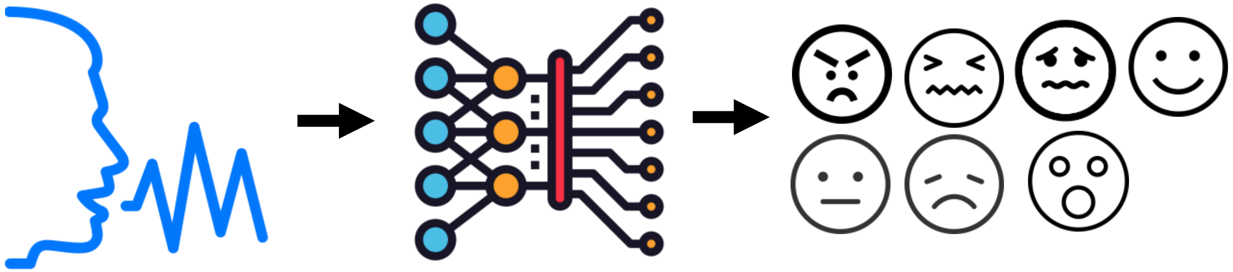
Speech Emotion Recognition Using 1D CNN

Name: Md. Intikhab Shahriar Hasan
Roll No.: 0424062503
Session: April 2024

Speech Emotion Recognition (SER)

What's Emotion Recognition?

- **Emotion Recognition** is the ability to precisely infer human emotions
- Utilizing facial expressions, body language, speech patterns, and text.



Emotion Recognition using speech signals

Why it's important?

Human-Computer Interaction (HCI) & Robotics

→ More intuitive and responsive to user emotions

Mental Health and Well-being

→ Detecting signs of stress, anxiety, or depression, enabling timely interventions

Security & Surveillance

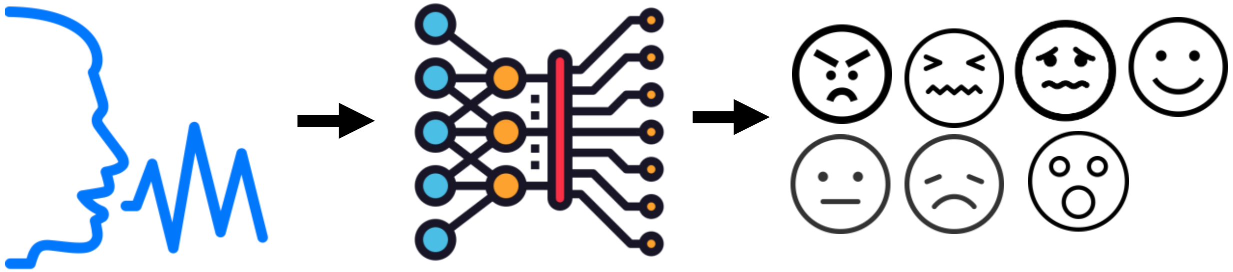
→ Unusual or aggressive behaviors from emotional cues

Marketing and Customer Experience

→Gauge consumer reactions to products, services,

Speech Emotion Recognition (SER)

Advantages of Emotion Recognition from speech signals:



Emotion Recognition using speech signals

Non-Intrusive and Privacy-Friendly

→ Only audio input is required

→ Cameras capture visual data, intruding on privacy

Effective in Low-Visibility Conditions

→ Works in low-light, dark, or visually obstructed environments

→ Suitable for phone calls or virtual meetings without video

Unobstructed by Physical Appearance or Expression Limitations

→ Can detect emotions even if a person's face is obscured, hidden, or masked, eg. in online meetings or while wearing face masks

→ Can detect emotions that are not strongly expressed in facial features, eg. sounds angry but the facial expression is neutral

Works in Unstructured and Natural Conversations

→ Text-based approaches struggles with sarcasm, irony, or subtle emotions that aren't conveyed in words

→ SER works while emotions are conveyed through tone, pitch, and vocal intonation, even if the words themselves are neutral.

Emotional Speech Dataset

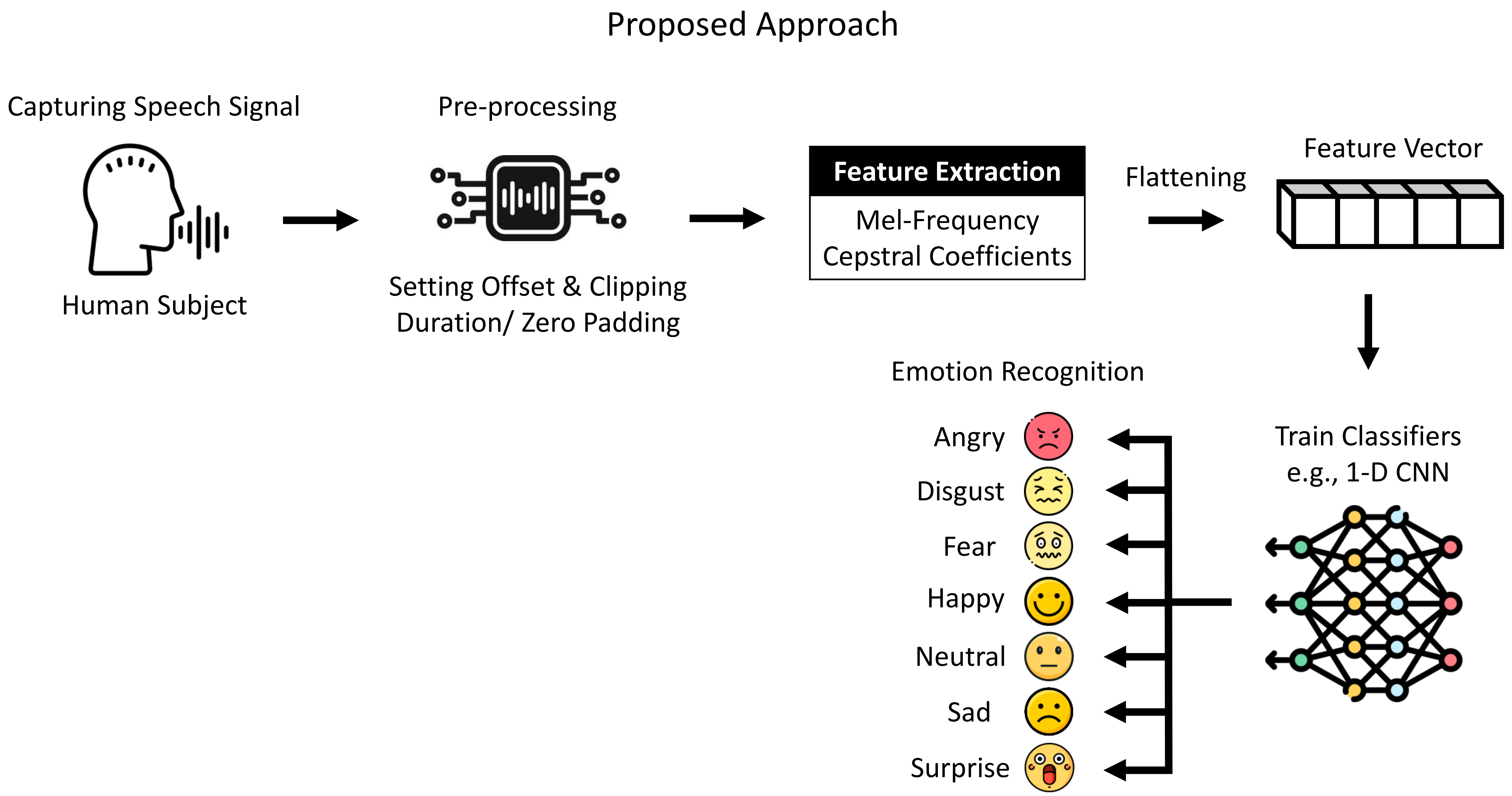
Toronto Emotional Speech Set (TESS)

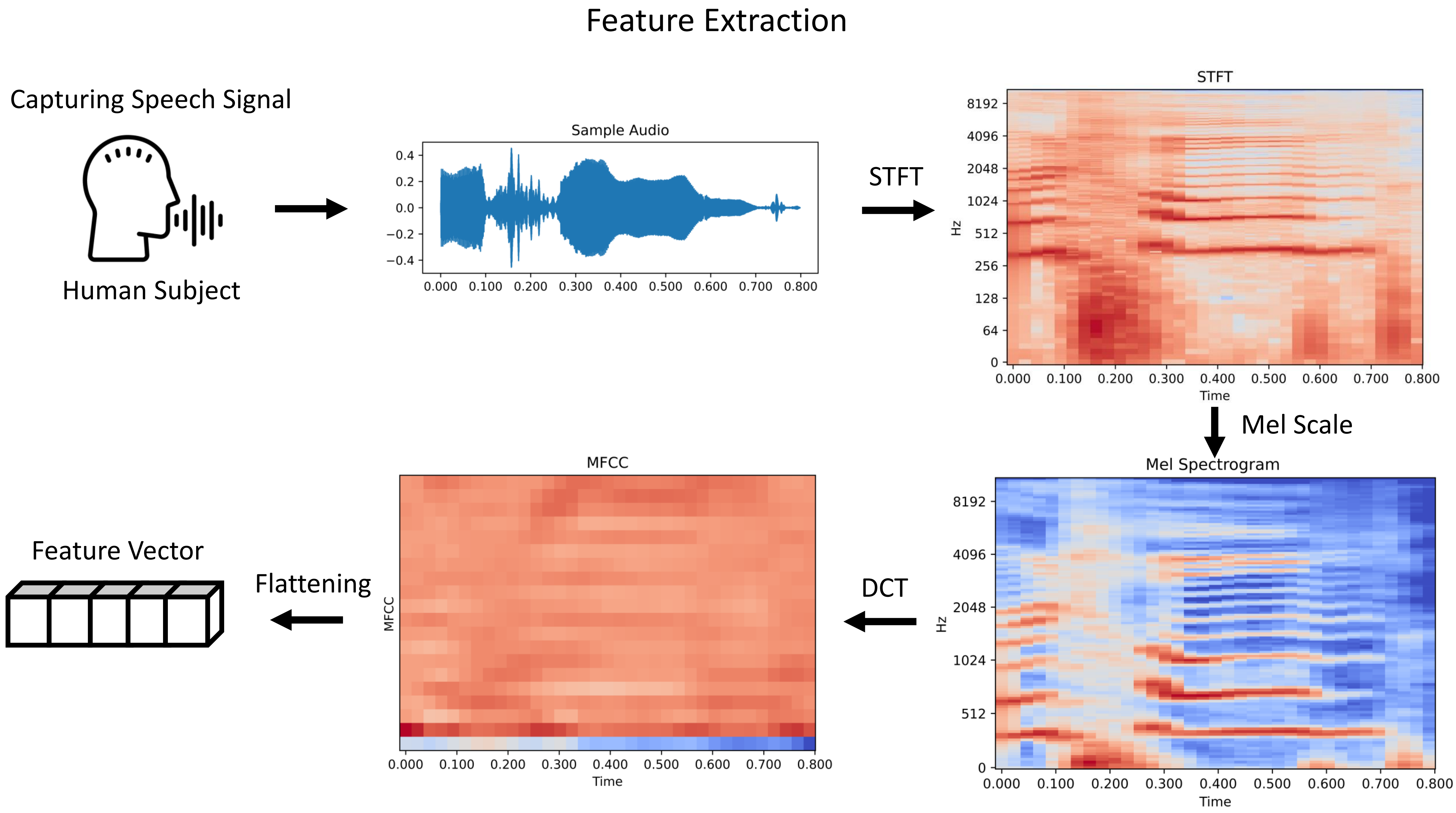
→ 2800 speech samples
→ 2 actors, 1400 samples per speaker
→ 7 different emotion classes
→ 400 sample audios for each of the classes

Prior Works

References & Year	Features Extracted	Architecture Used	Performance
[6], 2024	Multiple Time & Freq domain Features	Ensembling A (CNNs), B (BiLSTM-FCN), C (BiLSTM-FCN with transformer) Networks	99.857 %
[7], 2023	MFCC Spectrogram	CNN+LSTM+Attention	99.81 %

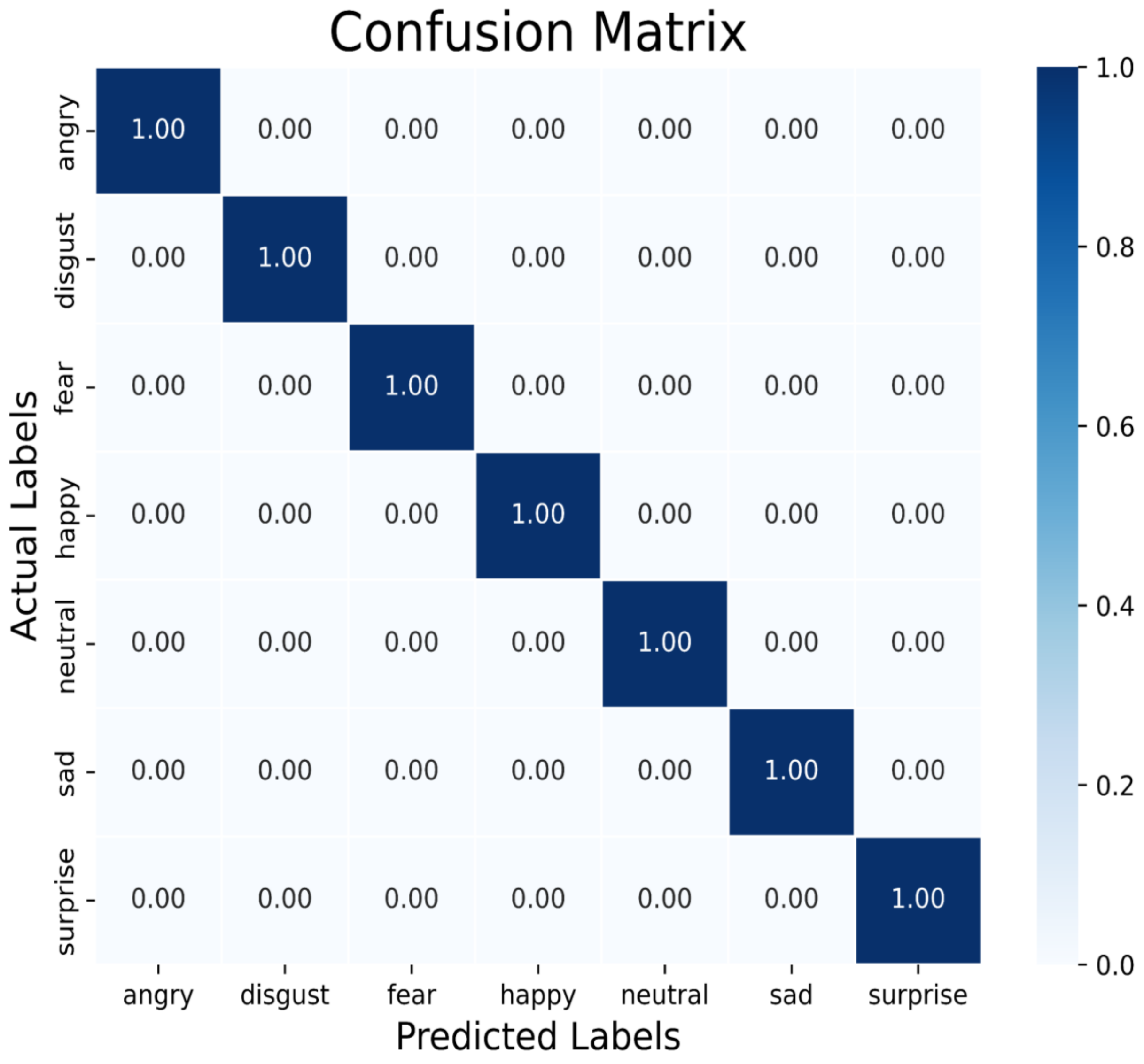
[6] Mengsheng Wang, Hongbin Ma, Yingli Wang, Xianhe Sun, Design of smart home system speech emotion recognition model based on ensemble deep learning and feature fusion. https://doi.org/10.1016/j.apacoust.2024.109886
[7] Singh J, Saheer LB, Faust O. Speech Emotion Recognition Using Attention Model. *International Journal of Environmental Research and Public Health*. 2023; 20(6):5140. https://doi.org/10.3390/ijerph20065140


Proposed Approach
Capturing Speech Signal
Human Subject
Pre-processing
Setting Offset & Clipping
Duration/ Zero Padding
Feature Extraction
Mel-Frequency
Cepstral Coefficients
Flattening
Feature Vector
Train Classifiers
e.g., 1-D CNN
Emotion Recognition
Angry
Disgust
Fear
Happy
Neutral
Sad
Surprise


Feature Extraction
Capturing Speech Signal
Human Subject
Sample Audio
0.4
0.2
0.0
−0.2
−0.4
0.000 0.100 0.200 0.300 0.400 0.500 0.600 0.700 0.800
STFT
STFT
8192
4096
2048
1024
512
256
128
64
0
Hz
0.000 0.100 0.200 0.300 0.400 0.500 0.600 0.700 0.800
Time
Mel Scale
Mel Spectrogram
8192
4096
2048
1024
512
0
Hz
0.000 0.100 0.200 0.300 0.400 0.500 0.600 0.700 0.800
Time
DCT
MFCC
MFCC
0.000 0.100 0.200 0.300 0.400 0.500 0.600 0.700 0.800
Time
Flattening
Feature Vector

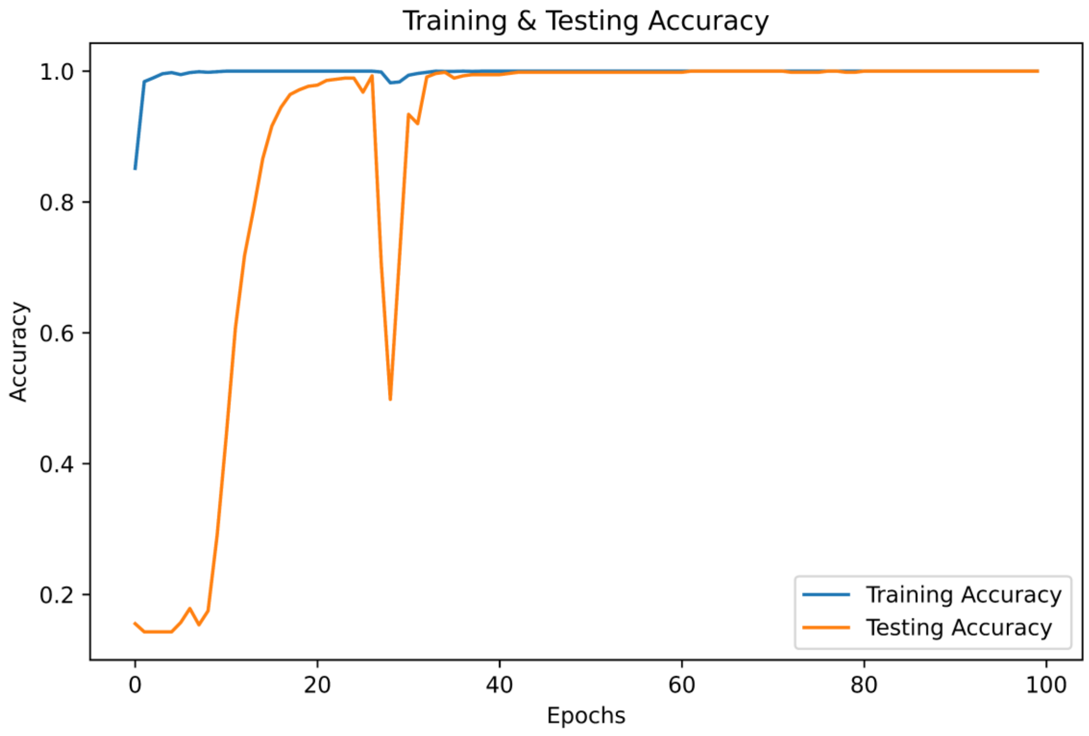
Performance Analysis

Classifier	Result (%)
Decision Tree	78.39
Random Forest	97.68
SVM	98.21
XGBoost	96.96
1D CNN	100

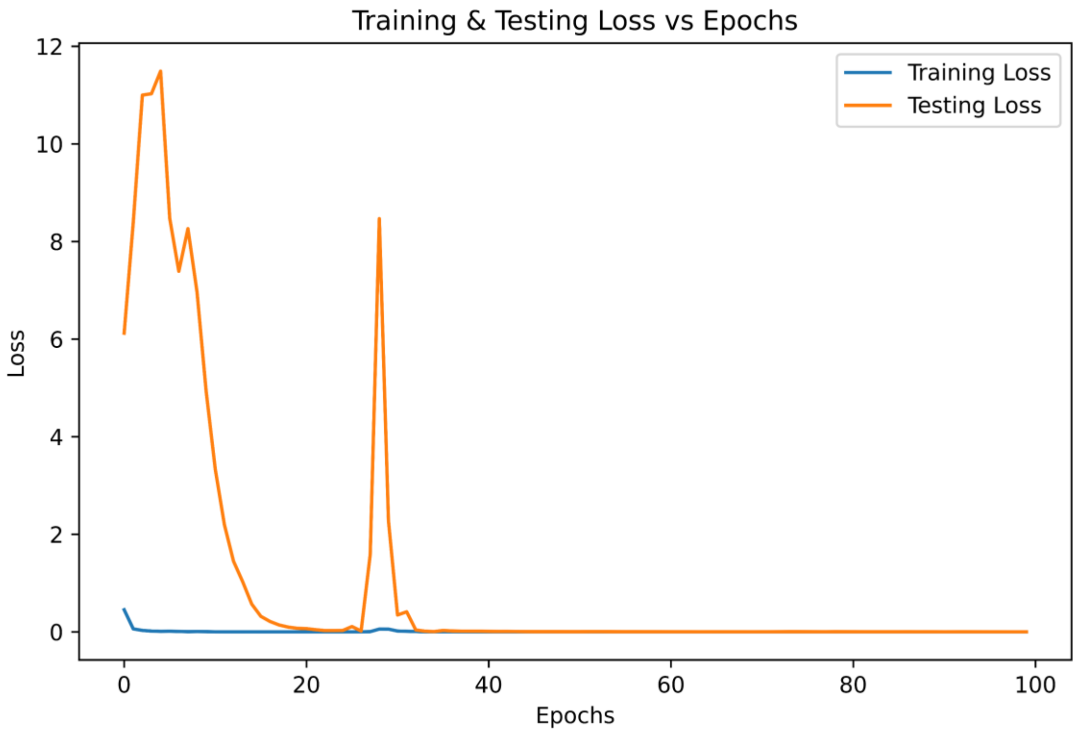


Results obtained using 1D CNN

Performance Analysis



Training & Testing Accuracy vs Epochs



Training & Testing Loss vs Epochs

Performance Analysis

References & Year	Features Extracted	Architecture Used	Performance
[6], 2024	Multiple Time & Freq domain Features	Ensembling A (CNNs), B (BiLSTM-FCN), C (BiLSTM-FCN with transformer)	99.857 %
[7], 2023	MFCC Spectrogram	CNN+LSTM+Attention	99.81 %
This Work	MFCC (Flattened)	1D CNN	100 %

1D CNN
Conv1d (filters=128, kernel_size=5)
BN, Relu, MaxPool
Conv1d (filters=128, kernel_size=5)
BN, Relu, MaxPool, *Dropout (0.2)*
Conv1d (filters=256, kernel_size=5)
BN, Relu, MaxPool
Conv1d (filters=256, kernel_size=3)
BN, Relu, MaxPool, *Dropout (0.2)*
Conv1d (filters=256, kernel_size=3)
BN, Relu, MaxPool
Conv1d (filters=512, kernel_size=3)
BN, Relu, MaxPool, *Dropout (0.2)*
Flatten, Dense (256), BN, Dense (7, softmax)

1D CNN Architecture