Advanced Data-Driven Framework for Real-Time Vehicular Entity Detection in Connected Mobility Ecosystems

P. Santosh Kumar Patra[1*], I. V. Prakash[2], Pathipaka Manasa[2], C. Nandhini[2], Vutkuru Venu[2], Peddi Shiva[2], P. Pavan[2]

[1]Professor and Group Director, Department of Computer Science and Engineering, Sree Dattha Institute of Engineering and Science, Sheriguda, Ibrahimpatnam, 501510, Telangana, India

[2]Department of Electronics and Communication Engineering, Sree Dattha Institute of Engineering and Science, Sheriguda, Ibrahimpatnam, 501510, Telangana, India

*Correspondence: P. Santosh Kumar Patra (drpskpatra@gmail.com)

ABSTRACT

The in-vehicle environment was found to generate a diverse range of acoustic signals originating from driver activities, engine operations, road interactions, and surrounding traffic conditions. These audio events carry significant information related to driver behaviour and vehicle safety. With the advancement of intelligent transportation systems, the need for automated and real-time monitoring of driver alertness using in-vehicle audio has gained considerable importance. Traditional approaches relied on manual observation or basic signal processing techniques, which were limited in handling complex acoustic patterns and large-scale audio data generated in modern vehicles. To overcome these limitations, a deep learning-based in-vehicle audio event detection framework was developed for enhancing driver alertness monitoring and safety. The proposed system utilized the Waveform Language Model (WavLM), a transformer-based architecture, to extract deep acoustic feature representations from raw audio signals. These features captured both temporal and contextual information, enabling effective modelling of complex in-vehicle sound patterns. The extracted feature vectors were used to train multiple machine learning classifiers, including Categorical Boosting Classifier (CBC), Histogram Gradient Boosting Classifier(HGBC) , Extra Trees Classifier (ETC), and a proposed Tree-Based Generalized Additive Model (TGAM). The system was designed to perform both main-class and sub-class audio event classification to accurately identify driver-related events such as distractions, alerts, and environmental sounds. The performance of the models was evaluated using standard metrics such as accuracy, precision, recall, and F1-score. Experimental results demonstrated that the proposed TGAM model achieved superior performance, with an accuracy of 99.92% for main-class classification and 99.58% for sub-class classification. The integration of deep learning-based feature extraction with advanced classification techniques significantly improved the reliability and efficiency of in-vehicle audio event detection systems, thereby contributing to enhanced driver alertness monitoring and overall road safety.

Keywords: Acoustic Signal Intelligence, Multilevel Audio Event Detection, Spectro-Temporal Feature Modeling, Real-Time Audio Analytics, Edge-Ready Audio AI Systems

1. INTRODUCTION

Everyday acoustic signals such as honking, sirens, and engine sounds were found to play a critical role in enabling drivers to make rapid and informed decisions, thereby ensuring safety. The in-vehicle soundscape was observed to contain diverse auditory cues that enhance driver awareness and environmental understanding. Similarly, autonomous driving systems were identified to benefit significantly from integrating such acoustic information for improved situational perception. However, existing autonomous vehicles predominantly relied on visual sensors such as cameras, along with RADAR and LiDAR technologies, for environment sensing [1]. Despite their effectiveness, these systems were limited to "seeing" the surroundings while lacking the capability to "hear" important auditory events. Studies indicated that sound signals were less affected by adverse weather conditions and object occlusions, making them reliable for environmental perception [2]. Therefore, incorporating Acoustic Event Detection (AED) was considered an effective approach to complement existing sensor systems, enhance redundancy, and improve overall safety and navigation performance.

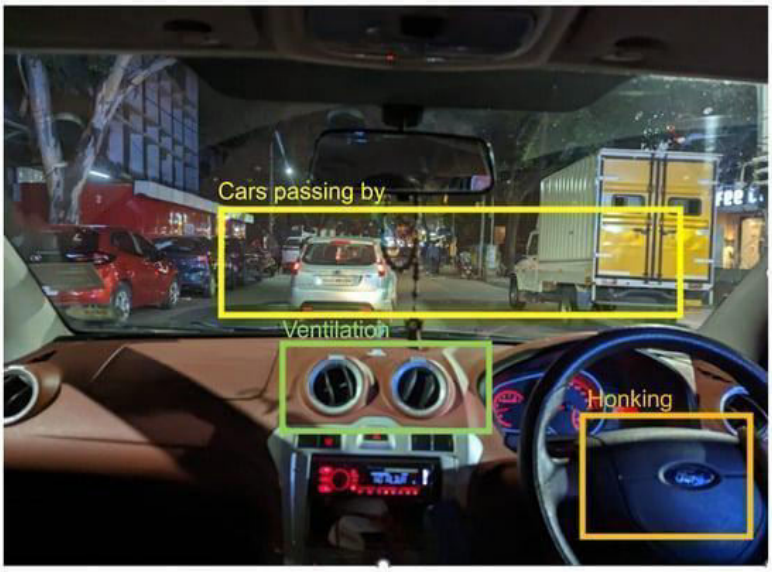


Fig. 1. Illustration of acoustic event detection in a vehicle environment, showing multiple in-car and surrounding sound sources.

AED, also known as the Sound Event Detection (SED) system, detects different sound events in an input audio along with the temporal instances at which they occur [3]. Depending on the extent of the overlap of events, an AED system can be monophonic or polyphonic. A monophonic AED system refers to a system that does not detect overlapping events, while a polyphonic system can detect overlapping events [4].

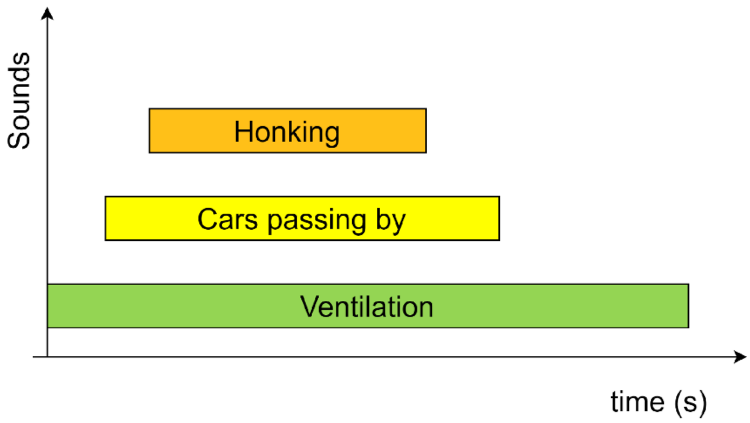


Fig. 2. Timeline representation of detected acoustic events in a vehicle environment.

Polyphonic Acoustic Event Detection (AED) was identified as more complex than monophonic AED due to the difficulty in distinguishing features of multiple simultaneous sounds. Real-world acoustic environments were inherently polyphonic, containing overlapping audio signals with variations in amplitude, pitch, and timbre. The study focused on developing a polyphonic AED system capable of detecting multiple concurrent sound events within an in-vehicle environment [5]. Consider an auditory scene that can happen inside a car as shown in Figure 1.1. The goal of the proposed AED system is to detect these events along with the time of occurrence as described. Predicting the activity of multiple sound classes for short audio segments can produce such an audio scene description. As shown in fig 1.2 this implies that the AED problem can be seen as a multi-label classification problem on short audio segments. If the system provides predictions for the duration of the whole audio, it is referred to as tagging, while if the predictions describe the activity pattern of the audio classes, it is referred to as detection. Vehicle environments produce diverse acoustic signals originating from engines, mechanical components, road conditions, and external traffic sources. These audio signals often overlap and vary significantly depending on environmental conditions, vehicle speed, and operating situations. Traditional monitoring approaches rely on manual inspection or simple signal processing techniques that are not capable of effectively handling large and complex audio datasets. As a result, accurately identifying and categorizing different audio events becomes challenging. The increasing volume of sensor data in modern vehicles further complicates the analysis process and requires efficient automated methods for interpreting acoustic information.

2. LITERATURE SURVEY

2.1 Audio-Based Vehicle Monitoring and Acoustic Event Detection

Recent studies have emphasized the importance of audio signals in enhancing vehicular safety and fault diagnosis. Rashed *et al.* [6] highlighted the potential of sound-driven diagnostics by analyzing internal vehicle sounds such as engine and brake noise, along with external emergency signals like sirens. They also introduced a specialized dataset comprising vehicle faults and environmental sounds, combined with a multi-model classification framework. Similarly, Antony *et al.* [9] developed an acoustic event detection system using transformer-based BEATs and neural networks to identify automotive sound events along with their temporal occurrence. Banchero *et al.* [10] further advanced this domain by proposing an intelligent system capable of detecting and localizing emergency sounds such as sirens and horns using deep learning architectures including CNNs and transformer models. Additionally, Almaadeed *et al.* [12] demonstrated that integrating audio surveillance systems significantly improves the detection of hazardous events like crashes and tire skidding, particularly in visually challenging conditions where traditional vision-based systems may fail.

2.2 Driver Behavior, Safety Monitoring, and Human Factors

Driver behavior analysis and safety monitoring remain critical aspects of intelligent transportation systems. Huang *et al.* [7] investigated the influence of different signal modalities on driver response using virtual reality, revealing that visual cues are more effective than auditory signals in reducing reaction time. Visconti *et al.* [8] provided a comprehensive overview of driver monitoring systems, covering aspects such as drowsiness detection, alcohol impairment, and vehicle condition monitoring using machine learning techniques. Knapik *et al.* [11] proposed a thermal imaging-based system for driver state detection, enabling robust monitoring under varying lighting conditions by analyzing facial cues such as yawning and head movements. Furthermore, Das *et al.* [14] introduced a CNN-LSTM-based deep learning model for detecting driver fatigue using temporal video data, demonstrating improved accuracy over traditional methods. Krstačić *et al.* [13] examined the safety implications of in-vehicle infotainment systems, identifying issues such as cognitive overload and distraction, while highlighting safer alternatives like speech-based interfaces.

2.3 Multi-Sensor Integration and AI-Driven Safety Systems

The integration of multiple sensing modalities has been explored to enhance perception and safety in vehicular environments. Yang *et al.* [15] conducted a comprehensive review of AI tools and sensor technologies used for driver monitoring and safety-critical event analysis, emphasizing the role of advanced machine learning models in reducing driver errors and improving road safety. Their study also identified gaps between academic research and industrial applications, suggesting future directions for bridging these differences. Collectively, these works indicate that combining audio, visual, and sensor-based data with AI-driven models significantly improves the robustness and reliability of intelligent transportation systems.

3. PROPOSED METHODOLOGY

The research methodology followed a structured and data-driven framework to analyse in-vehicle audio signals and detect acoustic events related to driver alertness and safety, as illustrated in Fig. 3. The process began with the acquisition of an in-vehicle audio dataset containing various driving and environmental sound events, followed by preprocessing to standardize and clean the raw audio signals. A deep learning-based feature extraction approach was then employed using the WavLM model to transform audio signals into rich and meaningful representations. These extracted features were further utilized to train and evaluate multiple classification models for accurate detection of audio events.

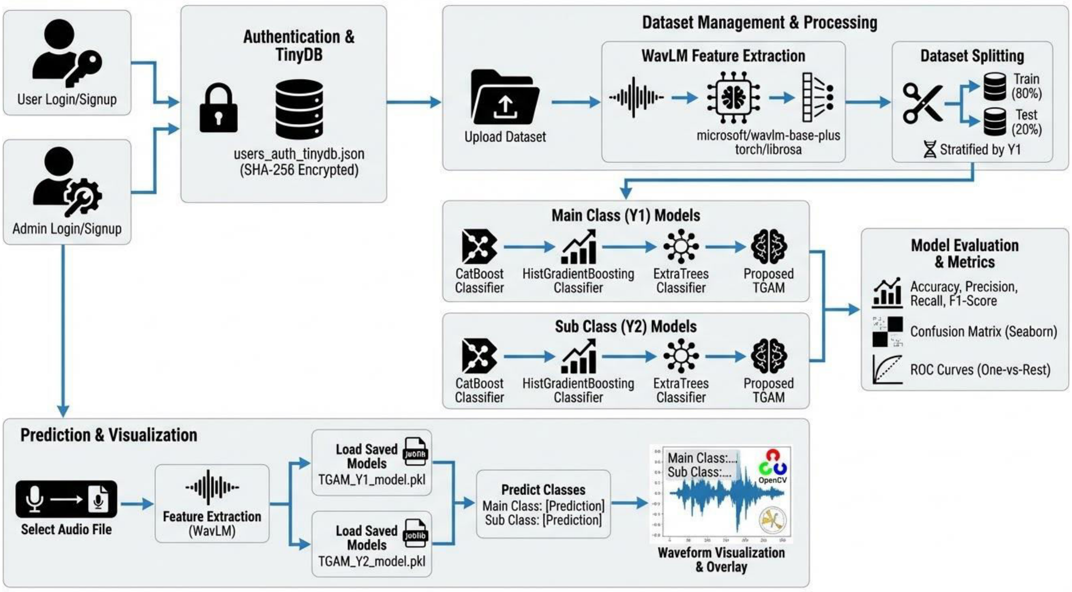


Fig. 3. Proposed system architecture of in-vehicle audio event classification framework.

The analytical pipeline integrated feature extraction, data preparation, model training, performance evaluation, and real-time prediction modules to ensure reliable detection of driver-related acoustic events. The system was designed to support both main-class and sub-class classification for detailed analysis of in-vehicle sounds. Additionally, a user-friendly graphical interface was incorporated to facilitate dataset loading, feature extraction, model training, and prediction visualization. Through this comprehensive architecture, the methodology aimed to enhance driver alertness monitoring and improve overall vehicle safety by effectively analysing complex in-vehicle acoustic environments.

The system is designed as an interactive and intelligent audio analysis platform supported by a Tkinter-based graphical user interface that enables seamless interaction for both administrators and general users. Administrators can upload structured audio datasets, initiate feature extraction, perform dataset preparation, train machine learning models, and visualize evaluation results, while users can upload new audio samples to obtain classification outputs along with waveform visualization. A lightweight TinyDB-based authentication module ensures secure access control by storing encrypted credentials and managing role-based permissions. The input dataset consists of categorized vehicle audio recordings organized into hierarchical folders, which are processed using the transformer-based WavLM model to extract high-dimensional feature representations capturing both temporal and spectral characteristics. These features are structured into a dataset with corresponding labels and split into training and testing sets for model development. Baseline classifiers such as CBC, HGBC, and ETC are trained to establish comparative performance, while the proposed TGAM model leverages an ensemble learning approach based on Random Forest decision trees to achieve accurate and interpretable classification. Model performance is evaluated using metrics including accuracy, precision, recall, and F-score, along with confusion matrices and ROC curves for detailed analysis. During prediction, new audio inputs undergo the same feature extraction process, and the TGAM model predicts both primary and subclass labels, displaying results through the interface. Finally, all trained models are stored using joblib for efficient reuse, ensuring reduced computational cost and enabling continuous, scalable audio event analysis.

TGAM model

As depicted in given fig. 4. the TGAM performs classification using a tree-based learning framework designed to capture complex relationships between input features and target classes. In this study, TGAM receives the feature vectors extracted from vehicle audio signals and analyses them to identify different acoustic event categories. The model learns patterns by constructing multiple tree-based structures that examine how different features contribute to classification decisions. Through this process, TGAM can recognize meaningful feature interactions and generate reliable predictions. In addition to classification, the model also produces simplified rule representations that help understand the reasoning behind the generated predictions.

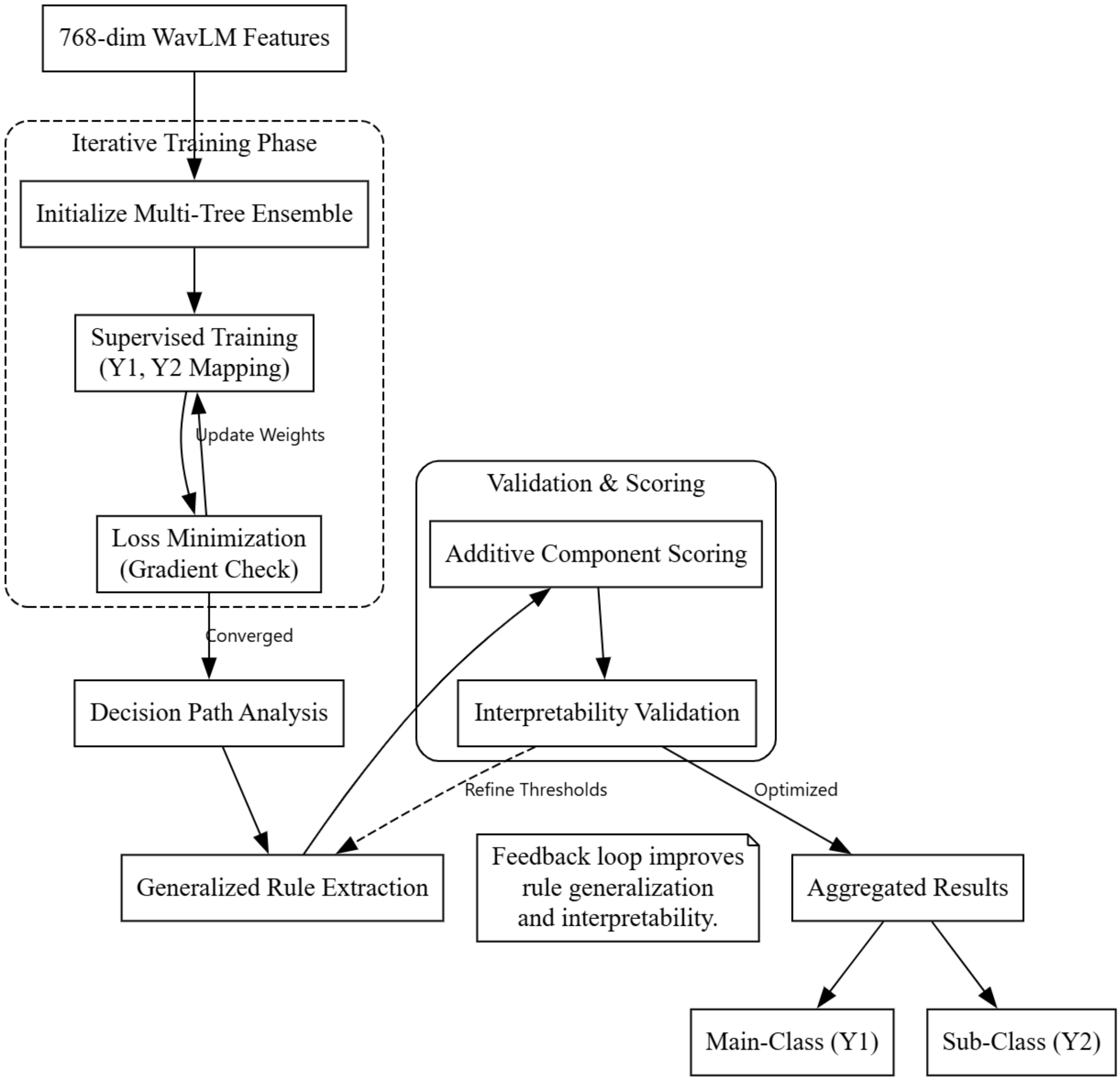
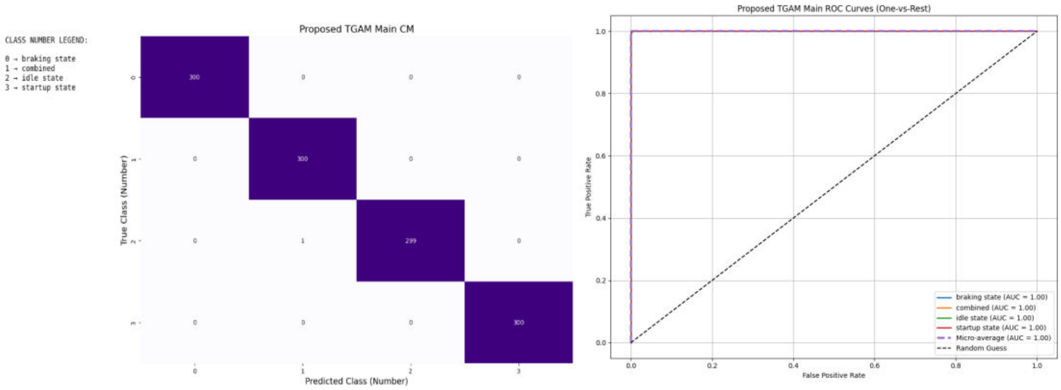


Fig. 4. Internal Operation Diagram of Proposed TGAM.

The process begins by providing the extracted audio feature vectors as input to the TGAM. These features represent important acoustic characteristics obtained from the feature extraction stage. Each feature vector is associated with a corresponding class label that represents the audio category. In this stage, the TGAM analyzes the relationships between input features and the corresponding class labels. The model examines how variations in feature values influence the classification outcome. This step allows the algorithm to identify meaningful feature interactions within the dataset. The TGAM builds multiple tree-based decision structures that partition the feature space into smaller regions. Each branch in the tree represents a decision rule based on feature values. These trees collectively learn patterns that distinguish one audio event category from another. After building the tree structures, the model extracts simplified rules that represent the learned decision boundaries. These rules describe how combinations of feature values lead to specific class predictions. This rule-based representation improves the interpretability of the model's decisions.

The predictions generated by the different tree structures are combined to produce a final classification decision. Each tree contributes to the prediction by evaluating the input feature vector according to its learned rules. The final output is determined by aggregating the decisions across all trees. Once the model is trained, it can analyze new audio feature vectors and classify them into the corresponding audio event categories. The trained TGAM evaluates the input features using the learned rules and tree structures. This process enables the system to accurately identify vehicle audio events from previously unseen data.

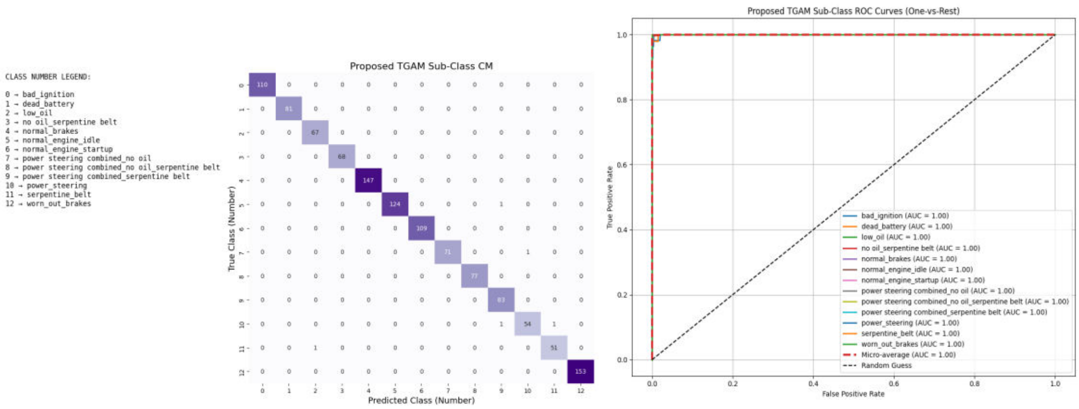
4. RESULTS DESCRIPTION



(a) (b)

Fig. 5 (a) (b) : Confusion matrix and ROC curve obtained using TGAM Model

Fig. 5 shows the main-class evaluation results of the proposed TGAMC model. Subfigure (a) presents a confusion matrix with nearly perfect diagonal accuracy, indicating that the model correctly identifies almost all samples across the primary engine states with minimal or no misclassification. This performance highlights the strength of the TGAMC framework in capturing high-level acoustic features. Subfigure (b) illustrates ROC curves for each class, all achieving an AUC of 1.00, suggesting flawless separation between categories regardless of threshold.



(a) (b)

Fig. 6 (a) (b). Confusion matrix and ROC curve obtained using TGAM model for sub class

Fig. 6. illustrates the TGAMC model's performance on sub-class classification, covering multiple detailed engine and system conditions. Subfigure (a) shows a confusion matrix with strong diagonal values, where most classes are predicted with exceptionally high accuracy, reflecting the model's ability to separate subtle differences in audio signatures. Misclassifications are minimal, indicating robust feature extraction and representation. Subfigure (b) presents ROC curves for all sub-classes, each achieving an AUC of 1.00, confirming excellent discriminative capability across highly diverse fault categories.

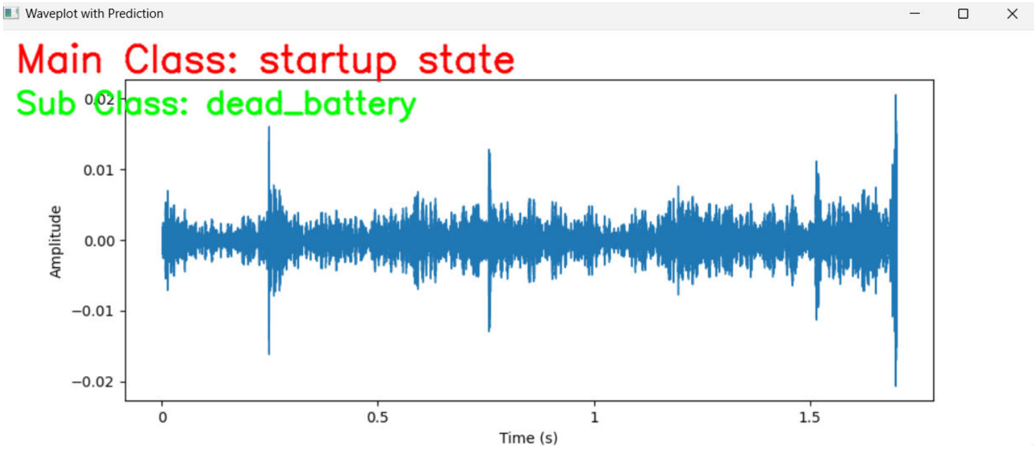


Fig. 7. Prediction output screen for dead battery

Fig. 7. illustrates the final prediction interface where the system displays both the main-class and sub-class outcomes for the uploaded audio sample. The figure includes a waveform plot of the selected signal, allowing users to visually inspect the amplitude variations over time. Above the waveform, the predicted main engine state and corresponding sub-condition are clearly shown, ensuring users can immediately understand the diagnostic result. This visualization links the raw audio input to the model's analytical output, offering transparency in how the prediction relates to the underlying signal.

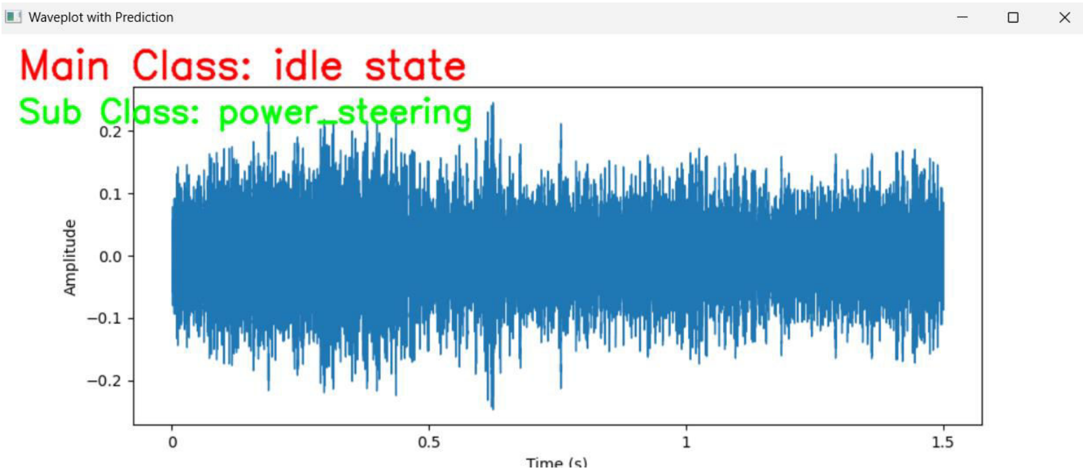


Fig. 8. Prediction output screen for power steering

Fig. 8. illustrates the prediction result generated by the system for an audio sample associated with a power steering condition. The waveform plot shows the amplitude variations of the selected signal across time, giving users a clear representation of the acoustic characteristics captured from the vehicle. The predicted main-class label, identified as idle state, and the sub-class label, detected as power steering, are prominently displayed above the waveform for easy interpretation. This output helps users understand both the high-level engine state and the specific component-related issue inferred from the audio.

5. CONCLUSION

The research demonstrates the effectiveness of machine learning techniques in analysing and classifying vehicle audio events using advanced feature extraction and classification methods. The system processes vehicle audio signals through the WavLM-based feature extraction mechanism, which converts raw audio data into meaningful numerical representations suitable for machine learning analysis. Using these extracted features, several classification models including CBC, HGBC, ETC, and TGAM were implemented and evaluated to identify different vehicle sound categories. The experimental results show that TGAMC significantly outperforms the other models in terms of classification performance. The comparative analysis indicates that TGAM achieved the highest accuracy of 99.92% for main class classification and 99.58% for sub class classification, along with very high precision, recall, and F1-score values. In contrast, CBC achieved moderate performance while HGBC and ETC produced comparatively lower classification accuracy. The superior performance of TGAM can be attributed to its ability to effectively capture complex feature relationships and decision patterns within the extracted audio features. The integration of deep audio feature extraction using the WavLM model further improves classification capability by providing rich acoustic representations of vehicle sounds. The research demonstrates that combining transformer-based audio feature extraction with advanced machine learning classification models significantly improves the accuracy and reliability of vehicle audio event recognition systems. The implemented framework provides an effective approach for analysing complex acoustic signals in vehicle environments and supports intelligent monitoring of vehicle sound events.

REFERENCES

[1] Gunadal, A. Koujalagi, S. Karavade, and S. Konnuri, ―Wireless black box using MEMS accelerometer and GPS tracking for accidental monitoring of vehicle,‖ research reference No. : 37S0430 college branch guide : s.g.balekundri institute of technology, belgaum, 2015.

[2] Chung-ChengChiu, Min-YuKu, Hung-Tsung, Chen Nat, "Motorcycle Detection and Tracking System with Occlusion Segmentation," Image Analysis for Multimedia Interactive Services. Santorini, vol. 2, pp. 32- 32, June 2007.

[3] G. Gurjar, N. Chandurkar, R. Nagpure, S. Ganorkar, and J. Shelke, (March 2017) 'Wireless Black Box using MEMS Accelerometer and GPS Tracking for Accidental Monitoring of Vehicles', IJESC, vol 7(No 3).

[4] Musa, E. Alozie, S.A. Suleiman, J.A. Ojo, and A.L. Imoize, "A Review of Time-Based Solar Photovoltaic Tracking Systems," Information (Switzerland), vol. 14, no. 4, pp. 211, 2023. [Online]. Available: http://www.mdpi.com/journal/information/. doi: 10.3390/info14040211.

[5] N. Purushotham and J. Madan Kumar, ―Wireless black box using MEMS accelerometer and GPS tracking for accidental monitoring of vehicles,‖ International Journal of Scientific Engineering and Technology Research Volume.03, IssueNo.03, March-2014, Pages:0403-0408.

[6] Rashed, A.; Abdulazeem, Y.; Farrag, T.A.; Bamaqa, A.; Almaliki, M.; Badawy, M.; Elhosseini, M.A. Toward Inclusive Smart Cities: Sound-Based Vehicle Diagnostics, Emergency Signal Recognition, and Beyond. Machines 2025, 13, 258. https://doi.org/10.3390/machines13040258

[7] Huang, A.; Derakhshan, S.; Madrid-Carvajal, J.; Nosrat Nezami, F.; Wächter, M.A.; Pipa, G.; König, P. Enhancing Safety in Autonomous Vehicles: The Impact of Auditory and Visual Warning Signals on Driver Behavior and Situational Awareness. Vehicles 2024, 6, 1613-1636. https://doi.org/10.3390/vehicles6030076.

[8] Visconti, P.; Rausa, G.; Del-Valle-Soto, C.; Velázquez, R.; Cafagna, D.; De Fazio, R. Innovative Driver Monitoring Systems and On-Board-Vehicle Devices in a Smart-Road Scenario Based on the Internet of Vehicle Paradigm: A Literature and Commercial Solutions Overview. Sensors 2025, 25, 562. https://doi.org/10.3390/s25020562.

[9] Antony, A.; Theimer, W.; Grossetti, G.; Friedrich, C.M. Acoustic Event Detection in Vehicles: A Multi-Label Classification Approach. Sensors 2025, 25, 2591. https://doi.org/10.3390/s25082591

[10] Banchero, L.; Vacalebri-Lloret, F.; Mossi, J.M.; Lopez, J.J. Enhancing Road Safety with AI-Powered System for Effective Detection and Localization of Emergency Vehicles by Sound. Sensors 2025, 25, 793. https://doi.org/10.3390/s25030793

[11] Knapik, M.; Cyganek, B.; Balon, T. Multimodal Driver Condition Monitoring System Operating in the Far-Infrared Spectrum. Electronics 2024, 13, 3502. https://doi.org/10.3390/electronics13173502

[12] Almaadeed, N.; Asim, M.; Al-Maadeed, S.; Bouridane, A.; Beghdadi, A. Automatic Detection and Classification of Audio Events for Road Surveillance Applications. Sensors 2018, 18, 1858. https://doi.org/10.3390/s18061858

[13] Krstačić, R.; Žužić, A.; Orehovački, T. Safety Aspects of In-Vehicle Infotainment Systems: A Systematic Literature Review from 2012 to 2023. Electronics 2024, 13, 2563. https://doi.org/10.3390/electronics13132563

[14] Das, S.; Pratihar, S.; Pradhan, B.; Jhaveri, R.H.; Benedetto, F. IoT-Assisted Automatic Driver Drowsiness Detection through Facial Movement Analysis Using Deep Learning and a U-Net-Based Architecture. Information 2024, 15, 30. https://doi.org/10.3390/info15010030

[15] Yang, G.; Ridgeway, C.; Miller, A.; Sarkar, A. Comprehensive Assessment of Artificial Intelligence Tools for Driver Monitoring and Analyzing Safety Critical Events in Vehicles. Sensors 2024, 24, 2478. https://doi.org/10.3390/s24082478.