

# Acoustic Vehicle Classification Using Mel-Frequency Features with Long Short-Term Memory Neural Networks

Ahmad Ihsan Yassin<sup>1</sup>, Khairul Khaizi Mohd Shariff<sup>1</sup>, Mustapha Awang Kechik<sup>2</sup>, Adli Md Ali<sup>3</sup>, Megat Syahirul Megat Amin<sup>1</sup>

<sup>1</sup>*Microwave Research Institute, Universiti Teknologi MARA, Shah Alam, Malaysia*

<sup>2</sup>*Department of Physics, Faculty of Science, Universiti Putra Malaysia, Serdang, Malaysia*

<sup>3</sup>*Kulliyah of Science, International Islamic University Malaysia, Kuantan, Malaysia*

**Abstract** – Monitoring vehicle traffic at a large scale is a challenging task for authorities, particularly considering the high cost of traffic sensors such as vision cameras. To meet the growing demand for more accurate traffic monitoring, the use of traffic sounds has become a popular approach, as it provides insight into the types of traffic present. This paper reports on an approach to vehicle classification based on acoustic signals, using the Mel-Frequency Cepstral Coefficients (MFCC) and the Long Short-Term Memory (LSTM) networks. This study exhibited classification accuracy scores of 82-86.2% across four vehicle categories: motorcycle, car, truck, and no traffic. The results demonstrated that large-scale, low-cost acoustic processing can be effectively used for vehicle monitoring.

**Keywords** – Acoustic vehicle classification, long short-term memory (LSTM), acoustic traffic noise, mel-cepstral frequency features (MFCC), Machine learning.

## 1. Introduction

Acoustic vehicle classification can be referred to as identification of vehicles class i.e., motorcycle, cars, and lorries based on acoustic emissions of passing vehicles. This approach is an alternative to video-based vehicle classification. It has a potential to ease some of the restriction of associated with video recording such as privacy concern, expensive signal processing and bulky systems [1]. As a result, they are well-suited for large-scale deployments in urban environments such as the wireless acoustic sensor networks (WASN) [2],[3],[4]. Acoustic emission of passing vehicle can be a form of signature. For example, an ambulance sirens sound may be immediately distinguished because it alternates between low and high frequencies. For common vehicles such as cars, identifying them is more difficult since the sound signature depends on many processes occurring on the vehicle. These include sound generated from exhaust tube emissions, engine vibrations, road-tyre friction and air stream drag [5]. These audio characteristics produced by common road vehicles have been examined by previous researchers [6],[7],[8].

The use of Machine Learning (ML) framework for acoustic traffic classification has been extensively studied, wherein features are extracted from raw traffic noise data and used as inputs of the ML algorithm. Upon training, the ML algorithm is applied to validate and test data to determine the type of vehicles. Many acoustic features, including hand-crafted, i.e., zero-crossing and spectrogram image have been employed to develop vehicle class models. On the other hand, long short-term memory (LSTM) algorithms based on deep learning have achieved impressive outcomes on many acoustic-related applications, including automatic speech recognition [9], and acoustic scene classification [10].

---

DOI: 10.18421/TEM123-29

<https://doi.org/10.18421/TEM123-29>

**Corresponding author:** Khairul Khaizi Mohd Shariff  
*Microwave Research Institute, Universiti Teknologi MARA, Shah Alam, Malaysia*


**Email:** [khairulkhaizi@uitm.edu.my](mailto:khairulkhaizi@uitm.edu.my)

*Received: 09 April 2023.*

*Revised: 14 May 2023.*

*Accepted: 19 July 2023.*

*Published: 28 August 2023.*

 © 2023 Ahmad Ihsan Yassin et al; published by UIKTEN. This work is licensed under the Creative Commons Attribution-NonCommercial-NoDeriv 4.0 License.

The article is published with Open Access at <https://www.temjournal.com/>

In comparison, LSTM models offer a much lower number of parameters than convolutional neural networks (CNNs) models, with reasonable performance maintained.

Recent studies have aimed to enhance the accuracy of acoustic vehicle classification; however, there has been limited research on utilising speech-based feature extraction and deep learning for traffic noise. This paper proposes the usage of Mel Frequency Cepstral Coefficients (MFCC). In particular, we investigate if only 13 of MFCC features, which dominantly represents the low-frequency features or envelope of spectra present in traffic noise is sufficient to provide good classification accuracy. Moreover, we used LSTM networks since it has been found to generate excellent outcomes in multiple speech classification works. Therefore, LSTM is deemed to be suitable for this task. This paper is organized as follows. Section 2 presents related works in acoustic vehicle classification, followed by Section 3 which explains the theory of MFCC and LSTM used in this work. Section 4 outlines the experiment methodology, Section 5 displays the results, and finally, Section 6 gives the conclusion.

## 2. Related Works

Previously, ML-based classification was focused on supervised machine learning approach, consisting of two stages: extracting 'hand-crafted' features from audio signals, followed by classifying the features using a classifier algorithm. Commonly used features include Mel Frequency Cepstral Coefficients (MFCC) [11], [12], other less studied features includes harmonics components [13], [14], and spectral based features such as zero-crossing rate [15], pitch frequency [16] while k-nearest neighbour (k-NN) [11] support vector machine (SVM) [17], and artificial neural-network (ANN) [18] are the commonly used classifier algorithms. However, this approach can be problematic due to the potential bias and uncertainty of the expert creating the features, as well as the difficulty of acquiring prior knowledge of optimal features from large datasets.

On the other hand, several studies have considered the application of deep learning algorithm for vehicle classification and showed promising results. CNN based on AlexNet is used to detect vehicles with modified loud exhaust achieving 96% accuracy [19]. Alternatively authors of study [19] used autoencoder neural networks to classify car and trucks with accuracy of 87%.

## 3. Theory on MFCC and LSTM

### MFCC

The MFCCs are obtained by applying the Fourier Transform to a windowed signal and then converting the resulting frequencies to the Mel Scale. A regular frequency scale can be transformed into the Mel Scale as done by study [20]

$$f_m = 2595 \log_{10} \left( 1 + \frac{f_l}{700} \right) \quad (1)$$

where  $f_m$  is the scale of mel-frequency,  $f_l$  denotes the linear frequency scale. The MFCC is intended to translate the frequency that humans perceive to the frequency that is measured, as humans are more likely to recognize changes in pitch at lower frequencies rather than higher frequencies.

Figure 1 shows extraction MFCC from audio data. The process consists of five steps: signal segmentation, power spectrum generation, applying mel-filter to power spectrum, compute logarithmic values of the filter banks, and applying discrete cosine transform (DCT).

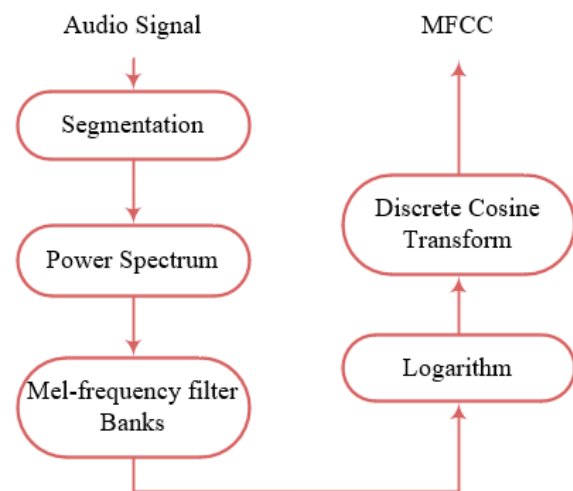


Figure 1. The process of extracting MFCC features

The audio signal is initially divided into frames in which it is assumed to remain constant. Short-term spectral measurements are taken over 25ms windows which are overlapped by 10ms with the subsequent frame. By overlapping the frames, the sound can be approximately centered on each frame. Usually, a Hamming or Hanning window function is applied to minimise spectral leakage in the FFT process.

Following this, the power spectrum is produced through the implementation of FFT, which is then multiplied by Mel triangular filters and the logarithm of the retained power is determined.

These triangular filters that have been empirically derived based on the human perception of pitch. Specifically, the Mel filter was initially developed for speech analysis, utilizing a non-linear representation of the speech signal analogous to the way in which the human ear perceives speech. This filter-bank typically consists of 13 triangular filters of varying center frequencies. The Mel-cepstrum coefficients are finally computed via the Discrete Cosine Transform [20]:

$$C_n = \sum_{m=0}^{M-1} \log D_m \cos\left(\frac{\pi n(k-0.5)}{m}\right); \quad (2)$$

$$n = 0, 1, \dots, C-1$$

where  $C_n$  denotes the MFCC coefficients,  $D_m$  is the magnitude of the Mel spectrum. In this work, we used  $n = 13$ , we exclude higher DCT coefficients since it degrades rapidly, and we wanted to keep dominant features from low frequencies only.

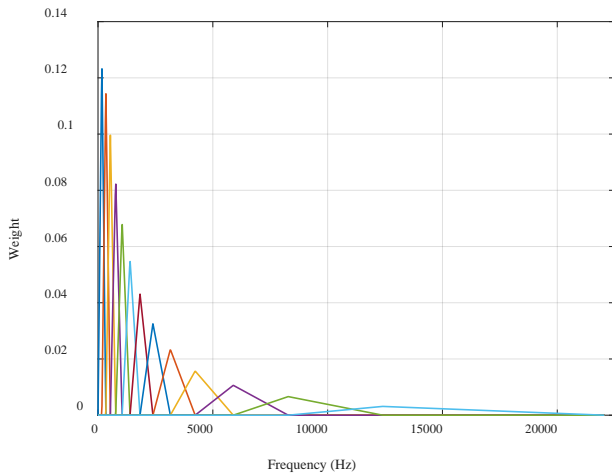


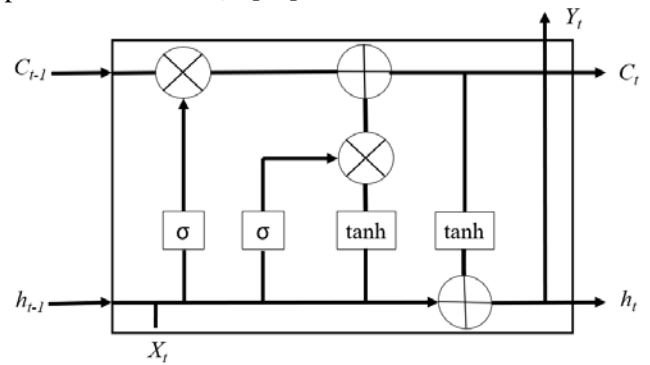
Figure 2. The 13 Mel-filter banks

### LSTM

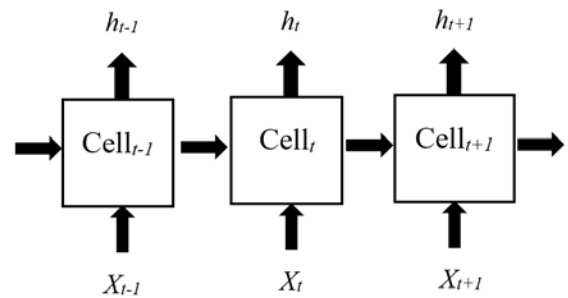
LSTM is a kind of recurrent neural network (RNN) that are widely used to model sequential data [9]. LSTM networks are designed to effectively tackle long-term memory tasks by mitigating the vanishing gradient problem. To achieve this, the RNN's hidden layer output is reused as input in the network, introducing a temporal element. Additionally, a unique internal memory state is generated and added to the processed input, which significantly reduces the diminishing effect of small gradients.

An LSTM system is composed of cells that have outputs that are altered throughout the network based on the memory content of previous cells. These cells have a shared cell state that allows them to maintain long-term dependencies across the entire chain of LSTM cells. Each memory cell  $t$  will process MFCC features at the corresponding time slot  $t$  and generate new feature  $h_t$ .

Each memory cell  $t$  receives the output from the lower layer,  $X_t$ , as well as information from the previous cell,  $Cell_{t-1}$  [21].



(a)



(b)

Figure 3. (a) Basic unit of LSTM, (b) LSTM network

### 4. Methodology

Figure 4 shows the overall structure of this work. The data processing consists of several tasks, including data pre-processing, data labelling, portioning, and feature extraction. The following paragraphs provide the details of each process.

#### Dataset

Table 1. Distribution of samples for each class of vehicles

Vehicle Type	Number of samples
Motorcycle	425
Car	2370
Truck	637
No Traffic	3047

In this work, data on traffic noise was obtained from the Fraunhofer Institute for Digital Media Technology IDMT [22]. Data was gathered from roads located in Ilmenau, Germany and recorded under four different conditions: one country road and three urban roads. The data was collected under both dry and wet states. The system for collecting audio was made up of two sets of microphones:

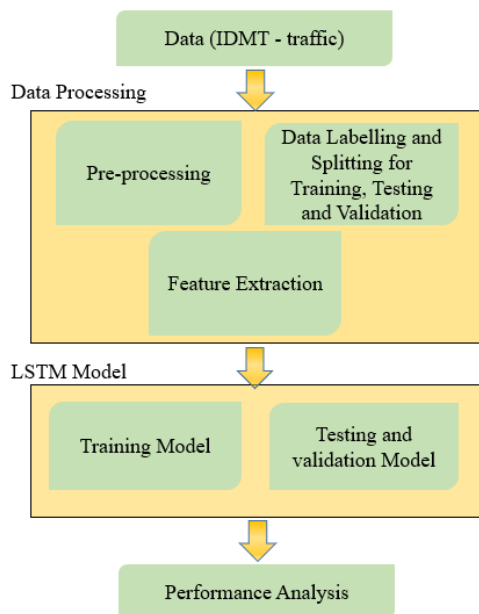


Figure 4. Experimental workflow

One set of condenser microphones from the SE electronics brand, model sE8, and another set of MEMS microphones. The microphones were positioned 50centimeters away from the road. All audio events were captured simultaneously at a rate of 48 kHz, with each sample of data lasting for two seconds.

### Data Processing

#### Pre processing

Stereo audio events were transformed into mono audio by taking the average of the left and right stereo channels. In order to decrease processing time, the audio signal was downsampled from 48 kHz to 22 kHz. All audio data was limited to duration of two seconds.

#### Data Labelling and Splitting

Each audio file was labeled according to the vehicle class it represented. The data was then separated into three different groups for the purpose of training, validating, and testing the LSTM model. To guarantee impartial results, the data were randomly split between these three sections. The proportions of data are 70:15:15 for training, validating, and testing, respectively.

#### Feature Extraction

This work uses MATLAB function *mfcc* to extract MFCC features from the audio signal. The following parameters are used for the feature extraction: Window type = Hamming, FFT length = 2048 samples and overlap = 512 samples. A single vector of MFCC is generated by Short-Time Fourier Transform (STFT). MFCCs are calculated over a 25ms frame with a 10ms interval between frames.

It is noteworthy to mention here that the optimal audio length for input remains an open question for vehicle classification; consequently, we decided to use two seconds since this is the average length of audio in the dataset.

### LSTM Model

Table 3 presents the overall structure of the LSTM in this work. The fully connected layer identifies the features through training and the softmax layer determines the probability of the four vehicle classes. Additionally, Table 4 presents the assigned hyperparameters. We used sigmoid and hyperbolic tangent as the activation function. Early stopping was used to prevent the model from overfitting.

The LSTM model was trained on a MATLAB on a personal computer with a GeForce GTX 1080 Ti GPU, and 64 MB RAM.

The LSTM model testing was repeated multiple times to ensure accurate results: three seed values were used to initialise distinct network weights and five different values of hidden layers were used, with the accuracy of each test being recorded.

Table 3. LSTM network structure

Layer	Specifications
<b>LSTM layer</b>	Vanilla (unidirectional) LSTM cells are arranged in one layer. 10, 20, 30, 40 and 50 units.
<b>Fully connected layer</b>	Four units based on the number of target classes (motorcycle, car, truck and no traffic)
<b>Classification layer</b>	Softmax activation function The four output vehicle types are mapped to a cross-entropy classification layer (motorcycle, car, truck and no traffic)

Table 4. LSTM training parameters

Training parameters	Specifications
<b>Initial weights</b>	Randomly generated according to the Mersenne-Twister pseudorandom number generator, with three initial seed values of 0, 100, and 200
<b>Learning rate</b>	0.001
<b>Optimiser</b>	Adaptive Moment Estimation (ADAM)
<b>Mini Batch Size</b>	128
<b>Maximum epochs</b>	500
<b>Dataset shuffle</b>	Every epoch
<b>Gradient threshold</b>	1

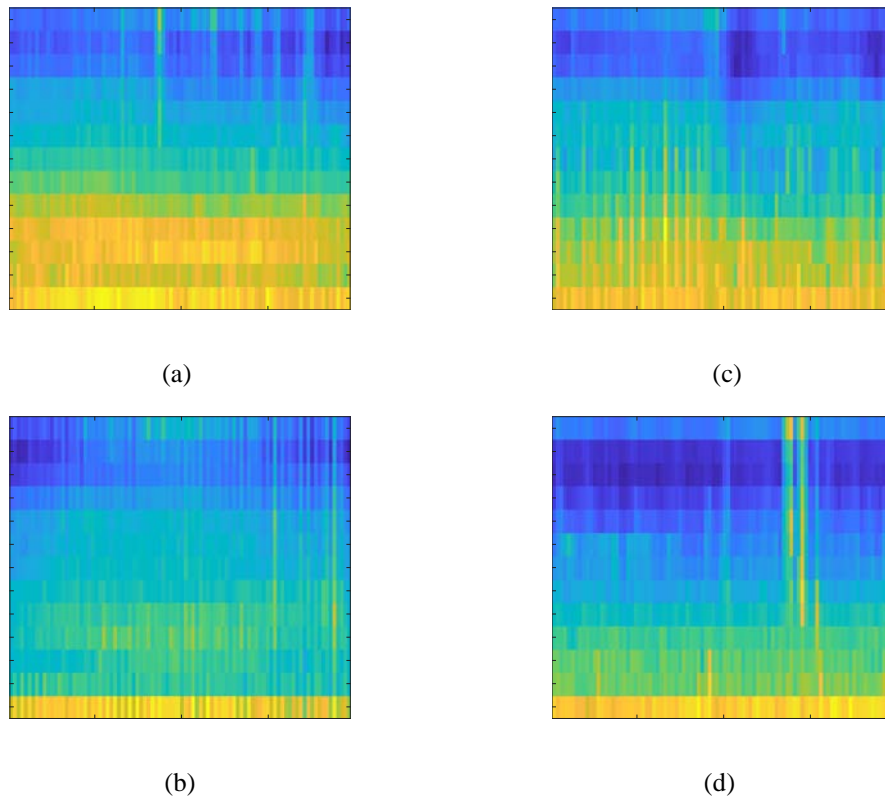


Figure 5. Examples of Mel-Cepstrum of audio signals. (a) Car, (b) Motorcycle, (c) Truck, and (d) No traffic

### 5. Results

In the first part of these results, the performance of the vehicle classification is judged by examining the Mel-cepstrums. Fig. 6 displays cepstrums for the 4 classes of vehicles. A strong yellow horizontal line is evident in all the spectra, indicating the presence of ambient noise below 300 Hz. The remaining upper frequency sounds have comparatively weaker lines due to their low power components. Nonetheless, they are visually different from each other, indicating that the spectra have discriminatory features for reliable classification.

In the second part of the study, a confusion matrix was used to evaluate the detection accuracy of the proposed method. The number and percentage of correctly classified classes are indicated by the diagonal elements of the matrix, while incorrect predictions are assigned to the incorrect class. The following parameters are established: is the chance of correctly predicting classes in relation to all predictions made. False Positive (FP) is the chance of incorrectly predicting classes in relation to all predictions made. True Negative (TN) is the chance of correctly predicting classes in relation to the total number of samples belonging to that class.

False Negative (FN) is the chance of incorrectly predicting classes in relation to the total number of samples belonging to that class, and while Sensitivity (TPR) is the proportion of samples belonging to that class that have been correctly classified as actual positives. A sensitivity of 100% indicates that there are no false negatives and all predictions made by the system are correct. Figure 6 shows the confusion matrix with LSTM parameters: hidden layer = 10 and batch size = 64. The obtained accuracy was 84.52%.

99.8% 984	31.8% 27	9.0% 95	81.9% 104
1.9% 21	50.6% 43	0.6% 6	4.7% 6
3.9% 42	11.8% 10	90.3% 948	7.1% 9
3.4% 37	5.9% 5	0.1% 1	6.3% 8

Figure 6. Confusion matrix

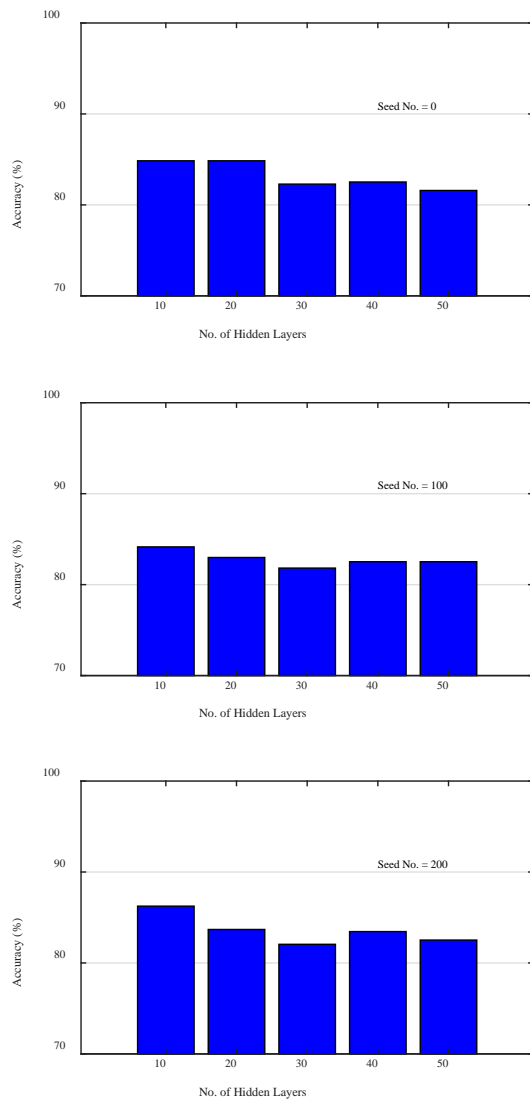


Figure 7. The classification accuracy recorded under different number of hidden layers and seed numbers.

Based on the confusion matrix, it was observed that the proposed method was highly accurate (>90%) in distinguishing between car and no traffic. However, the proposed method did very poorly in recognising truck and motorcycle sounds. Similar observation of poor accuracy for truck was found in [22]. They achieved an accuracy of only 50% using spectrogram images with CNN. Although the results of this study show less than optimal accuracy for all classes of vehicles, it has still demonstrated that the MFCC approach can be effectively employed to detect the binary conditions of traffic presence and absence with remarkable precision.

Moreover, we conducted evaluation of the proposed method with various LSTM parameters. As illustrated in Figure 7, the classification accuracy ranged from 82% to 86.2% when the hidden layer varied between 10 and 50 and three different seeds values were utilized. Consequently, the overall classification accuracy was satisfactory.

## 6. Conclusion

This paper presented an approach to vehicle classification based on acoustic signals using the Mel-Frequency Cepstral Coefficients (MFCC) and the Long Short-Term Memory (LSTM) networks. The study achieved classification accuracy scores of 82-86.2% across four vehicle categories: motorcycle, car, truck, and no traffic. However, the proposed method did poorly in recognising truck and motorcycle sounds. Although the results of this study show less than optimal accuracy for all classes of vehicles, it has still demonstrated that the MFCC approach can be effectively employed to detect the binary conditions of traffic presence and absence with remarkable precision. Despite the limitations, the results of this study contribute to the growing body of research on using acoustic signals for vehicle monitoring and pave the way for the development of more efficient and cost-effective traffic monitoring systems.

Further research is needed to enhance the accuracy of the proposed method, particularly in recognising truck and motorcycle sounds. Future studies could explore the combination of acoustic signals with other sensor modalities or employ more advanced deep learning architectures to improve classification performance.

## Acknowledgements

This study was fully funded by the MyRA-LESTARI grant, 600-RMC/MyRA, 5/3/LESTARI (017/2020). The authors would like to express their gratitude to the Research Management Institute (RMI) and Universiti Teknologi MARA for their support.

## References:

- [1]. Bernas, M., Płaczek, B., Korski, W., Loska, P., Smyła, J., & Szymała, P. (2018). A survey and comparison of low-cost sensing technologies for road traffic monitoring. *Sensors*, 18(10).
- [2]. Chmiel, W., et al. (2016). INSIGMA: An intelligent transportation system for urban mobility enhancement. *Multimedia Tools and Applications*, 75(17), 10529-10560.
- [3]. Franceschinis, M., Gioanola, L., Messere, M., Tomasi, R., Spirito, M. A., & Civera, P. (2009). Wireless sensor networks for intelligent transportation systems. In *VTC Spring 2009 - IEEE 69th Vehicular Technology Conference*, 1-5.
- [4]. Guerreiro, G., Figueiras, P., Silva, R., Costa, R., & Jardim-Goncalves, R. (2016). An architecture for big data processing on intelligent transportation systems. *An application scenario on highway traffic flows. In 2016 IEEE 8th International Conference on Intelligent Systems (IS)*, 65-72.
- [5]. Freitas, E., Pereira, P., de Picado-Santos, L., & Santos, A. (2009). Traffic noise changes due to water on porous and dense asphalt surfaces. *Road Materials and Pavement Design*, 10(3), 587-607.

- [6]. Braun, M. E., Walsh, S. J., Horner, J. L., & Chuter, R. (2013). Noise source characteristics in the ISO 362 vehicle pass-by noise test: Literature review. *Applied Acoustics*, 74(11), 1241-1265.
- [7]. Buratti, C., & Moretti, E. (2010). Traffic noise pollution: Spectra characteristics and windows sound insulation in laboratory and field measurements. *Journal of Environmental Science and Engineering*, 4, 1-9.
- [8]. Jonasson, H. (1999). *Measurement and modelling of noise emission of road vehicles for use in prediction models*. Swedish National Testing and Research Institute
- [9]. Oruh, J., Viriri, S., & Adegun, A. (2022). Long short-term memory recurrent neural network for automatic speech recognition. *IEEE Access*, 10, 30069-30079.
- [10]. Abeßer, J. (2020). A review of deep learning based methods for acoustic scene classification. *Applied Sciences*, 10(6), 2020.
- [11]. Chaudhary, M., Prakash, V., & Kumari, N. (2018). Identification Vehicle Movement Detection in Forest Area using MFCC and KNN. In *2018 International Conference on System Modeling & Advancement in Research Trends (SMART)*, 158-164.
- [12]. Kakade, P. V., & Roy, L. P. (2022). Fast Classification for Identification of Vehicles on the Road from Audio Data of Pedestrian's Mobile Phone. In *2022 IEEE 19th India Council International Conference (INDICON)*, 1-7.
- [13]. Liu, X., Huang, H., Peng, B., Zhou, M., & Li, Y. (2022). Advanced machine learning methods for autonomous classification of ground vehicles with acoustic data. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications IV*, 12113, 524-533. International Society for Optics and Photonics.
- [14]. Guo, B., Nixon, M. S., & Damarla, T. R. (2008). Acoustic information fusion for ground vehicle classification. In *2008 11th International Conference on Information Fusion*, 1-7.
- [15]. Wiczorkowska, A., Kubera, E., Słowik, T., & Skrzypiec, K. (2018). Spectral features for audio based vehicle and engine classification. *Journal of Intelligent Information Systems*, 50(2), 265-290.
- [16]. Dalir, A., Beheshti, A. A., & Masoom, M. H. (2015). Classification of Vehicles Based on Audio Signals using Quadratic Discriminant Analysis and High Energy Feature Vectors. *International Journal of Soft Computing*, 6(1), 53-64.
- [17]. Valero Gonzalez, X., & Alías, F. (2013). Automatic classification of road vehicles considering their pass-by acoustic signature. *Proceedings of Meetings on Acoustics*, 19(1), 040029.
- [18]. George, J., Mary, L., & S, R. K. (2013). Vehicle detection and classification from acoustic signal using ANN and KNN. In *2013 International Conference on Control Communication and Computing (ICCC)*, 436-439.
- [19]. Cheng, K. W., Wang, Y., Yang, Y., & Chen, C. H. (2023). Spectrogram-based classification on vehicles with modified loud exhausts via convolutional neural networks. *Applied Acoustics*, 205, 109254.
- [20]. Czyżewski, A., Kurowski, A., & Zaporowski, S. (2019, December). Application of autoencoder to traffic noise analysis. In *Proceedings of Meetings on Acoustics* (Vol. 39, No. 1). AIP Publishing.
- [21]. Abdul, Z. K., & Al-Talabani, A. K. (2022). Mel Frequency Cepstral Coefficient and its Applications: A Review. *IEEE Access*, 10, 122136-122158.
- [22]. Sherstinsky, A. (2020). Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network. *Physica D: Nonlinear Phenomena*, 404, 132306.