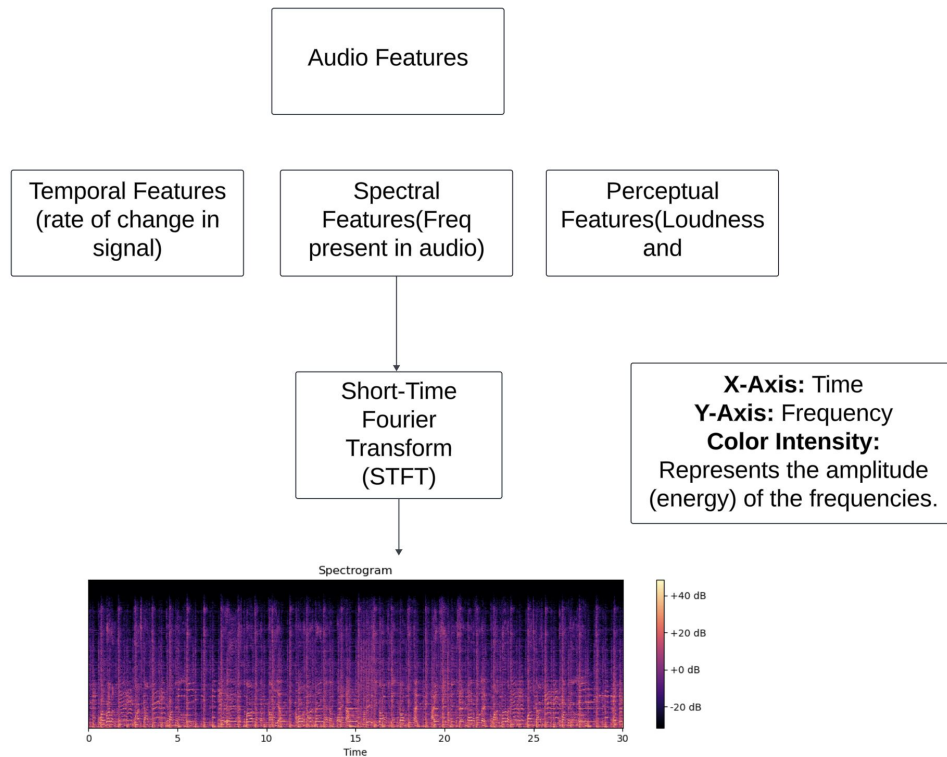# Music Genre Classification: comparing impact of feature extraction on the model prediction

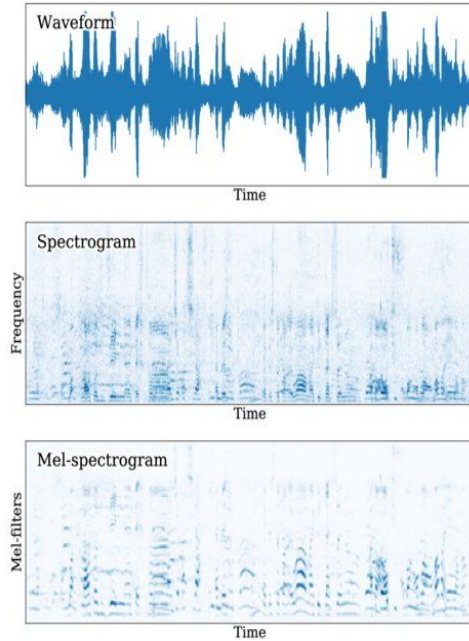Sulagna Saha

Deep Learning by Prof Bálint Gyires Tóth - AIT Budapest - Spring 2024
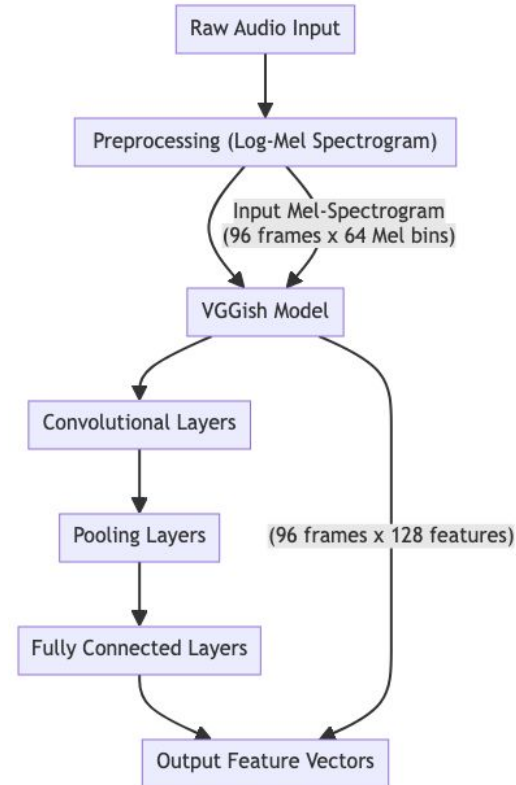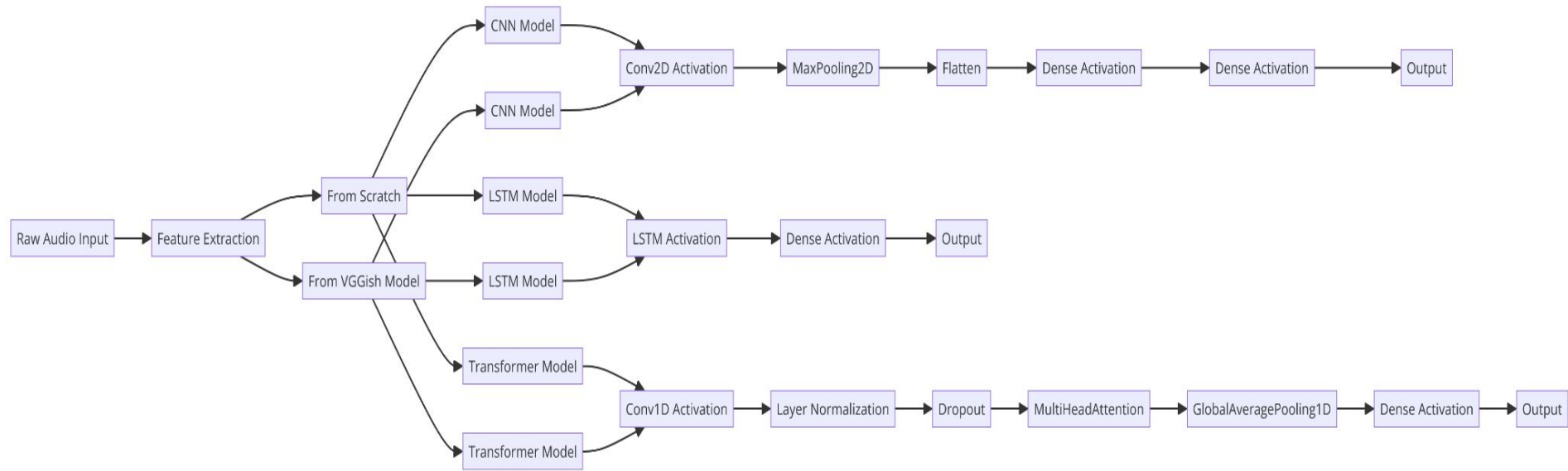
Features from audio(Scratch)

Audio Features

Temporal Features (rate of change in signal)

Spectral Features(Freq present in audio)

Perceptual Features(Loudness and

Short-Time Fourier Transform (STFT)

X-Axis: Time
Y-Axis: Frequency
Color Intensity: Represents the amplitude (energy) of the frequencies.

Spectrogram

+40 dB
+20 dB
+0 dB
-20 dB

0   5   10   15   20   25   30
Time

VGGish Model(pre-trained)

Waveform

Spectrogram

Mel-spectrogram

[Source](#)

Raw Audio Input

Preprocessing (Log-Mel Spectrogram)

Input Mel-Spectrogram
(96 frames x 64 Mel bins)

VGGish Model

Convolutional Layers

Pooling Layers   (96 frames x 128 features)

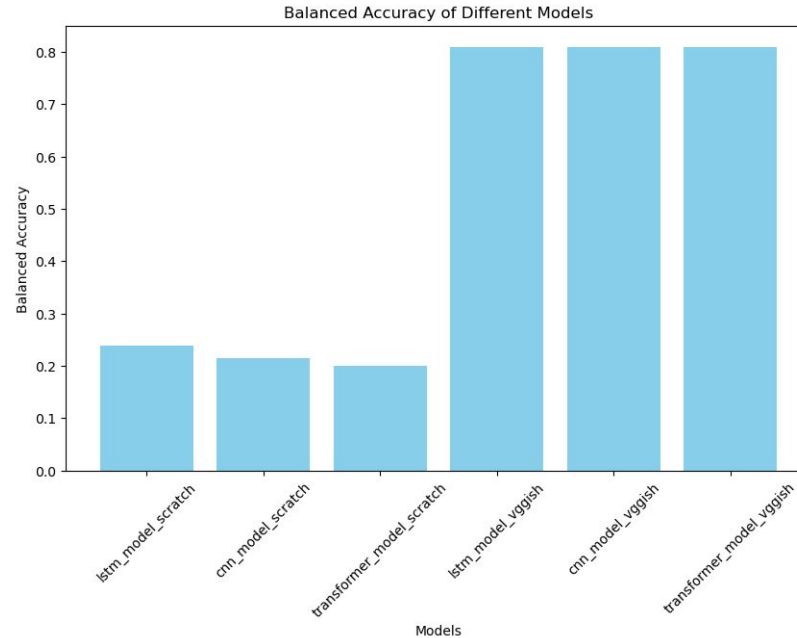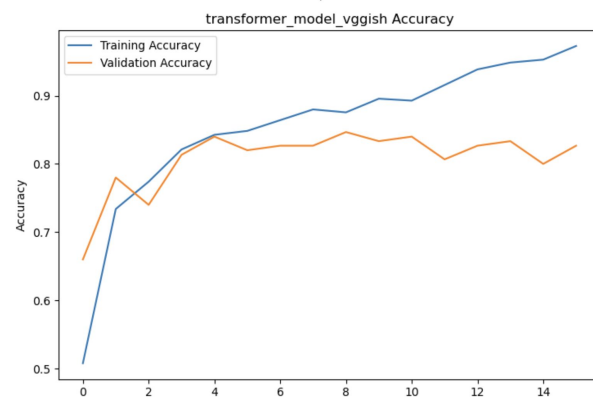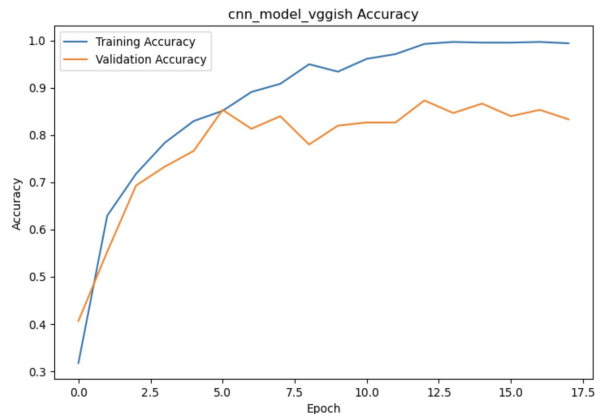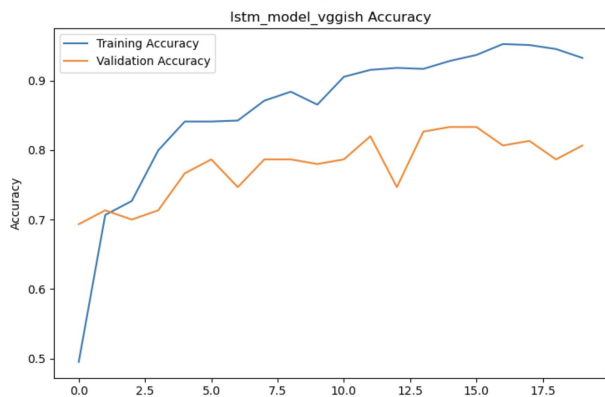Fully Connected Layers

Output Feature Vectors

Proposed Method

Results

Features from VGGish Model >> Features from Spectrogram

Best model: transformer_model_vggish

More stable and less overfitting compared to LSTM and CNN
Val accuracy ~ 85% which is more than the CNN model suggested earlier ~ 45%