OPEN ACCESS

Kashf Journal of Multidisciplinary Research
Vol: 01 Issue 12 (2024)
P-ISSN: 3007-1992 E-ISSN: 3007-200X
https://kjmr.com.pk

INTELLIGENT EMERGENCY VEHICLE SOUND CLASSIFICATION FOR PUBLIC SAFETY

Hira Farooq
Institute of Computer Science, Khwaja Fareed University of Engineering and IT, Rahim Yar Khan

Qamar Hafeez
Department of Computer Science, The Islamia University of Bahawalpur

Muhammad Shadab Alam Hashmi
Institute of Computer Science, Khwaja Fareed University of Engineering and IT, Rahim Yar Khan

Muhammad Mohsin
Institute of Computer Science, Khwaja Fareed University of Engineering and IT, Rahim Yar Khan

Talha Farooq Khan*
Department of Computer Science & IT, University of Southern Punjab Multan

Corresponding Author: Talha Farooq Khan (talhafarooq@isp.edu.pk)

Article Info

Abstract.

Traffic congestion in urban areas can be a nightmare for emergency vehicles as they are slowed down by the traffic condition, thus putting the patients' lives at risk for whom medical attention is urgently needed. Traditional visual and acoustic identification methods, such as flashing lights and sirens often fall short due to multiple factors like driver distraction, obstruction in the line of sight either by any vehicle or building, and even the advanced soundproofing features of modern vehicles could be a reason. This research is all about designing the correct and efficient real-time system that detects and distinguishes between emergency vehicle sounds so drivers, pedestrians, and also the management systems in their vicinity have prompt recognition and reactions to those sounds. To accomplish this, the proposed solution utilizes acoustic analysis along with sophisticated, cutting-edge algorithms by applying features extraction using Mel-frequency cepstral coefficients (MFCC). A wide range of machine learning algorithms, including Random Forest (RF), Support Vector Machine (SVM), Multilayer Perceptron (MLP), Extra Trees Classifier (ETC), and AdaBoost, were trained and tested using a comprehensive dataset consisting of emergency vehicle sirens and background traffic noise. Among them, the accuracy of Random Forest classifier is the highest, which reaches 99.17%, and AdaBoost classifier has similar performance. In this way, this system uses sound-based detection to enhance emergency response, public safety, and traffic management with innovative acoustic monitoring and analysis. The implementation of the system will streamline emergency operations and improve the efficiency and safety of urban traffic.

Keywords: Emergency Vehicle, MFCC, Audio Classification, AdaBoost

Introduction

With every passing year, the number of humans is multiplying exponentially, which in turn requires more transport vehicles to carry them. This places us in the problem of traffic congestion, especially during rush hours, where we can witness the distressing incidents of ambulances or other emergency vehicles stuck in lengthy queues or getting late to serve. These emergency vehicles help save the lives of persons at risk and also play a crucial role in controlling or preventing any kind of calamity. But these traffic jams can severely hinder their passage which worsens an already exacerbated situation.

Typically, the traditional method used in emergency vehicles is sirens. Their purpose is to emit sound signals at different wavelengths to alert drivers on the road and pedestrians to clear the way. However, there are many instances where drivers of private cars may not perceive nearby sirens due to multiple factors such as in-car audio interference, the soundproofing capabilities of modern cars, or other distractions. These factors can bring unintentional delays to emergency services and potential traffic accidents due to inadequate communication and cooperation.

In general, distress sounds are categorized as a subset of International Organization for Standardization (ISO)-standardized auditory danger signals (Tran & Tsai, 2020). ISO 7731 stipulates fundamental requirements for alarm noises. Nevertheless, standards and regulations regarding distress noises differ among nations. For instance, it is commonplace for the United States and New Zealand to utilize phasers and wails, whereas England predominantly employs the two-tone pneumatic horn. In Taiwan, the frequency range of fire truck sirens is continuously shifting from 650-750 Hz to 1450-1550 Hz, whereas ambulance sirens feature alternating tones: 650-750 Hz for the initial tone and 900-1000 Hz for the second.

Similarly, ambulance sirens in Japan are governed by specific regulations that mandate the use of two masses of 770 Hz and 960 Hz with a 1.3-second repetition period. Two tons of 392 Hz and 660 Hz are frequently employed by ambulances and fire vehicles in Europe, whereas two tons of 466 Hz and 622 Hz are utilized by police cars (Tran & Tsai, 2021). Thus, the traditional methods are now outdated, as self-driven automobiles are replacing the auto sector, and the technology for detecting emergency vehicles should be updated with time.

Literature Review

Since the emergence of self-driven automobiles like Tesla and the increase in population, we have been trying to shift to AI-based solutions for detecting ambulances and other emergency vehicles. A lot of research has already been done by scientists, and we are going to explore each one of them one by one.

For instance, Usaid et al. (2022) used acoustic-based detection methods to develop a Multi-Layer Perceptron (MLP) model using an emergency siren dataset. This system was extremely feasible because it achieved 90% accuracy with just 300 files. Similarly, Lisov et al. (2023) used the same method but with Convolutional Neural Networks (CNNs) to analyze spectrograms and achieved an accuracy of 93.3% in identifying emergency vehicle sounds with rapid recognition speeds.

Patel et al. (2022) proposed a system integrating neural network-based siren detection with 97.2% accuracy by utilizing IoT (Internet of Things) devices and GPS to create temporary emergency lanes. This research uses visual indicators on the route to alert traffic, ensuring faster ambulance transit. Likewise, some advanced AI techniques like CNNs with Mel-Frequency Cepstral Coefficients for sound conversion were used by Sathruhan et al. (2022), achieving 93% precision in siren detection. Tran & Tsai (2021) introduced an audio-visual detection system (AV-EVD) by combining YOLO-EVD for image-based detection and WaveResNet for sound-based emergency vehicle detection, achieving accuracy exceeding 95% and 98%, respectively.

Simplified and efficient models such as Extreme Learning Machines (ELMs) were leveraged by Islam & Abdel-Aty (2022) for audio-only detection at

signalized intersections. This approach struck a balance between simplicity and a high accuracy of 97% in quick real-time learning.

Mittal & Chawla (2023) developed an ensemble deep neural network model that was optimized configurations of CNN, Dense, and RNN models, resulting in an accuracy of 98.7%. Siren Net (Tran & Tsai, 2020) uses hybrid Wave Net and MLNet architectures to achieve the high accuracy for siren detection from a variety of traffic sounds. In noisy environments, this method maintains an accuracy of 98.24%.

For spectrogram-based localization, Marchegiani & Newman (2022) used image processing techniques on spectrograms for emergency vehicle detection and sound source localization. Their system achieved a 94% classification rate with minimal localization errors, even under noisy conditions. Also the, Cantarini et al. (2022) leveraged prototype networks with few-shot metric learning for siren detection. Using limited data and noise-filtering techniques, their system achieved SVM accuracy of 95% and CNN accuracy of 83–87%.Eventually , Dontabhaktuni et al., 2024 used the MFCC methodology but achieved low frequency in SVM as compare to us.

From this analysis, several gaps were identified, such as the limitation in the availability and diversity of datasets. Many studies relied on custom or specific datasets, which may not fully represent real-world scenarios and limit the generalizability of the models. The second concern is the potential for overfitting, where models may perform exceptionally well on the specific data they were trained on but struggle to generalize to new and unseen data. Therefore, we aim to address this void in the upcoming methodology.

Table 1 : Literature Review

Study	Model/Method	Dataset	Accuracy/Performance	Pros	Cons
(Usaid et al., 2022)	MLP	Emergency vehicle siren sounds and road noise	90% accuracy with 300 files	Pros: Simple model, reasonable accuracy	Cons: Limited dataset, may not generalize
(Lisov et al., 2023)	CNNs	Emergency Vehicle Siren Sounds dataset (siren sounds and city	Average accuracy of 93.3%, rapid recognition speed of 0.0004±5% seconds	Pros: High accuracy, fast recognition speed	Cons: Dataset specificity, potential overfitting
(Patel et al., 2022)	Neural network-based siren classifier, IoT devices	Custom dataset, GPS-based mobile app	Siren classifier accuracy of 97.2%, IoT devices for visual indicators and temporary emergency lanes	Pros: High accuracy, integration with IoT devices	Cons: Dataset collection challenges, reliance on mobile app
(Sathruhan et al., 2022)	CNN	Brief audio signals	93% precision	Pros: Good precision, efficient model	Cons: Limitation of information
(Tran & Tsai, 2020)	SirenNet (WaveNet and MLNet)	Traffic soundscape dataset	SirenNet accuracy of 98.24% for siren sounds	Pros: High accuracy for siren sounds, utilization of audio and ML models	Cons: Limited details, potential dataset bias
(Walden et al., 2022)	CNN	UrbanSound8k dataset	Classification rate of 97.82% for various audio categories	Pros: High classification rate, diverse audio categories	Cons: Dataset specificity, potential overfitting
(Fatimah et al., 2020)	Machine learning models (kNN, SVM, ensemble bagged	Siren sounds, traffic sound files, internet data	Accuracy of 98.49% using selected features	Pros: High accuracy, utilization of multiple models	Cons: Dataset diversity, potential
(Marchegiani & Newman, 2022)	Image processing with CNNs	Stereo signals converted to spectrograms	Average classification rate of 94%, median absolute error of 7.5° for sound source localization	Pros: Good classification rate, sound source localization	Cons: Limited information provided
(Tran & Tsai, 2021)	YOLO-EVD, WaveResNet	Custom dataset of images and audio	YOLO-EVD: mean average precision of 95.5%, WaveResNet: >98% accuracy, AV-EVD system with minimal misdetection rate of 1.54%	Pros: Comprehensive system, high accuracy	Cons: Complex architecture, dataset specificity
(Islam & Abdel-Aty, 2022)	ELM	Audio data	Estimated accuracy of 97%	Pros: Estimated high accuracy	Cons: Lack of specific performance metrics, limited details
(Mittal & Chawla, 2023)	Ensemble model (dense layer, CNN, RNN)	Google Audio set ontology dataset	Ensemble model accuracy of 98.7%	Pros: High accuracy, ensemble approach	Cons: Dataset specificity, potential complexity
(Cantarini et al., 2022)	Few-shot metric learning, prototypical	Publicly available or synthetic data	AUPRC scores of 0.86 (unfiltered) and 0.91 (filtered)	Pros: Effective few-shot learning, good AUPRC scores	Cons: Limited details, potential
(Cantarini et al., 2022)	Sound detection module (SVM), Deep Learning CNN	Audio data, camera images	SVM accuracy of 95%, CNN accuracy of 83-87%	Pros: Good SVM accuracy, integration of audio and visual data	Cons: Lower CNN accuracy, potential limitations in
Dontabhaktuni et al., 2024	SVM, RF, KNN, AdaBoost, LSTM	Emergency vehicle siren sounds and other audio sources	SVM: 99.5%, RF: 98.5%, KNN: 98.5%, AdaBoost: 96.0%, LSTM: 93.0%; Ensemble: 99.5%; LOOCV: SVM/RF - 98.5%	Pros: Combines temporal and spectral features, utilizes real-world augmentations (Doppler effect), high accuracy.	Cons: Computationally expensive, tuning required

Methodology

In order to analyze emergency vehicle sounds in traffic, we used a dataset from Kaggle that underwent preprocessing to ensure consistency and computational efficiency. The audio files were standardized using resizing techniques and transformed into structured representations through Mel-Frequency Cepstral Coefficients (MFCC), which were responsible for capturing the essential sound features of emergency vehicles, such as timbre and pitch.

These extracted features were organized into a structured dataset to split it further into the training and testing sets. After that different machine learning models which includes the Support Vector Machines (SVM), K-Nearest Neighbors (KNN), decision trees, random forests, and Multi-Layer Perceptron (MLP), were trained using the training set. Later the Hyperparameter optimization was employed using grid and random search to enhance model performance.

Eventually, the models were evaluated on the testing set, and the one with the highest classification accuracy was selected as the final model for emergency vehicle sound classification. A systematic flow, starting from data acquisition, followed by preprocessing, feature extraction, model training, hyperparameter tuning, and ending with evaluation, was followed to ensure reliable and replicable results. See methodology diagram in figure 1.

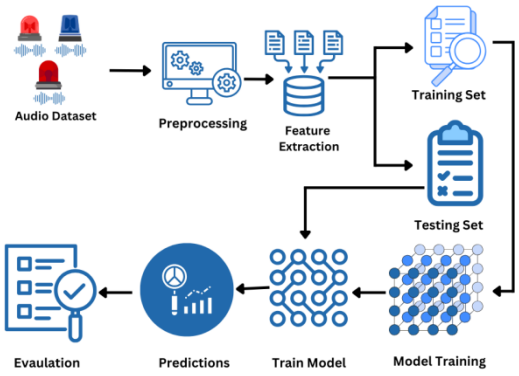

Figure 1: Methodology Diagram.

Dataset

Dataset which is being used here was publicly available on Kaggle collection with the name of "Emergency Vehicle Siren Sounds", having diverse collections of audio recordings, specifically for Ambulance and Firetrucks. Each file in dataset was in WAV format, with a fixed duration of 3 seconds. Also, it is divided into three categories: Ambulance, Firetruck, and Traffic. Each category contains 200 audio files, resulting in a balanced distribution of samples across the different classes as shown in Figure 2.

For each audio file, there are corresponding spectrogram images to have a visual representation of frequency content of the audio signals, resulting in a total of 200 spectrogram images per audio file. Here, we had the option to augment the audio files, similar to Hashmi et al., 2024. However, they worked on deep learning, whereas we focused on classic machine learning, so we had to drop that idea.

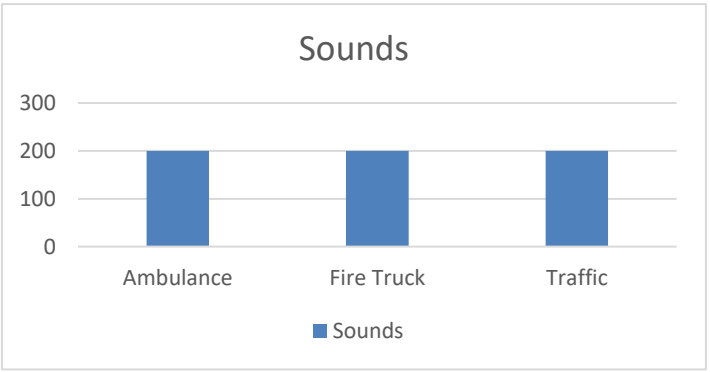

Figure Error! No text of specified style in document.**: Distribution of the Dataset across different classes.**

Preprocessing

Audio files often have varying samples rate which leads to inconsistencies during the data processing. In

order to cope with it , we apply resampling across all files which ensures consistent time resolutions and compatibility for analysis . The sample rate, measured in Hertz (Hz) represent the number of audio samples recorded per second. Standardizing this rate is crucial for extracting reliable features like MFCCs, which depend on uniform time scale. So, in order to do that, we used "librosa.resample", where audio signals are resampled to a target rate of 22,500 Hz , while preserving their integrity and essential characteristics , helping us to enhance the dataset consistency , improving the quality of feature extraction, modeling , and classifications task.

Feature Extraction

After the preprocessing comes the step of Feature Extraction for sound classification which is later used for the identification of the key attributes from Audio Signals. In acoustic analysis different techniques such as the wavelet analysis , chroma features, centroid-based methods, and Mel-Frequency Cepstral Coeffiecnt (MFCC) ensures the accurate analysis of audio datasets. In our study, features were extracted from emergency vehicle audio signals to facilitate classification tasks.

Here the Mel-Frequency Cepstral Coefficients (MFCCs) are used because of their ability to compactly represent the spectral shape of audio signals. The process involves transforming the audio Singal into a perceptual Mel scale and extracting coefficient using a combination of spectral analysis and the Discrete Cosine Transform. Major steps to extract MFCC features obtained from (Wu et al., 2018) are shown in Figure 3.

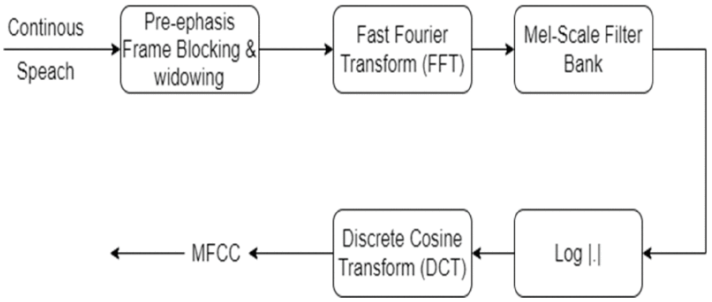

Figure Error! No text of specified style in document.**: Steps to Extract MFCC features.**

On the other hand, spectrograms provide a visual representation of an audio signal's frequency and amplitude over time. This study uses Librosa, a Python library, to dynamically generate Mel spectrograms by aligning the audio data with perceptual attributes on the Mel scale. This approach saved storage space by transforming the audio files into Mel spectrograms on the fly, whereas, in traditional methods, they were pre-stored on the disk for computation. There is no single formula in the literature for the Mel scale, but one of the formulas mentioned by O'Shaughnessy (1987) is presented here.

$$mel(f) = 2592 \times log10\left(1 + \frac{f}{100}\right) \quad \text{Equ: 1}$$

After acquiring the Mel scale, Mel filter banks were formed and multiplied with previously acquired spectrograms to produce Mel spectrograms (Rabiner & Schafer, 2010).

Classification Algorithms

This study utilizes five classification models based on their proven effectiveness after critically analysis it with the textual data available in the related literature.

The models chosen are Random Forest (RF), Extra Tree Classifier (ETC), Support Vector Machine (SVM), Multi-layer Perceptron (MLP), AdaBoost, and Logistic Regression (LR). These models were customized with hyperparameter tuning to achieve optimal accuracy.

For Instance, the Random Forest is an ensemble model that combine the predictions of multiple decision trees. Each tree makes a prediction and the final output is determined by majority voting. Here we utilized 100 decision trees with a maximum depth of 100. Similarly , the Extra Tree Classifier is an ensemble model that creates multiple decision trees . It differs from other tree-based methods by choosing random cut-points and using the entire training sample for tree construction. This randomness helps reduce over fitting and improve performance.

SVM, a powerful supervised learning algorithm, uses kernel functions to separate data points in an n-dimensional space and identifies the optimal boundary between classes. Likewise, the MLP is a type of neural network which learns complex patterns through multiple hidden layers and adjusts weights during training using backpropagation. Eventually we have AdaBoost which improves classification accuracy by combining weak learners, focusing on the errors of previous models and giving more weight to accurate predictions. All these models are optimized with hyperparameter tuning to enhance performance in the classification task.

Evaluation

The dataset for our research, which was downloaded from Kaggle, was split into training and testing sets using a 70:30 ratio, respectively. For features extraction during the feature extraction step: We used Mel-Frequency Cepstral Coefficients (MFCC). For the MFCC features, two sets were further generated—one with 20 coefficients and another with 40 coefficients. To evaluate the effectiveness and reliability of various machine learning classifiers, we mentioned before, these were trained and tested on three distinct datasets: 20 MFCC features, 40 MFCC features. Later, the model evaluation was conducted using multiple evaluation metrics which includes accuracy, precision, recall, and F1 score, supplemented by 10-fold cross-validation. This comprehensive and structured approach provided valuable insights into the classifiers' ability to differentiate and classify audio samples accurately based on the extracted features.

Eventually, the different machine learning classifiers were trained using 40 Mel-Frequency Cepstral Coefficient features in a grid search approach for hyperparameter optimization, which contains the following specification (see table 2):

Table 2: Hyperparameters of classifiers.

Model	Hyperparameters
RF	random state=142, max_depth=25, n_estimators=50
SVM	kernel='linear', C = 1.0, cache size=2000
MLP	hidden_layer_sizes=(100,), activation='relu', solver='adam', alpha=0.0001, max_iter=50
ETC	random_state=142, max_depth=50
AdaBoost	Extra Trees Classifier(n_estimators=50, max_depth=50, random_state=0),n_estimators=50

Then, to evaluate the model, a 10-fold cross-validation technique was utilized, and the end

results, including accuracy and other relevant metrics, are summarized below.

From Table 3, it is evident that the Random Forest (RF) classifier performed exceptionally well by achieving a threshold of 0.9917. Particularly because this model holds the ability to handle high-dimensional data and can capture complex relationships between features. Likewise, its precision, recall, and F1 score values also indicate strong reliability in distinguishing between classes and classifying them precisely.

Table 3: Classification Accuracy with 40 features.

Model	Accuracy	F1	Precision	Recall	K-fold Accuracy
RF	0.9917	0.99	0.99	0.99	0.97±0.02
SVM	0.9583	0.96	0.96	0.96	0.97±0.02
MLP	0.9833	0.98	0.98	0.98	0.98±0.01
ETC	0.9833	0.98	0.98	0.98	0.98±0.01
Ada Boost	0.9917	0.99	0.99	0.99	0.98±0.01

On the other hand, the Support Vector Machine (SVM) also performed well by achieving an accuracy of 0.9583. This machine learning classifier is known for handling non-linear data effectively by leveraging the kernel trick to separate classes and performing consistently across evaluation metrics. Additionally, it had a balanced F1 score, precision, and recall rate, which further confirm its effectiveness.

The Multi-Layer Perceptron (MLP) and Extra Trees Classifier (ETC) models achieved similar performance levels with precisely the same accuracy of 0.9833. However, the AdaBoost classifier competed with the RF model's accuracy of 0.9917. As an ensemble learning method that emphasizes misclassified instances during training, AdaBoost's performance reflects its ability to refine predictions and enhance accuracy. Its F1-score, precision, and recall values of 0.99 further validate its effectiveness in contrast to other models.

Thus, the results across all models demonstrated consistent reliability with minimal standard deviations ranging from ±0.01 to ±0.02 in K-fold validations. Results of testing can be seen in confusion matrixes (figure 4 to 8) where 0 represents Ambulance, 1 represents Fire Truck, and 2 Represents Traffic).

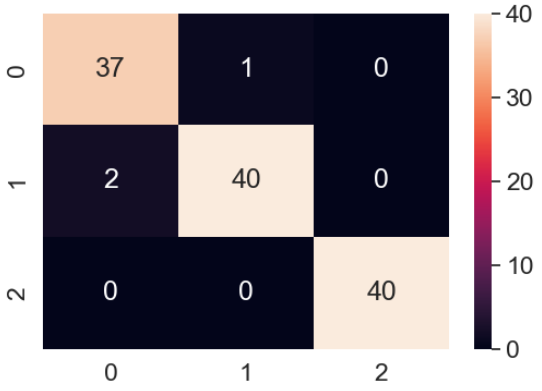

Figure Error! No text of specified style in document.**: Confusion Matrix of RF.**

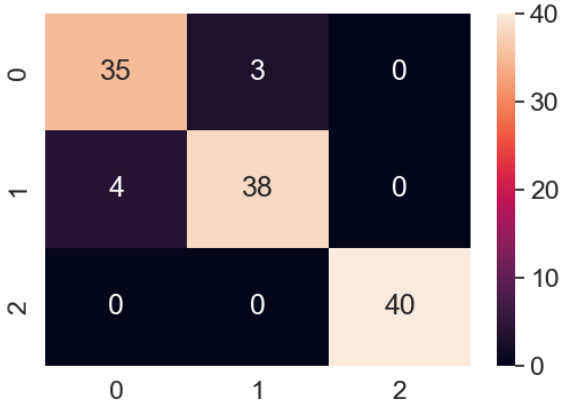

Figure 5: Confusion Matrix of SVM.

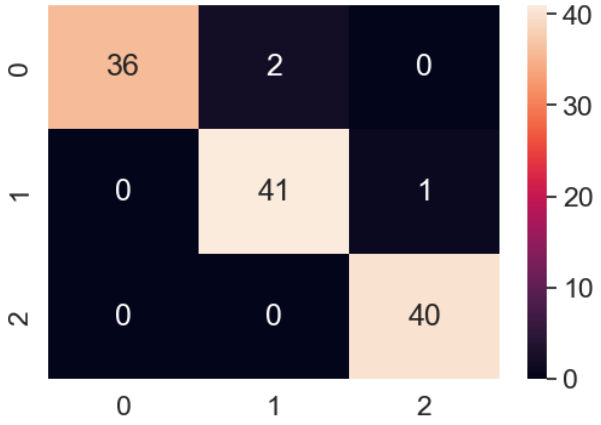

Figure 6: Confusion Matrix of MLP.

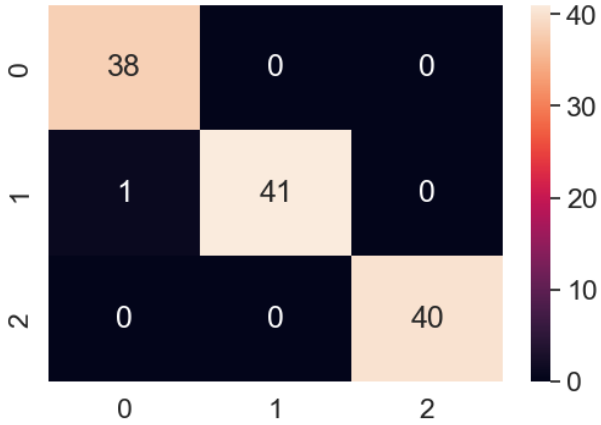

Figure 7 Confusion Matrix of ETC.

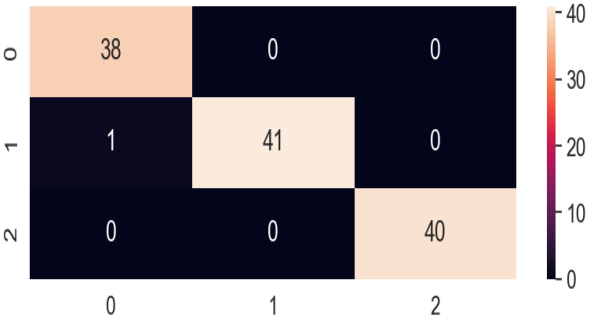

Figure 8: Confusion Matrix of ABD.

Result with 20 MFCC Features

Parallel to the 40 MFCC features, we trained our machine learning classifiers with 20 MFCC features as well, using the same hyperparameter optimization via the grid search method given in Table 4.1.

Again, a 10-fold cross-validation technique was applied to evaluate the models to ensure reliable performance assessments, the results of which are given in Table 4.

Table Error! No text of specified style in document.**: Classification Matrix of results with 20 MFCC features**

Classifier	Accuracy	Precision	Recall	F1	K-fold Accuracy
RF	0.95	0.95	0.96	0.95	0.96 ± 0.02
SVM	0.93	0.94	0.94	0.94	0.94 ± 0.03
MLP	0.975	0.97	0.98	0.98	0.96 ± 0.02
ETC	0.98	0.98	0.99	0.98	0.98 ± 0.02
ADB	0.958	0.96	0.96	0.96	0.98 ± 0.02

Let's decompose the results of the performance of classifiers with 20 MFCC features: Once again, Random Forest (RF) succeeded in providing excellent accuracy equal to 0.95 and F1-score of 0.95. Support Vector Machine (SVM) is not that behind in results with accuracy equal to 0.93 and an F1-score of 0.94. In case of MLP, results obtained 0.975 accuracy with 0.98 as the F1-score. The Extra Trees Classifier (ETC) matched the MLP level with 0.98 accuracy and F1-score. AdaBoost (ADB) also held its own, with 0.958 accuracy and a 0.96 F1-score.

Overall, MLP and ETC came out on top with slightly better numbers with 20 MFCC features as compare to 40 MFCC features, but RF, SVM, and ADB were no slouches either. All in all, it's clear these models can handle the task well, especially with the 20 MFCC features.

Results of testing can be seen in confusion matrixes (figure 9 to 13) where 0 represents Ambulance, 1 represents Fire Truck, and 2 Represents Traffic).

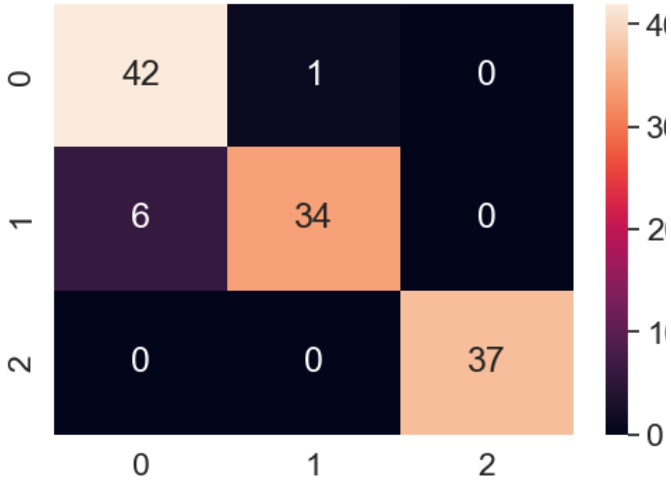

Figure 9: Confusion Matrix of RF.

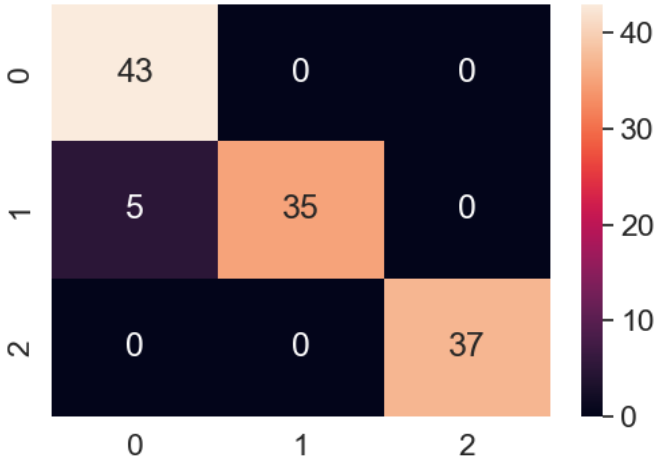

Figure 10 Confusion Matrix of SVM.

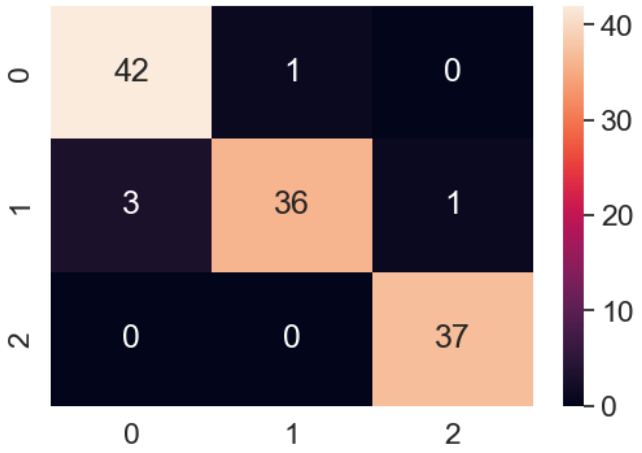

Figure 11: Confusion Matrix of MLP.

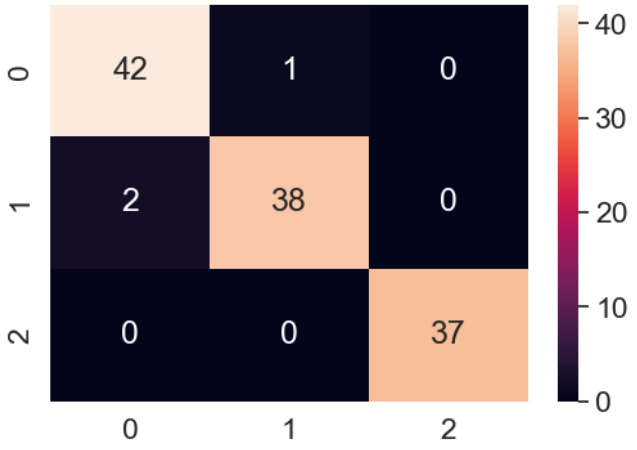

Figure 12: Confusion Matrix of ETC.

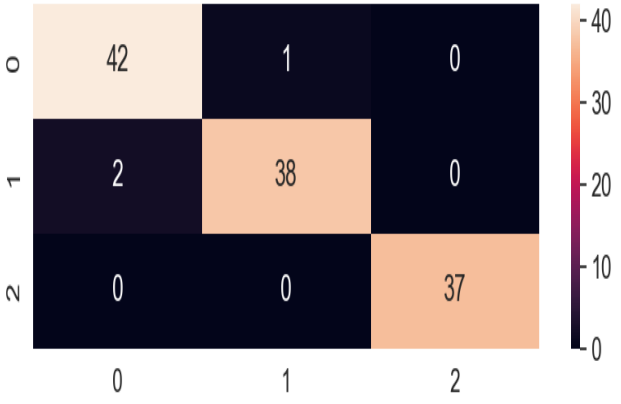

Figure Error! No text of specified style in document.**: Confusion Matrix of ADB.**

Conclusion

In this study, we have developed a real-time system for detecting and classifying emergency vehicle sounds in urban areas using acoustic analysis and machine learning. After following the entire methodology and summarizing the results, we found that the Random Forest classifier achieved the highest accuracy, followed by another model named AdaBoost. Both of these models demonstrate the potential of audio-based methods to complement traditional visual identification of emergency vehicles.

We can offer significant applications with this system, including the improvement of emergency response time, optimization of traffic management, enhancing public safety, and enabling advanced technological solutions. Since we have used a diverse dataset available over the internet, expanding the dataset in the future by adding a wider range of emergency vehicle sounds and diverse urban environments can improve the system's robustness. Also, integrating the visual and acoustic detection methods and testing the system in real-time urban traffic management will provide valuable insights into its practical effectiveness and vulnerabilities.

REFERENCES

Brownlee, J. (2020). *How to Develop an Extra Trees Ensemble with Python*. Machine Learning Mystery. https://machinelearningmastery.com/extra-trees-ensemble-with-python/#:~:text=The%20Extra%20Trees%20algorithm%20works,in%20the%20case%20of%20classification.

Caetano dos Santos, D. F., & Boccato, L. (2022, 2022). A Study of Emergency Siren Recognition on Resource-Constrained Devices.

Cantarini, M., Gabrielli, L., & Squartini, S. (2022). Few-Shot Emergency Siren Detection. *Sensors*, *22*(12), 4338. https://www.mdpi.com/1424-8220/22/12/4338

Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, *28*(4), 357-366.

Fatimah, B., Preethi, A., Hrushikesh, V., Singh, A., & Kotion, H. R. (2020, 2020). An automatic siren detection algorithm using Fourier Decomposition Method and MFCC.

Islam, Z., & Abdel-Aty, M. (2022). Real-time Emergency Vehicle Event Detection Using Audio Data. *arXiv preprint arXiv:2202.01367*.

Lisov, A. A., Kulganatov, A. Z., & Panishev, S. A. (2023). Using convolutional neural networks for acoustic-based emergency vehicle detection. *Modern Transportation Systems and Technologies*, *9*(1), 95-107.

Logan, B. (2000, 2000). Mel frequency cepstral coefficients for music modeling.

Marchegiani, L., & Newman, P. (2022). Listening for sirens: Locating and classifying acoustic alarms in city scenes. *IEEE Transactions on Intelligent Transportation Systems*, *23*(10), 17087-17096.

Mittal, U., & Chawla, P. (2023). Acoustic Based Emergency Vehicle Detection Using Ensemble of deep Learning Models. *Procedia Computer Science*, *218*, 227-234.

O'Shaughnessy, D. (1987). *Speech communications: Human and machine (IEEE)*. Universities press.

Patel, R., Mange, S., Mulik, S., & Mehendale, N. (2022). AI based emergency vehicle priority system. *CCF Transactions on Pervasive Computing and Interaction*, *4*(3), 285-297.

Rabiner, L., & Schafer, R. (2010). *Theory and applications of digital speech processing*. Prentice Hall Press.

Rupapara, V., Rustam, F., Shahzad, H. F., Mehmood, A., Ashraf, I., & Choi, G. S. (2021). Impact of SMOTE on imbalanced text features for toxic comments classification using RVVC model. *IEEE Access*, *9*, 78621-78634.

Sathruhan, S., Herath, O. K., Sivakumar, T., & Thibbotuwawa, A. (2022, 2022). Emergency Vehicle Detection using Vehicle Sound Classification: A Deep Learning Approach.

SETHI, A. (2020). *Support Vector Regression Tutorial for Machine Learning*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/03/support-vector-regression-tutorial-for-machine-learning/

Stevens, S. S., & Volkmann, J. (1940). The relation of pitch to frequency: A revised scale. *The American Journal of Psychology*, *53*(3), 329-353.

Stevens, S. S., Volkmann, J., & Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *The journal of the acoustical society of america*, *8*(3), 185-190.

Tiwari, V. (2010). MFCC and its applications in speaker recognition. *International journal on emerging technologies*, *1*(1), 19-22.

Tran, V.-T., & Tsai, W.-H. (2020). Acoustic-based emergency vehicle detection using convolutional neural networks. *IEEE Access*, *8*, 75702-75713.

Tran, V.-T., & Tsai, W.-H. (2021). Audio-vision emergency vehicle detection. *IEEE Sensors Journal*, *21*(24), 27905-27917.

Hashmi, M. S. A., Ibrahim, M., Bajwa, I. S., Siddiqui, H.-U.-R., Rustam, F., Lee, E., & Ashraf, I. (2022). Railway track inspection using deep learning based on audio to spectrogram conversion: An on-the-fly approach. *Sensors, 22*(5), 1983. https://doi.org/10.3390/s22051983

Umesh, S., Cohen, L., & Nelson, D. (1999). Fitting the mel scale.

Usaid, M., Asif, M., Rajab, T., Rashid, M., & Hassan, S. I. (2022). Ambulance Siren Detection using Artificial Intelligence in Urban Scenarios. *Sir Syed University Research Journal of Engineering & Technology*, *12*(1), 92-97.

Walden, F., Dasgupta, S., Rahman, M., & Islam, M. (2022). Improving the Environmental Perception of Autonomous Vehicles using Deep Learning-based Audio Classification. *arXiv preprint arXiv:2209.04075*.

Wu, J., Yang, Y., Li, E., Deng, Z., Kang, Y., Tang, C., & Sunny, A. I. (2018). A high-sensitivity MFL method for tiny cracks in bearing rings. *IEEE Transactions on Magnetics*, *54*(6), 1-8.

Ye, J., Kobayashi, T., Murakawa, M., & Higuchi, T. (2014, 2014). Robust acoustic feature extraction for sound classification based on noise reduction.

Ying, C., Qi-Guang, M., Jia-Chen, L., & Lin, G. (2013). Advance and prospects of AdaBoost algorithm. *Acta Automatica Sinica*, *39*(6), 745-758.

Dontabhaktuni, J., Modugu, K., Kollem, S., Samineni, P., Nadikatla, C., & Maturi, T. (2024). Emergency vehicle classification using combined temporal and spectral audio features with machine learning algorithms. Electronics, 13(3873). https://doi.org/10.3390/electronics13193873