Experimenting with Spectrograms and Windowing Techniques

Shreya Pandey (M24CSA030)

Abstract

This study analyzes the impact of Hann, Hamming, and Rectangular windowing on spectrograms, evaluating CNN, SVM, and Random Forest models for sound classification using UrbanSound8k. Spectrogram analysis reveals each window's effect on frequency resolution and leakage. Random Forest achieved the highest accuracy, highlighting the importance of window selection and demonstrating machine learning's effectiveness in sound recognition.

1 Introduction

This study examines the effects of three different windowing techniques—Hann, Hamming, and Rectangular—on spectrogram generation using the Short-Time Fourier Transform (STFT). It also compares the performance of three classification models—Convolutional Neural Networks (CNN), Support Vector Machines (SVM), and Random Forest—based on features extracted from the UrbanSound8k dataset [1].

2 Theory

2.1 Spectrogram and Windowing Techniques

A spectrogram is a visual representation of the frequency spectrum of a signal as it varies with time [2]. It is generated using the Short-Time Fourier Transform (STFT), where a window function is applied to segment the signal into short overlapping frames. The choice of window function significantly impacts the spectral resolution and leakage.

Types of Windowing Functions:

- **Hann Window:** Minimizes spectral leakage by smoothly tapering the signal at the edges [3].
- **Hamming Window:** Similar to the Hann window but slightly better frequency resolution [4].
- **Rectangular Window:** Abrupt truncation leads to high spectral leakage.

2.2 Machine Learning Models

Convolutional Neural Networks (CNN): CNNs are deep learning models specifically designed for processing image-like data [5]. In the case of spectrogram classification, CNNs learn hierarchical patterns from frequency representations.

- Strengths: Automatically extracts features, effective in handling spatial information.
- Weaknesses: Computationally expensive, requires large training datasets.

Support Vector Machines (SVM): SVMs are supervised learning models used for classification tasks, particularly effective in high-dimensional spaces [6].

- Strengths: Works well with small datasets, effective in separating complex data structures.
- Weaknesses: Computationally expensive for large datasets, requires feature engineering.

Random Forest (RF): Random Forest is an ensemble learning method that builds multiple decision trees to enhance classification accuracy [7].

- Strengths: Robust to noise, less prone to overfitting.
- Weaknesses: Can be less interpretable compared to other models.

3 GitHub repository

The complete assignment can be accessed from this repository. M24CSA030

4 Task A: Windowing Techniques and Classification Performance Observations

4.1 Implementation

The implementation can be found in this Google Colab notebook. M24CSA030

4.2 Windowing Techniques and Spectrogram Analysis

General Observations:

- **Frequency Range:** All spectrograms extend up to the Nyquist frequency (half the sampling rate).
- **Time Axis:** Represents the duration of the audio sample, varying across different sounds.
- **Color Intensity:** Bright colors correspond to higher energy levels at particular frequency-time points.

4.2.1 Sample-Wise Analysis

1. **Sample 1: Drilling**
    - Hann: Produces a clean spectrogram with distinct horizontal frequency lines and minimal background noise.
    - Hamming: Similar to Hann, but with slightly sharper frequency lines, potentially introducing mild artifacts.
    - Rectangular: Causes significant spectral leakage, resulting in smeared frequency components and a blurry spectrogram.

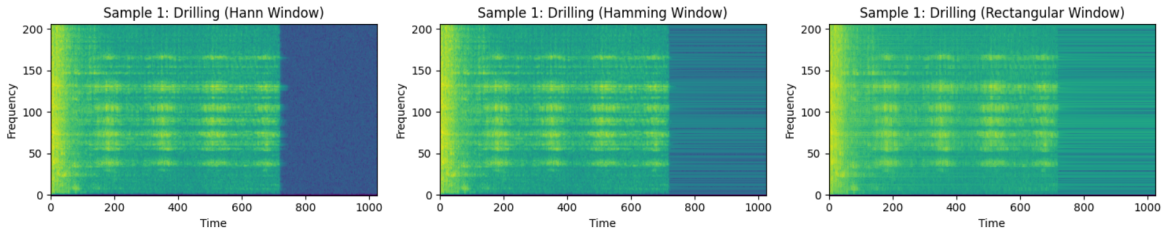

Figure 1: Spectrograms of Drilling with different windowing functions.

2. **Sample 2: Street Music**
    - Hann: Provides clear representation of musical components.
    - Hamming: Slightly improved frequency separation over Hann.
    - Rectangular: Displays heavy spectral leakage, reducing clarity.

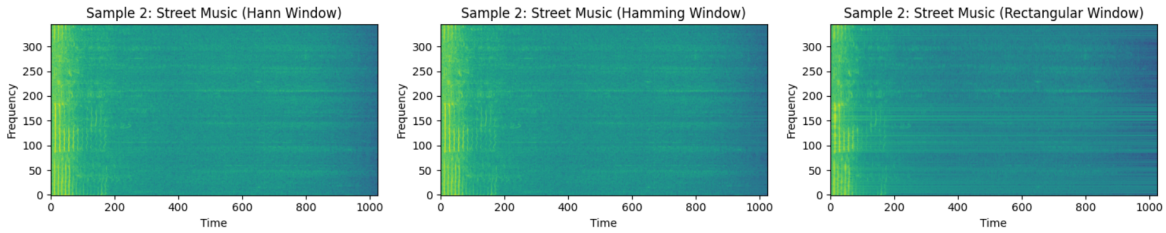

Figure 2: Spectrograms of Street Music with different windowing functions.

3. **Sample 3: Car Horn**

- Hann: Clearly captures the distinct frequencies of the car horn.
- Hamming: Similar to Hann, with possibly sharper frequency representation but potential artifacts.
- Rectangular: Displays significant spectral leakage, making fundamental frequencies hard to distinguish.

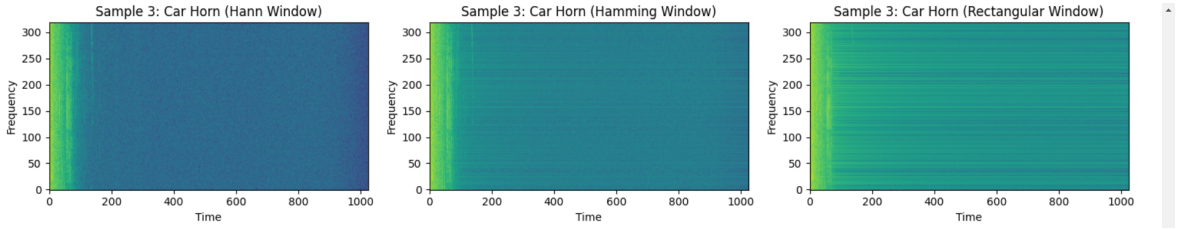

Figure 3: Spectrograms of Car Horn with different windowing functions.

4. **Sample 4: Siren**

- Hann: Accurately represents rising and falling tones in the siren.
- Hamming: Slightly better time resolution of frequency sweeps but some artifacts.
- Rectangular: Blurred frequency sweeps due to spectral leakage.

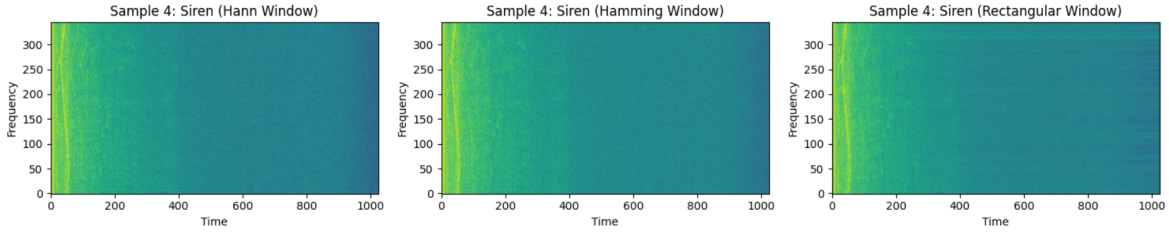

Figure 4: Spectrograms of siren with different windowing functions.

5. **Sample 5: Children Playing**

- Hann: Displays a clean spectrogram, capturing various voice frequencies distinctly.
- Hamming: Might provide better separation of individual frequencies, but artifacts may arise.
- Rectangular: Causes substantial spectral leakage, making individual voices hard to distinguish.

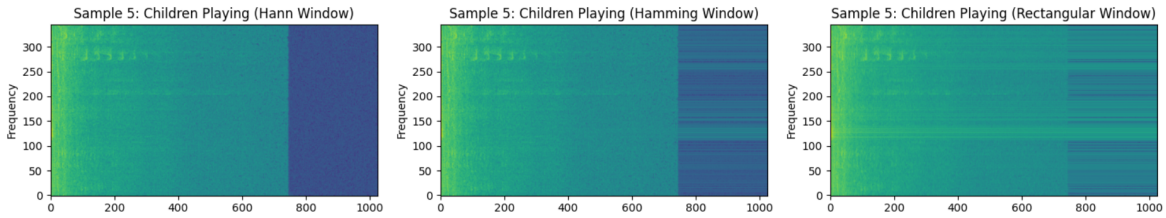

Figure 5: Spectrograms of Children Playing with different windowing functions.

4.3 Windowing Correctness

- **Hann and Hamming Windows:** Offer a good balance between frequency resolution and spectral leakage minimization.
- **Rectangular Window:** Generally ineffective for spectral analysis due to abrupt signal truncation.

4.4 Classification Performance

4.4.1 CNN

- Accuracy: 83.14
- Strengths: Learns hierarchical patterns in spectrograms.
- Weaknesses: Computationally expensive.

```
CNNModel(
  (conv1): Conv2d(16, 64, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
  (conv2): Conv2d(64, 128, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
  (pool): MaxPool2d(kernel_size=(2, 1), stride=(2, 1), padding=0, dilation=1, ceil_mode=False)
  (dropout): Dropout(p=0.1, inplace=False)
  (fc1): Linear(in_features=512, out_features=1024, bias=True)
  (fc2): Linear(in_features=1024, out_features=10, bias=True)
)
Accuracy: 83.14%
```

Figure 6: CNN analysis

4.4.2 SVM

- Accuracy: 58.96
- Strengths: Performs well in high-dimensional spaces.
- Weaknesses: Poor feature representation.

	precision	recall	f1-score	support
0	0.68	0.58	0.63	238
1	0.56	0.60	0.58	121
2	0.29	0.82	0.43	233
3	0.57	0.50	0.53	246
4	0.79	0.58	0.67	258
5	0.77	0.62	0.69	255
6	0.93	0.63	0.75	99
7	0.89	0.76	0.82	266
8	0.71	0.38	0.50	223
9	0.64	0.43	0.52	244
accuracy			0.59	2183
macro avg	0.68	0.59	0.61	2183
weighted avg	0.68	0.59	0.61	2183

Figure 7: SVM analysis

4.4.3 Random Forest

- Accuracy: 89
- Strengths: Highest accuracy, robust to noise.
- Weaknesses: May overfit.

	precision	recall	f1-score	support
0	0.93	0.98	0.96	238
1	0.99	0.68	0.80	121
2	0.77	0.90	0.83	233
3	0.88	0.80	0.84	246
4	0.93	0.92	0.93	258
5	0.95	0.97	0.96	255
6	0.91	0.82	0.86	99
7	0.92	0.96	0.94	266
8	0.89	0.93	0.91	223
9	0.81	0.80	0.81	244
accuracy			0.89	2183
macro avg	0.90	0.88	0.88	2183
weighted avg	0.89	0.89	0.89	2183

Figure 8: Random Forest analysis

5 Task B: Spectrogram Comparison Across Different Music Genres

5.1 Implementation

The implementation can be found in this Google Colab notebook. M24CSA030

5.2 Selected Songs

- **Fitoor.wav (Shamshera, 2022)** - Romantic Ballad
- **Saadda Haq.wav (Rockstar, 2011)** - Rock
- **Tere Naina.wav (My Name is Khan, 2010)** - Contemporary Classical
- **O Haseena Zulfonwale.wav (Teesri Manzil, 1966)** - Retro Bollywood

5.3 Spectrogram Observations

5.3.1 Fitoor.wav

- Sparse frequency content with energy concentrated below 500 Hz.
- Some diagonal patterns indicating smooth frequency shifts.

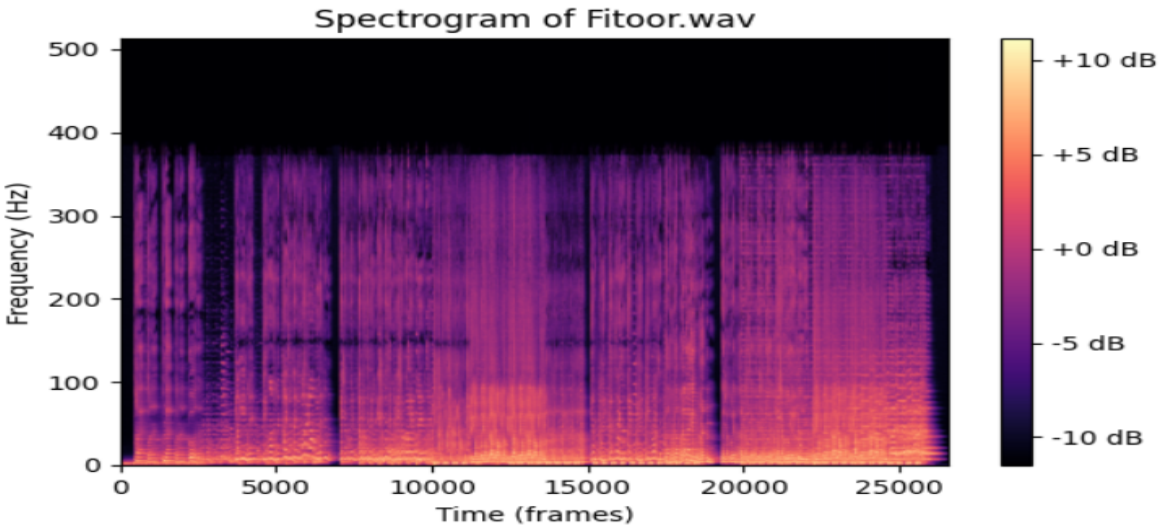

Figure 9: Spectrogram of Fitoor.wav

5.3.2 Saadda Haq.wav

- Broad frequency distribution with high-energy content.
- More high-frequency activity and visible formant-like structures.

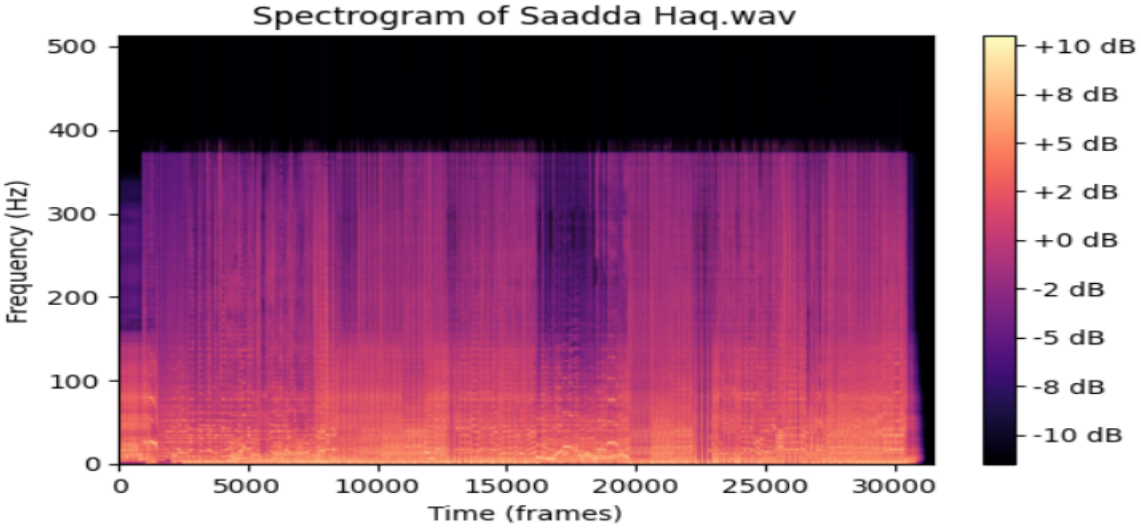

Figure 10: Spectrogram of Sadda Haq.wav

5.3.3 Tere Naina.wav

- Well-defined formants, emphasizing mid-frequency vocal elements.
- Periodic patterns indicating structured musical composition.

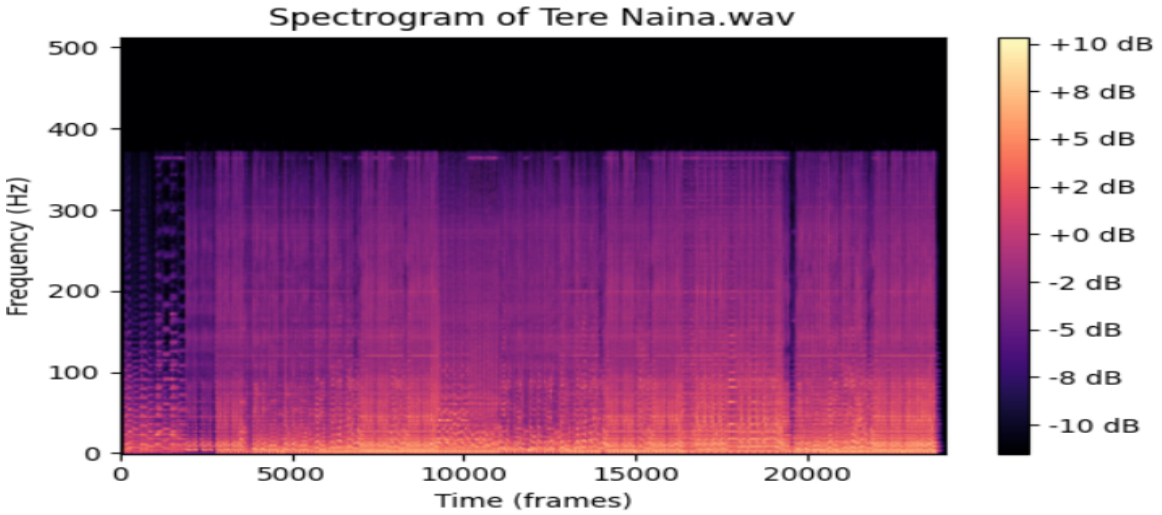

Figure 11: Spectrogram of Tere Naina.wav

5.3.4 O Haseena Zulfonwale.wav

- Highest complexity with overlapping frequency patterns.
- Broadband activity suggesting a rich instrumental composition.

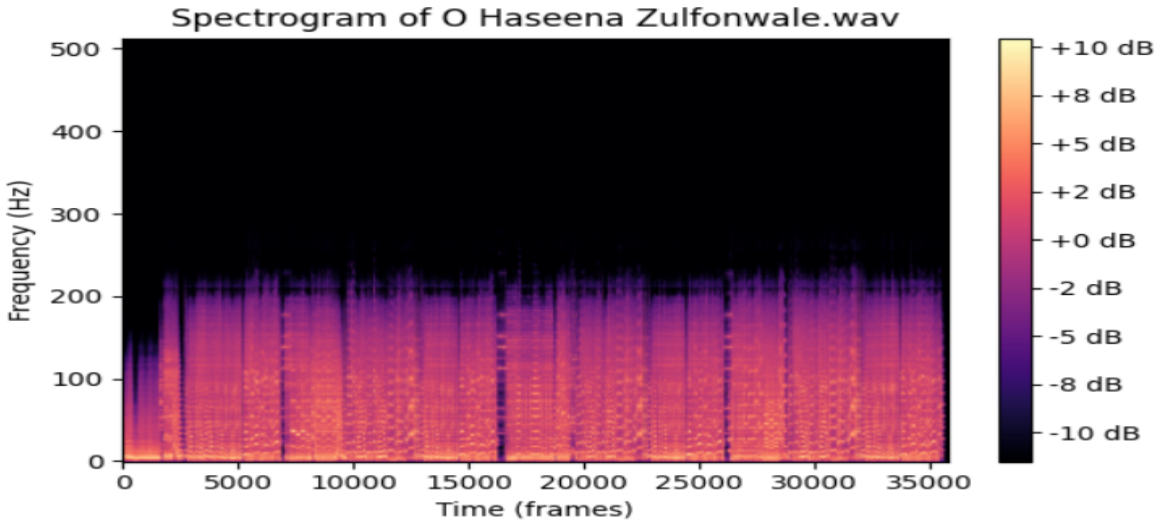

Figure 12: Spectrogram of O Haseena Zulfonwale.wav

6 Conclusion

This study highlights the importance of windowing techniques in spectral analysis and evaluates classifier performances for sound recognition. The Random Forest model demonstrated the highest accuracy, while CNN showed potential with further tuning. Spectrogram comparisons provided insights into genre-based spectral characteristics, offering a foundation for deeper music and sound classification research. Future work can explore advanced deep learning techniques to further improve performance.

References

[1] J. Salamon, P. Bello, G. Farnsworth, M. Montesinos, and M. D. Cohen, "A dataset and taxonomy for urban sound research," *Proceedings of the 25th ACM International Conference on Multimedia*, pp. 1041–1045, 2017.

[2] L. R. Rabiner and B.-H. Juang, "Theory and implementation of a speech recognition system," *Englewood Cliffs*, vol. 3, no. 4.2, 1993.

[3] F. J. Harris, "On the use of windows for harmonic analysis with the discrete fourier transform," *Proceedings of the IEEE*, vol. 66, no. 1, pp. 51–83, 1978.

[4] R. W. Hamming, *Digital filters*. Prentice-Hall, Inc., 1989.

[5] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[6] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[7] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.