Renesas Reality AI Project Report
Audio Challenge

Juejing Han
jjhan201707@gmail.com

Abstract—This study investigates the performance of two machine learning algorithms – Support Vector Machines (SVM) and K-Nearest Neighbors (KNN) – for audio classification with a balanced dataset comprising four categories: Air Conditioner, Car Horn, Engine Idling, and Siren. Three noise reduction techniques, including Wavelet Transform, Spectral Gating, and Bandpass Filtering, are applied to the original signals, with Bandpass Filtering demonstrating superior effectiveness compared to the others. Mel-Frequency Cepstral Coefficients (MFCCs) are extracted and scaled as features for model input. The performance of both algorithms is evaluated using 5-fold cross-validation, and SVM outperforms KNN in terms of accuracy, precision, recall, and F1-score. Both models achieve precise classification for the Air Conditioner class, suggesting this sound has distinct features with minimal overlap in feature space with other sound categories. Wall clock time analysis reveals that SVM has a higher computational cost than KNN due to its quadratic optimization process.

1 EXPERIMENT DESIGN

1.1 Dataset

The audio dataset has 4 categories (classes): Air Conditioner, Car Horn, Engine Idling, and Siren. Each class has 40 samples, most of which have a 4-second duration, and no missing data is identified. Therefore, the dataset is considered balanced.

1.2 Data Pre-processing (Noise Reduction & Feature Engineering)

Noise reduction techniques are applied to the original data before feature engineering. Common methods such as Wavelet Transform[1], Spectral Gating[2], and Bandpass Filtering[3] are explored for their effectiveness in audio processing. After comparing the results using an untuned linear SVM model (evaluated with 5-fold cross-validation), Bandpass Filtering is selected due to its superior performance (see Table 1). Given that the dataset is balanced, accuracy is used as the evaluation metric.

[1] https://www.ripublication.com/acst17/acstv10n10_15.pdf

[2] https://cic.du.ac.in/userfiles/downloads/Audio Denoising Using Spectral Gating.pdf

[3] https://ijtre.com/wp-content/uploads/2021/11/2014011122.pdf

Table 1—Noise Reduction Methods.

Denoising Methods	Spectral Gating	Wavelet Transform	Bandpass Filtering
Accuracy	0.79	0.88	0.96

Mel-Frequency Cepstral Coefficients (MFCCs) are used to capture the spectral characteristics of sound and are a widely adopted feature extraction technique in audio processing[4]. During the feature engineering, MFCCs are extracted, scaled, and used as model input.

1.3 Model & Setting

Two machine learning algorithms – Support Vector Machines (SVM) and K-Nearest Neighbors (KNN) – are employed in this study. The model selection is based on the characteristics of both the data and the algorithms. SVM is well-suited for small datasets with clear boundaries, and KNN is ideal for small datasets with distinguishable features. To ensure a more reliable evaluation, 5-fold cross-validation is applied. Additionally, Grid Search is utilized to optimize the models and enhance their performance.

2 RESULTS & ANALYSIS

2.1 Waveform & Spectrogram

2.1.1 *Air Conditioner*

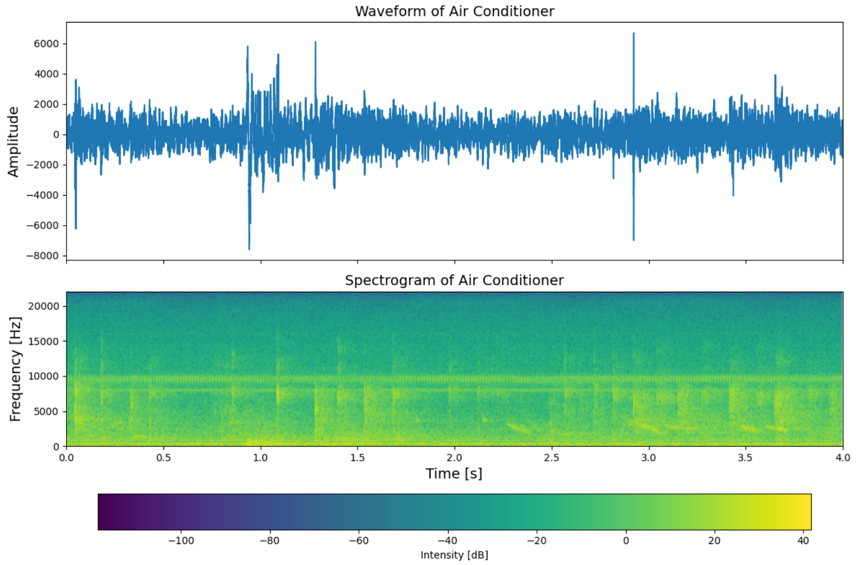

Figure 1—Plots for Air Conditioner

[4] https://repo.bg.wat.edu.pl/info/article/WAT23a0638100cc4f629fead136dc130268/

The waveform demonstrates a steady pattern with periodic spikes, and its amplitude ranges between -6,000 and +6,000.

The spectrogram reveals a consistent energy distribution, with most of the energy concentrated below 5,000 Hz, reflecting the steady nature of the sound. Occasional spikes in the high-frequency range extend to 10,000 Hz, indicating brief fluctuation in sound intensity. The intensity remains constant at low frequencies and reaches a maximum of around 40 dB, suggesting moderate loudness.

2.1.2 *Car Horn*

The waveform starts with a steady pattern and then a sudden peak, which reflects the sharp and loud nature of a car horn. Its amplitude ranges from -30,000 to +30,000.

The spectrogram demonstrates an increase in energy across the frequency range over time. The lower frequency bands (< 2,000 Hz) are especially prominent as the sound gets louder. High-frequency bands extend beyond 5,000 Hz with a maximum intensity exceeding 40 dB.

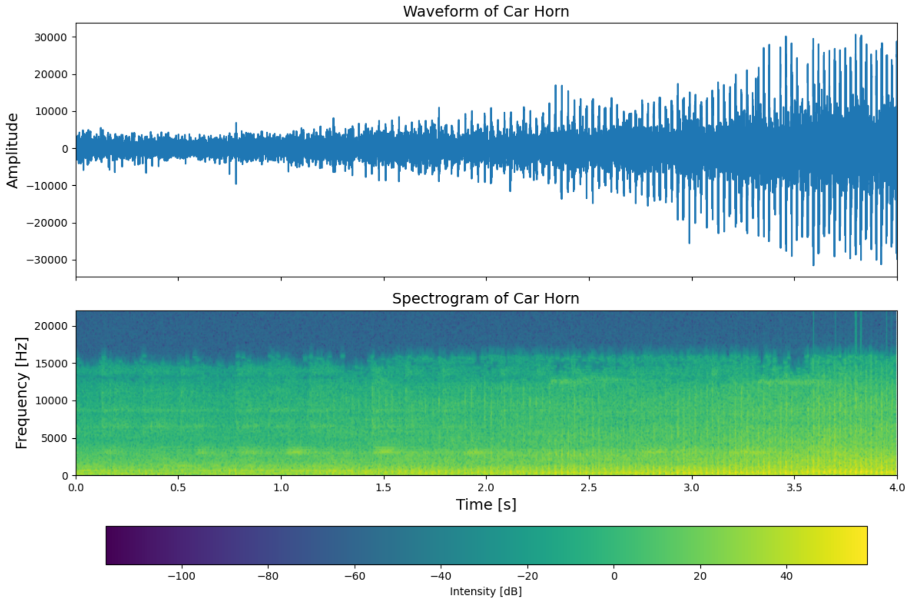

Figure 2—Plots for Car Horn

2.1.3 *Engine Idling*

The waveform displays a consistent and oscillatory pattern with peaks and valleys within the amplitude range of -800 to +800.

The spectrogram shows a steady distribution of energy in the low frequencies (< 1,000 Hz), indicating the rhythmic nature of engine idling, with maximum intensity around 20 dB.

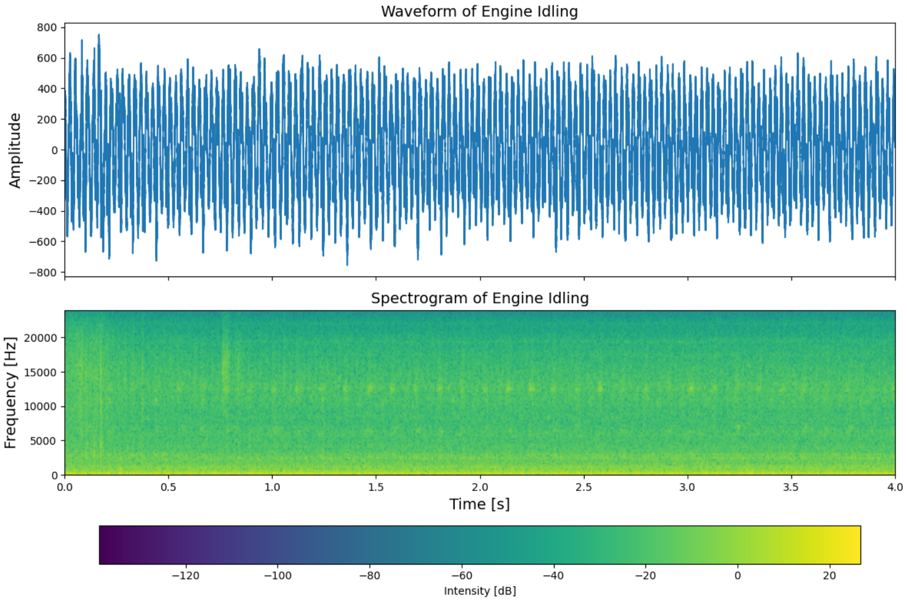

Figure 3—Plots for Engine Idling

2.1.4 *Siren*

The waveform demonstrates a rise and fall pattern with an amplitude range of -500 to +500, indicating the sound intensifies and diminishes in a repeated pattern.

The spectrogram reveals that most energy is concentrated below 2,000 Hz, with intensity peaking below 1,000 Hz. The maximum intensity is around 20 dB.

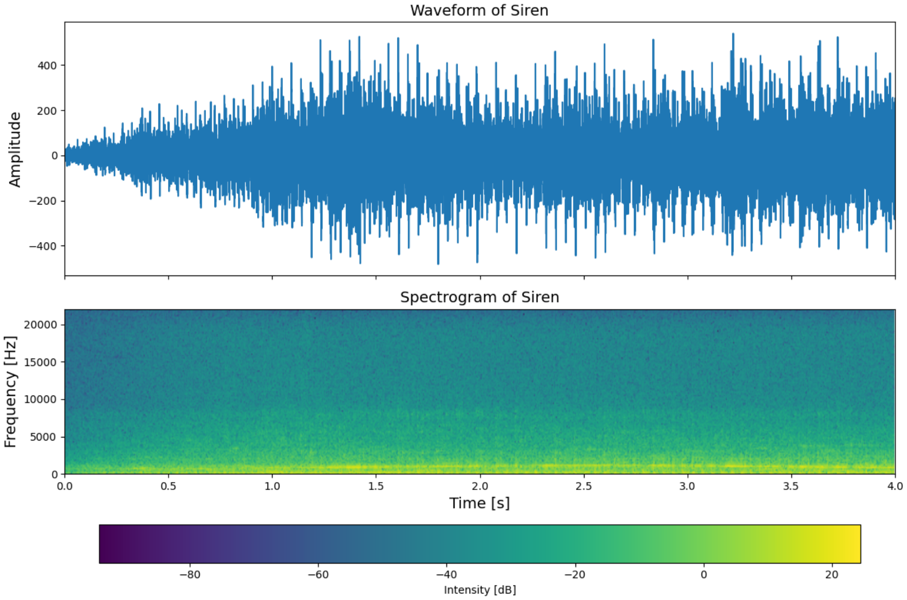

Figure 4—Plots for Siren

2.2 Model Performance

As discussed in Section 2.1, different classes exhibit distinct characteristics. Given that 1) the dataset is small (160 samples in total), and 2) the classes show clear and distinguishable patterns, SVM and KNN are well-suited for this classification task.

2.2.1 *SVM (Linear Kernel)*

As the sample size increases, the model's generalization performance improves, effectively reducing overfitting (Figure 5). After reaching 115 instances, the model attains a validation accuracy of approximately 0.95, with additional samples providing marginal improvement as the validation score plateaus.

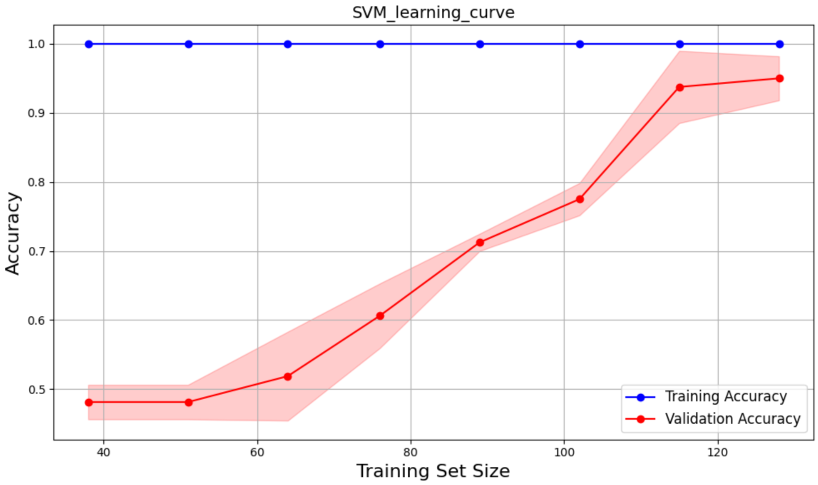

Figure 5—Learning Curve for SVM

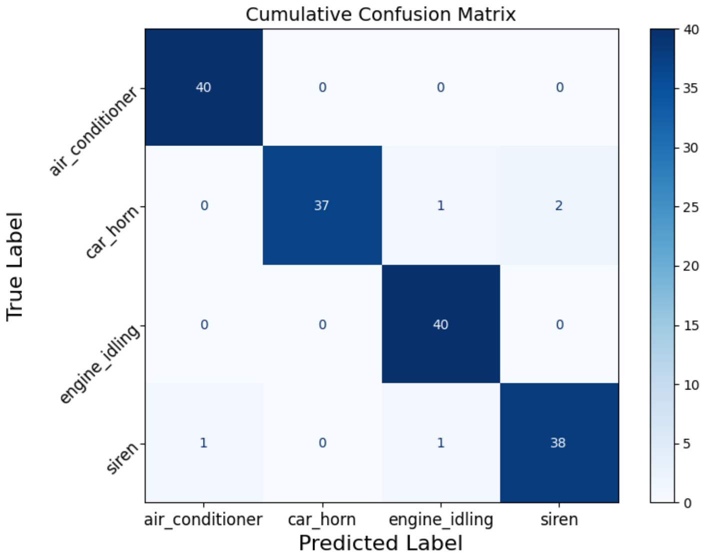

Figure 6—Cumulative Confusion Matrix for SVM

The cumulative confusion matrix across the 5 folds (Figure 6) demonstrates high performance, with an accuracy of 0.969, precision of 0.972, recall of 0.968, and an F1-score of 0.968, indicating strong overall model performance.

Both Air Conditioner and Engine Idling classes are classified correctly, indicating that these two classes have more distinct frequency patterns that the SVM model can easily distinguish. Car Horn samples are misclassified as 2 Siren and 1 Engine Idling, while Siren samples are misclassified as 1 Engine Idling and 1 Air Conditioner. This could be due to the feature set (MFCCs) failing to capture the distinctive characteristics of these sounds, or because the feature space may have a non-linear structure that the linear SVM model cannot fully capture.

2.2.2 *KNN*

As the sample size increases, the model's generalization performance improves. While more samples mitigate overfitting, the training score decreases (Figure 7). KNN is a lazy learner that stores the entire training set during the training phase. With a small training size, it only needs to memorize a few examples, leading to high training accuracy. However, as the training set grows, more diverse and potentially conflicting samples are introduced, making it harder for the model to accurately classify all the data points. On the other hand, a larger training set allows the model to learn the pattern better, leading to improved validation scores.

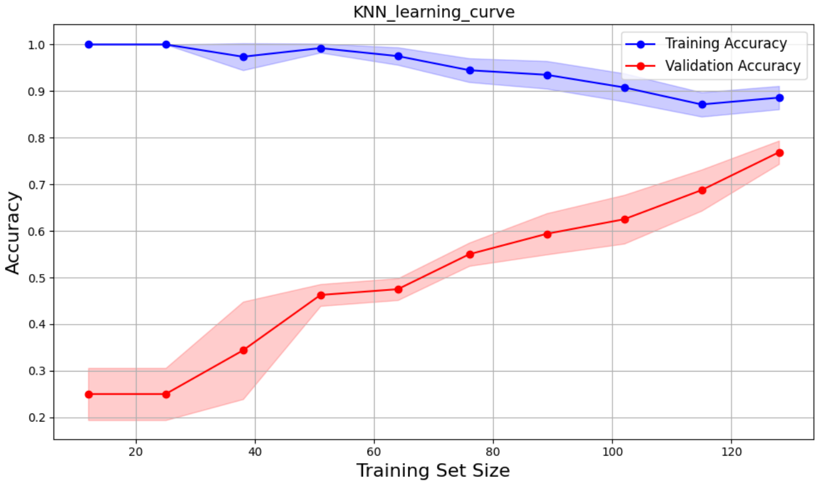

Figure 7—Learning Curve for KNN

The cumulative confusion matrix (Figure 8) also demonstrates the high performance of the KNN model, particularly with the Air Conditioner class. KNN yields an overall accuracy of 0.950, precision of 0.959, recall of 0.944, and an F1-score of 0.945. However, as an instance-based learning method, KNN slightly underperforms compared to SVM in this study.

KNN struggles to correctly identify Car Horn, resulting in 5 misclassifications. It misclassifies 2 Engine Idling samples as Car Horn and 1 Siren sample as Air Conditioner. KNN relies on proximity in feature space, and when the features of Car Horn and Engine Idling are not well-separated, it has difficulty distinguishing between them. Since KNN makes decisions based on the closest neighboring points, the presence of noise or outliers in the dataset can further contribute to these misclassifications.

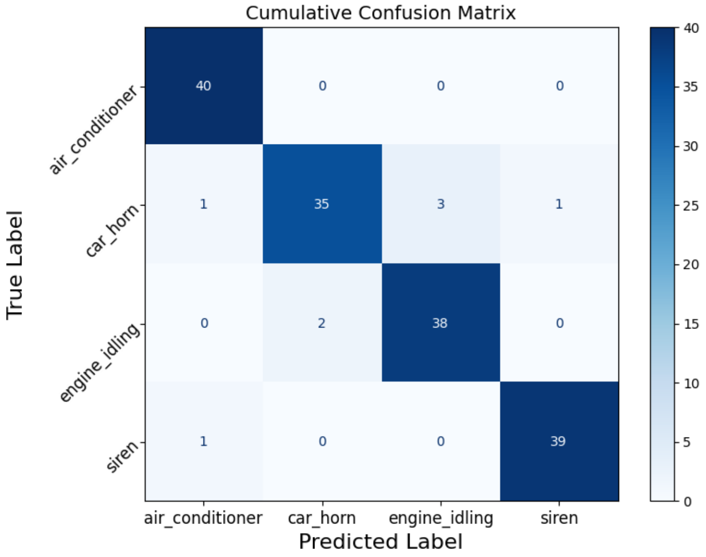

Figure 8—Cumulative Confusion Matrix for KNN

Both algorithms (SVM and KNN) achieve perfect classification for the Air Conditioner class, which may be attributed to two key factors: 1) the features (MFCCs) extracted from Air Conditioner sounds create a linearly separable feature space for SVM and a well-clustered feature space for KNN; 2) Air Conditioner sounds share fewer overlapping characteristics with other environmental sounds in the data, reducing the likelihood of misclassification.

2.2.3 *Wall Clock Time*

The overall computational cost (wall clock time) is compared between the two algorithms. SVM has a higher computational cost (3.44s) compared to KNN (1.17s). SVM must solve a convex quadratic optimization problem to identify the optimal hyperplane that separates the feature space. Even with a linear kernel, the training complexity of SVM can still be approximately O(n^2), where n is the training sample size. In contrast, KNN's training process only consists of storing the data, with most of its complexity arising during the testing phase.

3 CONCLUSIONS

This study evaluates the performance of SVM and KNN in classifying environmental sounds across four distinct categories. Both algorithms perform well, with SVM slightly outperforming KNN in terms of accuracy, precision, recall, and F1-score. The Air Conditioner class is precisely classified by both models, highlighting the distinct and easily separable characteristics of its features. SVM, while more computationally intensive due to its quadratic optimization, demonstrates more robust generalization, particularly with increasing training size. On the other hand, KNN displays faster training and testing, making it more efficient for smaller datasets, though slightly less effective in managing diverse and potentially conflicting samples. Overall, the study demonstrates the effectiveness of both algorithms for environmental sound classification, and SVM stands out as the more accurate and consistent model for this task.

4 FUTURE WORK

While this study successfully develops multi-class classifiers for predicting urban sounds, several areas of future work could further enhance its performance:

Resampling: Given varying sample rates of the audio files, resampling techniques could be introduced in future work to ensure consistency across different classes.

Improved Feature Extraction: Beyond MFCCs, techniques like Mel-Spectrograms and Chroma or combining features could boost performance.

Large and More Complex Dataset: Increasing the dataset size and diversity by incorporating sounds from various environments would improve model generalization and robustness in real-world applications.

Advanced Classifiers: Exploring non-linear SVMs, Convolutional Neural Networks (CNNs), or ensemble methods could address overlapping sound classes and improve performance.

Error Analysis and Confusion Resolution: A more in-depth analysis of misclassifications could guide hyperparameter tuning (or the use of hybrid models) to reduce errors.