SPEECH UNDERSTANDING

(CSL7770)

Experimenting with Spectrograms and Windowing Techniques

ASSIGNMENT TASK-A REPORT

(GitHub Link)

SUBMITTED BY

ARNAV SHARMA

(M24CSE004)

Department of Computer and Engineering

Indian Institute of Technology Jodhpur

Karwar 342030, Rajasthan, India

ABSTRACT

The objective of this study is to explore the effectiveness of different windowing techniques in the context of Short-Time Fourier Transform (STFT) for audio classification. The UrbanSound8K dataset is utilized, containing a diverse set of urban sound recordings. Three windowing functions—Hann, Hamming, and Rectangular—are applied to extract spectrogram features, which are subsequently used to train a convolutional neural network (CNN) classifier. The trained models' performance is evaluated, and a comparative analysis is conducted to determine the impact of each windowing technique on classification accuracy. Results indicate that the Hann window achieves the highest accuracy of 59%, followed by the Hamming window at 57%, and the rectangular window at 54%. This study highlights the importance of appropriate window selection in spectrogram generation for improved audio classification accuracy.

INTRODUCTION

Sound classification is a crucial task in machine learning, with applications spanning speech recognition, environmental monitoring, and security surveillance. The ability to correctly classify urban sounds can enhance smart city initiatives, assist in noise pollution monitoring, and improve assistive technologies.

One of the key steps in audio classification is feature extraction, where raw waveform data is transformed into meaningful representations. The Short-Time Fourier Transform (STFT) is commonly used for this purpose, as it captures both time and frequency information in an audio signal. However, the effectiveness of STFT heavily depends on the windowing function applied before the transformation. Different windowing functions impact the spectral representation by controlling leakage and resolution trade-offs.

This study investigates three widely used windowing techniques—Hann, Hamming, and Rectangular—in the context of urban sound classification. The goal is to evaluate how these windowing methods affect spectrogram generation and, consequently, the performance of a deep learning-based classifier. The UrbanSound8K dataset is used to train and test a convolutional neural network (CNN) using spectrogram features derived from each windowing function. A comparative analysis is conducted to determine which windowing function yields the best classification accuracy

The dataset used, UrbanSound8K comprises 8732 labelled sound excerpts, categorized into ten classes representing different urban sounds, such as sirens, drilling, and car horns. The dataset provides metadata, including fold information for cross-validation purposes. Audio samples are loaded and processed at a sample rate of 22050 Hz, with a target duration of four seconds per sample.

METHODOLOGY

The methodology employed in this study consists of several key steps, starting with the preprocessing of the UrbanSound8K dataset, applying different windowing techniques,

generating spectrograms, and training a convolutional neural network (CNN) for classification. The audio data is first loaded and resampled to a consistent sample rate of 22050 Hz. Since the recordings vary in duration, all waveforms are either truncated or zero-padded to a fixed length of four seconds, ensuring uniform input dimensions for subsequent processing.

To transform the waveform data into frequency-domain representations, the Short-Time Fourier Transform (STFT) is applied using three different windowing functions: Hann, Hamming, and Rectangular. The Hann window smooths signal edges, minimizing spectral leakage by gradually reducing amplitude toward the frame's boundaries. The Hamming window functions similarly but retains slightly more signal energy by modifying the tapering function. The Rectangular window, in contrast, does not apply any smoothing, leading to increased spectral leakage and potential artifacts in the spectrogram representation. The STFT is computed with a window size of 512 samples and a hop length of 256 to balance time and frequency resolution.

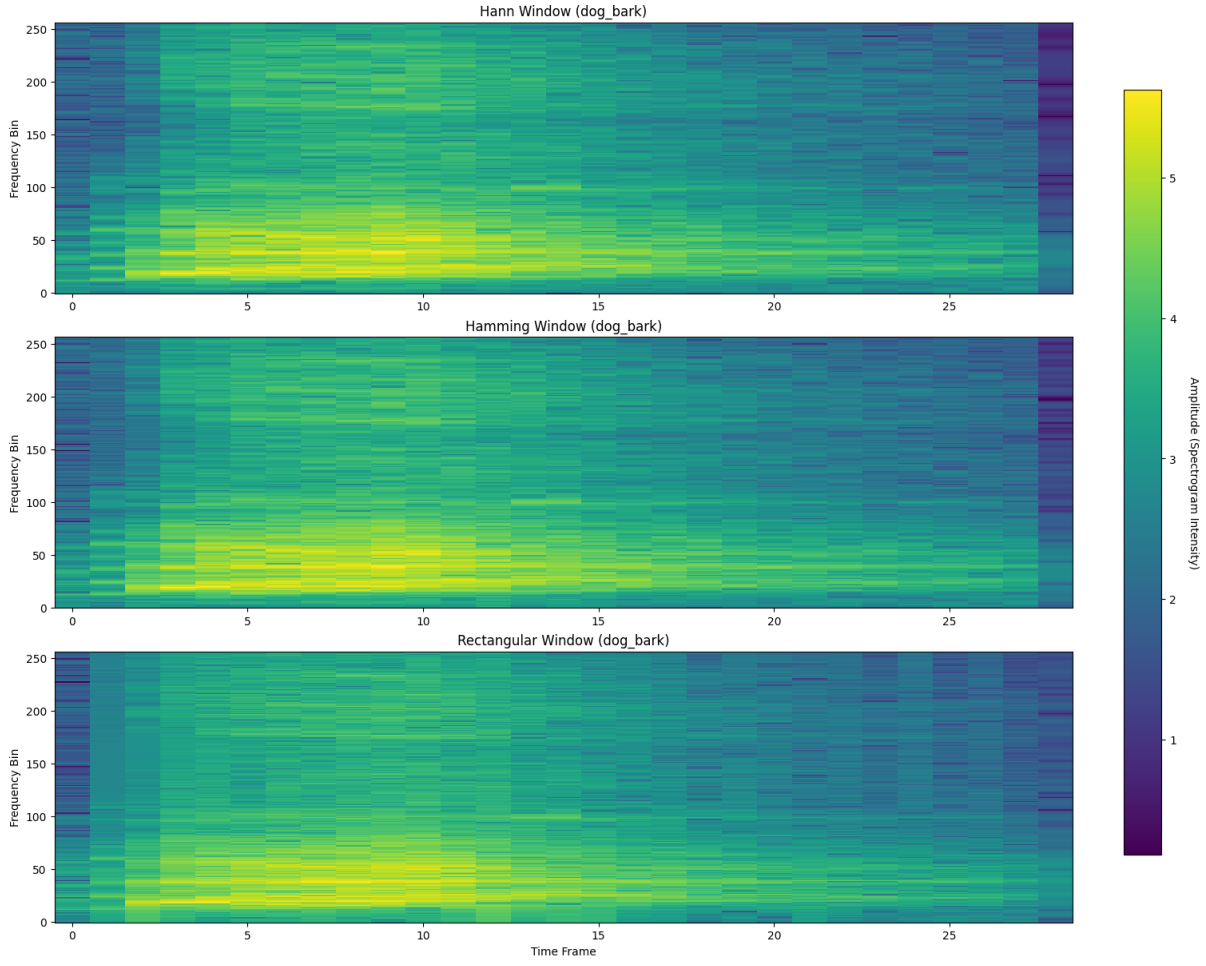

Figure 1 Spectrogram comparison of a 'dog bark' sound using Hann, Hamming, and Rectangular windows, illustrating the impact of windowing on spectral representation

The computed spectrograms are converted to logarithmic scale and normalized to enhance feature contrast. These spectrograms serve as input features for a convolutional neural network (CNN) classifier. The CNN architecture consists of two convolutional layers with 16 and 32 filters, respectively, followed by batch normalization and max pooling layers to reduce

dimensionality. Fully connected layers, incorporating dropout regularization, process the extracted features, culminating in a final SoftMax layer that classifies the input into one of the ten urban sound categories.

For model training and evaluation, the dataset is split into training (80%) and testing (20%) subsets. The Adam optimizer is used with a learning rate of 0.001, and cross-entropy loss serves as the objective function. The model undergoes multiple training epochs, and its performance is assessed using accuracy metrics and confusion matrices. The impact of different windowing techniques is analysed by comparing classification accuracies and inspecting spectrogram visualizations.

RESULTS AND DISCUSSIONS

The results of the study indicate that the Hann window achieves the highest classification accuracy at 59%, followed by the Hamming window at 57% and the rectangular window at 54%. The superior performance of the Hann window can be attributed to its effective reduction of spectral leakage, which enhances the quality of extracted spectrogram features. The Hamming window, while similar in design, retains slightly more energy, which contributes to its slightly lower accuracy compared to Hann. In contrast, the rectangular window, which does not apply smoothing, results in increased spectral leakage and the introduction of unwanted frequency artifacts, negatively impacting classification performance.

Spectrogram analysis further supports these findings. The Hann and Hamming window-generated spectrograms exhibit smoother and more distinct frequency representations, while the rectangular window spectrograms appear noisier with abrupt transitions. This suggests that smoother windowing functions aid in capturing more meaningful frequency features, which are crucial for CNN-based classification.

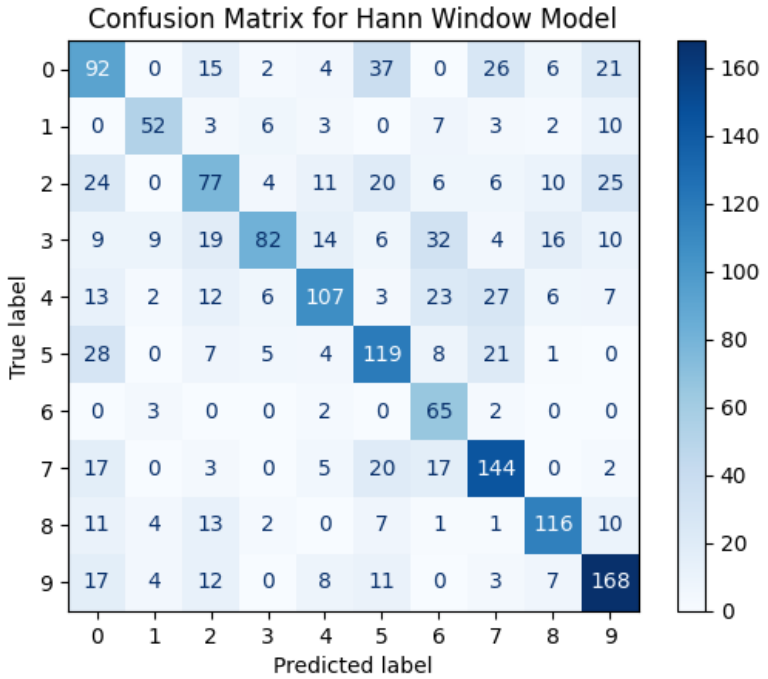

Figure 2 Confusion Matrix for Hann Window model

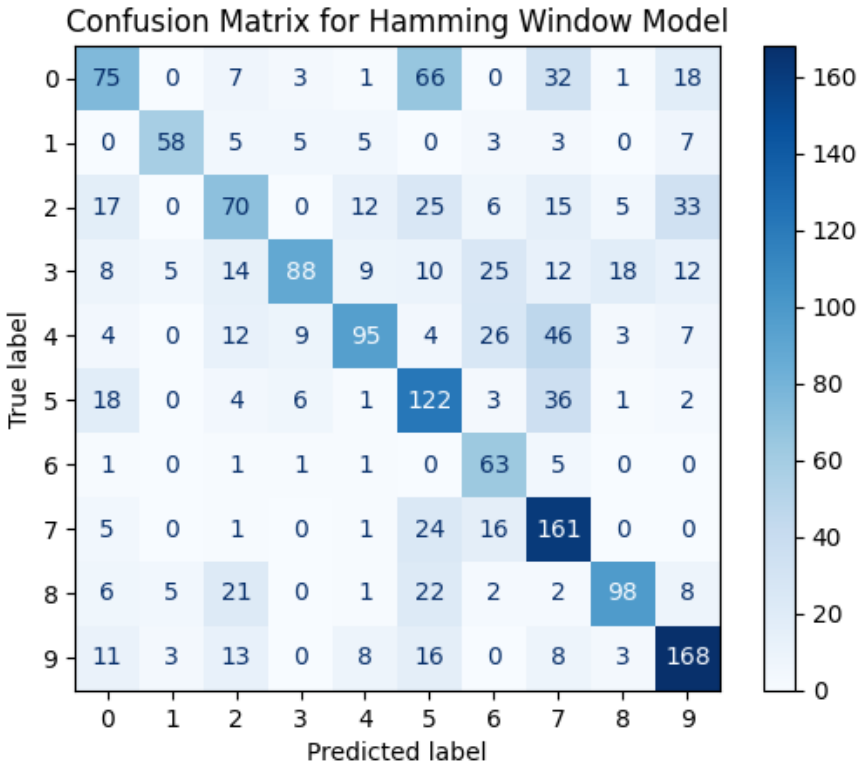

Figure 3 Confusion Matrix for Hamming Window model

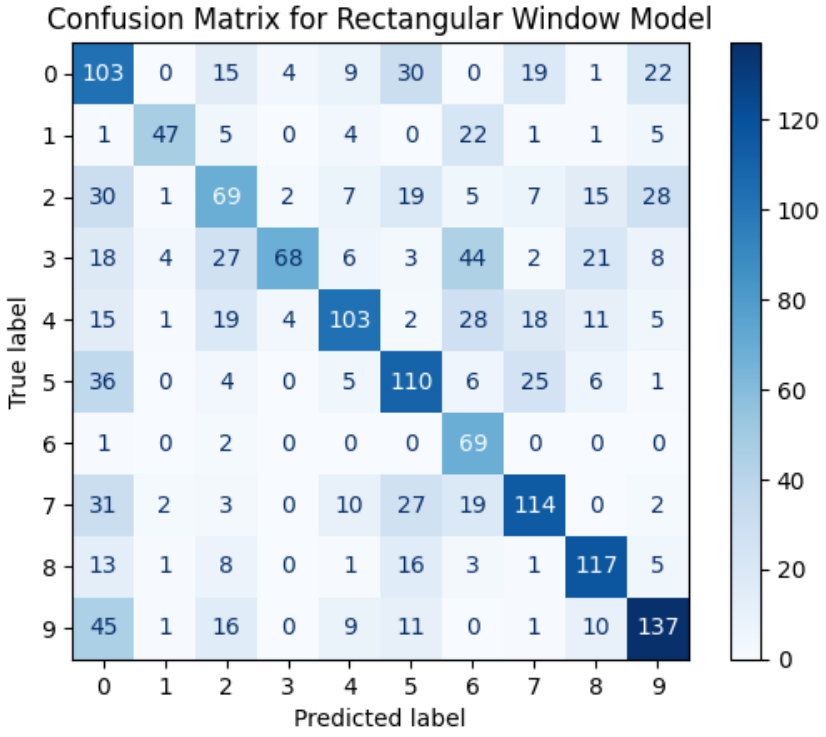

Figure 4 Confusion Matrix for Rectangular Window model

The confusion matrix analysis provides deeper insights into model performance. The Hann window model shows fewer misclassifications, particularly in distinguishing sounds with overlapping frequency components, such as sirens and drilling noises. The Hamming window model exhibits similar trends but with slightly higher misclassification rates. The Rectangular window model, due to its increased spectral distortion, struggles more with differentiating between acoustically similar sound classes.

These findings emphasize the importance of selecting an appropriate windowing function for spectrogram generation in audio classification tasks. While all three window types produce viable classification models, the Hann window emerges as the most effective in minimizing spectral leakage and enhancing model performance. Future work may explore additional windowing techniques, hybrid approaches, or adaptive windowing strategies to further optimize classification accuracy in real-world applications.

CONCLUSIONS

This study demonstrates that the choice of windowing function significantly impacts spectrogram quality and, consequently, classification accuracy in an urban sound classification task. The Hann window yields the best performance, making it the preferred choice for such applications. The findings highlight the importance of careful feature extraction choices when working with audio data, particularly in deep learning applications. Future research can extend this work by investigating additional windowing functions, such as Blackman or Kaiser windows, to determine their effectiveness in similar tasks. Additionally, experimenting with different deep learning architectures, including recurrent neural networks (RNNs) and transformers, could further enhance urban sound classification performance. Exploring real-world deployment strategies and optimizing computational efficiency for large-scale applications would also be valuable directions for future work.