Assignment Question 2

Abhishek Sahu

February 2025

1 Introduction

This report is submitted for question two of the Speech Understanding assignment. Where I have analyzed the spectrogram and windowing techniques, also I have trained the classifier for the UrbanSound8k dataset. GitHub

2 Task A: Window techniques

Windowing techniques are essential in signal processing, particularly when working with Short-Time Fourier Transform (STFT) or spectrogram computation. Windowing helps reduce spectral leakage by tapering the signal at the edges.

2.1 Hann Window

The Hann Window is a smooth window function that tapers the signal at the edges. Its smooth tapering at the edges, reduces spectral leakage effectively and commonly used in audio processing. It is defined by the following formula:

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

2.2 Hamming Window

The Hamming Window is like the Hann window but has a slightly different shape. Its slightly flatter than the Hann window, reduces spectral leakage but not as much as the Hann window and often used in speech processing. It is defined by the following formula:

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

2.3 Rectangular Window

The Rectangular Window (also called the Boxcar Window) is the simplest window function. It is not tapering at the edges, introduces significant spectral leakage and rarely used in practice but serves as a baseline. It does not taper the signal at all and is defined as:

$$w(n) = 1, \quad 0 \leq n \leq N-1 \tag{3}$$

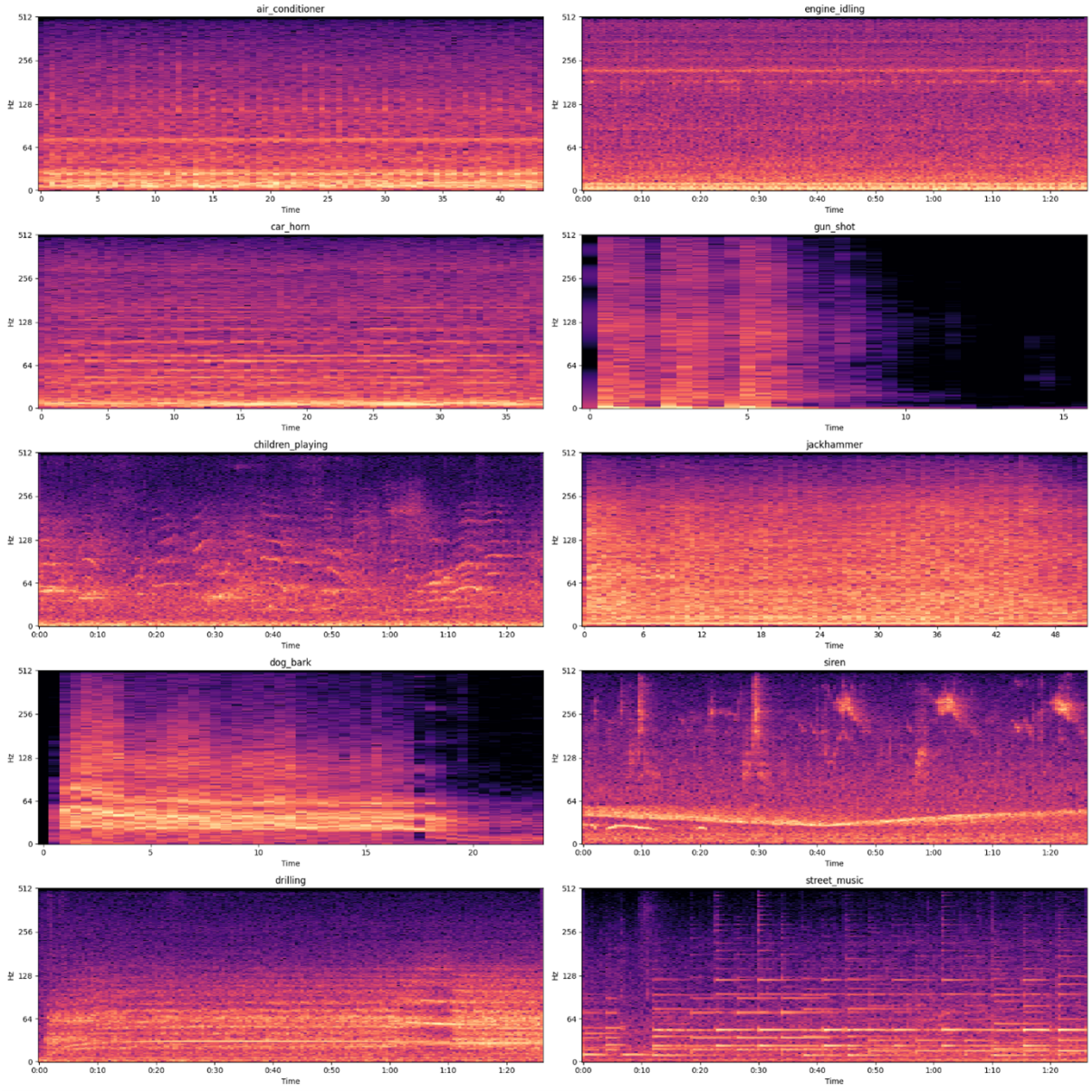

Figure 1: STFT with Hann Window for all audio types

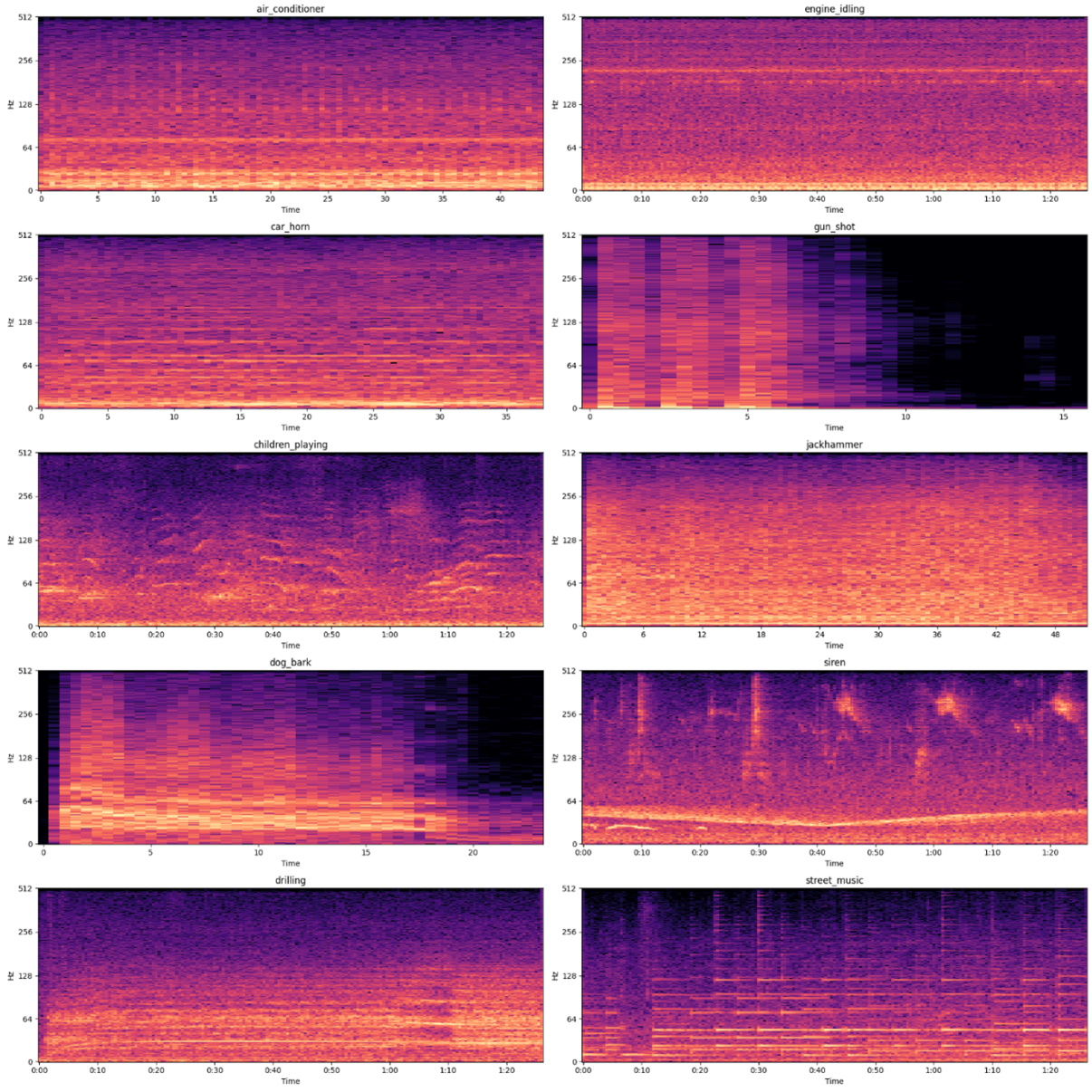

Figure 2: STFT with Hamming Window for all audio types

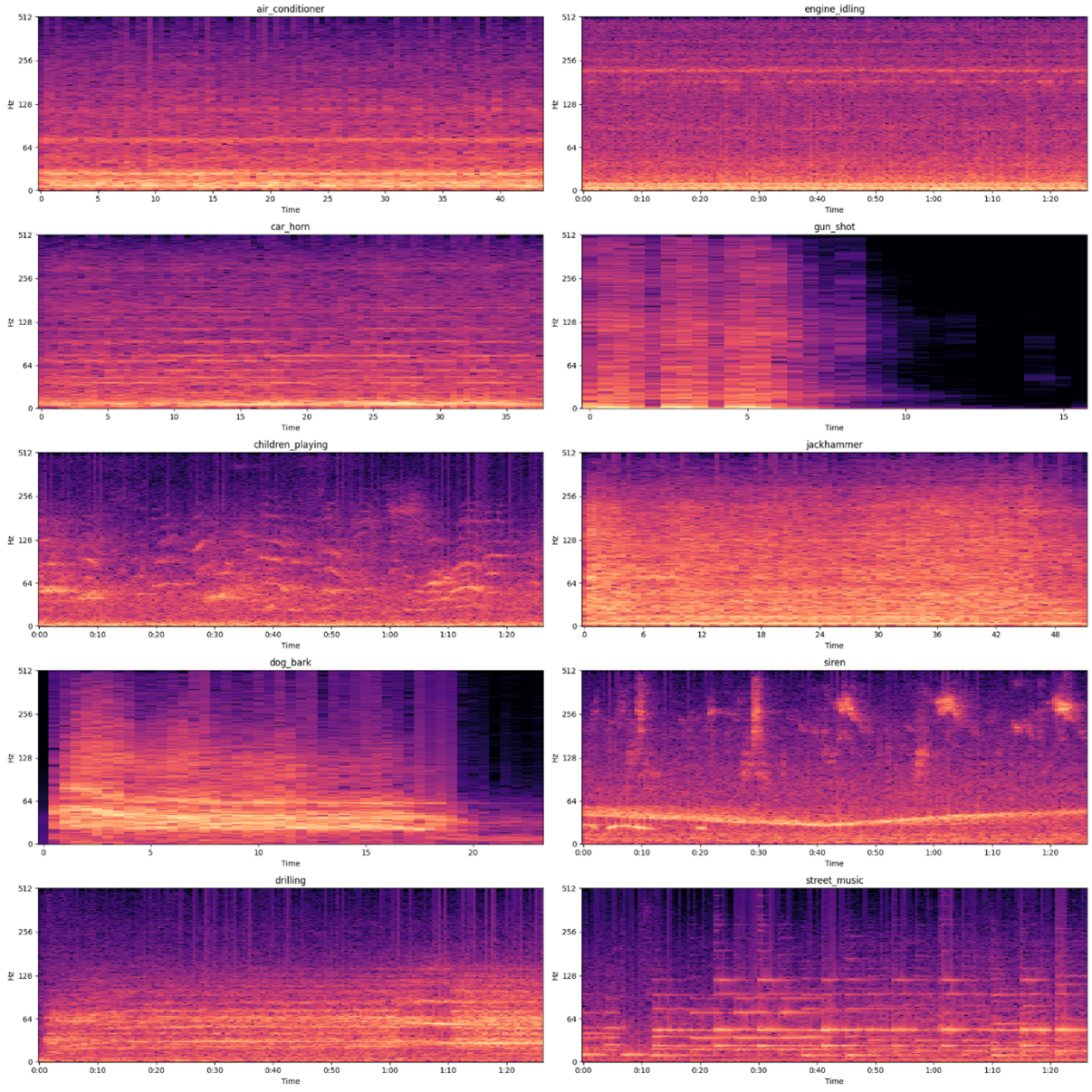

Figure 3: STFT with Rectangular Window for all audio types

3 Comparison of Windowing Techniques

- **Class-Specific Characteristics:** Transient sounds (e.g., car horn, gunshot) are better preserved by Hann and Hamming windows. Continuous sounds (e.g., air conditioner, engine idling) are clearer with Hann and Hamming windows.
- **Spectral Leakage:** The Rectangular Window introduces significant leakage, making it difficult to distinguish between frequency components in all classes. The Hann and Hamming Windows reduce leakage effectively, with the Hann Window performing slightly better for most classes.
- **Frequency Resolution:** Hamming Window provides slightly better resolution than the Hann window for transient classes. The Rectangular window has poor resolution due to leakage, making it unsuitable for analysing urban sounds.

4 Correctness of Windowing Techniques

- The **Hann Window** is the best choice for general-purpose audio analysis on the UrbanSound8K dataset, as it provides a good balance between spectral leakage reduction and frequency resolution.
- The **Hamming Window** can be used when slightly better frequency resolution is needed, especially for classes with sharp transients.
- The **Rectangular Window** should be avoided for practical applications due to its high spectral leakage.

5 Training and Classification of UrbanSound8k Dataset

	slice_file_name	fsID	start	end	salience	fold	classID	class
0	100032-3-0-0.wav	100032	0.0	0.317551	1	5	3	dog_bark
1	100263-2-0-117.wav	100263	58.5	62.500000	1	5	2	children_playing
2	100263-2-0-121.wav	100263	60.5	64.500000	1	5	2	children_playing
3	100263-2-0-126.wav	100263	63.0	67.000000	1	5	2	children_playing
4	100263-2-0-137.wav	100263	68.5	72.500000	1	5	2	children_playing

Figure 4: First five row of CSV file

class	slice_file_name
air_conditioner	1000
car_horn	429
children_playing	1000
dog_bark	1000
drilling	1000
engine_idling	1000
gun_shot	374
jackhammer	1000
siren	929
street_music	1000

dtype: int64

Figure 5: Classes and their count

5.1 Statistics description of the data

We can observe in the following tables that the data has been recorded and digitalized in different ways. The Random Forest Classifier with the Hann windowing technique achieved the highest accuracy.

- It has been mostly recorded using 2 channels in almost all the samples (stereo).
- The sample rates go from 8kHz to 192kHz (mostly 44kHz, 48Khz)
- The length of the audios goes from 0.0008s to 4s (mostly 4s)
- The bits per sample used go from 4 to 32 (mostly 24 bits)
- The data will need to be standardized before to be fed to a machine learning model.

	salience	classID	length	bitrate	channels	sample_rate	bits_per_sample
count	8732.000000	8732.000000	8732.000000	8.732000e+03	8732.000000	8732.000000	8732.000000
mean	1.347000	4.592877	3.603644	4.495311e+05	1.915369	48456.979272	18.780119
std	0.476043	2.894544	0.980913	5.480813e+05	0.278348	15300.080707	4.227168
min	1.000000	0.000000	0.000816	1.102500e+04	1.000000	8000.000000	4.000000
25%	1.000000	2.000000	4.000000	3.528000e+05	2.000000	44100.000000	16.000000
50%	1.000000	4.000000	4.000000	3.528000e+05	2.000000	44100.000000	16.000000
75%	2.000000	7.000000	4.000000	5.292000e+05	2.000000	48000.000000	24.000000
max	2.000000	9.000000	4.000000	4.515840e+07	2.000000	192000.000000	32.000000

Figure 6: Statistics description of the data

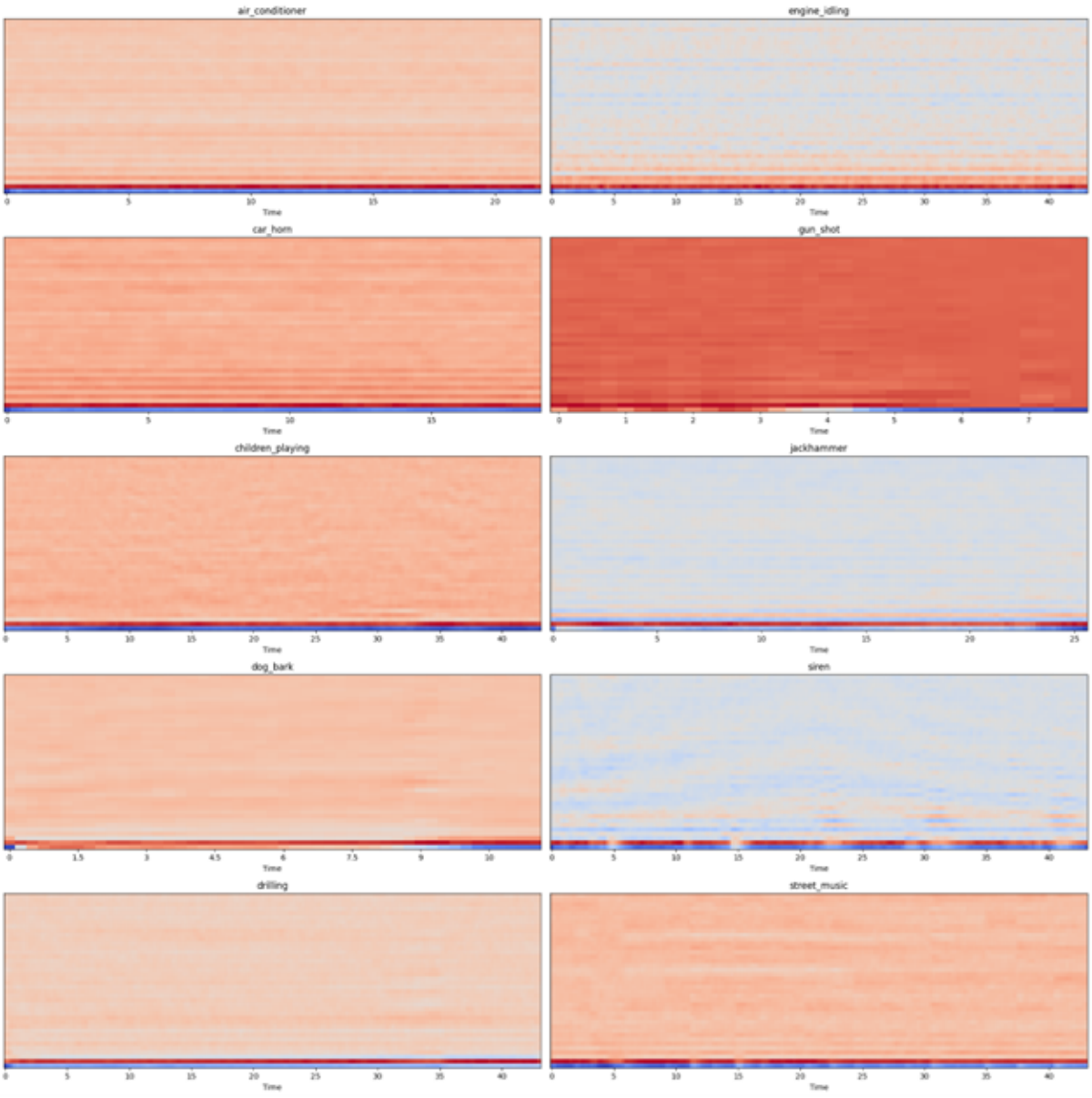

Figure 7: MFCCs for all audio types

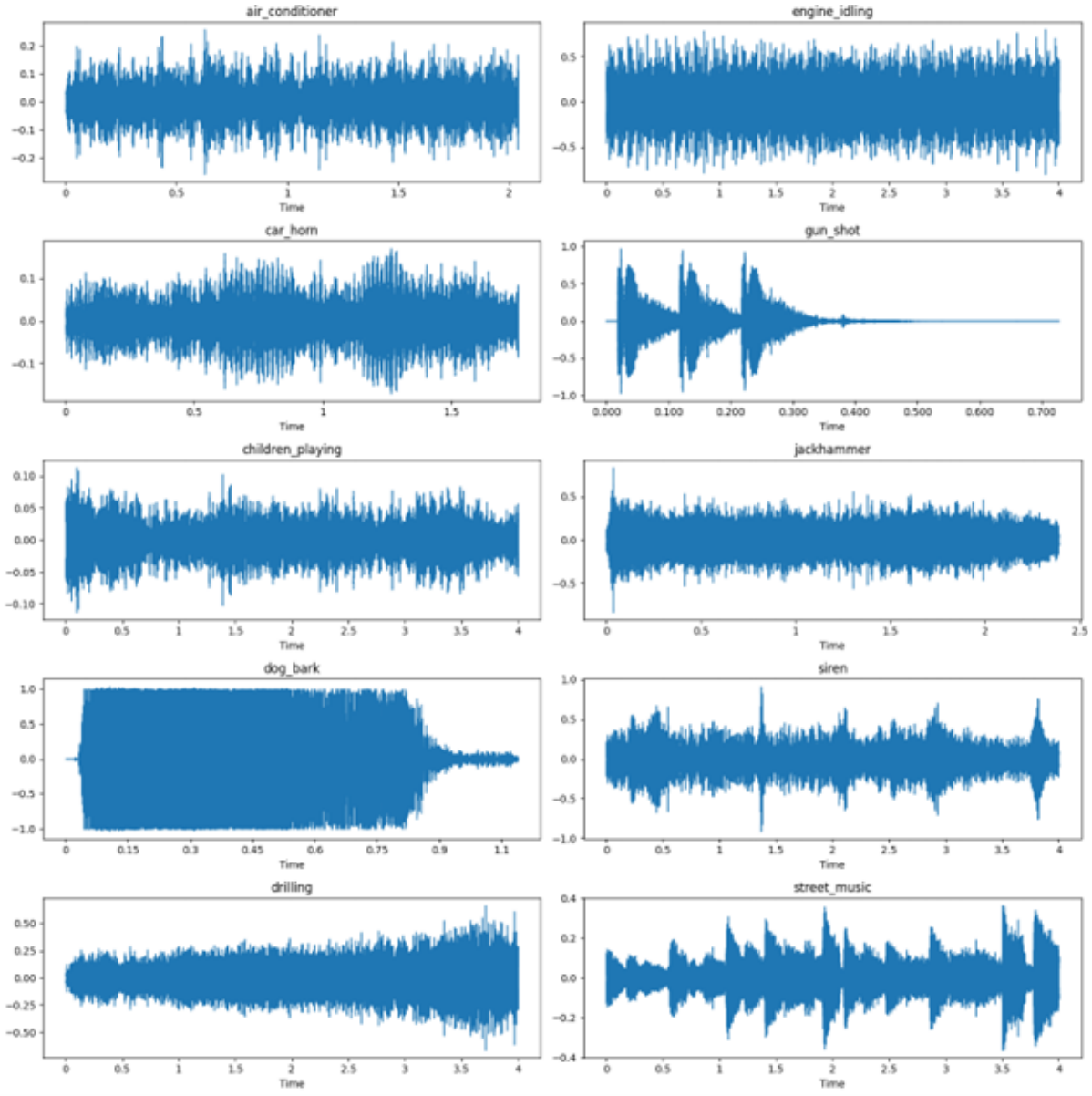

Figure 8: Waveform for all audio type

5.2 Model Training

I have used eight ML Models to train UrbanSound8k dataset with all three Windowing techniques.

	Classifier	Accuracy	Precision	Recall
2	RandomForestClassifier	0.905552	0.912604	0.884557
0	KNeighborsClassifier	0.888380	0.877331	0.861805
7	XGBClassifier	0.884946	0.883357	0.862772
6	MLPClassifier	0.760160	0.755922	0.739530
1	DecisionTreeClassifier	0.742988	0.725742	0.722150
4	GradientBoostingClassifier	0.732685	0.728923	0.702525
3	AdaBoostClassifier	0.329708	0.262373	0.291135
5	SGDClassifier	0.286777	0.211803	0.258168

Figure 9: The Model Precision and Accuracy with Hann Window Technique

	Classifier	Accuracy	Precision	Recall
2	RandomForestClassifier	0.902118	0.910943	0.880582
0	KNeighborsClassifier	0.882084	0.869757	0.859449
7	XGBClassifier	0.880366	0.877637	0.861561
6	MLPClassifier	0.764167	0.757403	0.740169
4	GradientBoostingClassifier	0.728105	0.733855	0.699170
1	DecisionTreeClassifier	0.712078	0.694766	0.693962
3	AdaBoostClassifier	0.308529	0.308340	0.272352
5	SGDClassifier	0.245564	0.233704	0.223356

Figure 10: Model Precision and Accuracy with Hamming Window Technique

	Classifier	Accuracy	Precision	Recall
2	RandomForestClassifier	0.854035	0.850886	0.827730
0	KNeighborsClassifier	0.846594	0.838888	0.827451
7	XGBClassifier	0.830567	0.828301	0.804807
6	MLPClassifier	0.734402	0.730025	0.720442
4	GradientBoostingClassifier	0.706354	0.718329	0.675735
1	DecisionTreeClassifier	0.674871	0.652905	0.652680
3	AdaBoostClassifier	0.323412	0.264910	0.286894
5	SGDClassifier	0.279336	0.255524	0.263937

Figure 11: Model Precision and Accuracy with Rectangular Window Technique

5.3 Highest Accuracy

As Random Forest Classifier is giving the highest accuracy with Hann Widowing Technique, I'm considering the best ML model for UrbanSound8k dataset. Below the Confusion matrix.

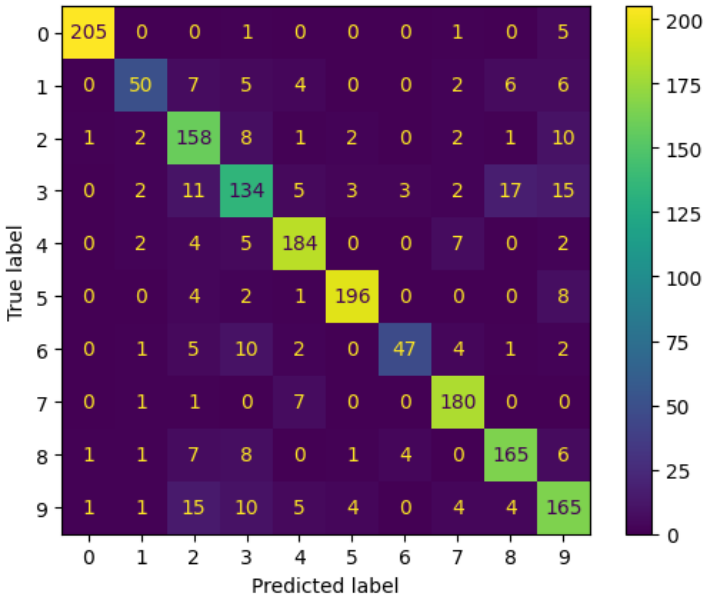

Figure 12: Confusion Matrix for Random Forest Model

6 Task B: Spectrogram Analysis for Different Music Genres

6.1 Spectrograms for Electronic Songs

- **Spectrogram Features** Electronic music, especially progressive house like this song will show significant energy in both low and high frequencies due to the bass and synths. The spectrogram will have strong low-frequency elements from the kick drum and bassline, as well as sharp, rhythmic patterns from synths and hi-hats in the higher frequency range.

- **Analysis** The low-end frequencies (around 20-100 Hz) will have consistent, powerful energy, and you might see the characteristic "sawtooth" waveforms from synths, particularly in the buildup or drop sections. There will be less variation in mid-range frequencies compared to rock or classical music, as the emphasis is often on bass and highs.

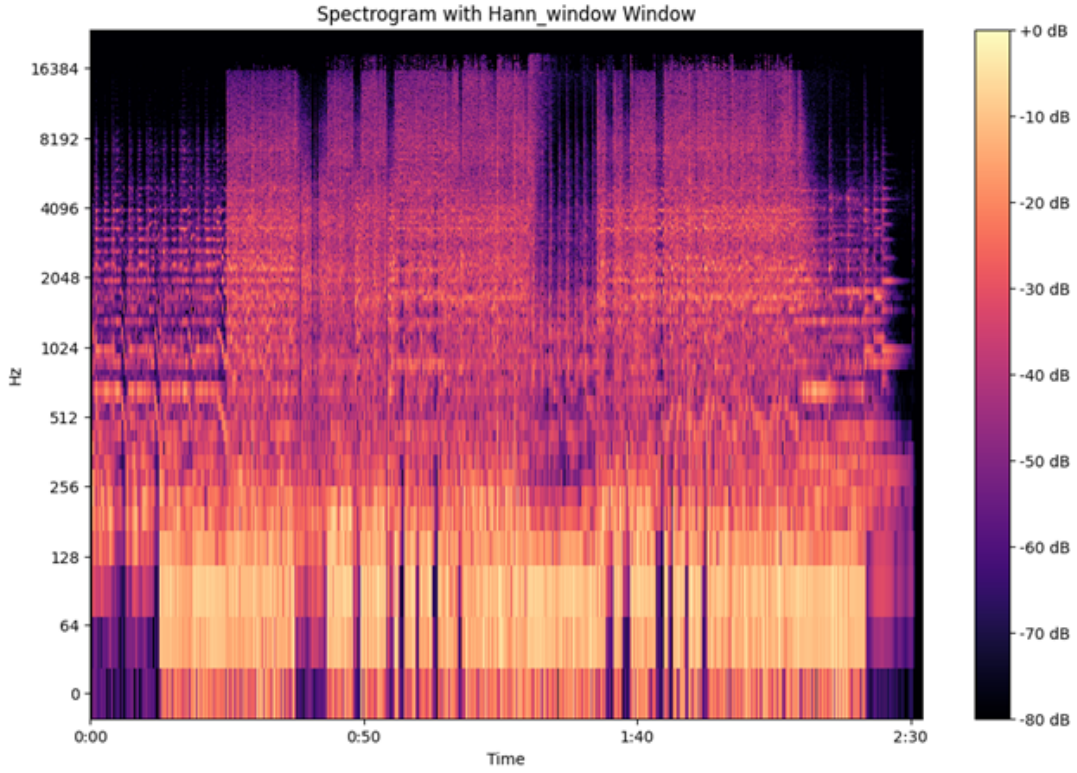

Figure 13: Spectrogram for Electronic Music With Hann Window

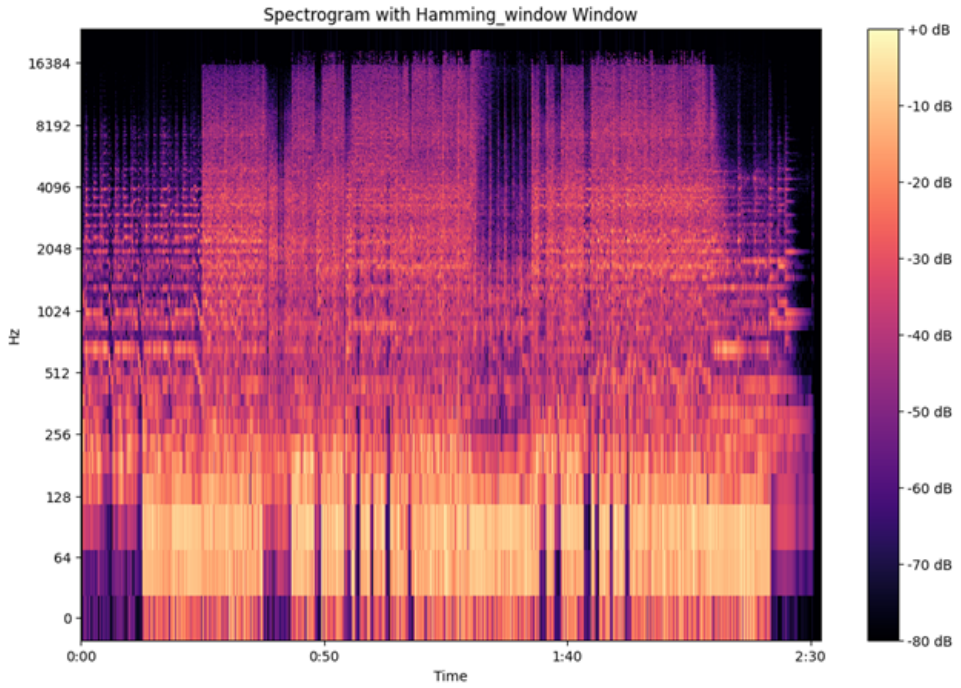

Figure 14: Spectrogram for Electronic Music With Hamming Window

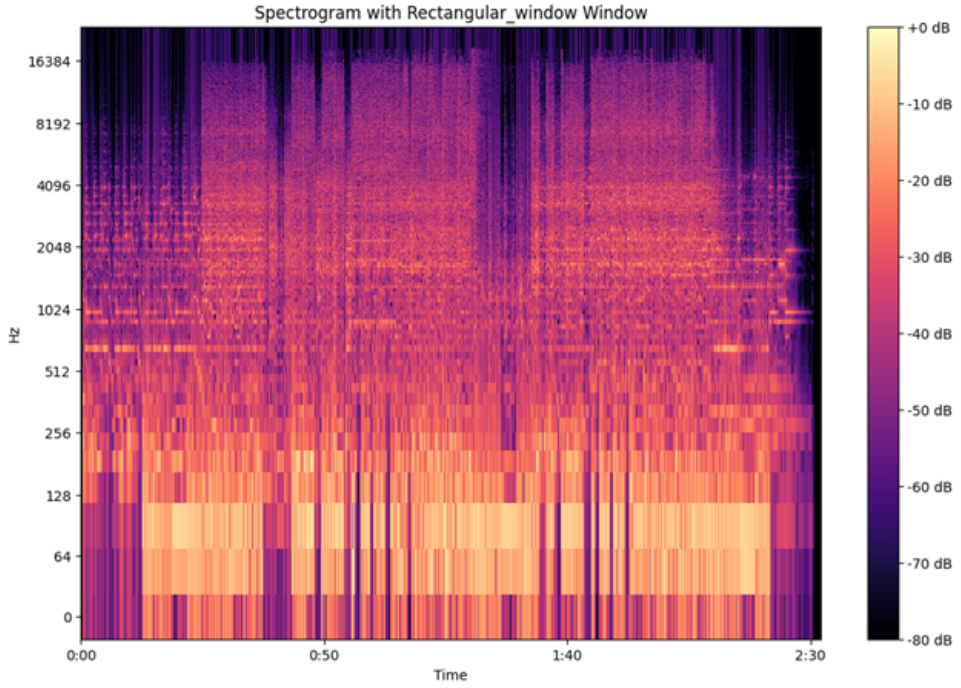

Figure 15: Spectrogram for Electronic Music With Rectangular Window

6.2 Spectrograms for Pop Songs

- **Spectrogram Features** Pop music typically has a consistent beat with electronic instrumentation, vocals, and synths. The spectrogram for this song will likely show a regular pattern, especially in the mid-range frequencies where the vocals and

synthesizers sit. You'll notice sharp, narrow peaks corresponding to beats, with smoother, broader bands representing the synths and vocals.

- **Analysis** There might be distinct high-frequency energy from the synths, and you'll see a rhythmic pattern from the percussion. Vocals will occupy a central frequency range with clear, well-defined peaks.

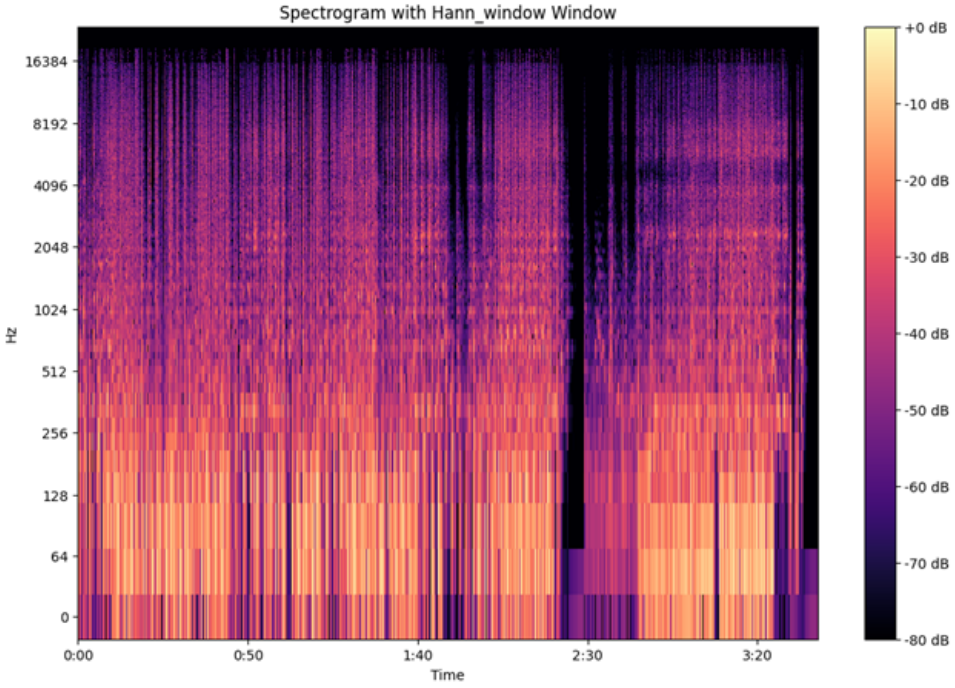

Figure 16: Spectrogram for Pop Music With Hann Window

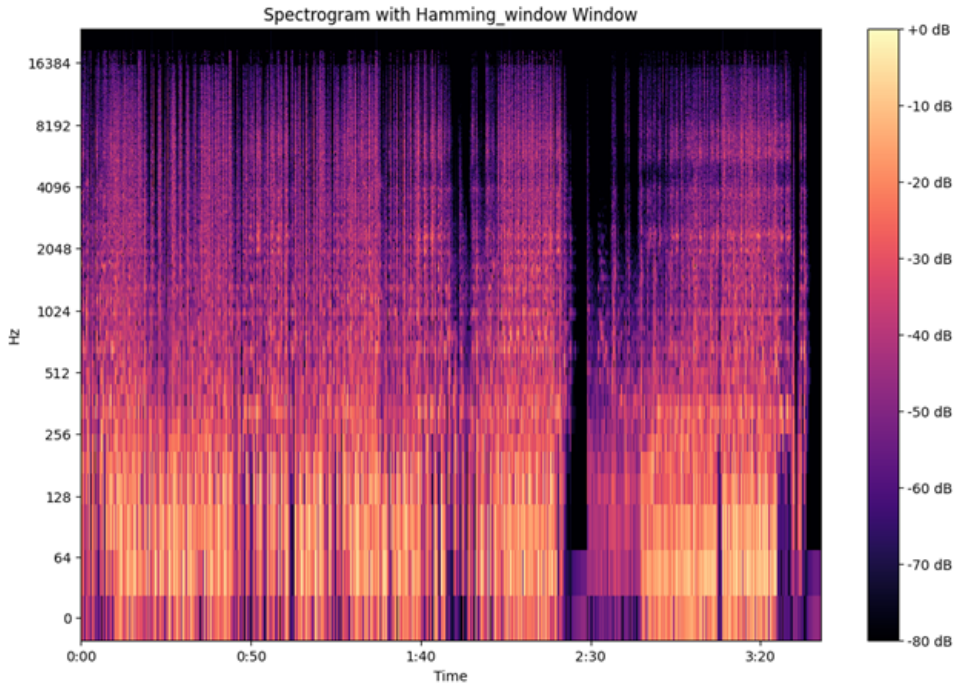

Figure 17: Spectrogram for Pop Music With Hamming Window

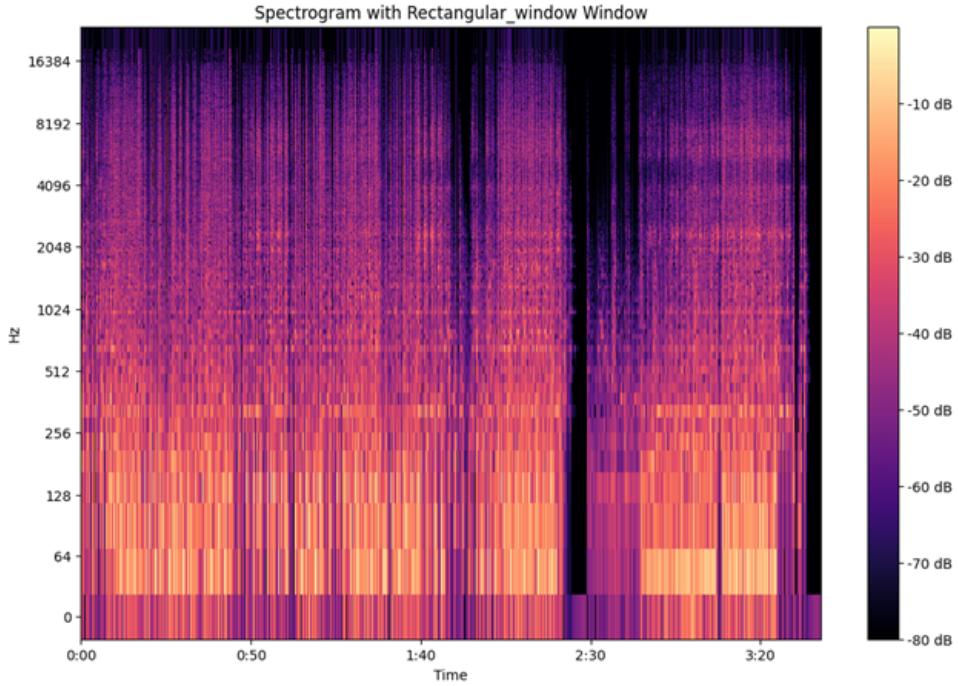

Figure 18: Spectrogram for Pop Music With Rectangular Window

6.3 Spectrograms for Classical Songs

- **Spectrogram Features** Classical piano pieces like this song tend to have a more organic, flowing pattern. The spectrogram will likely show a broad distribution of frequencies across the low, mid, and high ranges, with many subtle variations.

Since it's a solo piano piece, you'll see peaks where the notes are played, with softer, continuous energy in between.

- **Analysis** The piano's harmonic overtones will create dense clusters in the spectrogram. The dynamics of the piece will be reflected in the intensity and concentration of frequency bands. Softer passages will show as less concentrated areas, while louder notes or chords will create more defined peaks.

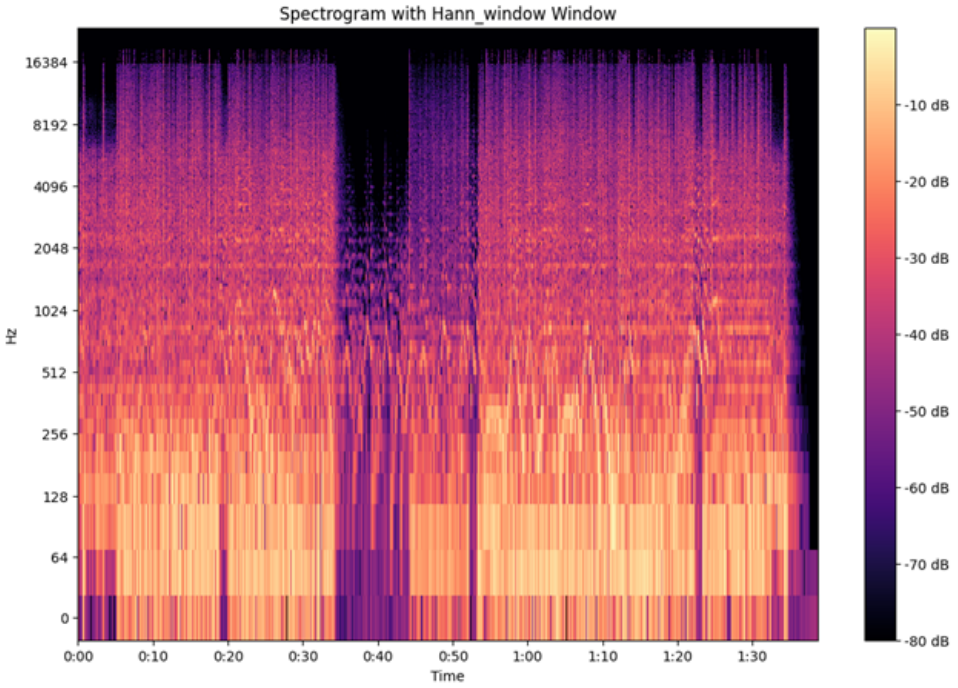

Figure 19: Spectrogram for Classical Music With Hann Window

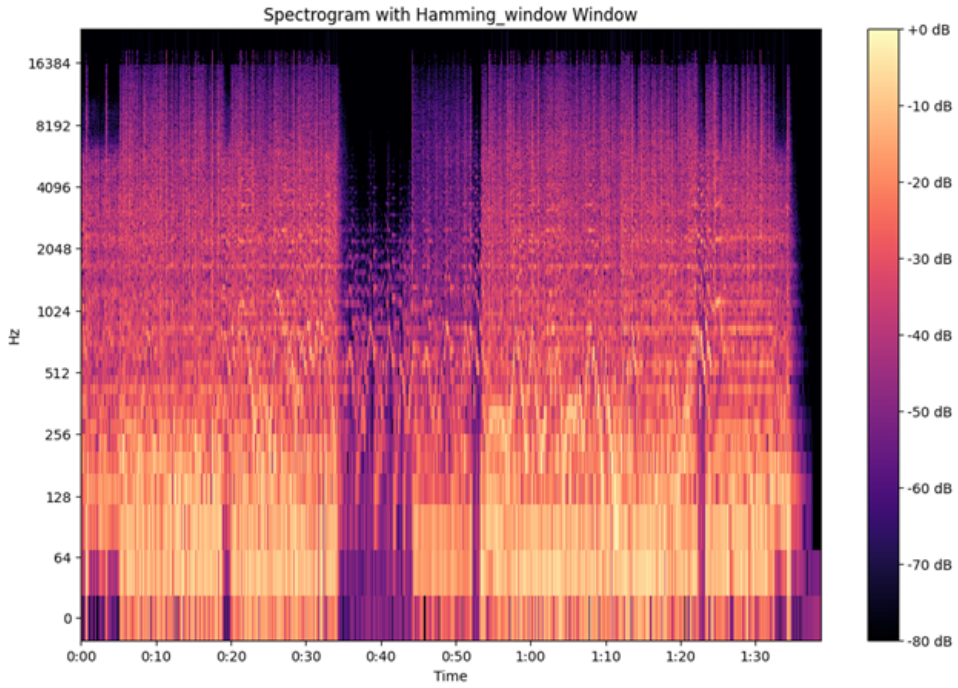

Figure 20: Spectrogram for Classical Music With Hamming Window

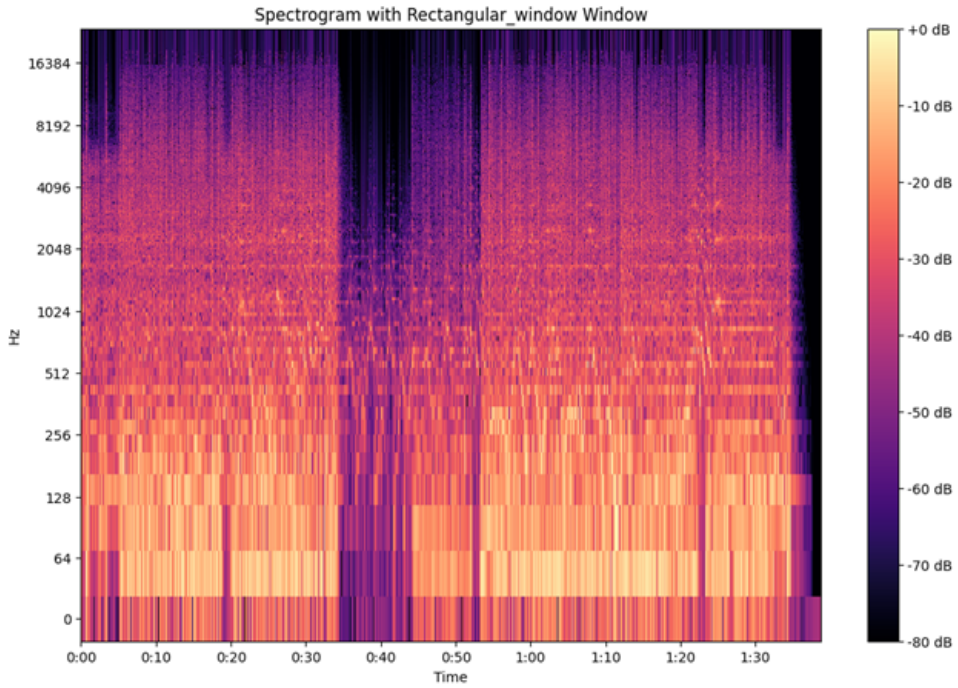

Figure 21: Spectrogram for Classical Music With Rectangular Window

6.4 Spectrograms for Rock Songs

- **Spectrogram Features** Rock music is often more dynamic and complex with layered instruments like electric guitars, drums, and vocals. The spectrogram will likely show a variety of frequency bands, especially in the high mid-range and treble

for guitars, as well as more low-end energy from the bass and drums. Vocals will have a prominent presence in the mid range.

- **Analysis** You'll notice a more varied spectrum with different sections of the song showing distinct patterns. The guitar solos will show sharp, higher frequency content, and the dramatic changes in tone and tempo will be visible as shifts in the frequency bands.

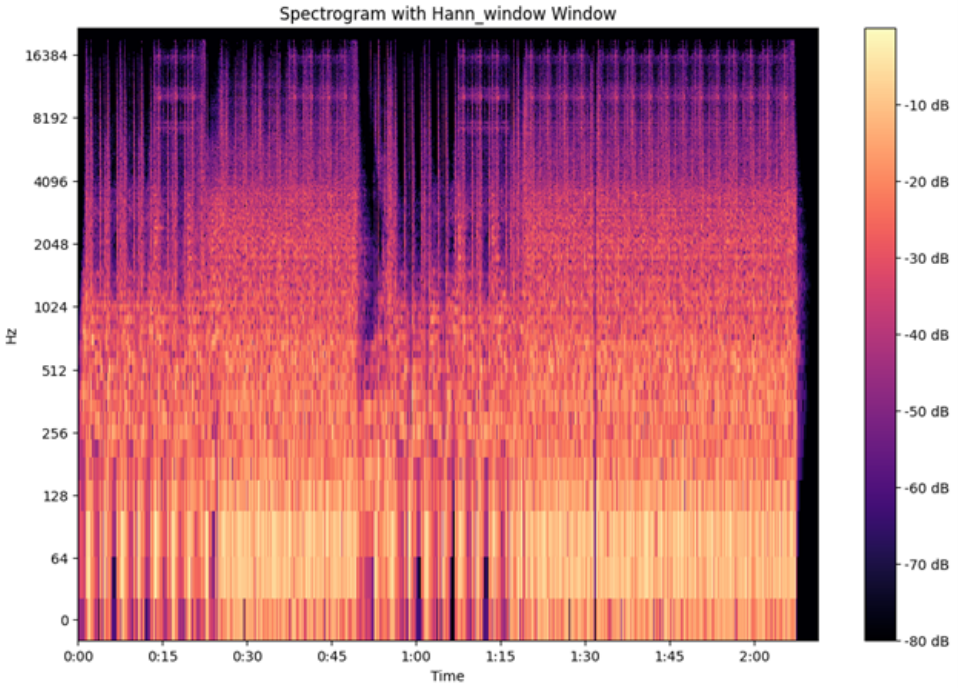

Figure 22: Spectrogram for Rock Music With Hann Window

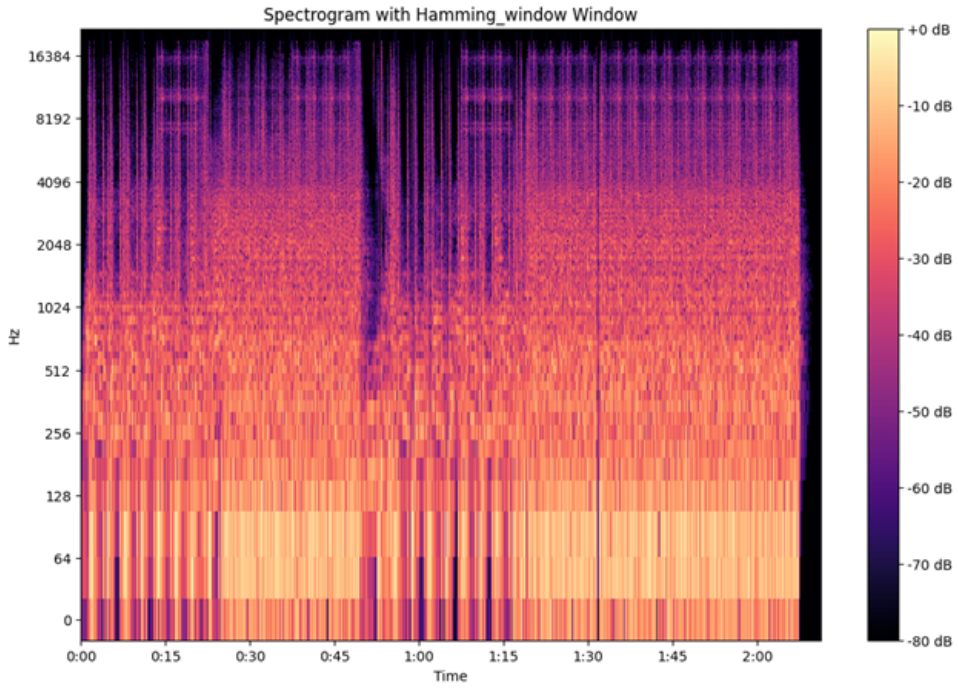

Figure 23: Spectrogram for Rock Music With Hamming Window

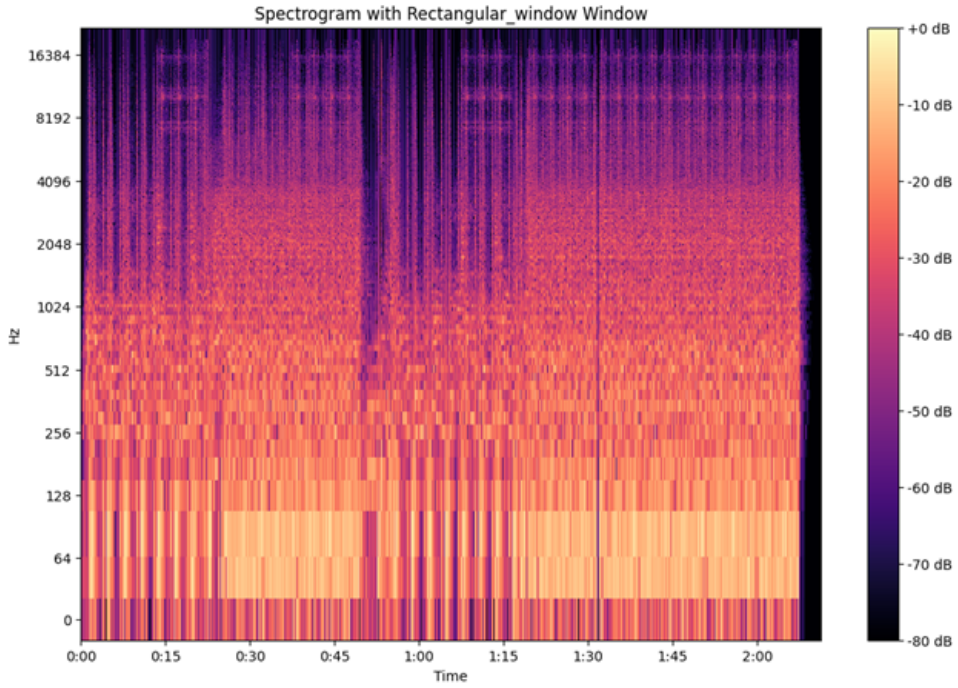

Figure 24: Spectrogram for Rock Music With Rectangle Window

7 Comparative Analysis

- **Frequency Range:** Classical music will likely have the broadest frequency range with gentle, flowing transitions between frequencies. Pop and rock songs have a

more concentrated mid-range, with electronic music focusing heavily on low-end and high-frequency energy.

- **Dynamic Range:** The dynamic range in classical music will be more fluid and subtle compared to the punchier, more abrupt shifts in rock and pop songs. Electronic music has a more predictable dynamic range, focusing on build-ups and drops.
- **Complexity:** The rock song will probably show the most complex spectrogram due to its many layered instruments and varied sections. Pop songs land electronic will have cleaner, more consistent patterns, with pop focusing on smooth, regular intervals and electronic music on repetitive, bass-driven energy. Classical will be more organic with gentle changes.