Vehicle Sound Recognition Assistance in IoT Systems for Hearing-Impaired Drivers

Osman Salem, Ahmed Mehaoua, and Raouf Boutaba

Abstract

Hearing-impaired drivers face significant challenges in detecting critical auditory cues, such as emergency vehicle sirens, essential for safe driving. This article presents an advanced IoT-based sound recognition system designed to enhance situational awareness for these drivers. Audible signals are recognized and transformed into alerts displayed in the dashboard. Our approach involves preprocessing audio data to extract 23 features. We normalize these features and evaluate multiple Machine Learning and Deep Learning models for their classification performance. The top five models, selected based on their performance metrics, are then combined into an ensemble model using majority voting to improve accuracy and robustness. Our dataset comprising 1500 audio samples enabled us to achieve a final accuracy of 94.2% with the ensemble voting approach. These results demonstrate a significant performance in sound classification accuracy compared to individual models, indicating the effectiveness of our ensemble approach. This research provides a valuable step towards developing more accessible and safer driving assistance systems for individuals with hearing impairments.

Introduction

Hearing impairment is a significant global health issue, with the World Health Organization (WHO) projecting that approximately 2.5 billion people will experience some degree of hearing loss by 2050. This issue is compounded by the fact that many young adults are unaware of the risks associated with unsafe listening practices [1]. Hearing loss can have profound effects on individuals, leading to psychological distress, anxiety, and depression. The suicide rate among deaf or hard-of-hearing adolescents aged 10 to 20 is reported to be twice as high as that of their hearing peers. Moreover, adults with hearing impairments often face considerable challenges in the workforce and daily life, struggling with issues of independence and integration. Several research works and products have been developed to assist people with such disability, including smart rings for sign language translation for example [2].

One of the most critical areas affected by hearing loss is driving. Effective driving requires constant situational awareness, including the ability to perceive auditory signals such as honks and emergency vehicle sirens. While individuals with hearing impairments may develop heightened visual acuity, they still face significant difficulties detecting and responding to these crucial auditory cues. This can be particularly dangerous as it may lead to delayed reactions to important warnings, increasing the risk of accidents [3].

Modern vehicles often feature sound-isolation technologies and entertainment systems that can further hinder sound detection. Distractions such as loud music and in-car conversations exacerbate the issue, making it even more challenging for hearing-impaired drivers to detect and respond to critical sounds. Studies have shown that such distractions contribute to an elevated risk of traffic accidents for drivers with hearing impairments [4]. The inability to hear emergency vehicle sirens or traffic signals can result in dangerous delays and even collisions, highlighting the urgent need for innovative solutions to enhance driving safety for these individuals.

In response to these challenges, we propose an advanced sound recognition system based on the Internet of Things (IoT) to assist drivers with hearing impairments. Our system employs four sensors installed at the front and rear of the vehicle to capture and analyze environmental sounds. Our approach classifies and interprets various vehicle sounds, including emergency sirens, horns, and engine noises with both Machine Learning (ML) and Deep Learning (DL) algorithms such as K-Nearest Neighbors (KNN), Adaptive Boost (AB), Extra Trees (ET), Random Forest (RF), and Convolutional Neural Network (CNN). We preprocess the audio data to extract 23 distinct features, which we then use to improve sound recognition accuracy and provide timely alerts to drivers through a visual display on the dashboard. To enhance the reliability and performance of our system, we employ an ensemble learning strategy that combines the outputs of multiple models using Majority Voting (MV). This method not only improves classification performance but also increases robustness against varying environmental conditions.

Our prototype for real-time experimentation includes a Raspberry Pi 4, microphones, and a 3.5-inch screen. The system captures and preprocesses environmental sounds, which are then classified based on a training dataset of 1500 audio samples. The classification results are translated into text and displayed on the screen to assist hearing-impaired drivers by identifying and alerting them to surrounding noises. This study

Digital Object Identifier: 10.1109/IOTM.001.2400140

Osman Salem and Ahmed Mehaoua are with the Université Paris Cité, France; Raouf Boutaba is with the University of Waterloo, Canada.

2576-3180/25/$25.00 © 2025 IEEE

extends our previous work in [5] on IoT-based sound recognition systems for drivers with hearing impairments which only had a dataset of 600 samples and 4 classes (ambulance, police cars, horns, and other sounds).

This article makes several contributions to the field of sound recognition systems for drivers:

- We present an advanced system that integrates multiple sensors and ML techniques to detect and classify critical vehicle sounds.
- Building upon our previous research, we have expanded the dataset with more samples including additional sound classes, such as fire trucks and engine noises, and performed extensive experiments to refine the system's performance.
- We employ an ensemble learning method that combines the outputs of KNN, AB, ET, RF, and CNN models through MV. This method enhances classification performance, robustness, and security.
- A prototype system was developed using a Raspberry Pi 4. The system was evaluated with a dataset of 1500 audio samples, achieving a final classification accuracy of 94.2% and an average response time of 0.25 seconds, demonstrating its effectiveness in real-world scenarios and the possibility of implementing our approach in any microcontroller.

The remainder of this article is organized as follows. We review recent related work in the field. We detail the components and methodology of our proposed approach for converting detected sounds into visual messages on the dashboard. We present the experimental results and performance analysis of our system. We discuss the challenges encountered in practical implementation and their potential solutions. Finally, we offers concluding remarks and discusses future research directions.

Related Work

The intersection of auditory perception and vehicular safety has garnered increasing attention in recent years, particularly with the advent of advanced technologies such as ML and the IoT. This section reviews related research that informs and contextualizes our work in sound recognition systems for drivers with hearing impairments.

Sound recognition has been extensively explored across various domains, including environmental monitoring and safety systems. In the context of automotive applications, sound recognition systems are employed to enhance driver safety and convenience. Early work in sound classification focused on using traditional signal-processing techniques. Suman *et al.* in [6] used acoustic signal processing to detect mechanical malfunctions in vehicles. They introduced a smart device that can be installed inside the vehicle, equipped with both a microphone and vibration sensors to capture relevant signals. The device employs a microcontroller to process these signals using a proposed algorithm that combines Kalman filtering and Mel-Frequency Cepstral Coefficients (MFCCs) to identify mechanical faults. The proposed Kalman adaptive filter enhances acoustic signals by reducing noise to detect faults in rotating equipment.

More recent advances leverage ML and DL to achieve higher accuracy and robustness. Shabbir *et al.* in [7] determined that acoustic data analysis plays a crucial role in smart traffic management systems, especially in distinguishing road noises and emergency vehicle sirens to enhance emergency response times and traffic flow. This study proposed using stacking ensemble DL techniques to classify emergency vehicle sirens from background noises. The proposed model utilizes Multi-Layer Perceptron (MLP) and Deep Neural Network (DNN) as base learners, with a Long Short-Term Memory (LSTM) model serving as a meta-learner. The stacking ensemble achieved an accuracy of 99.12% and F1 scores around 98%. However, there were only two classes, and different emergency vehicles were not distinguished.

Usaid *et al.* in [8] applied MLP to detect the siren of an ambulance on the road. Their model achieved 90% detection accuracy with a dataset of only 300 files, but their model is limited to two classes: siren and noise. Islam *et al.* in[9] applied Extreme Learning Machines (ELM) for the detection of emergency vehicles. Their experimental results on a dataset of 2000 audio clips showed a detection accuracy of 97% during classification into two categories: emergency vehicles and urban sounds. Cantarini *et al.* in [10] proposed an emergency siren detection system using a low computational complexity algorithm based on CNN, with Short-Time Fourier Transform (STFT) spectrograms as features, and a harmonic percussive source separation technique to improve the accuracy of the classification. Jonnadula *et al.* in [11] compared different classification methods and features for emergency vehicle detection. They found that an Artificial Neural Network (ANN) with three hidden layers presents higher accuracy compared to one layer. They used Google Audio Dataset with noises like people talking and horns in their experiments.

Otoom *et al.* in [12] propose an assisting device to help deaf drivers receive GPS directions using voice recognition and speech-to-vibration. The spirit of their work is like ours. They map each voice navigation to a vibrotactile stimulus (vibrator motors) mounted on a bracelet, where the vibrations are translated by deaf drivers into 6 instructions. They extract 13 features from audible signals and classify them into one of six classes (turn left, turn right, slight right, slight left, straight, and silence) using ML algorithms. They compare the accuracy of Naïve Bayes (NB), KNN, Support Vector Machine (SVM), and RF. They found that KNN with $K = 1$ outperforms the five others.

Gourisaria *et al.* in [13] emphasized the superior potential of DL models for sound classification in diverse contexts such as on the road, at home, or in parks. It leveraged algorithms such as MFCCs and STFT to manage and analyze audio data. The study demonstrated that DL models, particularly ANN, exhibit remarkable efficiency and accuracy in classifying audio signals, outperforming other models with an accuracy of 91.41% and 91.27% on different datasets. This work highlights the effectiveness of DL techniques in handling complex and noisy audio environments, in comparison to traditional ML models such as SVM, Decision Tree (DT), RF, and KNN.

In particular, CNN classifiers have been privileged for the recognition of sounds as they achieve better accuracy. Nithya *et al.* in [14] presented a sophisticated approach to vehicle sound

Sound recognition has been extensively explored across various domains, including environmental monitoring and safety systems. In the context of automotive applications, sound recognition systems are employed to enhance driver safety and convenience.

The system is designed to provide timely visual alerts of important auditory cues, such as emergency sirens, horns, and engine noises, through a dashboard display.

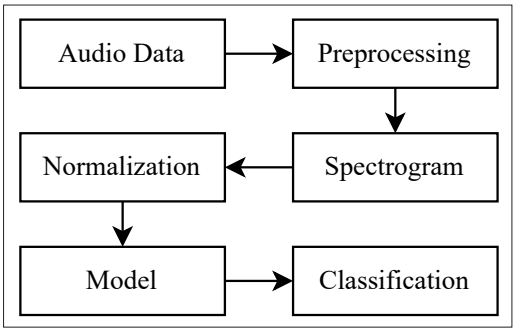

FIGURE 1. Workflow from audio input to sound classification.

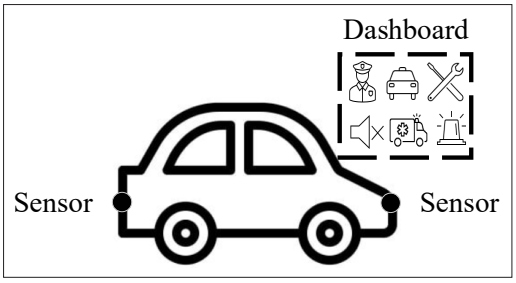

FIGURE 2. Sound recognition system overview.

identification and classification, focusing on emergency vehicles. A key highlight of the article is the introduction of a Triangular Bluestein (TB) MFCCs multifuse feature, which combines advanced techniques in data augmentation and feature extraction to enhance model performance. This involves augmenting the audio data through noise injection, stretching, shifting, and pitching to increase dataset diversity and combat overfitting. The study employs a Multi-stacked CNN (MCNN) integrated with an Attention-based Bidirectional LSTM (A-BiLSTM) network, thus improving prediction accuracy. The results show that this combined approach achieves a classification accuracy of 98.66%, demonstrating CNN's potential for precise sound classification.

Shams *et al.* in [15] introduced a DL model called the Self-Attention Layer within a Convolutional Neural Network (SACNN), designed for detecting acoustic data in extensive datasets containing emergency vehicle sirens and road noise. The dataset includes 3-second audio files categorized into three classes: fire truck, ambulance, and traffic sounds, with 200 sound files and corresponding spectrogram images per class. SACNN combines EfficientNet and One-Dimensional CNN (1D-CNN) to enhance detection accuracy and efficiency. Experimental results demonstrated that SACNN achieved a perfect accuracy, surpassing the performance of EfficientNet, which had an accuracy of 94.16%, and CNN, which achieved 96.66%. SACNN also shows the highest computational efficiency with an average of 0.21 seconds per sample. These findings highlight the superior capabilities of CNN-based models in sound classification tasks and their potential for real-time acoustic data detection systems.

A review of the existing literature reveals that most research in the field focuses on detecting general categories of emergency vehicles, without distinguishing between specific types, such as police cars, ambulances, and fire trucks. In contrast, our research aims to provide a more granular classification by differentiating among these types of emergency vehicles and additional classes such as engine sounds, horns, and other road noises. Additionally, many of these studies focus primarily on algorithmic development and do not extend to real-world application and testing. Our work addresses this gap by implementing a practical IoT-based system using a Raspberry Pi for real-time experiments. This approach not only demonstrates the feasibility of deploying such systems but also provides valuable insights into the system's performance and reliability in real-world scenarios.

Proposed Approach

To address the challenges faced by hearing-impaired drivers, we propose an advanced IoT-based sound recognition system that enhances situational awareness by detecting and interpreting critical vehicle sounds. Our approach, illustrated in Fig. 1, involves several key components: sensor deployment, audio data preprocessing, feature extraction, model training, and ensemble learning. The system is designed to provide timely visual alerts of important auditory cues, such as emergency sirens, horns, and engine noises, through a dashboard display.

As shown in Fig. 2, the system consists of four sensors installed on the vehicle, 2 in the front and 2 in the back to capture ambient audio signals and their directions to draw the attention of the driver. These sensors are connected to a central processing unit, which preprocesses the audio data before classification. Preprocessing includes noise reduction to eliminate background interference and normalization to standardize the audio signals, ensuring consistency for accurate feature extraction and classification.

In our sound recognition system, feature extraction is a crucial step that transforms raw audio data into a structured form suitable for ML models. To enhance classification efficiency and ensure comprehensive sound capture, we split each audio file into segments of 0.5 seconds, with an overlap of 0.1 seconds between consecutive segments. This segmentation approach helps ensure that the system captures complete sounds, allowing for more precise classification. By processing these smaller audio segments, we optimize the model's performance and improve response speed, which is vital for timely alerts to the driver. This approach also enhances the system's ability to process data safely and swiftly, leading to a more efficient sound recognition process.

From each audio segment, we extract 23 distinct features to capture its essential characteristics. These features include the Zero Crossing Rate (ZCR), Spectral Centroid (SC), Spectral Roll-Off Point (SROP), and 20 MFCCs. Together, these features enable the differentiation between the six sound classes: Ambulance, Police Car, Fire Truck, Horn, Engine, and Other. This structured representation of audio data ensures that our models can accurately classify each sound, leading to a reliable and effective sound recognition system for hearing-impaired drivers.

The ZCR is a time-domain feature that measures the rate at which the audio signal crosses the zero-amplitude line. It reflects the noisiness

and the frequency content of the signal. Sounds with more abrupt changes, such as horn sounds, typically exhibit a higher ZCR due to their sharp, transient nature. In contrast, ambulance and police sirens, which have smoother waveforms, tend to have a lower ZCR.

SC is a frequency-domain feature that indicates the center of mass of the spectrum. It is associated with the perceived brightness or sharpness of a sound. Higher SC values suggest that the sound has a higher frequency content. For example, fire truck sirens, which often contain high-frequency components, tend to have higher SC values. On the other hand, engine sounds, characterized by lower frequency content, usually exhibit lower SC values.

Another key frequency-domain feature is the SROP. This feature defines the frequency below which a specified percentage of the total spectral energy is contained. SROP measures the right-side asymmetry of a sound spectrum. It is used for example to distinguish sound from normal speech. It provides insights into the spectral shape and distribution of energy across frequencies. Sounds such as sirens often have higher SROP values, reflecting their broad frequency spectrum, whereas engine sounds, which concentrate their energy in lower frequencies, show lower SROP values.

The MFCCs correspond to a sinusoidal transformation of the power of a signal and capture the short-term power spectrum of a sound signal. They are derived from the Mel scale, which approximates the human ear's nonlinear response to different frequencies. MFCCs are crucial for capturing the timbral texture of sounds. Each of the six sound classes has distinctive MFCC patterns that reflect their unique acoustic properties. For instance, horn and siren sounds may display pronounced MFCC variations due to their harmonic and tonal characteristics, while engine sounds often have more stable MFCC patterns due to their consistent frequency content. We used 20 MFCCs to derive the shape of the signal.

Then, we classify the extracted features into distinct sound categories using several ML and DL models including KNN, AB, ET, RT, and CNN. Each model offers unique strengths and weaknesses, making them suitable for different aspects of the classification task.

The KNN algorithm is a straightforward, instance-based learning method that classifies data into one of the target classes by looking at the values of the k nearest neighbors in the feature space. KNN is particularly effective at capturing local patterns and can be highly accurate when the feature space is well-defined. However, its performance may degrade with large datasets due to increased computational costs and sensitivity to irrelevant or redundant features.

AB is an ensemble method that combines multiple weak classifiers to form a strong classifier. It assigns weights to misclassified instances, focusing on improving their classification in subsequent iterations. AB is robust to overfitting and can achieve high accuracy, but its performance may decline in the presence of noisy data, as it tends to focus heavily on difficult-to-classify instances.

The ET classifier is an ensemble learning technique based on randomized decision trees. It builds multiple trees from random subsets of the data and features, which helps reduce variance and improves robustness. ETs are computationally efficient and offer good performance with large datasets, but they may require careful tuning to achieve optimal results.

RF is another ensemble method that aggregates predictions from numerous decision trees, each built on a random subset of the data. It provides high accuracy and resilience to overfitting by averaging the predictions of individual trees. However, RF can be computationally expensive and may become unwieldy with very large datasets.

Finally, the CNN is a DL model specifically designed to learn hierarchical features from raw data automatically. CNNs are highly effective at capturing complex patterns in sound signals, making them ideal for sound classification tasks. They can generalize well with large amounts of training data but require significant computational resources and careful design to prevent overfitting. The CNN classification can be divided into two main parts: feature learning and classification. In the first part, a series of convolutional layers learn appropriate representations by extracting useful features from the input. In the second part, the fully connected layers act as a classifier, which processes the extracted features and assigns probabilities to each class for prediction.

To optimize both the accuracy and robustness of sound classification, we employ an ensemble learning strategy. This approach combines the outputs of our five models using a MV mechanism. By integrating these models, ensemble learning capitalizes on their strengths and mitigates their weaknesses, resulting in a system that performs better than any single model.

The core advantage of ensemble learning lies in its ability to aggregate predictions from multiple models, thereby achieving higher accuracy. Each model in the ensemble brings a unique perspective to the classification task. For instance, while decision tree-based models like RF and ET excel at capturing complex decision boundaries, KNN is adept at recognizing local patterns in the feature space. AB, with its focus on misclassified instances, contributes to refining the decision-making process. Meanwhile, CNNs offer DL capabilities and are adept at identifying intricate patterns in the audio signals.

This diversity in model approaches enhances the ensemble's ability to generalize to new, unseen data. By pooling the predictions, the ensemble reduces the variance associated with individual models, leading to a more stable and reliable system. It also diminishes the impact of noisy or outlier data points, as different models may react differently to such anomalies, allowing the ensemble to provide a consensus decision that is less likely to be swayed by errors present in any one model.

In addition to improved accuracy and robustness, ensemble learning provides increased resilience to overfitting. While single models might tailor their learning too closely to the training data, the ensemble's diverse perspectives help balance this tendency, ensuring that the final model retains its effectiveness across different datasets.

Moreover, the ensemble approach enhances the system's security against adversarial attacks. Manipulating a single model to produce erroneous results can be challenging, but deceiving multiple

The core advantage of ensemble learning lies in its ability to aggregate predictions from multiple models, thereby achieving higher accuracy. Each model in the ensemble brings a unique perspective to the classification task.

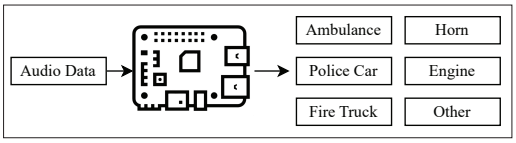

FIGURE 3. Sound processing and classification on Raspberry Pi.

Model	Acc	Pre	Rec	F1	AUC	Time Taken	
						Train[1]	Test[2]
KNN	0.868	0.876	0.868	0.867	0.88	0.01	29.50
AB	0.916	0.912	0.916	0.902	0.93	5.80	30.40
ET	0.912	0.894	0.912	0.899	0.92	3.50	18.10
RF	0.926	0.920	0.926	0.913	0.95	5.00	26.50
CNN	0.922	0.926	0.922	0.922	0.97	106.21	208.14
MV	0.942	0.971	0.971	0.971	0.99	106.21	208.14

[1] In seconds.
[2] Time for a single prediction. In microseconds.

TABLE 1. Performance metrics after feature extraction from cnn.

models simultaneously is significantly more complex. This makes the ensemble more robust against potential threats, ensuring that it maintains high performance even in adversarial environments.

The use of ensemble learning in our system is a strategic choice that balances the need for high classification accuracy, robustness to data variations, and resilience to external manipulations. By integrating the strengths of diverse models, our ensemble offers a comprehensive solution that significantly enhances the reliability of sound recognition for hearing-impaired drivers.

To demonstrate the feasibility and effectiveness of our approach, we developed a prototype system using a Raspberry Pi 4, high-sensitivity microphones, and a 3.5-inch screen. As depicted in Fig. 3, the Raspberry Pi serves as the central processing unit, responsible for capturing audio signals, performing feature extraction, and executing the ensemble model for classification. Once the sounds are classified, the results are displayed as text on the screen, providing clear and timely visual alerts to the driver about the surrounding environment.

Experimental Results

This section presents the experimental evaluation of our IoT-based sound recognition system designed to assist drivers. We evaluated the system using a dataset that we developed ourselves, consisting of 1500 audio samples across six distinct classes: Ambulance, Police Car, Fire Truck, Horn, Engine, and Other. As explained earlier, each audio file was segmented into 0.5-second clips with an overlap of 0.1 seconds. We included external noise, such as rain and lightning, in the audio samples to enhance the dataset's realism and robustness. From these clips, we then extracted 23 features. The dataset was carefully balanced to ensure that each class was equally represented, which helped mitigate potential bias across different models.

The data normalization process involves scaling the spectrograms to ensure that the model receives inputs in a standardized range. The input data (spectrograms) is normalized by dividing the values by the amplitude range to scale the data between 0 and 1. This normalization step helps to reduce the effect of varying audio signal intensities and ensures stable training. The dataset is split into 80% training data and 20% test data, and this ensures that 80% of the data is used for training and 20% for testing the model's performance.

To assess the performance of our sound recognition models, we employed several key metrics: accuracy, precision, recall, F1 score, and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). These metrics provide a comprehensive evaluation of the models' ability to classify sounds accurately. Additionally, we analyzed the computational efficiency of each model by recording training times and the duration for a single prediction.

The CNN architecture used in this model consists of several layers designed to extract features from audio data. It includes three convolutional layers, each followed by max-pooling layers for downsampling the feature maps. The convolutional layers utilize ReLU activation functions to introduce non-linearity. The model incorporates dropout layers (with a dropout rate of 0.3) after the second pooling layer and before the fully connected layers to prevent overfitting. There are no batch normalization layers used in this architecture. The fully connected layers consist of 64 and 32 neurons, also employing ReLU activations, and the final output layer has 4 neurons with a softmax activation for multi-class classification. This setup is designed to efficiently process audio spectrograms and classify them into one of 6 classes.

The CNN model is trained using the Adam optimizer, which is chosen for its adaptive learning rate and efficiency in handling sparse gradients. The loss function used is categorical crossentropy, appropriate for multi-class classification tasks. The model is trained for 50 epochs with a validation set to monitor performance during training. The batch size is set to 32 to ensure gradient updates and avoid overfitting.

Table 1 provides a summary of the performance metrics for each model to reveal the impact of CNN feature extraction on traditional ML models. RF particularly outperforming other models. While CNN demonstrates strong classification performance with an accuracy of 92.2% and the highest AUC of 97%, the inclusion of its features into RF boosts RF's accuracy to 92.6% and its AUC to 95%, narrowing the gap between the two models. This enhancement highlights the effectiveness of integrating DL features with traditional algorithms.

The RF model demonstrates a more efficient trade-off in computational time, requiring significantly less training time (5.00 seconds) compared to CNN's 106.21 seconds. This indicates that RF, combined with CNN features, is a viable alternative for applications requiring near-CNN-level performance but with constraints on training computational resources.

The ensemble model, combining the five models into a majority vote, outperformed all individual models, achieving an accuracy of 94.2%. The ensemble approach harnesses the strengths of all five models, resulting in enhanced classification performance and robustness. The precision, recall,

and F1 scores for the ensemble were all 97.1%, while the AUC was an impressive 0.99. This indicates that the ensemble model provides excellent discrimination between the sound classes.

In the ensemble voting mechanism, all models train and make predictions simultaneously. As a result, the training and testing times for the ensemble model match those of the CNN model which is the longest of the individual models. This simultaneous execution ensures that the benefits of the ensemble approach are achieved without a significant increase in computational overhead beyond the most complex component model.

Figure 4 shows the comparison between the training and validation loss over the epochs for the MV model, highlighting how the ensemble method performs during training. In Fig. 5, the training and validation accuracy curves demonstrate how the MV model's performance improves as it integrates the predictions from the five algorithms, offering a robust classification approach. These plots offer a clear visualization of the MV model's learning behavior and its ability to generalize effectively to new, unseen data by exploiting the strengths of each individual algorithm.

The ROC curves presented in Fig. 6 highlight the trade-off between sensitivity and specificity, providing insight into the MV model in diagnostic abilities. In addition to the individual ROC curves, we plotted the micro and macro ROC curves, which offer aggregated views of the model's performance across all classes. The micro-average ROC curve is calculated by aggregating the contributions of all classes to compute the average metric, effectively treating each element of the confusion matrix equally. This approach provides an overall sense of performance by considering each instance independently. Conversely, the macro-average ROC curve computes the metric independently for each class and then takes the average, treating all classes equally, regardless of their records in the dataset.

In our results, the micro and macro ROC curves are almost identical, which is expected due to the balanced nature of the dataset. This indicates that each class contributes equally to the overall performance, reflecting the system's balanced and unbiased classification capability. As illustrated in Table 1, the ensemble learning model using MV surpasses the performance of all individual models, demonstrating its superior classification capability. However, even these models are outperformed by the MV ensemble approach, it leverages the strengths of each model to achieve superior overall results.

The experimental results demonstrate the efficacy of our proposed IoT-based sound recognition system for assisting hearing-impaired drivers. The ensemble learning model outperforms individual models, providing high accuracy and robustness in classification. The balanced dataset and effective feature extraction techniques contribute significantly to this performance, ensuring the system can reliably identify critical auditory signals in real time.

The computational efficiency of the system, particularly the rapid prediction times, ensures that drivers receive timely alerts, which is crucial for maintaining safety on the road. Furthermore, the system's robustness to variations and high resilience against adversarial conditions make it a viable solution for practical deployment. These results underscore the potential of integrating ML and DL techniques within IoT frameworks to enhance the safety and autonomy of hearing-impaired drivers.

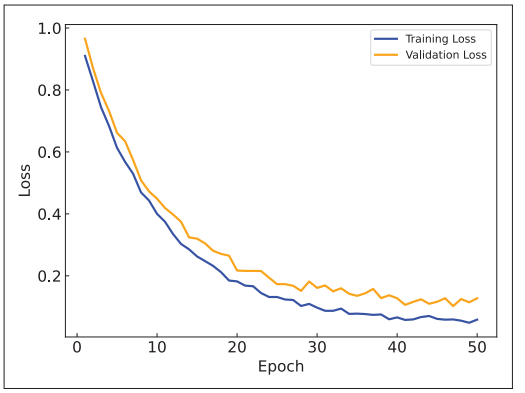

FIGURE 4. Training and Validation Loss for MV.

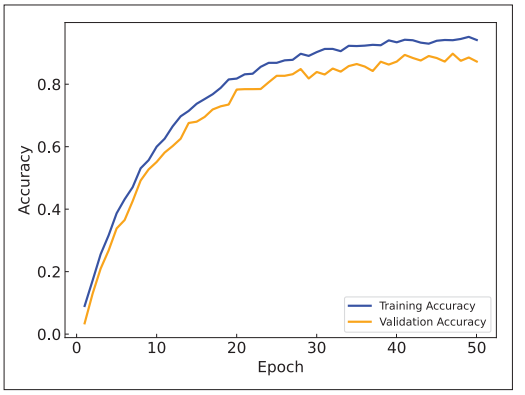

FIGURE 5. Training and Validation Accuracy for MV.

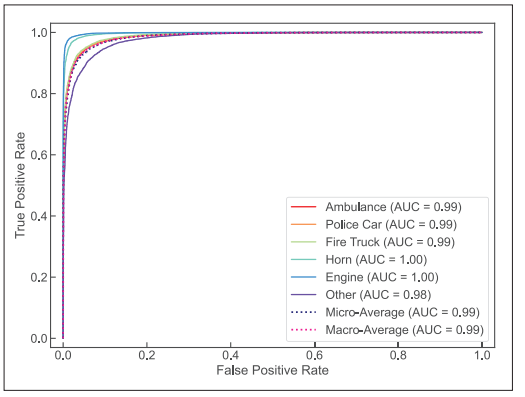

FIGURE 6. ROC Curves for MV.

Challenges in Practical Implementation

This section addresses the key considerations for integrating the proposed sound recognition system into vehicles.

Cost: The system leverages a Raspberry Pi platform, which is a relatively low-cost embedded computing solution. However, the system would still require additional components like microphones, processing unit, display unit and housing. We recognize the need to balance affordability with high-performance hardware and sensors to ensure accessibility for a wide range of users.

The micro-average ROC curve is calculated by aggregating the contributions of all classes to compute the average metric, effectively treating each element of the confusion matrix equally.

Strategies to minimize the per-unit cost, such as volume purchasing or partnering with automotive suppliers, should be explored to make the system economically viable for widespread deployment.

Compatibility: The system should be designed with the goal of seamless integration into vehicle electrical and computing architectures. Compatibility with common in-vehicle communication protocols (e.g., CAN bus, Ethernet) and data interfaces will be crucial. Modular designs that allow the noise classification module to be easily integrated alongside existing audio/infotainment systems would facilitate widespread adoption.

User Training: For the system to be effective, drivers and passengers will need to be trained on its capabilities and proper usage. This could involve educational materials, tutorials, and potentially intuitive user interfaces that minimize the learning curve. Ensuring the system provides clear and actionable insights to the user will be important for driver adoption and acceptance.

We recognize that conducting user testing with hearing-impaired individuals is vital for validating the system's effectiveness and usability. Such testing would allow us to gather insights on the clarity, responsiveness, and intuitiveness of the system's alerts, as well as identify potential areas for improvement in real-world scenarios. Unfortunately, due to time and resource constraints, this step was not conducted during the current phase of development. However, we plan to address this limitation in future work by conducting structured user testing sessions. These sessions will focus on evaluating the system's usability, accessibility, and perceived reliability through direct interaction with hearing-impaired individuals, along with qualitative feedback.

Conclusion

In this article, we presented an IoT-based sound recognition system designed to assist drivers. Our approach leverages ML and DL techniques to identify various vehicle-related sounds. By implementing a comprehensive feature extraction process, which includes 23 distinct features from each audio sample, we ensured that our system effectively differentiates between six sound classes: Ambulance, Police Car, Fire Truck, Horn, Engine, and Other. To maximize classification accuracy and robustness, we employed ensemble learning, which combines the outputs of multiple models through a MV approach. This ensemble model harnesses the strengths of individual classifiers, such as KNN, AB, ET, RF, and CNN. The results demonstrate that the ensemble model outperforms each model, achieving an impressive accuracy of 94.2%.

Our dataset was custom-built to reflect real-world conditions, consisting of 1500 audio samples. We included background noise in the samples to ensure that the system is robust and performs well in realistic driving environments. This attention to detail in dataset creation contributes significantly to the system's overall effectiveness and reliability.

The system's real-time capability was validated through the development of a prototype utilizing a Raspberry Pi 4, microphones, and a 3.5-inch display. This setup enables the system to provide timely alerts to drivers by translating detected sounds into visual alerts on the dashboard, enhancing situational awareness and safety for hearing-impaired individuals. Moreover, our solution is not limited to assisting hearing-impaired drivers. It also benefits drivers who may be distracted by environmental noise, listening to loud music, or otherwise unfocused, thus broadening the scope of its applicability.

While the proposed approach demonstrates competitive performance, several directions for future improvements have been identified. First, expanding the dataset with additional labeled samples across diverse environments and conditions is crucial. This would enhance the generalizability of the model and ensure its robustness in real-world scenarios. Real-world tests will also be conducted to evaluate the model's performance under practical deployment conditions, such as varying background noise levels and device-specific audio distortions.

Moreover, incorporating more advanced deep learning architectures offers promising avenues for performance improvement. Architectures such as Transformer-based models (e.g., Audio Spectrogram Transformer) and Convolutional Recurrent Neural Networks (CRNN) could provide a better understanding of temporal dependencies and spatial features in audio signals. These models, known for their ability to capture long-term dependencies and hierarchical feature representations, may outperform the current CNN-based framework.

In addition, integrating self-supervised learning techniques, which pretrain models on unlabeled audio data, could help mitigate the challenge of limited labeled data. Similarly, ensemble techniques, such as stacking or boosting using neural networks, could be explored to further refine performance metrics.

Finally, considering techniques such as attention mechanisms and fine-tuning pre-trained audio models (e.g., Wav2Vec, HuBERT) can provide a more nuanced approach to extracting features from complex audio datasets. These methodologies are expected to address current limitations and push the boundaries of audio classification accuracy in this domain.

References

[1] World Health Organization, "Deafness and Hearing Loss;" https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearingloss, Feb. 2024, accessed: July 2024.

[2] D. Martin *et al.*, "Fingerspeller: Camerafree Text Entry Using Smart Rings for American Sign Language Fingerspelling Recognition," *Proc. 25th Int'l. ACM SIGACCESS Conf. Computers and Accessibility*, 2023, pp. 1–5.

[3] M. Mohammadiyan *et al.*, "Association of Hearing Health with Traffic Accidents Among Heavy Vehicle Drivers," *Int'l. J. Environmental Health Engineering*, vol. 12, no. 2, 2023, p. 6.

[4] S. Tokić, D. Sumpor, and M. Z. Zeba, "Degradation of the Performance of Road Vehicle Drivers Due to the Influence of Cabin Distractions," *Acta Technica Napocensis-Series: Applied Mathematics, Mechanics, and Engineering*, vol. 65, no. 3S, 2023.

[5] O. Salem, A. Mehaoua, and R. Boutaba, "The Sight for Hearing: An IoT-Based System to Assist Drivers with Hearing Disability," *2023 IEEE Symp. Computers and Communications (ISCC)*, Los Alamitos, CA, USA: IEEE Computer Society, July 2023, pp. 1305–10.

[6] A. Suman, C. Kumar, and P. Suman, "Early Detection of Mechanical Malfunctions in Vehicles Using Sound Signal Processing," *Applied Acoustics*, vol. 188, 2022, p. 108578.

[7] A. Shabbir *et al.*, "Smart City Traffic Management: Acoustic-Based Vehicle Detection Using Stacking-Based Ensemble Deep Learning Approach," *IEEE Access*, 2024.

[8] M. Usaid *et al.*, "Ambulance Siren Detection Using Artificial Intelligence in Urban Scenarios," *Univ. Research Jour. of Eng.*

& Technology, vol. 12, no. 1, 2022, pp. 92–97.
[9] Z. Islam and M. Abdel-Aty, "Real-Time Emergency Vehicle Event Detection Using Audio Data," arXiv preprint arXiv:2202.01367, 2022.
[10] M. Cantarini *et al.*, "Acoustic Features for Deep Learning-Based Models for Emergency Siren Detection: An Evaluation Study," *12th Int'l. Symp. Image and Signal Processing and Analysis (ISPA)*, 2021, pp. 47–53.
[11] E. P. Jonnadula and P. M. Khilar, "Comparison of Various Techniques for Emergency Vehicle Detection Using Audio Processing," *Cloud Security*, CRC Press, 2021, pp. 64–75.
[12] M. Otoom, M. A. Alzubaidi, and R. Aloufee, "Novel Navigation Assistive Device for Deaf Drivers," *Assistive Technology*, vol. 34, no. 2, 2022, pp. 129–39.
[13] M. K. Gourisaria *et al.*, "Comparative Analysis of Audio Classification with MFCC and STFT Features Using Machine Learning Techniques," *Discover Internet of Things*, vol. 4, no. 1, 2024, p. 1.
[14] T. Nithya *et al.*, "TB-MFCC Multifuse Feature for Emergency Vehicle Sound Classification Using Multistacked Cnn–Attention BILSTM," *Biomedical Signal Processing and Control*, vol. 88, 2024, p. 105688.
[15] M. Y. Shams, T. Abd El-Hafeez, and E. Hassan, "Acoustic Data Detection in Large-Scale Emergency Vehicle Sirens and Road Noise Dataset," *Expert Systems with Applications*, vol. 249, 2024, p. 123608.

Biographies

Osman Salem (osman.salem@u-paris.fr) received the M.Sc. and Ph.D. degrees in Computer Science from Paul Sabatier University, Toulouse, France, in 2002 and 2006, respectively, and the Habilitation à Diriger des Recherches (HDR) degree from Université Paris Cité, France, in 2016. Since September 2008, he has been an Associate Professor at Université Paris Cité, France. He has extensive expertise in cybersecurity, particularly in the domains of secure communication protocols, threat analysis, and anomaly detection. His research focuses on addressing challenges in security and anomaly detection within medical wireless body area networks, ensuring data privacy and reliability in critical healthcare applications. He has actively participated in several research projects involving cybersecurity and privacy in Internet of Things (IoT) environments. These projects often integrate interdisciplinary approaches, leveraging artificial intelligence and machine learning techniques for detecting and mitigating emerging threats.

Ahmed Mehaoua (ahmed.mehaoua@u-paris.fr) received the M.Sc. and Ph.D. degrees in computer science from the University of Paris, France, in 1993 and 1997, respectively. He is currently a Full Professor of computer networking at University of Paris, and the Head of the Artificial Intelligence for Data Science and Cybersecurity Group at the CNRS BORELLI Research Center, a governmental mathematics and computer science research center in Paris, France. His research interests include security and resource management in wireless medical sensor networks, wireless body area networks design and optimization, and quality of service management in IP multimedia networks.

Raouf Boutaba [F] (rboutaba@cs.uwaterloo.ca) received the M.Sc. and Ph.D. degrees in computer science from Sorbonne University in 1990 and 1994, respectively. He is currently a University Professor and the Director of the David R. Cheriton School of Computer science at the University of Waterloo (Canada). He also holds a University Research Chair at Waterloo and the Rogers Chair in Network Automation. His research interests fall in the areas of computer networking and distributed systems. Dr. Boutaba served as the founding Editor-in-Chief of the *IEEE Transactions on Network and Service Management* (2007–2010) and the Editor-in-Chief of the *IEEE Journal on Selected Areas in Communications* (2018–2021). He is a fellow of the Engineering Institute of Canada, the Canadian Academy of Engineering, and the Royal Society of Canada.