IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 20 December 2023, accepted 20 February 2024, date of publication 26 February 2024, date of current version 12 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3370867

RESEARCH ARTICLE

Smart City Traffic Management: Acoustic-Based Vehicle Detection Using Stacking-Based Ensemble Deep Learning Approach

AHSAN SHABBIR[1], AMMARA NAWAZ CHEEMA[2], INAM ULLAH[1], IBRAHIM M. ALMANJAHIE[3], AND FATIMAH ALSHAHRANI[4]

[1]Faculty of Computing and Artificial Intelligence, Department of Creative Technology, Air University, Islamabad 44000, Pakistan
[2]Department of Mathematics, Air University, Islamabad 44000, Pakistan
[3]Department of Mathematics, College of Science, King Khalid University, Abha 62529, Saudi Arabia
[4]Department of Mathematical Sciences, College of Science, Princess Nourah bint Abdulrahman University, P.O Box 84428, Riyadh 11671, Saudi Arabia

Corresponding author: Ammara Nawaz Cheema (ammara.cheema@mail.au.edu.pk)

This work was supported by Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2024R358), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia, and the Deanship of Scientific Research at King Khalid University through the Research Groups Program under grant number R.G.P. 2/406/44.

ABSTRACT Acoustic data analysis has emerged as a critical area of exploration for the detection of different events for quick actions in smart traffic management systems, particularly in traffic management and safety as a step toward smart cities. A specific challenge is to precisely classify road noises and emergency vehicles using sound, which is essential for speeding up emergency response times and improving traffic flow management. While existing solutions address this problem, there is opportunity for enhancement in terms of precision and accuracy to enhance the traffic flow in a sustainable smart city through a demanding and innovative technique. In this study, we suggest stacking ensemble deep learning techniques to intelligently classify emergency vehicle sirens from various background noises using a data set of traffic collected on roads via microphone sensors. The ensemble model incorporates Multi-Layer Perceptron (MLP) and Deep Neural Network (DNN) as base-learners, with an LSTM model as a meta-learner. This approach not only optimizes model efficiency but also facilitates advanced feature engineering to extract useful features including Mel Frequency Cepstral Coefficients (MFCC), Z-score, root mean square (RMS), spectral centroids, spectral flux, mel spectrogram, chroma, contrast, and Tonnetz. Using these features, our proposed Stacking Ensemble LSTM successfully classified traffic noises and emergency vehicle sirens with the highest efficiency. Upon evaluation of the test set of data for our proposed model, it has gained an accuracy of 99.12% with F1 scores ranging from 98%. This significant improvement highlights the dominance of our proposed model approach over prior research. Our proposed model presents assurance in advancing traffic control and safety statutes, demonstrating potential applicability in daily intelligent transportation systems.

INDEX TERMS Data-driven urban planning, smart city technologies, intelligent transportation systems, smart urban infrastructure, sustainable urban development, smart traffic management, urban mobility solutions, urban sensor networks.

I. INTRODUCTION

Smart The National Crime Records Bureau estimates that 24,012 individuals die every day because of improperly delayed medical care, often due to the slowing of emergency vehicles [1]. There is a need to be more study or discussion on this topic. Emergency vehicles must be allowed as much time and room as possible to navigate through traffic because of their crucial role in safeguarding the safety of general population's safety. Consequently, it is necessary to create algorithms to recognize and classify these vehicles based on their characteristic acoustic sound [2]. One of the exciting uses of acoustics and sound research is the classification of vehicle sounds, especially the classification

The associate editor coordinating the review of this manuscript and approving it for publication was Xueqin Jiang.

of emergency vehicles [2]. For convenience, smart cities aim to improve the quality of life of their residents by using technology and data. The integration of acoustic-based emergency vehicle detection into smart city infrastructure has significant potential to improve emergency response times, enhance traffic management, and contribute to overall public safety. The studies on acoustics and good analysis in [1], [2], and [3] are well demonstrated.

This study offers a significant new understanding of the methods and techniques used in audio signal feature extraction, which is a critical area of our inquiry. This issue may be resolved by the efficiency of classifying vehicles according to their sounds, making it feasible to recognize emergency vehicles without human assistance. In an attempt to create a trustworthy and helpful system for categorizing emergency vehicles according to their acoustic features, the study in [4] expands on these prior results. It does this by using advancements in sound classification and acoustic analysis using machine learning techniques. By doing this, the researchers pave the way for creating technologies that will make it easier for emergency vehicles to travel through congested areas quickly and safely, improving public safety. We want to determine whether a system could work better if it contained features particular to a specific configuration or area. The Emergency Vehicle Detection system pipeline based in segmented data, Features Extraction and Machine Learning Model building for Classification [5].

We have proposed a comprehensive and valuable solution for the detection of emergency vehicles based on sound to further the broader objective of reducing emergency response times and saving countless lives. This way, we can enhance traffic management and ensure emergency vehicles can pass through roads without any disturbance from other vehicles in smart cities. Our research can be used in many other fields, such as automated sound recognition in smart home systems, classification of wildlife sounds for biodiversity monitoring, Patients monitoring in Healthcare and many more. Our work can serve as a strong foundation for future developments in the acoustic analysis, Particularly in traffic management and safety, acoustic analysis has emerged as a significant approach for detecting events and actions. The appropriate classification is required for particular issues like road noise and the sounds made by emergency vehicles. Due to its potential to reduce road congestion and speed up emergency response times, this issue is technically fascinating and practically essential. The effectiveness of the existing solutions to this issue could have been better, prompting the need for a novel strategy. Our study's has following contributions, Enlarging Feature-set for better Illustration of acoustic data to model and Stacking based Ensemble Approach in Emergency Vehicle Detection with Acoustic Data. Our Proposed model has achieved Highest Accuracy. Our proposed approached is shown in Figure 1, depicting the flow of our proposed system. With this proposed system we were able to solve the Emergency vehicle detection system with acoustic data with highest Efficiency.

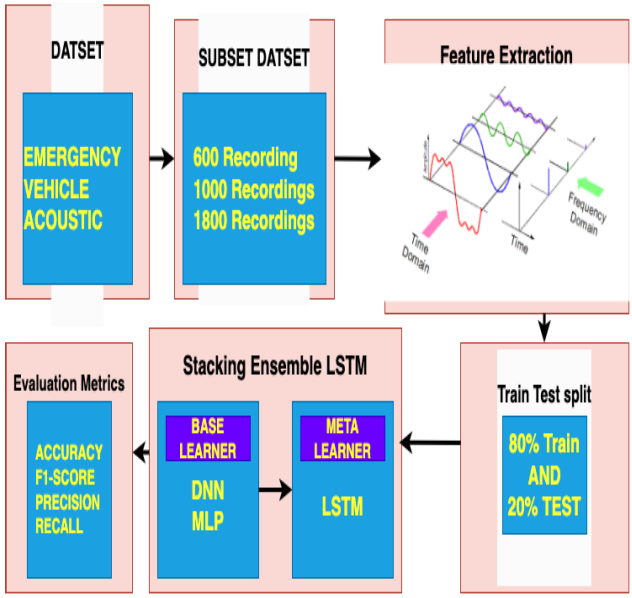

FIGURE 1. Proposed approached acoustic based emergency vehicle detection.

II. LITERATURE REVIEW

The Detection of emergency siren sounds is said to have received far less scientific attention than other acoustic issues [6]. In their study, researchers looked at LCS (Longest Common Sub sequence) based ambulance siren detection [2]. In their suggested model, they used Mel Frequency Cepstral Coefficients. The suggested model was predicted to have an accuracy of 85%. Research [7] used two distinct methods to locate sirens. a neural network with several layers that uses sinusoidal modeling. Speech recognition was accomplished using MNN technology. The sinusoidal model approach used warning tones and gathered data from background noise to minimize noise interruption. Each approach was tested on a small sample, and both models produced results that were quite accurate. PBMs were developed by [8] to differentiate sirens in congested traffic while considering the spectro-temporal domain. PBMs were first used in computer vision. When trained on MFCC or log-mel characteristics, hidden Markov models (HMMs) outperformed PBMs in comparison. However, only around 80% of the attempts were successful. Research has developed a two-stage detection method for audio-based categorization in self-driving cars [9]. The initial step was detection an uncommon sound, and the subsequent steps were noise categorization, noise reduction, and noise elimination. The idea in [9] was built upon image processing. A segmentation technique was used to recover and extract the required signal from each incoming input's spectrogram. According to [10], siren detection was carried out using DSP methods to auditory data, namely computing frequency components within a predetermined frequency range. The KNN method was used once the noise was minimized, and it produced an accuracy of 83%. SVM and feature selection techniques were used to categorize alarm sounds [11]. It demonstrated an

accuracy of more than 90% on a small sample. The initiative's major flaw was the duration of the feature engineering. A dataset from Investigation [12] may be used with AI algorithms to distinguish between the sirens of emergency vehicles and other road sounds. The results were more accurate since we used this dataset in our research and the methods we offered. The high-resolution dataset might speed up emergency response times, reduce traffic congestion, and manage traffic. The paper discusses the methods used for data collecting, model assessment, and pre-processing. The article provides a dataset that may be used to artificially distinguish between traffic and other types of road noise, such as emergency vehicle sirens. Data from the high-resolution collection includes both text and audio. The audio recordings are in WAV format, while the text data is in CSV format. Once it was obtained, the dataset underwent several pre-processing procedures to ensure its reliability and quality. Real-world traffic noises and emergency vehicle sounds are among the varied and distinctive data in the obtained dataset used for testing and training. The article covers pre-processing approaches, including qualitative human examination of the audio data and quantitative assessment to support the dataset's parametric description. Two labeled classes—one for traffic sounds and one for emergency vehicle sirens are used to categorize the dataset. This research used a multi-layer perceptron artificial neural network to generate an ensemble model. For all 1800 samples, it achieved 97% accuracy; for 1000 wav samples, 94% accuracy, and for 600 audio samples, 90% accuracy. The precision of our suggested method was 99.97%. An emergency vehicle detection (EVD) system based on siren emissions has been suggested by researchers using sound collection and processing [13]. This study examines how well deep layers of CNN and RNN in DNN models can identify the sounds of emergency vehicles. Additionally, the authors provide an ensemble model that combines the top models. The proposed ensemble model's accuracy is 98%, compared to the RNN model's accuracy of 94.5%. The efficiency of several machine learning models, including decision trees, the SVM, and simpler models like the Perceptron, is compared in the study. The study focuses on sound recording and processing in the time- and frequency domains to assess and categorize the sounds produced by emergency vehicles. Convolutional layers extract high-level features and shift-invariant data in the time-frequency domain. The authors used mel-frequency cepstral coefficients (MFCC) to extract features from a dataset created from the Google Audio set ontology. Three deep neural network (DNN) designs were examined: dense layer, CNN, and RNN models. An ensemble model was created by combining the top-performing models after adjusting hyper-parameters and conducting tests. The accuracy of the ensemble model was 98.7%, which was higher than the 94.5% accuracy of the separate RNN models. A comparison is also made between the effectiveness of deep learning models and traditional machine learning models like Perceptron, SVM, and decision trees. Convolutional neural networks (CNNs) are being investigated for use in detecting emergency vehicles based on their audio inputs [14]. The authors provide a paradigm for detection and classifying emergency vehicles that are based on CNN. They use feature fusion algorithms to combine high-level and low-level data to distinguish between various car sizes. To improve speed, the network architecture was built using convolutional layers rather than fully connected ones. The authors evaluated their suggested network against cutting-edge detectors using the JiangSu Highway Dataset (JSHD). Their network outperformed the opposition in terms of mean average accuracy (mAP) and recognition of automobiles of all sizes. The article does not mention the limitations and issues associated with using acoustic-based detection methods, such as the potential for false positives or the impact of background noise on system performance. CNN is well-known and widely used for audio detection tasks, including music Detection [15], automatic speech recognition (ASR), and ambient sound Classification. Researchers used GoogleNet and Alexnet, two well-known image recognition networks, to identify ambient sound in the study [15]. The spectrogram and Mfcc serve as inputs for these models. They created a model that demonstrated the potential of the [16] technique with up to 90% accuracy. Models for neural network-based ambient sound detection were reported in papers [5] and [17]. The models in [5] and [17] produced less than 80% accuracy with nearly the same precision when trained with log-mel spectrogram data. Reference [18] used a two-step process to identify emergency vehicles. After creating the border boxes, classification was carried out in two stages. The study [19] suggested audio and vision-based methods for detection of emergency vehicles use Wave-ResNet for sound processing and YOLO for image processing. The two main issues with the work done so far in Emergency Vehicle Detection are feature selections that could be more task-specific, which reduces efficiency, and a lack of enthusiasm for building an effective model at the model level. We suggested combining the MLP and DNN base models with the core LSTM to prevent this. After being trained using the predictions of the fundamental model, the final model performed at 99.12%. In research [20], the authors proposed using semantic segmentation to treat the spectrograms of stereo data entry as images. An Unet architecture is utilized to achieve this, separating the target sound from the background noise. To establish the kind of alarm sound, they also use a multi-task learning technique where they categorize acoustic events in addition to signal denoising. Using the denoised data, the audio source is ultimately found on the horizon plane. This is done by using the convolutional neural network (CNN) architecture to regress the direction of the sound's arrival. The system evaluated had an average classification rate of 94%, a median absolute localization error of 7.5° for audio frames running at 0.5s, and a median absolute localization error of 2.5° for frames running at 2.5s. The technology worked effectively even in challenging circumstances with a lot of noise.

III. RESEARCH METHODS

We want to enhance the performance of our Deep learning model in our proposed strategy by enlarging the Features on the audio and assembling numerous Deep Learning models. These Features, which originate from the cepstral, temporal, frequency, and harmonic (also known as pitch) domains, were chosen with great care. The goal is to increase model performance by simplifying the signals for the model to grasp. Taking into account timing and periodicity, features are also required to record the signals' temporal and spectral information. These qualities could aid the model's comprehension of how signals change over time and at various frequencies. The inclusion of harmonic or pitch information is supported by the notion that pitch plays a significant role in detection of audio signals based on vehicle signals and non vehicle signals. Strong feature selection and fusion strategies are required since it is possible that pitch and MFCC performance on noisy speech signals will need to be improved. We used a thorough feature selection technique to increase the accuracy of our model. This strategy is essential since it might influence how properly and meaningfully the problem is communicated. By removing redundant data, feature selection reduces over-fitting, increases accuracy, and accelerates training. Notably, our technique for feature selection does not rely on heuristics. Instead, we used different ''views'' of our data to build models, which we then combined with individual forecasts to form an ensemble. This strategy often yields better outcomes. Finally, it makes sense to increase our feature set to enhance our model's functionality. We seek to build an accurate and efficient model by combining spatial, temporal, frequency, and harmonic domain data with a rigorous feature selection process.

A. DATA COLLECTION

For our Analysis, we have used a Large-scale audio dataset for emergency vehicle sirens and road noises compiled in [13]. This collection contains recordings of road noise and emergency vehicle sirens on the streets of Karachi, Pakistan. The data is uniformly dispersed and comprises 1800 samples of vehicle sirens and non-vehicle sirens at a frequency of 22kHz. We selected this dataset because of its unique qualities. The traffic noises and emergency vehicle siren sounds that were captured perfectly capture Karachi's distinctive metropolitan ambiance. Due to its complexity and diversity, this dataset is a strong option for testing and improving our machine-learning model. We raised the frequency of these samples from 22k Hz to 44100 Hz to improve our model's performance. With the help of deep learning algorithms, the ''audio super-resolution'' technique raises the sample rate of audio recordings to increase audio quality.

B. SUBSET CREATION

In this part of the Study, We divided the Data set in the Three Data set First with 600 Recording we picked 300 from Emergency vehicle recordings and 300 from Road noise the reason behind was to keep data set balanced so the model must not be biased. In second division we took 1000 Recordings 500 from Emergency vehicle sirens and 500 from Road Vehicle Sirens. In Third Division we keep the Full data set 900 from Emergency vehicle siren and 900 from the Road Vehicle Siren. The Logic behind was to see the Performance of model with small to Large data set.

C. ACOUSTIC FEATURES

The following features were taken into consideration in our research to represent the signal and serve as an input to the model: MFCCs are a feature that is widely used in audio processing. They provide a tiny representation of a signal's power spectrum that is near the audible range. To compute MFCCs, a time-domain sign must first be transformed into Fourier form, the power spectrum must then be mapped to the Mel scale, which mimics the response of the human auditory system. Lastly, the log power spectrum on the Mel scale must be transformed into Discrete Cosine Transform (DCT). This method generates a set of coefficients that serve as a condensed illustration of the spectral structure of the signal. The discrete cosine transforms (DCT), the Mel filter bank (M(f), and the Power Spectrum of the movement (P(f)) are used to construct the MFCCs mathematically, as indicated in Equation 1.

$$\mathrm{MFCC} = \mathrm{DCT}[\log(M(f))] \tag{1}$$

The rate at which a signal flips from positive to negative (or vice versa) is known as the zero-crossing rate (zcr). It is often used to categorize speech, music, and musical genres. One may calculate the zero-crossing rate by counting the number of zero crossings that occur in each frame of a signal and dividing that number by the frame length. When N is the frame length and x[n] is the signal, equation (2) gives the zero-crossing rate Z:

$$Z = \frac{1}{2N}\sum_{n=1}^{N} |\mathrm{sgn}(x[n]) - \mathrm{sgn}(x[n-1])| \tag{2}$$

The ''loudness'' of a sound is often assessed using the audio processing method known as Root Mean Square (RMS), which gauges the magnitude of an audio wave. By squaring the signal values, summing these squares across the frame, and taking the square root of the result, one may get the RMS value for each frame of a signal. The RMS value R provided in equation 3 is computed using the signal's x[n] value and the frame length N.

$$R = \sqrt{\frac{1}{N}\sum_{n=1}^{N} x[n]^2} \tag{3}$$

A measurement used in digital signal processing to describe a spectrum is called the spectral centroid. It stands for the ''center of mass'' of the spectrum and is often connected to how ''bright'' a sound is perceived to be. The mean frequency weighted by the magnitude, divided by the total of the magnitudes provided by equation 4, is used to get the

spectral centroid C for a signal with spectrum X(f).

$$F = \sqrt{\sum (P(f,t) - P(f,t-1))^2} \tag{4}$$

Mel Spectrogram: Based on a nonlinear Mel scale of frequency and a linear cosine transform of a log power spectrum, a Mel Spectrogram is a depiction of the short-term power spectrum of a sound. It is often used for music genre detection and voice recognition. A 2D time-frequency representation is produced by the Mel spectrogram, which is calculated identically to the MFCCs but without the final DCT step, as shown by equation 5 and the Spectogram plot of the emergency vehicle and Road Noises is presented in Figures 2 and 3.

$$\text{Mel-Spectrogram} = \log(M(f)) \tag{5}$$

Chroma features: The 12 distinct semitones (or chroma) of the musical octave are represented by the 12 bins created by chroma characteristics, which divide the whole spectrum into. For purposes like chord and key recognition, this is a potent representation of music sounds. Equation 6's method of mapping frequencies to the 12 chroma bins and adding the magnitudes inside each bin yields the Chroma vector C for a signal with spectrum X(f):

$$C[i] = \sum_{f \text{ in bin } i} |X(f)| \text{ for } i = 1 \text{ to } 12 \tag{6}$$

D. MODEL ARCHITECTURE

A Multi-Layer Perceptron (MLP) and a deep learning ensemble made up of a Deep neural network and a Long Short-Term Memory (LSTM) network have been suggested by us as an ensemble of two models on the training set of data; the MLP is initially trained. MLP-Classifier uses a fully connected neural network with a maximum repetition limit of 1000 and a defined random state for repeatability. The training data is then applied to make predictions using the MLP To ensure a fair picture of the class distribution in each fold, we then create a 5-fold cross-validation. The DNN and LSTM models will be trained and tested using this cross-validation. We build, train, and test a DNN model for every fold in the cross-validation. Figure 2, shows the DNN model Architecture. DNN is a sequence model with two thick layers; the first has 32 neurons and a ReLU activation function, and the second has a single neuron and a sigmoid activation function. We apply the DNN model to make results on the thresholded training data for classification after training it. The Stacked Ensemble LSTM model's input is then made by adding the results from the MLP and DNN. The stacked forecasts are changed to a 3D form since the LSTM expects 3D input. Next, the Stacked LSTM model is created, Trained, and evaluated. Figure 5, Shows the Stacked LSTM Model. Stacked LSTM is a sequence model with two layers: a dense layer with a sigmoid activation function, an LSTM layer with 32 neurons, and a tanh activation function.

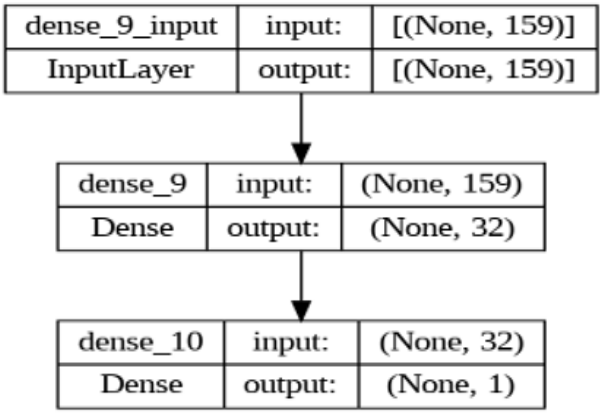

FIGURE 2. DNN model's architecture.

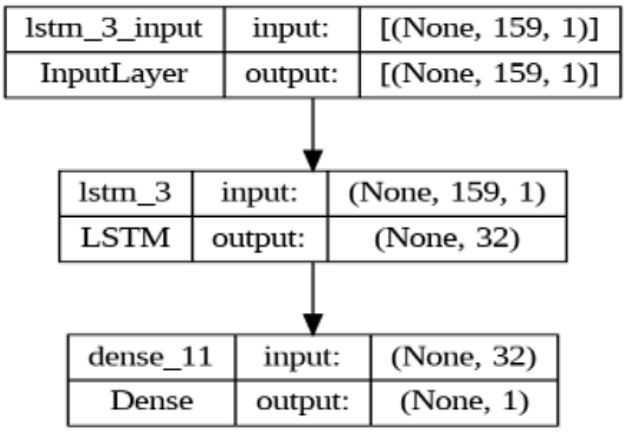

FIGURE 3. Stacked ensemble LSTM architecture.

IV. RESULTS

The ability to distinguish between the sounds of emergency vehicles and everyday road noises using a stacked ensemble deep learning approach has demonstrated great potential. By integrating the Learning of MLP, DNN, and stacking to LSTM model, the ensemble model shows outstanding accuracy and precision as an outputs. The use of advanced feature engineering, which was used to extract domain-specific features like Mel Frequency Cepstral Coefficients (MFCC), Z-score, root mean square (RMS), spectral centroids, spectral flux, Mel spectrogram, chroma, contrast, and Tonnetz, has been credited for the model's successful performance. The test results significantly outperformed past studies in this area, with an accuracy of 99.12% and Precision-Recall F1 scores between 98% and 100%. This technique might significantly impact the improvement of safety standards and traffic management. Making it feasible for emergency vehicles to be recognized and reacted to more promptly may improve traffic flow management and drastically reduce emergency response times. Throughout the study, multiple assessments on subsets of the available audio data (1800, 1000, and 600 audios) were conducted to ensure the consistency and reliability of the model's performance. The stacked LSTM models consistently outperformed the DNN and MLP models

across all test data sets. An unexpected observation was that the DNN models' test accuracy somewhat decreased when the audio input volume increased. This demonstrates that the learning rate and other hyperparameters may need to be accurately adjusted for larger datasets. Future research and advancements in this area are conceivable. The MLP models did not perform better than the LSTM and DNN models, but they did maintain consistent performance across different volumes of audio data. This is probable because MLP does not share LSTM Machine Learning Mastery's inability to manage temporal relationships and patterns in the audio input. Overall, it has been shown that the stacked ensemble approach, which combines the strengths of the MLP, DNN, and LSTM models, is a successful method for Classification of acoustic data. This work provides a solid foundation for further study and application in traffic safety and control. Further we used very basic models with few Hidden Layers to not to make model complex the real time detection of the the emergency vehicle sirens can be detected by more Efficently.

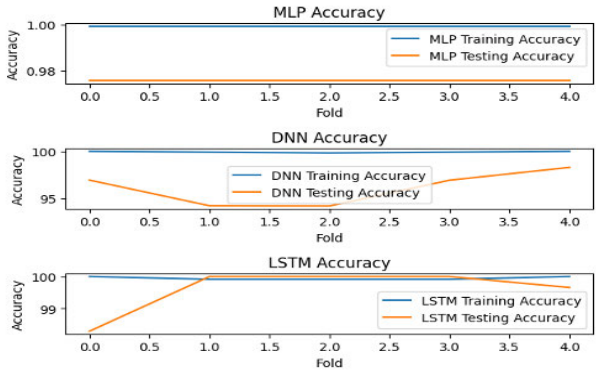

FIGURE 4. Accuracy plot with 5-Folds 1800 audios.

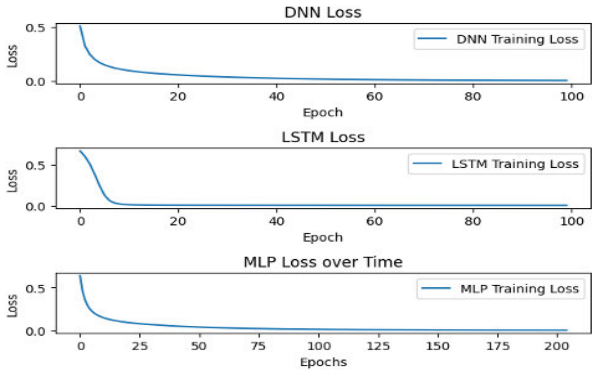

FIGURE 5. Loss with respect to training Epochs 1800 audios.

The performance of the LSTM, DNN, and MLP models on a dataset of 1000 audio recordings is shown in 2. The MLP model comes in at 99% accuracy, followed by the DNN model at 97.5%, and the LSTM model at 98.56%. The model loss during training is seen in Figure 5. Each model's loss is shown to decrease over time for each epoch, demonstrating how the model is doing better as it gains knowledge from the training set of data.

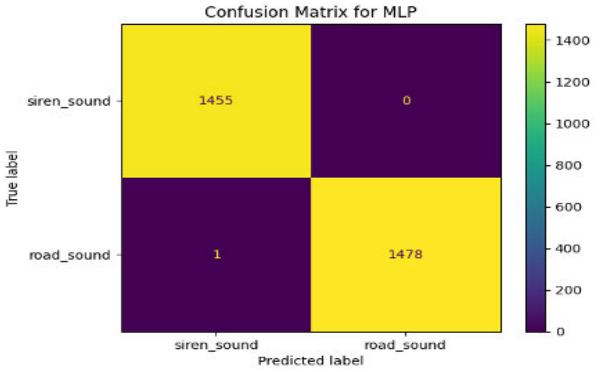

FIGURE 6. MLP confusion matrix of complete dataset (1800 Audios).

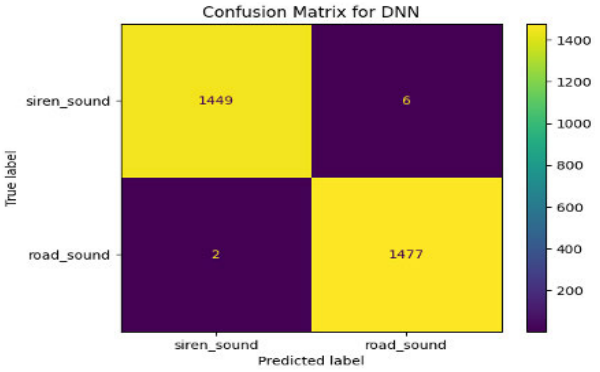

FIGURE 7. Confusion matrix of DNN complete dataset (1800 Audios).

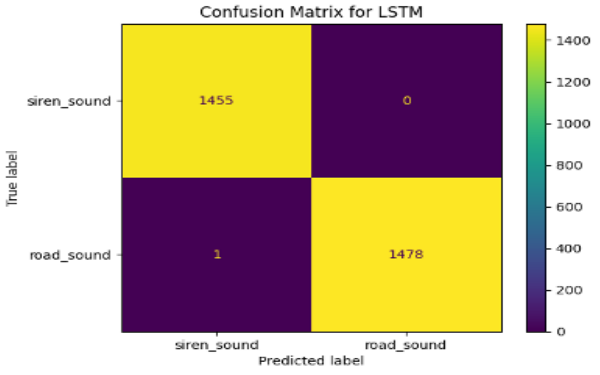

FIGURE 8. Confusion matrix LSTM complete dataset 1800 audios.

The model Accuracy and loss during training are presented in Figure 4 and Figure 5 respectively. Each model's loss is shown to decrease over time for each epoch, demonstrating how the model is doing better as it gains knowledge from the training set of data. The confusion matrix for the multi-layer perceptron (MLP) model is shown in Figure 6. The performance of the model is broken out in great depth in this matrix, which displays the quantity of true positives, false positives, true negatives, and false negatives. Only 1 False positive of road sound from 1500 training recordings is

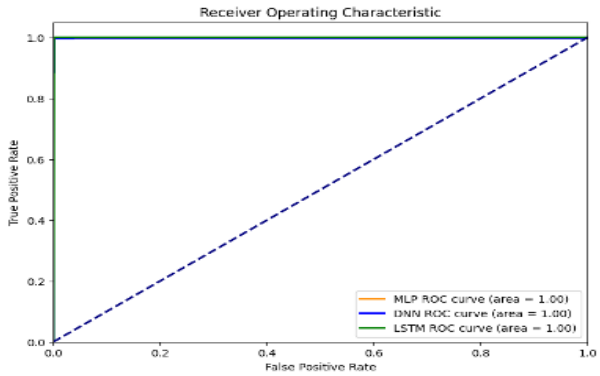

FIGURE 9. Au-roc curve of complete Dataset 1800 audios for Dnn, Mlp and Lstm.

present conveying the effectiveness of model. The confusion matrix for the DNN model is shown in Figure 7. It offers insights into the models, much like the prior picture, showing a 2 False positive of road sound from 1500 training recordings. The confusion matrix for the stacked LSTM model is shown in Figure 8 showing a 3 False positive of road sound from 1500 training recordings. The performance of the model is shown in detail in this matrix, which also shows the true positives, false positives, true negatives, and false negatives. The model's AU-ROC curve is shown in Figure 9. The trade-off between the true positive rate and the false positive rate at various classification levels may be seen on this graph, which measures performance for classification issues.

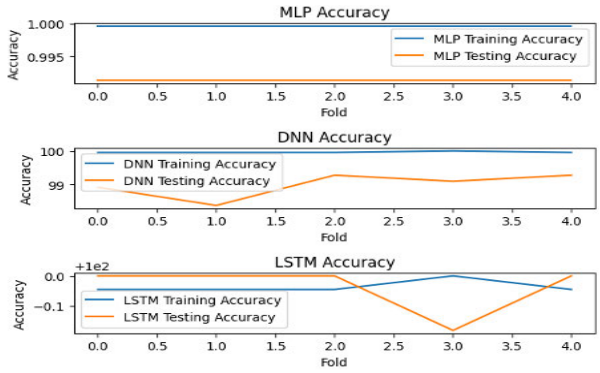

FIGURE 10. Accuracy with respect to 5-Folds 1000 audios.

A. SUBSET OF 1000 IMAGES RESULT PLOT

The accuracy of the LSTM, DNN, and MLP models is shown in Table 3, using a dataset of 1000 audio recordings. The MLP model comes in at 99% accuracy, followed by the DNN model at 98.54%, and the LSTM model at 99.12%. Figure 11 the model loss, which illustrates how the models have improved throughout training as the loss goes down.

The confusion matrices for the MLP, DNN, and stacked LSTM models are shown in Figures 12, 13, 14 respectively. These matrices thoroughly assess each model's performance, indicating how many true positives, false positives, true negatives, and false negatives there were overall. Non of

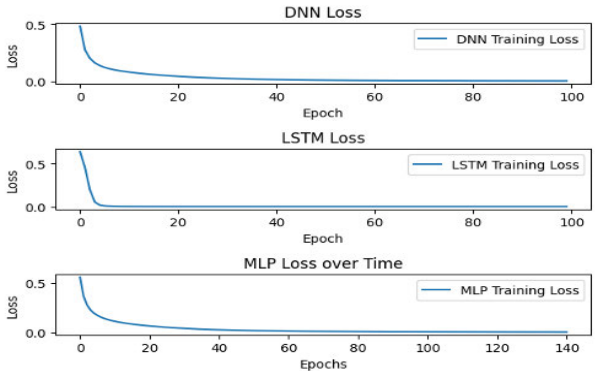

FIGURE 11. Model losses with respect to epochs 1000 audios.

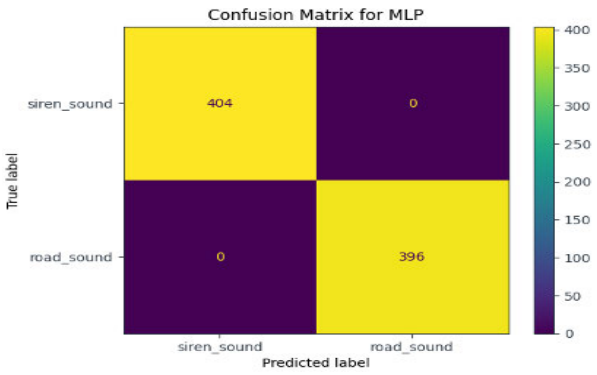

FIGURE 12. MLP confusion Matrix of complete Dataset (1000 Audios.

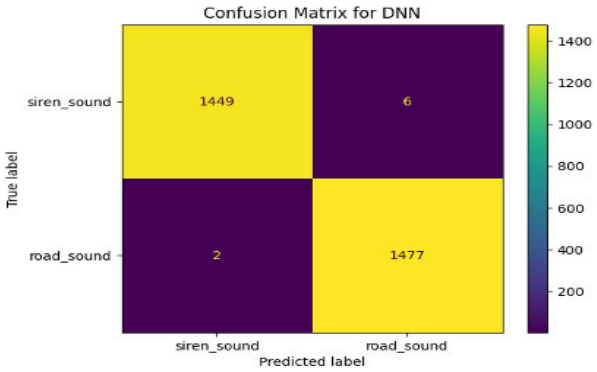

FIGURE 13. Confusion matrix of DNN complete dataset (1000 Audios).

the miss classification with MLP and Stacking LSTM and just 3 False postive records in Road noises. The AU-ROC curve for each model is shown in Figure 15. This curve offers a performance indicator for categorization issues at different threshold settings. It reveals how well a model can differentiate across classes.

B. SUBSET OF 600 IMAGES RESULTS

On a dataset of 600 audio files, Table 4 shows the precision of the LSTM, DNN, and MLP models. The MLP model comes in second with 98.2% accuracy, closely followed by the LSTM model with 98.46% accuracy and the DNN model with 97% accuracy. The model loss during training is shown in Figure 17. This graph demonstrates how the loss for each

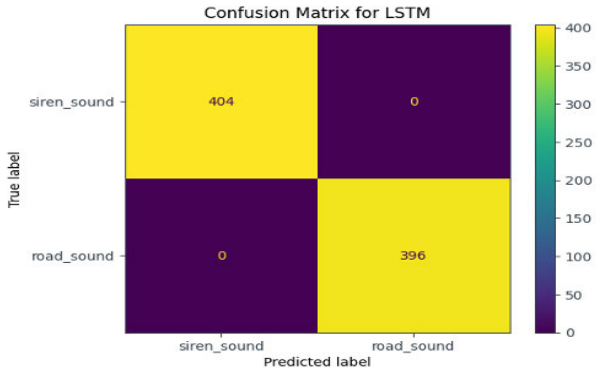

FIGURE 14. Confusion matrix stacked LSTM complete dataset 1000 Audios.

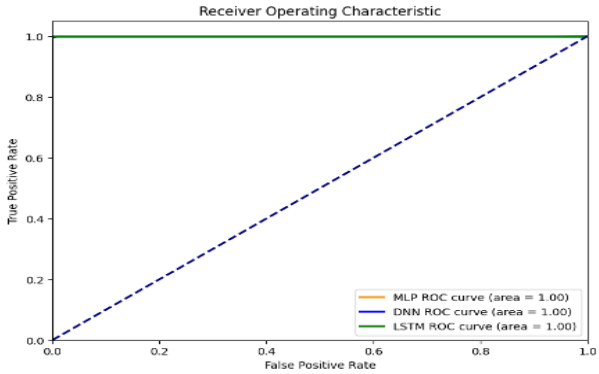

FIGURE 15. Au-roc curve of 1000 audios for Dnn, Mlp and LSTM.

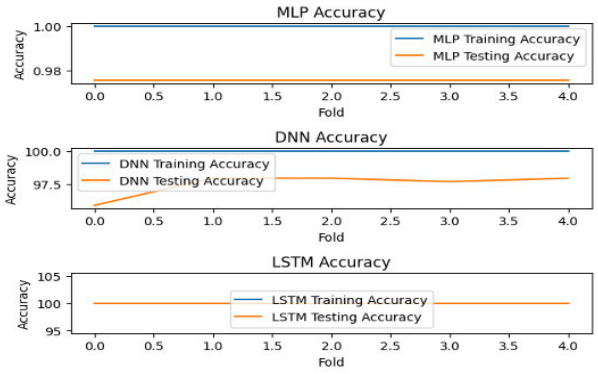

FIGURE 16. Accuracy with respect to 5-Folds 600 audios.

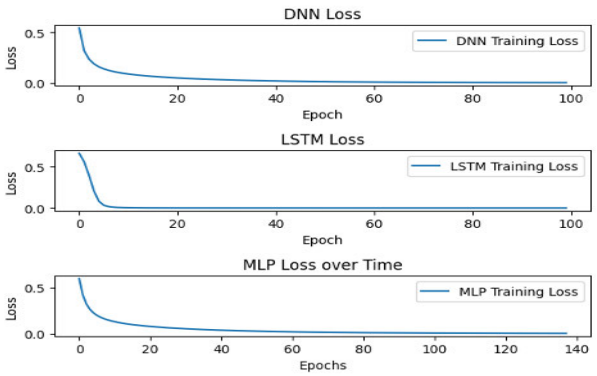

FIGURE 17. Model losses with respect to epochs 600 audios.

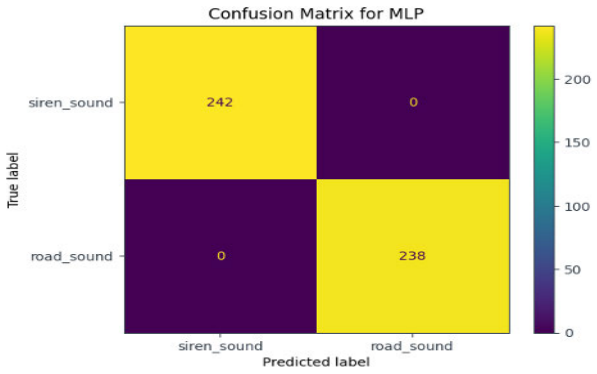

FIGURE 18. MLP confusion matrix of complete dataset (600 Audios).

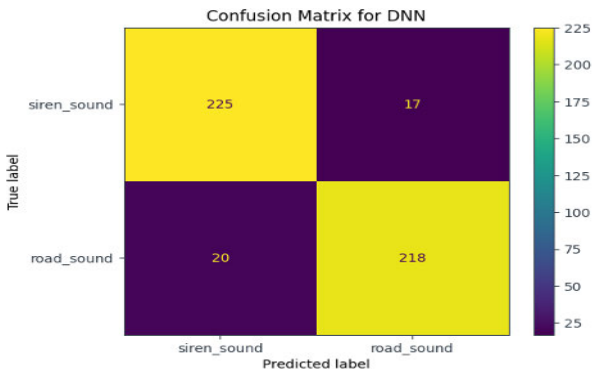

FIGURE 19. Confusion matrix of DNN complete dataset (600 Audios).

model lowers across epochs, demonstrating how the models become better as they gain knowledge from the training data.

The confusion matrices for the MLP, DNN, and stacked LSTM models are shown in Figures 19, 20 and 21, respectively. By displaying the quantity of true positives, false positives, true negatives, and false negatives, these matrices provide a thorough overview of the model's performance. They shed light on the kinds of mistakes the models make.

The suggested stacking LSTM model significantly outperforms the work by Asif et al. [13] on identical audio datasets in terms of accuracy. The suggested model obtains an accuracy of 98.56% for the 1800 audio files dataset, which is much higher than the 97% accuracy reported by Asif et al. The suggested model outperforms Asif et al.'s study's 94% accuracy for the 1000 audio files dataset, with an accuracy of 99.12%. For the 600 audio dataset, where the suggested model's accuracy is 98.46% vs. Asif et al.'s 90%, the difference is much more noticeable. The suggested stacking LSTM model's improved performance

TABLE 1. Result in comparison with an already present study on this data.

Model	1800 Audios	1000 Audios	600 Audios
our study Stacking LSTM	98.56%	99.12%	98.46%
Asif et al [12] Accuracy	97%	94%	90%

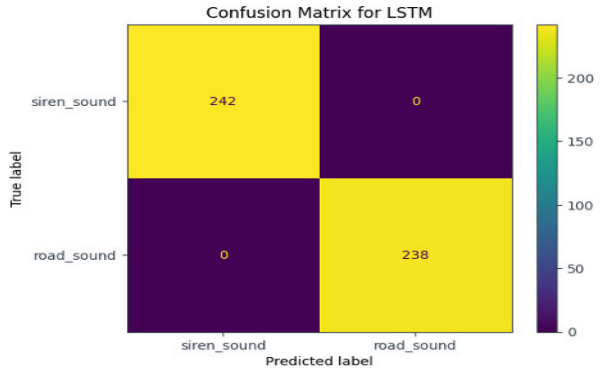

FIGURE 20. Confusion matrix stacked LSTM complete dataset 600 Audios.

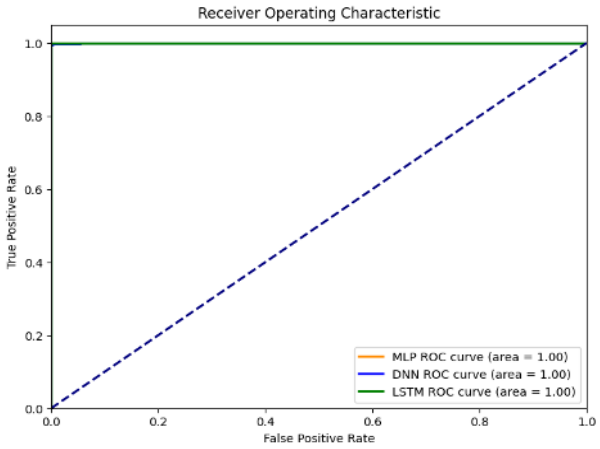

FIGURE 21. Au-roc curve of 600 audios for Dnn, Mlp and LSTM.

might be due to the LSTM models' prowess in handling sequential data well, as mentioned in the stacking of LSTM layers, which has been shown to boost model performance in difficult scenarios, may also have improved the model's performance. These results highlight the suggested stacking LSTM model's potential for audio categorization applications.

TABLE 2. 1800 audios results.

Models	Accuracy	Precision	Recall	F1-Score
LSTM	98.56%	98%	97%	98%
DNN	97.5%	98%	95%	98%
MLP	98.1%	98%	97%	98%

Table 2 shows the performance of three machine learning models LSTM, DNN, and MLP tested on a dataset of 1800 audio recordings. The LSTM model exceeds the other two models, and its accuracy rating of 98.56% is the highest. The LSTM and MLP models have a slightly greater recall of 97% compared to the DNN model's 95%, but all three models have the same accuracy and F1-score of 98%. Despite these minor variations, all three models perform well on this dataset. The LSTM model, however, could be the best option for this job given its maximum accuracy and balanced precision and recall, assuming that all metrics are regarded as equally essential for this particular work.

TABLE 3. 1000 audios results.

Models	Accuracy	Precision	Recall	F1-Score
LSTM	99.12%	100%	96%	98%
DNN	98.54%	100%	97%	97%
MLP	99%	100%	98%	98%

Table 3 shows the performance evaluation findings of three machine learning models, LSTM, DNN, and MLP, using a dataset of 1000 audio recordings. The LSTM model performs well, with a precision of 100% and an accuracy of 99.12%. While the DNN model's accuracy is also 100%, its precision is much lower 98.54%. The MLP model works well, with a precision of 100% and an accuracy of 99%. Although all models are quite accurate, the optimal selection may depend on how critical a certain metric is for the work.

TABLE 4. 600 audios results.

Models	Accuracy	Precision	Recall	F1-Score
LSTM	98.46%	99%	96%	98%
DNN	97%	98%	99%	97%
MLP	98.2%	99%	98%	98%

Table 4 is showingThe three machine learning models—LSTM, DNN, and MLP—are compared in the table based on how well they performed on a dataset of 600 audio recordings. The evaluation metrics employed are the F1-Score, accuracy, precision, and recall. The LSTM model, which has the highest accuracy 98.46% and F1-score 98%, is the best model for this issue. It assumes equal weight for all measurements. Even though the DNN model has a higher recall 99.12%, it performs slightly worse than the LSTM. The MLP model and LSTM are comparable in terms of performance metrics.

V. CONCLUSION AND DISCUSSION

To conclude this study we want to present our suggested approach performance on the Road emergency siren detection with the large acoustic dataset recorded in real world setting, Our Approach has achieved significant results with of Stacking LSTM achieved 98.56% Accuracy when trained and test on 1800 audios. With 1000 recording our Proposed has gained an Accuracy of 99.12% and with the 600 audios we got 98.46% Accuracy. These results depicting the strong base of our study with Multiple features and Meta Learning we can make more Efficient Model. Further when comparing with Latest study on this Dataset for Siren detection we have gained 2% more accuracy than latest study on this dataset. We have compared our Suggested Method Results with our Peer study in [12]. The application f this study can be deployed in real world and can help in creating Smart cities Traffic management more responsive and effective.

ACKNOWLEDGMENT

The authors thank and extend their appreciation to the funders of this work.

REFERENCES

[1] L. Lu, H.-J. Zhang, and H. Jiang, "Content analysis for audio classification and segmentation," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 7, pp. 504–516, Oct. 2002, doi: 10.1109/TSA.2002.804546.

[2] Z. Azkorra, G. Pérez, J. Coma, L. F. Cabeza, S. Bures, J. E. Álvaro, A. Erkoreka, and M. Urrestarazu, "Evaluation of green walls as a passive acoustic insulation system for buildings," *Appl. Acoust.*, vol. 89, pp. 46–56, Mar. 2015, doi: 10.1016/j.apacoust.2014.09.010.

[3] G. Sharma, K. Umapathy, and S. Krishnan, "Trends in audio signal feature extraction methods," *Appl. Acoust.*, vol. 158, Jan. 2020, Art. no. 107020, doi: 10.1016/j.apacoust.2019.107020.

[4] D. Bonet-Solà and R. M. Alsina-Pagès, "A comparative survey of feature extraction and machine learning methods in diverse acoustic environments," *Sensors*, vol. 21, no. 4, p. 1274, Feb. 2021, doi: 10.3390/s21041274.

[5] Y. Su, K. Zhang, J. Wang, and K. Madani, "Environment sound classification using a two-stream CNN based on decision-level fusion," *Sensors*, vol. 19, no. 7, p. 1733, Apr. 2019, doi: 10.3390/s19071733.

[6] A. Zbiciak and T. Markiewicz, "A new extraordinary means of appeal in the Polish criminal procedure: The basic principles of a fair trial and a complaint against a cassatory judgment," *Access Justice Eastern Eur.*, vol. 6, no. 2, pp. 25–42, Mar. 2023, doi: 10.33327/ajee-18-6.2-a000209.

[7] U. Mittal and P. Chawla, "Acoustic based emergency vehicle detection using ensemble of deep learning models," *Proc. Comput. Sci.*, vol. 218, pp. 227–234, Jan. 2023, doi: 10.1016/j.procs.2023.01.005.

[8] V.-T. Tran and W.-H. Tsai, "Acoustic-based emergency vehicle detection using convolutional neural networks," *IEEE Access*, vol. 8, pp. 75702–75713, 2020, doi: 10.1109/ACCESS.2020.2988986.

[9] J. Schroder, S. Goetze, V. Grutzmacher, and J. Anemuller, "Automatic acoustic siren detection in traffic noise by part-based models," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Process.*, May 2013, pp. 493–497, doi: 10.1109/ICASSP.2013.6637696.

[10] L. Marchegiani and I. Posner, "Leveraging the urban soundscape: Auditory perception for smart vehicles," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2017, pp. 6547–6554, doi: 10.1109/ICRA.2017.7989774.

[11] O. Karpis, "System for vehicles classification and emergency vehicles detection," *IFAC Proc. Volumes*, vol. 45, no. 7, pp. 186–190, 2012, doi: 10.3182/20120523-3-cz-3015.00037.

[12] S. Sathruhan, O. K. Herath, T. Sivakumar, and A. Thibbotuwawa, "Emergency vehicle detection using vehicle sound classification: A deep learning approach," in *Proc. 6th SLAAI Int. Conf. Artif. Intell. (SLAAI-ICAI)*, Dec. 2022, pp. 1–6, doi: 10.1109/SLAAI-ICAI56923.2022.10002605.

[13] M. Asif, M. Usaid, M. Rashid, T. Rajab, S. Hussain, and S. Wasi, "Large-scale audio dataset for emergency vehicle sirens and road noises," *Sci. Data*, vol. 9, no. 1, p. 599, Oct. 2022, doi: 10.1038/s41597-022-01727-2.

[14] Z. Islam and M. A. Abdel-Aty, "Real-time emergency vehicle event detection using audio data," 2022, *arXi:2202.01367*.

[15] A. Baghel, A. Srivastava, A. Tyagi, S. Goel, and P. Nagrath, "Analysis of Ex-YOLO algorithm with other real-time algorithms for emergency vehicle detection," in *Proc. 1st Int. Conf. Comput., Commun., Cyber-Secur. (IC4S)*. Singapore: Springer, 2020, pp. 607–618.

[16] K. Choudhury and D. Nandi, "Review of emergency vehicle detection techniques by acoustic signals," *Trans. Indian Nat. Acad. Eng.*, vol. 8, no. 4, pp. 535–550, Sep. 2023, doi: 10.1007/s41403-023-00424-9.

[17] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," 2016, *arXiv:1608.04363*.

[18] J. Lee, T. Kim, J. Park, and J. Nam, "Raw waveform-based audio classification using sample-level CNN architectures," 2017, *arXiv:1712.00866*.

[19] B. Fatimah, A. Preethi, V. Hrushikesh, A. Singh B., and H. R. Kotion, "An automatic siren detection algorithm using Fourier decomposition method and MFCC," in *Proc. 11th Int. Conf. Comput., Commun. Netw. Technol. (ICCCNT)*, Jul. 2020, pp. 1–6, doi: 10.1109/ICCCNT49239.2020.9225414.

[20] L. Marchegiani and P. Newman, "Listening for sirens: Locating and classifying acoustic alarms in city scenes," 2018, *arXiv:1810.04989*.

AHSAN SHABBIR was born in Pakistan, in 1998. He received the degree in computer science from the Government College University of Faisalabad, in 2020, and the master's degree in data science from Air University, Islamabad, in 2023. He is currently a Research Officer with Pakistan Air Force. His main research interests include acoustics and AI with its applications in intelligent homes, smart cities, bio acoustics, and marine acoustics.

AMMARA NAWAZ CHEEMA received the M.Sc. and M.Phil. degrees in statistics from Quaid-i-Azam University (QAU), Islamabad, Pakistan, in 2007 and 2009, respectively, and the Ph.D. degree in statistics from Riphah International University, Islamabad. She was with Pakistan Bureau of Statistics, National Accounts, Islamabad, for more than one year. She has more than 20 years of teaching and research experience. She has been an Assistant Professor with the Department of Mathematics, Air University, Islamabad, since 2010. Her research interests include Bayesian statistics, engineering probability, statistical learning, applied statistics, survey methodology, financial statistics, and econometrics and data analysis. She received the Gold Medal during the master's degree.

INAM ULLAH received the B.S. degree in electronics engineering from International Islamic University, in 2016, and the M.S. degree in artificial intelligence from Air University, Islamabad, in 2023. He is a certified data scientist and machine learning engineer. He has a profound expertise in delving into machine learning, deep learning, information retrieval, natural language processing, and computer vision, dedicated to pushing the boundaries of AI and technology. He is an emerged prominent figure in the field, possessing a rich educational background, robust experience, and a portfolio of exceptional projects.

IBRAHIM M. ALMANJAHIE was born in Saudi Arabia, in 1979. He received the B.Sc. degree in mathematics from King Khalid University, Saudi Arabia, in 2002, and the M.Sc. degree in mathematical and statistical science and the Ph.D. degree in probability and statistical modeling from The University of Western Australia, Australia, in 2008 and 2015, respectively. He is currently a Professor with the Department of Mathematics, College of Science, King Khalid University. He is also the President of the Saudi Association for Statistical Sciences. His main research interests include modeling and analysis of ion channel data, hidden Markov models, the EM algorithm, finite mixture models, MCMC, computational methods in statistics, time series modeling, applied statistics, sampling, and functional statistics.

FATIMAH ALSHAHRANI received the bachelor's degree in mathematics from Princess Nourah bint Abdulrahman University, Riyadh, the Master of Science degree from The University of Queensland, in 2010, and the Ph.D. degree in statistics from Michigan State University, in 2020. She has been an Assistant Professor of statistics and probability with Princess Nourah bint Abdulrahman University, since 2021, where she was a Lecturer in 2010. She has published papers in the field of statistics and probability, especially in functional data analysis, probability theory, and applied statistics.