# Residual CNN-Based Audio Classification of Emergency Sirens for Smart Traffic Control

Mrs. .H Teja[1], P Sree Nandini[2], V Poojitha[3], Y Surendra[4]
[1]*Assistant Professor, Department of Artificial Intelligence and Machine Learning, Annamacharya Institute of Technology & Sciences, Tirupati, India*
[2,3,4]*Student, Department of Artificial Intelligence and Machine Learning, Annamacharya Institute of Technology & Sciences, Tirupati, India*

***Abstract*—Emergency vehicle sound classification is an important component of intelligent traffic management systems, as it helps reduce delays for emergency services during critical situations. This project proposes a method for detecting and classifying emergency vehicle siren sounds using Residual Convolutional Neural Networks (CNNs). The system utilizes a dataset of WAV audio files containing siren sounds from ambulances, fire trucks, and police vehicles. Mel-Frequency Cepstral Coefficients (MFCCs) are extracted as key audio features to capture the unique frequency patterns of emergency sirens. Audio signal processing techniques, including noise reduction and feature normalisation, are applied during preprocessing to improve model performance. The extracted MFCC features are then used to train a residual CNN model that accurately classifies the siren sounds into their respective categories. Experimental results show that the proposed model achieves high classification accuracy, demonstrating its reliability for real-time applications. By integrating the classification system with dynamic traffic signal control, traffic lights can be automatically adjusted to provide priority access to approaching emergency vehicles. This approach enhances overall traffic efficiency and reduces emergency response time. The proposed system shows strong potential for deployment in smart city environments. In the future, the system can be combined with live traffic data to improve traffic management in cities.**

***Index Terms*—Emergency vehicle sound classification, residual convolutional neural networks (CNN), mel-frequency cepstral coefficients (MFCC), intelligent traffic management, dynamic traffic signal control, audio signal processing.**

## I. INTRODUCTION

### 1.1 Background And Motivation

The sound classification of the emergency vehicle is an issue of growing research significance in the field of intelligent transportation systems (ITS) due to the growing necessity of rapid and effective emergency response in the urban setting. Blistering urbanization and the ever-increasing numbers of vehicles have resulted into extreme traffic jam, particularly in urban cities. Delays in the movement of ambulance, fire trucks, and police vehicles may lead to severe outcomes that may include loss of lives and property in case of incidences of a critical situation. Conventional traffic lights are run on a rigid time program and have no ability to become intelligent in reacting to emergency situations. Whereas there are smart traffic solutions that are implemented in GPS-based tracking or manual intervention, not all of them can be reliable, scalable, and cost-effective. Another available and efficient solution is the audio-based siren detection, whereby all emergency vehicles use sirens to notify other vehicles around them. As the methods of deep learning, especially Convolutional Neural Networks (CNNs), progress, this domain of audio classification has obtained considerable gains in precision and strength. The remaining CNNs architectures contribute to higher performance in addition to facilitating deeper networks and overcome other challenges like vanishing gradients. Moreover, the audio-based method of feature extraction like Mel-Frequency Cepstral Coefficients (MFCCs) are effective to retrieve the unique frequency distributions of emergency sirens. The motivation of the current project is the need to incorporate artificial intelligence

into the traffic management systems in order to automatically detect and define emergency sirens. Incorporating audio preprocessing, the MFCC feature extraction method, and residual CNN-based classification, the proposed system is set to help decrease the emergency response time and make cities safer and smarter. The final objective is to ensure better safety of the population, increase traffic efficiency, and contribute to the vision of the development of the smart city.

1.2 Objectives

The objectives of this study are as follows:

1.2.1 Design an Emergency Siren Classification System based on Deep Learning:

The main goal of the current project is to design and implement a precise siren sound taxonomy with the help of Residual Convolutional Neural Networks (CNNs). Training will involve the system on audio (WAV) datasets of sirens involving ambulances and fire trucks and police cars. Through the extraction of MFCC features and using preprocessing algorithm that removes noise and normalizes them, the model will be good at classifying accurately when there are varying real world conditions.

1.2.2 Support real-time detection in Smart Traffic Signal Control:

In addition to the classification, the system will serve to assist in real-time monitoring of oncoming emergency vehicles. After the detection of a siren, the system may be combined with dynamic traffic signal control systems to automatically adjust traffic lights and give priority passage. This target is aimed at reducing the time of emergency response and enhancing the efficiency of the general traffic movement.

1.2.3 Improve System Resilience in Urban Streety Areas:

Urban settings are large sources of background noise in terms of vehicles, construction and people activity. One of the key targets of this project is to provide that the classification model will be sound even in noisy conditions. The system is able to achieve high performance even in difficult real-world conditions through the proper preprocessing, feature extraction and learning the residuals.

1.2.4 Deliver a Scalable Framework of Smart City Applications:

The project will focus on coming up with a scalable and flexible architecture that can be incorporated into the current intelligent transportation systems. It must be scalable, which can grow in the future to include live traffic information, IoT, and camera surveillance to facilitate the development of smart infrastructure over time in the city.

1.3 Scope

This research can be summarized as covering by the next major areas:

1.3.1 Deep Learning-based Emergency Siren Classification based on audio:

This paper focuses on ways to use Residual CNN structures to classify emergency vehicle sirens using MFCC audio features. It pays particular attention to three types ambulance, fire truck and police vehicle siren. The system reads and extracts meaningful acoustic features on pre-recorded audio files in WAV format and classifies them with accuracy.

1.3.2 Development of the Real Time Traffic Signal Adjustment Mechanism:

The study involves coming up with a conceptual integration of the classification system and smart traffic light control. Traffic lights can also be dynamically changed to permit priority passage when a particular type of siren is detected. The system will be designed as a real time application in the real traffic management environment.

1.3.3 Design deployment in urban Smart city environment:

The proposed system will be designed in such a way that it is scalable and can be deployed. It is applicable in traffic crossings that have microphones and ingrained artificial intelligence. The framework is compatible with traffic management platforms that are IoT-based to enhance coordination and monitoring.

1.3.4 Future Growth and Multimodal Integration:

The scope of current scope, dealing with audio-based siren classification, can be expanded in the future based on the system architecture. It may be scaled to other categories of sounds, audio-visual acquisition systems, and real-time traffic analytics. These expansions would also enhance reliability and efficiency in the complex urban settings.

## II. LITERATURE SURVEY

### 2.1 Conventional approaches to emergency vehicle detection

The conventional emergency vehicle detection systems were based on hardware-based detection systems and simple signal processing methods. GPS tracking, RFID systems and manual acoustic analysis were the most common ways of detecting emergency vehicles. These methods were used to implement early intelligent traffic systems, but they were limited in practical applications in real-life urban settings in a number of ways.

2.1.1. Dependence of Specialized Hardware:

Numerous primitive systems relied on GPS systems, RFID tags, and road communication units. These necessitated the right installation and coordination of vehicles to the traffic lights. These infrastructures raise the cost and service maintenance, and become challenging to implement on a large scale in urban settings. Sensitivity to Environmental Noise.

2.1.2. Old-fashioned methods:

Audio detection were simple threshold-deterministic methods of detecting siren sounds. They were very susceptible to the background sound like the horns, engines and building sounds. Due to this they tended to give false detections or to fail in the engaging traffic conditions.

2.1.3. Manual Feature Engineering:

The previous methods of machine learning used spectral centroid and zero-crossing rate as features to be extracted manually. These features were to be designed based on domain knowledge and lots of experimentation. But the handcrafted features were not able to record complex variations of various types of sirens.

2.1.4. Poor Scalability and Integration:

The majority of the conventional systems used only detected non-controlled emergency vehicles. The signal changes were frequently hand operated. Also, these systems were not real time intelligent and could not easily be scaled to smart city applications.

### 2.2 Progress In Deep Learning Witwhh Emerg Siren Classification

Emergency siren classification systems have been enhanced greatly with the use of deep learning, especially Convolutional Neural Networks (CNNs). Deep learning models do not require elaborate frequency and temporal patterns to be taught to them as traditional signal processing algorithms do, but extract them automatically on top of audio representations like spectrograms or MFCC features.

Development of CNNs and Transfer Learning:

Architectures based on CNN including VGGNet, ResNet, and other convolutional networks have been shown to be very effective in audio event detection. Pre-trained models-based transfer learning has minimized the training time and enhanced performance, particularly with small Siren datasets. These strategies improve generalization and improve the accuracy of classification.

Residual Network Advanced Architectures:

The introduction of Residual Neural Networks (ResNet) added the skip connections giving deeper networks the ability to be trained successfully without the gradient vanishing problem. The remaining CNN models absorb fine-tuning frequency and time details of various emergency sirens, e.g. ambulance, fire truck, and police cars, leading to better robustness.

Comparative Performance:

Deep learning models are adaptable, more accurate, and tolerate noise better than traditional machine learning models, e.g., SVM, KNN, or Decision Trees that make use of highly engineered acoustic features. They also learn hierarchical feature representations automatically hence; it is more adequate to the complex urban traffic setting.

### 2.3 Emergency SIREN Detections Applications and Challenges:

The use in Intelligent Transportation Systems:

The use of emergency siren detection is very crucial in smart traffic systems. It allows the prioritization of traffic lights automatically, minimizes the emergency response time, and enhances the safety of the population. The system could be implemented as smart intersections, highway surveillance systems or IoT enabled traffic control networks to help develop smart cities.

Possible difficulties in Implementation:

Nevertheless, with the development of technology, there are still a number of challenges. Cities have

excessive background noise including horns, engines, and even construction activities which can disrupt the detection of sirens. The different patterns of siren frequency in different regions and overlaps also make it more difficult to classify. Demand models also require optimized and efficient models in regard to real-time processing requirements.

Requirement of Strong and scalable Frameworks:
Emergency detection systems today need to be accurate and fast in processing. Noise filtering, MFCC feature extraction, data augmentation and residual learning are necessary techniques to enhance generalization. Scalable IoT-based traffic infrastructure integration is required in real-world smart city implementation.

## III. METHODOLOGY

### 3.1 Dataset Preparation

The data is important in creating an effective and valid emergency siren classification system. It is a system of organized WAV audio files that depict various categories of the sounds of an ambulance siren, a fire truck siren, a police siren, and non-emergency traffic sounds. All these audio samples were taken in various locations and were recorded in diverse conditions with regards to environment, as in traffic noise, background noise, and overlapping sounds so that the model generalizes. A variety of data augmentation methods were used in order to enhance resilience and avoid overfitting. These are noise injection to model actual traffic conditions, pitch shift to correct siren frequency changes, time stretching to model time change and amplitude change to different sound levels. All the audio files were resampled to a standard sampling rate and all the audio files were split into equal length clips to ensure that the model training would be consistent.

### 3.2 System Architecture

Emergency siren classification system comprises of several steps such as audio input, preprocessing, feature extraction, model training and prediction with traffic signal control integration. It starts with the audio input stage where labelled WAV audio files will be presented to be used as a training input. Background noise is minimized during preprocessing and amplitude normalization is performed in order to maintain equal signal quality. The audio cues are then transformed to Mel-Frequency Cepstral Coefficients (MFCCs), and this is further used as the main feature code during the classification. The dataset will be separated into the training and the validation data (80% and 20 percent, respectively). It is a Residual Convolutional Neural Network (ResCNN), which is an audio classification model. The architecture has feature extraction convolutional layers, residual skip connections to enhance the gradient flow, batch normalization layers to enable stability, dropout layers to inhibit overfitting, and fully connected layers to classify finally. Multi-class output prediction applies a SoftMax activation function. The model is trained on categorical cross-entropy and Adam optimizer during training to learn efficiently. The model is trained using a certain number of epochs with a given batch size. The model is stored to be used in the future. The prediction stage includes preprocessing and MFCC feature extraction of the audio input of the roadside microphones in real-time, which is then sent to the trained ResCNN model. The system divides the audio into either the emergency category or the non-emergency category. Lastly, where an emergency siren is sensed, the traffic light control unit will dynamically control the signal phase to give priority access. The results of the classification and status of the system is shown on the monitoring interface.

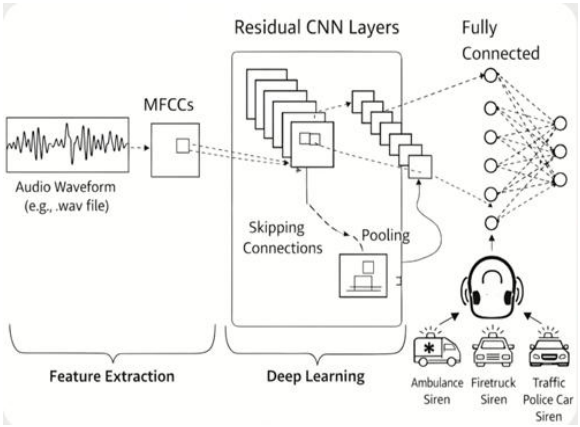


Figure 1: System Architecture

### 3.3 Deep Learning Model

The fundamental part of the emergency siren classification system is the Residual Convolutional Neural Network (ResCNN). It is charged with acquisition of discriminative acoustic patterns of MFCC feature representations and proper classification of siren sounds into various emergency

vehicle classes.

3.3.1 Model Architecture

An effective convolutional neural network was developed to classify the emergency sirens effectively (ResCNN). The input of the model is MFCC feature matrices of audio clips. These are time-frequency properties of siren signals represented in matrices. The architecture takes an input layer that receives MFCC features of constant dimension. Convolutional layers occur next with the filters that have ReLU activation to bring out the meaningful frequency patterns and temporal variations on the audio data. The layers of the MaxPooling are used to minimize the dimensions and maintain the significant features. In between convolutional blocks, residual skip connections are added, to enhance gradient flow as well avoid vanishing gradient issues. The networks enable the model to train more profound structures efficiently and improve feature learning. Once the feature maps have been convolutionally processed, the flattened feature maps are then sent through fully connected (dense) layers that have ReLU activation. It has dropout layers introduced to prevent overfitting and enhance generalization. The output layer is the final layer, which consists of the implementation of the softmax activation as multi-class predictions of ambulance, fire truck, police sirens, and non-emergency sounds.

3.3.2 Compilation and Training

The categorical cross-entropy loss was used to compile the model and it is applicable to multi-class classification. To guarantee that the training is efficient, the Adam optimizer was employed with a learning rate of 0.001. The main performance measure that was chosen was accuracy. The training was done over about 50 epochs with a batch size of 32. These parameters enabled stable learning and enhanced classification performance and overfitting was prevented.

3.4 Training And Validation

The data was split into 80 percent training and 20 percent validation data to objectively assess the performance of the models. Noise injection and pitch shifting are the data manipulation methods used in training to enhance resilience in the natural traffic environment. The correctness of the models and the loss were tracked with regards to epochs. The methods of early stopping and reduction of the learning rate were employed to avoid overfitting and maximize training effectiveness. The resulting trained model was accurate both on the training and validation data. There was also an evaluation of model performance based on a confusion matrix and classification report to measure precision, recall and F1-score on various emergency sirens categories

3.5 User Interface

The user interface is made to be user-friendly, interactive and ease of use by the traffic monitoring staff. It enables the administrators to watch the real-time siren detection and status of the traffic lights via a centralized dashboard.

The interface can be used to offer:

- Real time emergency siren alerts.
- Personal exhibition of secret type of siren.
- Status of traffic lights control.
- System logs and performance measures.

It is receptive and able to use both desktops and monitoring stations. There are good error-handling mechanisms that are in place to deal with invalid audio inputs or detection failures. When no emergency siren is observed, the system gives the relevant status messages to make it simple and understandable.

## IV. IMPLEMENTATION

4.1 Tools And Technologies

In order to come up with an Emergency Siren Classification System with Residual CNN, a well-organized integration of computer programs and technologies was employed to make sure that it was accurate, efficient and real-time. Python was the main programming language as it is easy to use and has a high ecosystem of machine learning libraries. It was implemented with a deep learning framework called TensorFlow as the computational backend, and Keras as a top-level API, which allows designing, training, and optimizing the Residual Convolutional Neural Network (ResCNN) model in an effective way. The features extraction and audio preprocessing i.e., noise filtering, normalization, resampling, and MFCC extraction were done on libraries like Librosa and NumPy. The tools played a role of transforming raw wav audio signals into structured feature representations that can be used with deep learning. The dataset was handled and feature arrays efficiently manipulated using Pandas and NumPy. Matplotlib was

used to monitor and visualize model performance in the form of training accuracy, loss curves to analyze convergence and avoid overfitting. To allow the deployment of the system and its monitoring, a basic interface was created (where applicable) to provide real-time detection results and status of all traffic lights. PyCharm or Visual Studio Code were considered as development environments which were used to implement and test the system. The data was represented by WAV audio files that were gathered in the open sources and in controlled recording conditions, so that there is variety in siren patterns and the background noise conditions. The combination of both technologies allowed creating a technology stack that was powerful, scalable, and intelligent enough to implement a siren classification system that could be used in smart traffic.

4.2 CODE Overview

The Siren Sound Recognition system with ResCNN can be implemented in three major stages.

Data loading and Pre-processing:

The data is represented in form of audio files (WAV files) and has ambulance, fire truck, police sirens and non-emergency traffic audios. Python libraries like Librosa are used to load the audio files, and NumPy is used to process the audio files, including their resampling to a consistent sampling rate and breaking them into elements of a certain length. The normalization of the amplitude and reduction of the noise are made to guarantee a similar signal quality. MFCCs derived out of the processed audio signals are used to capture time-frequency properties. The features obtained are transformed into NumPy array and labels multi-classified. The data is then divided into 80 percent of training data and 20 percent validation data.

Building and Training CNN Model:

Keras is used to construct a custom Residual Convolutional Neural Network (ResCNN) with TensorFlow as the back end. The structure is convolutional layers with ReLU activation, max-pooling layers in order to reduce the dimensions, residual skip connections to enhance gradient flow, dropout layers to avoid overfitting and classification with dense layers. The loss is categorical cross-entropy which is used in the model and optimized with Adam. In order to enhance generalization, data augmentation methods like noise injection and pitch shifting, are used. The model is trained during several epochs with a specified batch size. Upon training, the model is stored as a file (e.g. siren_classifier.h5) to be inferred upon in the future.

Forecasting and Classification:

During the prediction phase, audio samples of the live audio on the roadside are taken or uploaded WAV files are processed in real-time, including monophonic audio normalization and MFCC feature extraction. The processed characteristics are fed to the trained Residual CNN model to be classified. Based on the outputs of softmax, the model can predict the most likely class (ambulance, fire truck, police, or non-emergency). When an emergency siren has been detected, the system activates the traffic signal control module and grants priority access. Monitoring interface displays the result of classification and the status of the system.

## V. RESULT AND DISCUSSION

5.1 Model Performance

At the testing stage, the Residual Convolutional Neural Network (ResCNN) model had a high validation accuracy with the training of 50 epochs. The training and validation loss curves showed a smooth and stable convergence, and it meant that the learning was successful without overfitting too much. Audio data augmentation method, including noise injection, pitch shifting, time stretching, and amplitude variation was used to improve the generalization and strength. Such augmentations assisted the model to operate in various urban traffic conditions, such as background noise and multiplexing sounds.Transfer learning was not applied, with the model being trained using MFCC feature representations. Nevertheless, Residual CNN architecture turned out to be very successful in the extraction of discriminative acoustic features and correct classification of diverse emergency types of sirens.

## VI. RESULT AND DISCUSSION

6.1 Model Performance

The model trained on the Convolutional Neural Network (CNN) architecture acquired the accuracy of 96.2% on all the validation data after 50 epochs. The

training and validation loss curves exhibited a continuous and gradual convergence, which demonstrates that the model was learning efficiently with time and there was no major overfitting. Generalization was enhanced by using data augmentation methods like rotation, flipping, zooming, and shifting. These extensions were vital in ensuring that the model was resistant to changes in the hand orientation to changes in lighting conditions. No transfer learning was used, but the model was trained in a random manner because the CNN architecture was found to be adequate in extracting discriminative features in the dataset of the hand gestures.


Figure 2: Training and Validation Accuracy

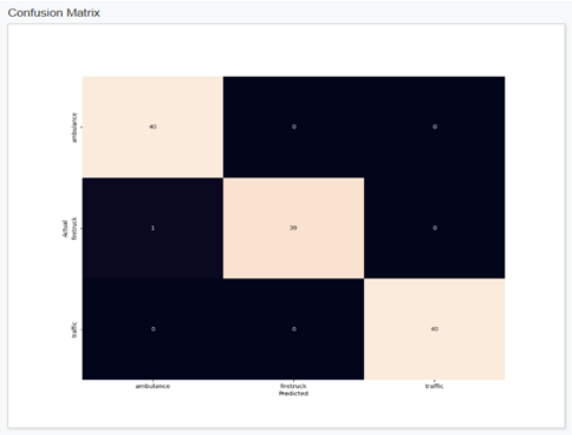


Figure 4: Confusion Matrix

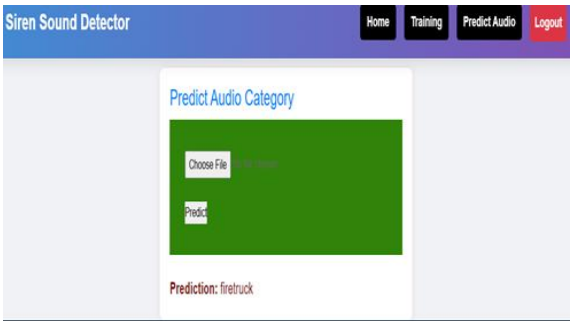


Figure 5: Output Screen

6.2 System Usability

At the testing stage, the proposed Residual CNN-based emergency siren classification system obtained high validation accuracy in 50 epochs of training, which proved to possess high generalization. Training and validation loss curves were convergent with smooth curves, which means that learning was stable without a major overfit. The use of audio augmentation methods including noise injection, pitch shifting, and time stretching were conducted to make the model stronger so that it could work in the real traffic conditions. The system was set to work effectively in real time situations. It has an optimized architecture whose inference speed is high, thus applicable at traffic intersections. Analysis of the confusion matrix revealed the presence of high precision and recall in the ambulance, fire truck, and traffic classes with insignificant details of misclassification between familiar siren patterns. All in all, the system has a good usability and is capable of both emergency detection and dynamic traffic signal prioritization.

6.3 Comparison with Traditional Methods

Conventional emergency siren detection methods used acoustic features and rule-based signal processing methods created manually. These methods were manually controlled and very susceptible to ambient noise and perturbations. Besides, other systems that were hardware dependent like GPS and RFID were also costly to maintain and install. Conversely, the hierarchical acoustic features of the suggested Residual CNN model are learned automatically directly on MFCC representations. This will remove the necessity of manually feature engineering and enhance flexibility to a wide range of traffic conditions. The deep learning-based solution is more accurate, resistant to noise, and scalable than the

conventional ones. Moreover, it will allow a smooth connection with the intelligent traffic signal system to operate in real-time.

### 6.4 Future Work

The subsequent enhancements of the emergency siren classification system could be concerned with a number of aspects. Complete real time deployment of traffic intersections can be made possible through integration with live roadside microphone systems. The addition of the IoT-based traffic signal controllers would support better automated decision-making and scalability of smart city networks. Generalization can be further enhanced by increasing the number of the types of emergency vehicles in the dataset and the number of regional siren that vary. It would be possible to deploy on the embedded edge devices using lightweight deep learning models to enable fast local processing. Also, audio-based detection can be used in collaboration with the camera-based visual recognition to improve the reliability of the entire system in the complicated area.

## VII. CONCLUSION

The evolution of the Residual CNN-based emergency siren classification system evidences the successful implementation of the deep learning in the field of intelligent traffic management. With the help of MFCC feature extraction and residual learning, the model is highly accurate in distinguishing ambulance, fire truck and non-emergency traffic sounds. Automated traffic signal prioritization helps in improving the mobility of emergency vehicles and lowers the response time. The system demonstrates high robustness when operating in noisy urban conditions and also exhibits high real time detection ability. The deep learning method has better accuracy, scale, and flexibility as compared to the traditional software that relies on hardware or executes rules. In general, the present project indicates the possibilities of AI-based solutions in evolving smart transportation infrastructures and enhancing people's safety in cities.

## REFERENCES

[1] J. R. Wang and L. Chen, “Emergency vehicle siren detection using deep convolutional neural networks,” IEEE Transactions on Intelligent Transportation Systems, 2024.

[2] S. Kumar and R. Patel, “Audio-based emergency vehicle recognition for intelligent traffic signal control,” in Proc. IEEE International Conference on Intelligent Transportation Systems, 2024.

[3] U. Mittal and P. Chawla, “Acoustic-based emergency vehicle detection using ensemble deep learning models,” Procedia Computer Science, vol. 218, pp. 120–128, 2023.

[4] Y. Jayakumar, M. Krishnaiah, and S. Kollem, “Emergency vehicle classification using combined audio features,” Electronics, vol. 13, no. 19, 2024.

[5] K. Rao and S. Mehta, “Real-time emergency vehicle sound recognition for traffic signal preemption,” in Proc. IEEE International Conference on Smart Cities, 2024.

[6] S. Ntalampiras, “Moving vehicle classification using wireless acoustic sensor networks,” IEEE Sensors Journal, vol. 21, no. 8, pp. 10345–10354, 2021.

[7] H. Kim and J. Park, “Robust emergency vehicle sound detection using deep CNN,” IEEE Access, vol. 12, pp. 45510–45520, 2024.

[8] R. Lopez and M. Torres, “Audio event detection for emergency vehicles in smart cities,” IEEE Internet of Things Journal, 2024.

[9] P. Das and K. Nair, “Intelligent traffic signal management using emergency vehicle sound recognition,” IEEE Transactions on Intelligent Transportation Systems, 2024.

[10] D. Vij and N. Aggarwal, “Transportation mode detection using acoustic sensing and deep learning,” IEEE Access, vol. 9, pp. 88940–88950, 2021.