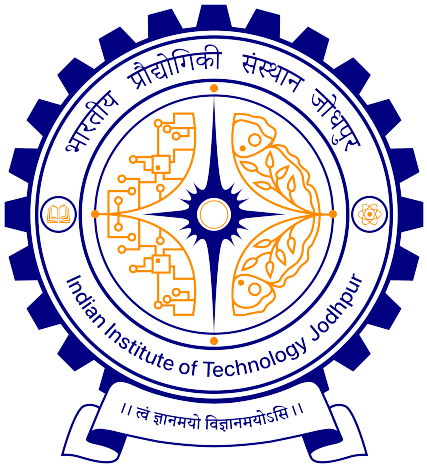
Speech Understanding: Assignment 1

Spectrograms and Windowing Techniques



K. K. N. Shyam Sathvik

[1]Indian Institute of Technology, Jodhpur, Jodhpur 342037, India
b22ee036@iitj.ac.in
February 2, 2025

Abstract

This report presents the work undertaken by Shyam Sathvik for Assignment 1 in the Speech Understanding course. The methodology, analysis, and results pertain to Question 2 of the assignment. The first part focuses on analyzing the UrbanSound8k dataset, exploring various windowing techniques for the Short-Time Fourier Transform (STFT) to generate spectrograms, and leveraging these spectrogram features to train a neural network for a classification task. The second part examines spectrogram differences across music genres, delving into their unique acoustic features and characteristics. Additionally, the repository for the work can be found here.

Contents

1 Task A: Training a Neural Network on Spectrogram Features

This report focuses on the analysis of the UrbanSound8k dataset, which contains audio samples from 10 different urban sound classes. The primary objective is to understand and implement various windowing techniques, generate spectrograms using the Short-Time Fourier Transform (STFT), and train a simple classifier using features extracted from these spectrograms. The report also includes a comparative analysis of spectrograms generated using different windowing techniques and their impact on classification performance.

2 Methodology

This section focuses on the methodology used for experimenting with spectrograms and different windowing techniques on the UrbanSound8K dataset.

2.1 Dataset

The UrbanSound8k dataset contains audio recordings of urban sounds categorized into different classes. The dataset was preprocessed using the librosa library by loading audio samples at a sampling rate of 22050Hz, ensuring a uniform duration of 3 seconds.

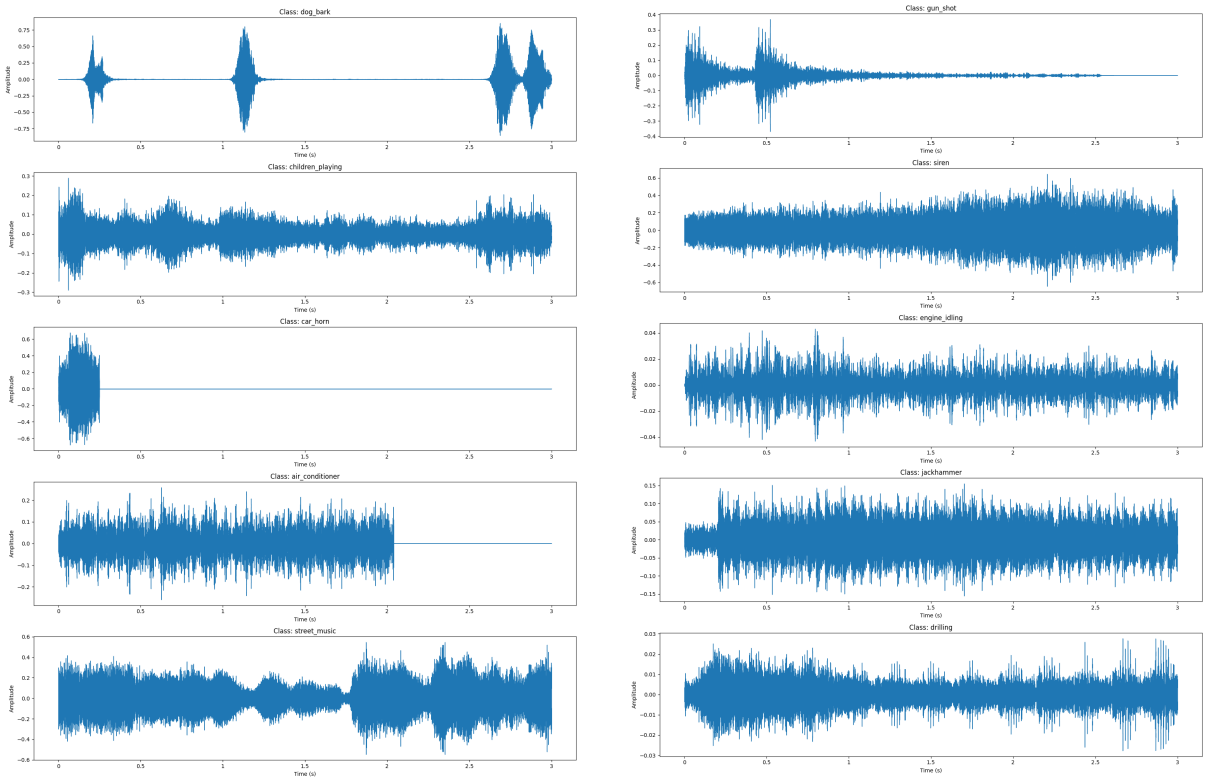


Figure 1: Waveform represntation of different classes present in the UrbanSound8k dataset

2.2 Windowing Techniques

Windowing methods are essential in signal processing for minimizing spectral leakage during time-frequency analysis using the Short-Time Fourier Transform (STFT). Spectral leakage arises from abrupt discontinuities at the boundaries of a finite signal segment, causing frequencies to spread into neighboring bins. To address this, various window functions are applied to the signal before transforming it into the frequency domain. Among the most commonly used windows are the Rectangular, Hann, and Hamming windows, each offering unique benefits and compromises.

2.2.1 Rectangular Window

The Rectangular window is the most straightforward window function, assigning equal weight to all samples within the window. Its mathematical representation is:

$$w(n) = 1, \quad 0 \leq n \leq N-1 \tag{1}$$

where N N is the total number of samples in the window. This window preserves the signal's original amplitude without any modification, making it ideal for scenarios where signal integrity is paramount. However, its sharp transitions at the edges introduce substantial spectral leakage, causing high-frequency components to spread across the spectrum. Consequently, it is generally not suitable for applications demanding precise frequency resolution.

2.2.2 Hann Window

The Hann window, also referred to as the Hanning window, employs a smooth tapering function that gradually reduces the amplitude at the edges, effectively minimizing spectral leakage. Its mathematical formulation is:

$$w(n) = 0.5\left(1 - \cos\left(\frac{2\pi n}{N-1}\right)\right), \quad 0 \leq n \leq N-1 \tag{2}$$

This window significantly reduces spectral leakage compared to the Rectangular window, making it a preferred choice for frequency analysis. However, the tapering effect slightly diminishes the signal's amplitude, potentially leading to a loss of information. Additionally, its broader main lobe results in reduced frequency resolution compared to the Rectangular window. Despite these drawbacks, the Hann window is widely used in signal processing due to its balanced trade-off between leakage suppression and resolution.

2.2.3 Hamming Window

The Hamming window is an enhanced version of the Hann window, designed to further reduce spectral leakage while better preserving the signal's amplitude. It is defined as:

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \leq n \leq N-1 \tag{3}$$

Compared to the Hann window, the Hamming window offers improved leakage suppression and better amplitude retention. However, this comes at the expense of a slightly wider main lobe, which marginally reduces frequency resolution. Due to its effectiveness in minimizing leakage while maintaining signal strength, the Hamming window is frequently employed in speech and audio processing, where preserving signal fidelity is critical.

2.2.4 Comparison and Practical Use

The selection of a windowing method depends on the specific needs of the application:

- The **Rectangular window** is advantageous when maintaining the original signal amplitude is crucial, though it suffers from significant spectral leakage.
- The **Hann window** strikes a balance between frequency resolution and leakage reduction, making it suitable for general spectral analysis tasks.
- The **Hamming window** excels in minimizing spectral leakage while preserving signal amplitude, making it ideal for applications such as speech and audio processing.

By carefully choosing an appropriate window function, one can enhance the accuracy of frequency domain representations and optimize signal analysis. The accompanying figure demonstrates the impact of these windowing techniques on an audio sample, highlighting their effects on spectral leakage and frequency resolution.

2.3 Spectrogram Generation

To analyze the frequency content of audio signals, we employed the Short-Time Fourier Transform (STFT) with different windowing techniques. The STFT was applied to each audio sample, and the resulting spectrograms were converted to a decibel (dB) scale using logarithmic scaling to enhance visualization. All spectrograms were resized to a uniform dimension of 128x128 pixels.

2.4 Spectrograms for Each Class

The UrbanSound8k dataset comprises various audio classes, each representing distinct environmental sounds. Below, we present the spectrograms for each class, generated using the STFT with a Hann window. These spectrograms provide a detailed frequency-domain representation of the audio signals, enabling us to observe the unique spectral patterns associated with each sound class.

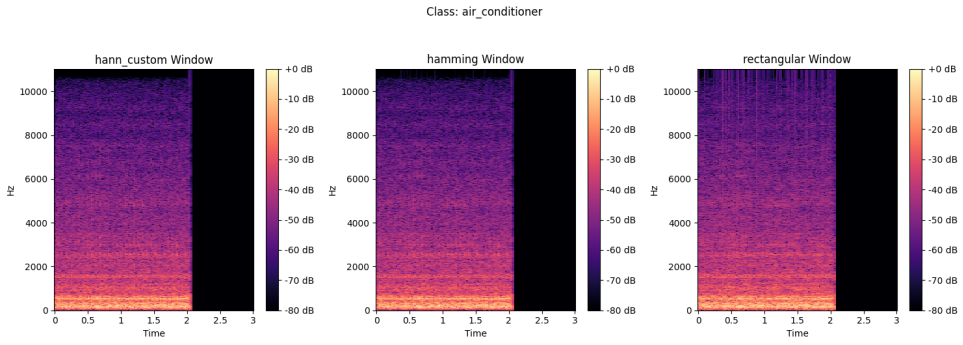


Figure 2: Spectrogram for the air conditioner class.

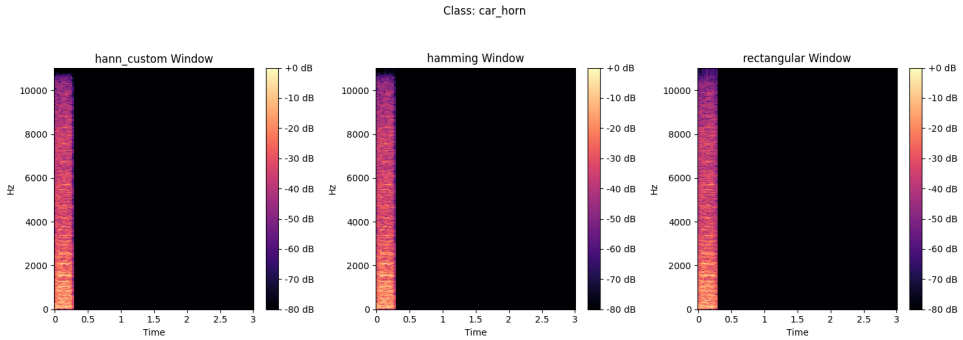


Figure 3: Spectrogram for the car horn class.

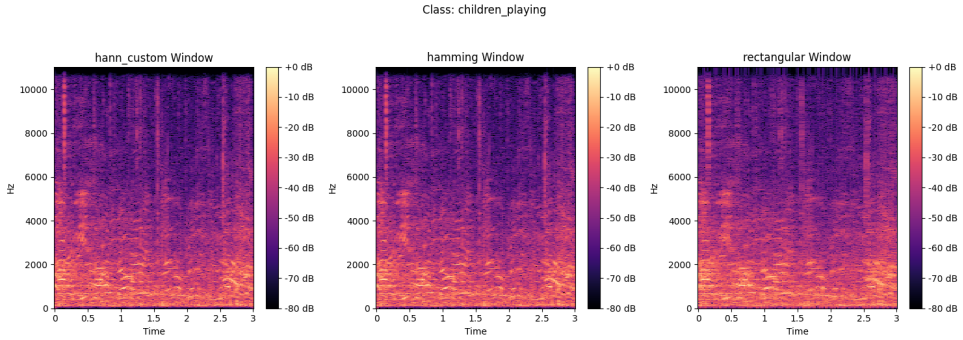


Figure 4: Spectrogram for the children playing class.

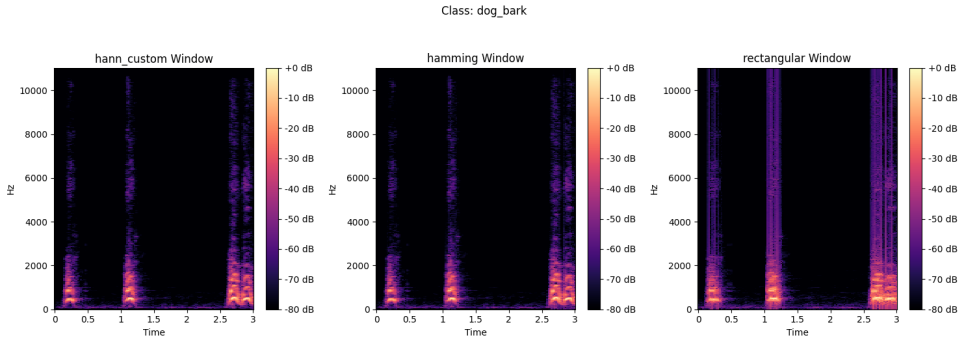


Figure 5: Spectrogram for the dog bark class.

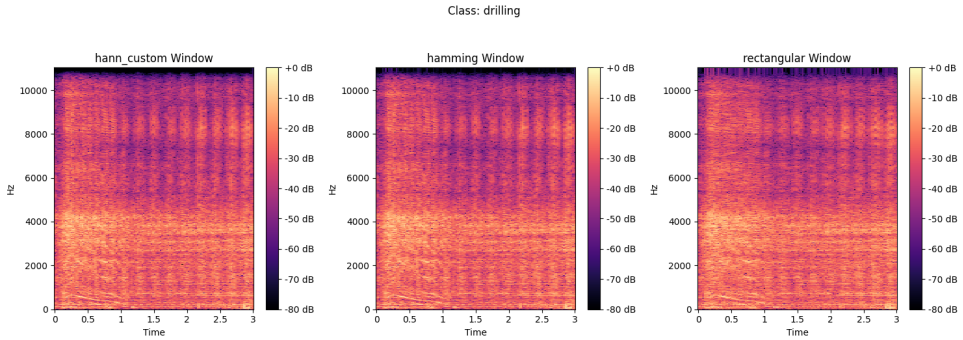


Figure 6: Spectrogram for the drilling class.

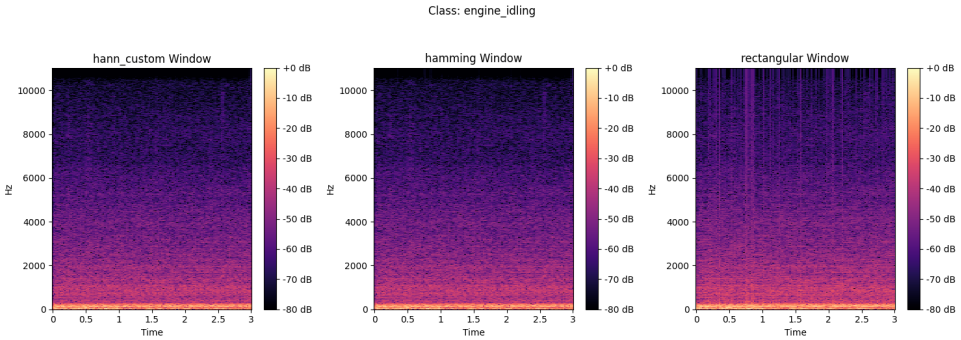


Figure 7: Spectrogram for the engine idling class.

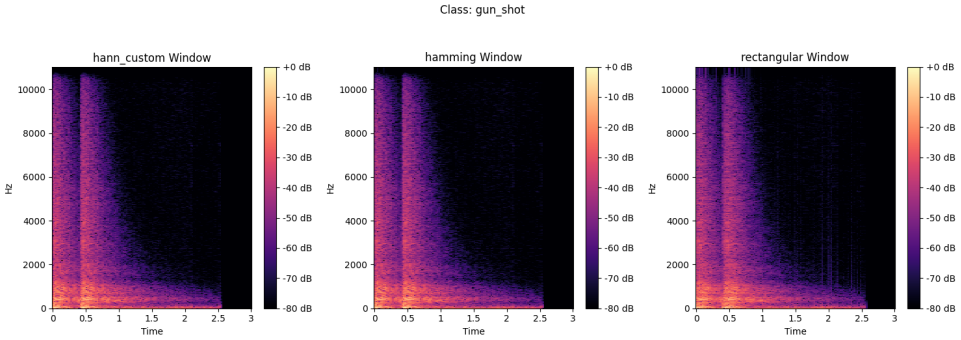


Figure 8: Spectrogram for the gunshot class.

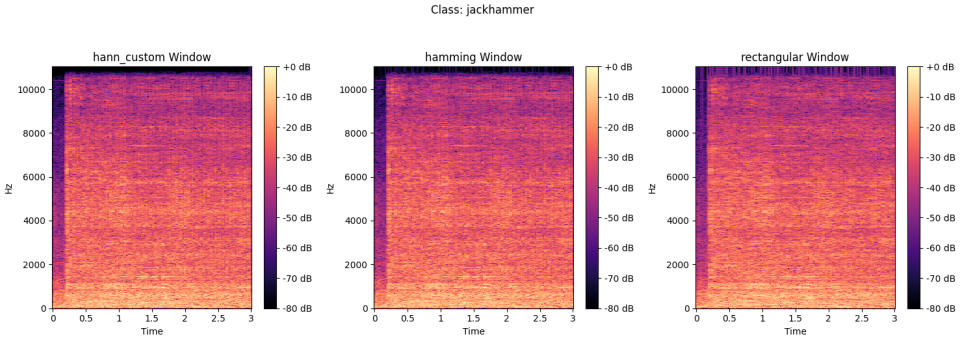


Figure 9: Spectrogram for the jackhammer class.

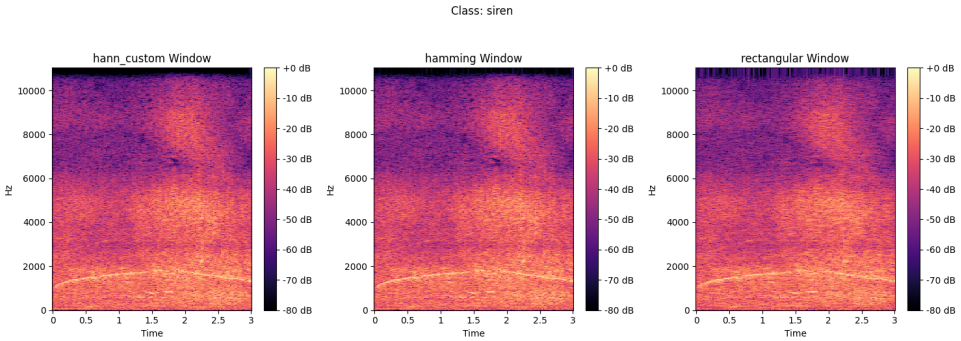


Figure 10: Spectrogram for the siren class.

2.5 Observations

The spectrograms reveal distinct frequency patterns for each class, which can be used to differentiate between the sounds. Below are some key observations:

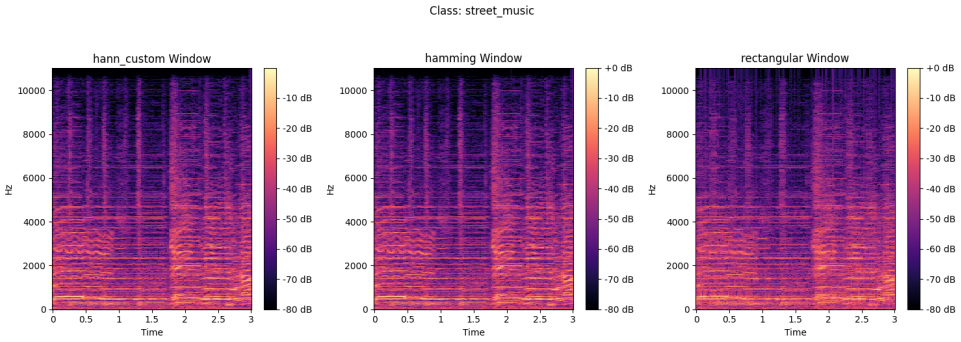


Figure 11: Spectrogram for the street music class.

- **Air Conditioner**: The spectrogram shows a consistent, low-frequency hum with occasional variations, indicative of the steady operation of an air conditioner.
- **Car Horn**: A sharp, high-frequency spike is observed, corresponding to the abrupt and loud nature of a car horn.
- **Children Playing**: The spectrogram displays a wide range of frequencies with intermittent bursts, reflecting the varied and dynamic sounds of children playing.
- **Dog Bark**: A series of high-frequency bursts are visible, representing the sharp and repetitive nature of a dog's bark.
- **Drilling**: The spectrogram exhibits a strong, continuous high-frequency component, characteristic of the mechanical noise produced by drilling.
- **Engine Idling**: A low-frequency rumble with minor fluctuations is observed, consistent with the sound of an idling engine.
- **Gunshot**: A sudden, high-intensity spike across a broad frequency range is evident, capturing the explosive nature of a gunshot.
- **Jackhammer**: The spectrogram shows a repetitive, high-frequency pattern, reflecting the rhythmic impact of a jackhammer.
- **Siren**: A sweeping frequency pattern is visible, corresponding to the oscillating sound of a siren.
- **Street Music**: The spectrogram reveals a complex mix of frequencies, indicative of the diverse and layered sounds of street music.

2.6 Classifier

A Convolutional Neural Network (CNN) model was implemented to classify the extracted spectrograms. The model consisted of multiple convolutional layers with batch normalization and dropout to prevent overfitting. The model was trained using the Adam optimizer and cross-entropy loss.

2.7 Experimental Setup

The experiments performed were 10-fold cross-validation to evaluate the performance of each windowing technique. The dataset was split into training and validation sets, where the training set has 9 out of the 10 folds present in the dataset, and the leave one out fold was used as validation set, and accuracy, plus loss metrics were recorded for comparison.

3 Results and Analysis

3.1 Training Results

We analyzed the training and validation accuracy for each windowing technique. The accuracy plots for each windowing technique are presented below:

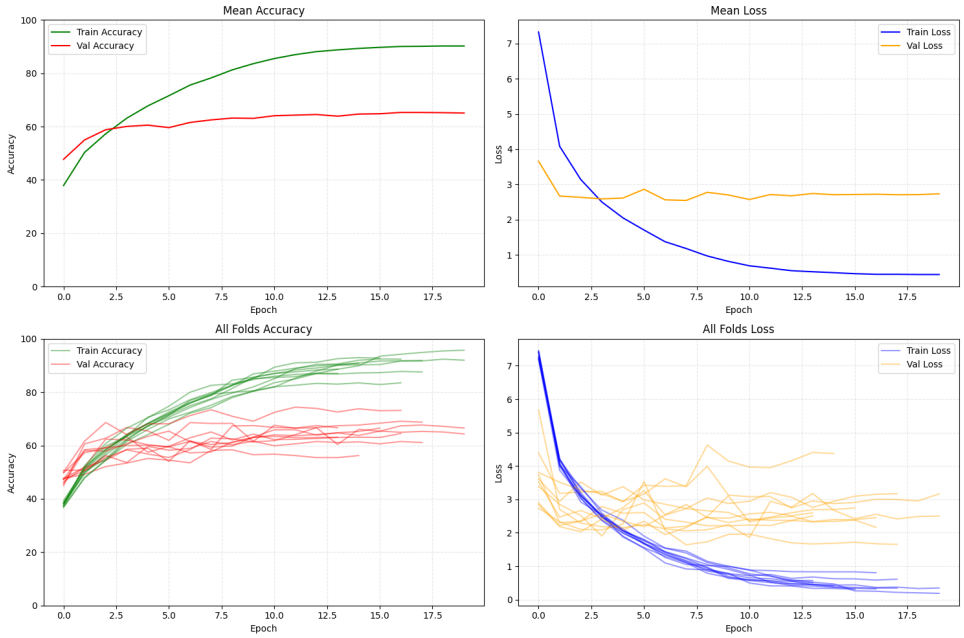


Figure 12: Accuracy plot for Hann window

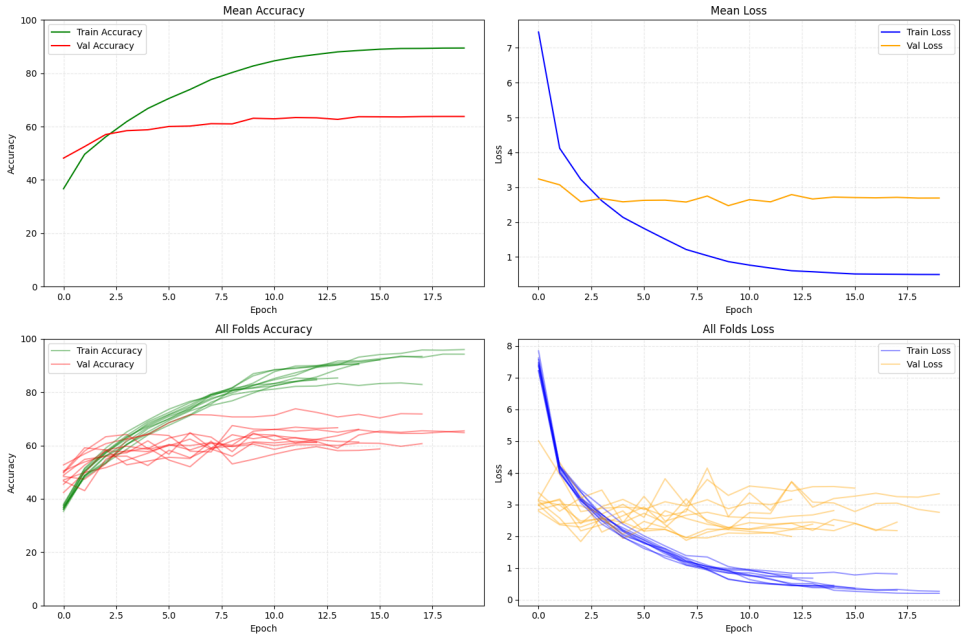


Figure 13: Accuracy plot for Hamming window

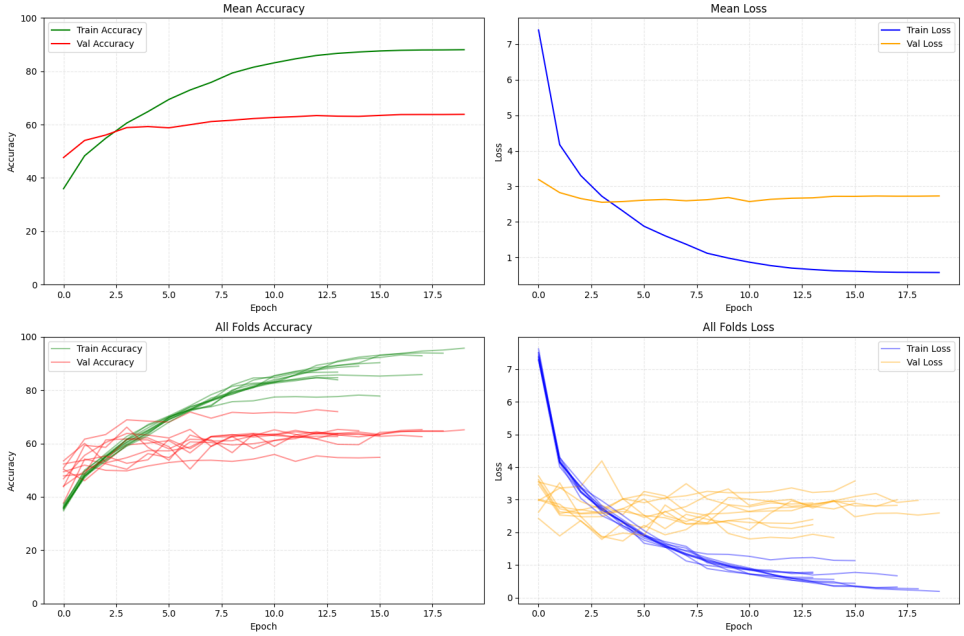


Figure 14: Accuracy plot for Rectangular window

3.2 Comparison of Windowing Techniques

The validation accuracy for each windowing technique was computed over the 10 cross fold validation, and the mean accuracies along with their standard deviations are summarized in Table 1.

Window Type	Mean Accuracy	Std Accuracy
Hann	65.09%	4.53%
Hamming	63.80%	3.90%
Rectangular	63.85%	4.17%

Table 1: Comparison of windowing techniques in terms of accuracy.

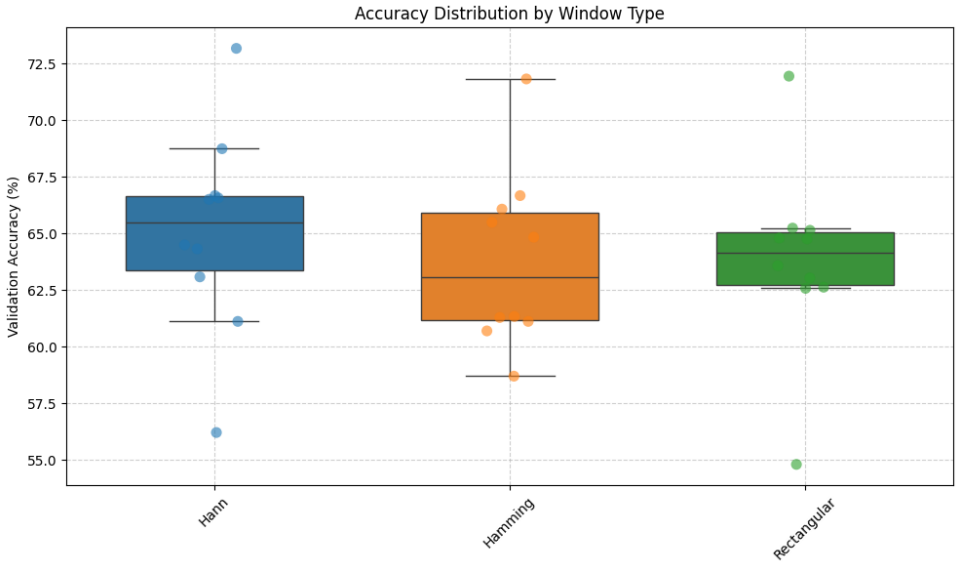


Figure 15: Box plot comparing different windowing techniques.

3.3 Conclusion

Based on the results, the Hann window performed the best with the highest mean accuracy. This is likely due to its better spectral resolution and lower spectral leakage. The Hamming window, while similar to Hann, resulted in slightly lower accuracy. The Rectangular window performed the worst due to its high spectral leakage, leading to less discriminative spectrogram features.

4 Task B: Genre-Based Spectrogram Analysis

4.1 Song Selection

I chose four of my favorite songs from different genres which closely represent completely different styles of music production. The spectrograms of these songs were generated using both the librosa library and the Audacity software, after which they were compared. The songs that I chose for this task and their corresponding genres are as follows:

- **Death Grips - "Hacker"** (Experimental Hip-Hop)
- **Joji - "Glimpse of Us"** (R&B/Lo-fi)
- **Martin Garrix - "Animals"** (EDM)
- **Prince - "Soft and Wet"** (Funk/R&B)

4.2 Technical Details

All audio files are in stereo format with a sample rate of 44100 Hz and 32-bit float encoding while generating the waveform and spectrogram in the Audacity software. While using the librosa library however, sampling rate of 22050 Hz was conventionally followed in both the tasks in question 2. The spectrograms were generated using the Short-Time Fourier Transform (STFT) with a Hann window.

Figure 16: Selected songs for Task 2.

4.3 Spectrograms and Waveforms

The spectrograms were generated using both audacity and python. Audacity gave generally better results with zoomability.

4.4 Analysis of Spectrograms

The spectrograms and waveforms were analyzed to understand the differences in frequency content, intensity, and temporal structure across the genres.

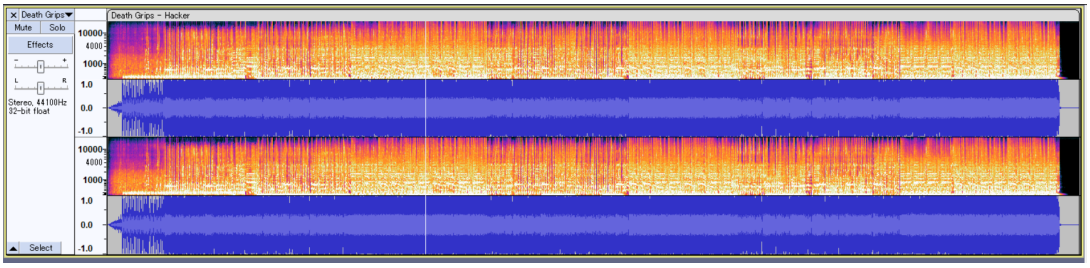


Figure 17: Combined Spectrogram and Waveform for Death Grips - "Hacker" (Stereo).

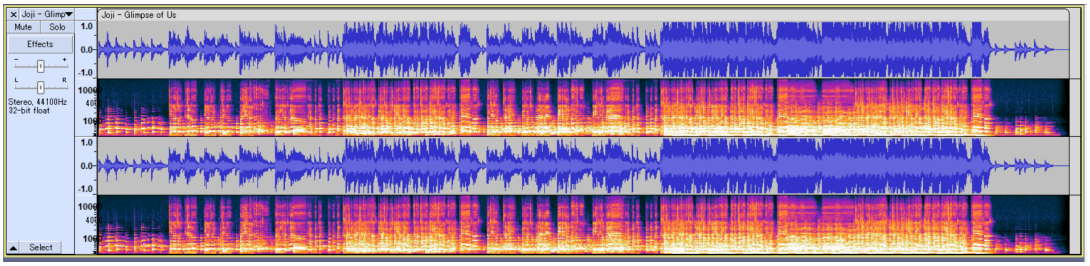


Figure 18: Combined Spectrogram and Waveform for Joji - "Glimpse of Us". (Stereo)

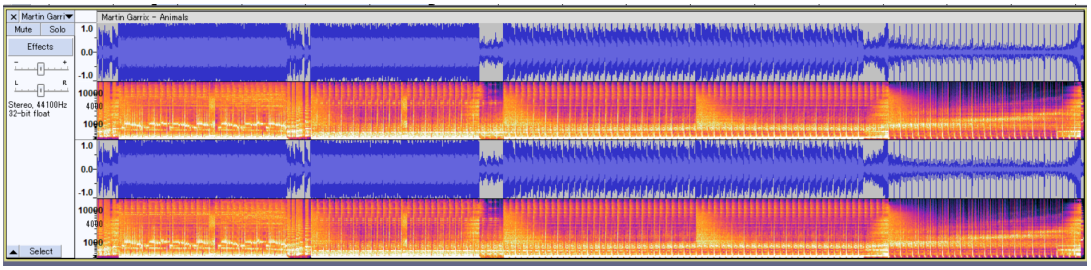


Figure 19: Combined Spectrogram and Waveform for Martin Garrix - "Animals". (Stereo)

Figure 20: Combined Spectrogram and Waveform for Prince - "Soft and Wet". (Stereo)

- **Death Grips - "Hacker"**: The spectrogram shows chaotic patterns with a wide frequency range and high-intensity spikes, reflecting the experimental and industrial nature of the song.
- **Joji - "Glimpse of Us"**: The spectrogram shows smooth, continuous frequencies with softer intensity, typical of R&B and lo-fi genres. There are areas of quiet with not much happening which slowly lead in to the highs of the song.
- **Martin Garrix - "Animals"**: The spectrogram shows repetitive, high-energy patterns with strong bass frequencies, characteristic of Electronic Dance Music.

The spectrogram shows very precise control over the frequencies of the notes in the audio.

- **Prince - "Soft and Wet"**: The spectrogram shows a mix of rhythmic, groovy patterns with a wide frequency range, reflecting the funk and R&B characteristics. The spectrogram has a very analog nature to it which reflects the old school style of analog music.

4.5 Comparative Analysis

The following table summarizes the comparative analysis of the spectrograms:

Aspect	**Death Grips**	**Joji**	**Martin Garrix**	**Prince**
Frequency Range	Wide, chaotic	Mid-range, smooth	Strong bass, synthetic highs	Wide, dynamic
Intensity	High, irregular spikes	Soft, continuous	High, repetitive	Varied, groovy
Temporal Structure	Chaotic, unpredictable	Smooth, sustained	Repetitive, rhythmic	Rhythmic, dynamic
Harmonic Content	Minimal, dissonant	Clear, harmonic	Synthetic, rhythmic	Rich, harmonic

Table 2: Comparative analysis of spectrograms for the selected songs.

4.6 Conclusions

- Spectrograms are a powerful tool for analyzing and comparing the characteristics of different music genres.
- The analysis highlights the importance of frequency content, intensity, and temporal structure in defining the unique characteristics of each genre.
- The comparative analysis provides insights into how different genres utilize frequency and intensity to create distinct musical experiences.

References

- Audacity Manual
- UrbanSound8k Dataset
- Librosa Documentation
- Kaggle Notebook