IJERST

International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 2(3) (2026)

ijerst.editor@gmail.com

editor@ijerst.com

ISSN 2319-5991 www.ijerst.com
© 2026 IJERST. All Rights Reserved
Peer Reviewed, Referred & Indexed Journal

Research Paper

A Low-Latency Machine Learning Architecture for Scalable Vehicle Recognition in Networked Mobility Ecosystems

K. Vijay Kumar[1], Kathoju Sravani[1], Keshaboina Manichandra[1], B. Vishnuvardhan Reddy[1], M. Manikanth Goud[1]

[1]Department of Computer Science and Engineering, [1]Sree Dattha Institute of Engineering and Science, Nagarjuna Sagar Road, Sheriguda, Ibrahimpatnam, Rangareddy Dist, 501510, Telangana, India.

ABSTRACT

The rapid advancement of intelligent transportation systems and automotive monitoring has increased the demand for automated and reliable detection of vehicle conditions using acoustic signals. In-vehicle audio data contains rich information about mechanical states and faults such as braking anomalies, engine idle variations, startup failures, and combined fault conditions. However, the primary challenge lies in the unstructured and high-dimensional nature of audio signals, where subtle variations correspond to different operational behaviors. Earlier approaches rely on manual feature extraction techniques and basic Machine Learning (ML) models, which fail to capture complex temporal dependencies and hierarchical relationships between primary vehicle states and fine-grained fault categories, resulting in reduced accuracy and poor generalization. To address these limitations, this work proposes a transformer-driven hierarchical audio classification framework. The system utilizes Waveform Language Model (WavLM) to extract deep and contextual acoustic representations from raw audio signals. These features are used for dual-level classification, where the first level predicts the main class (Y1) and the second level identifies the sub-class (Y2). Multiple ensemble models, including Categorical Boosting Classifier (CBC), Histogram-Based Gradient Boosting (HGB), Extra Trees Classifier (ETC), and a proposed Tree-Based Generalized Additive Model (TGAM), are employed to enhance prediction performance. The TGAM model further improves interpretability by generating simplified rule-based insights from tree structures. The proposed system achieves improved accuracy, robustness, and interpretability, making it suitable for real-time vehicle diagnostics and intelligent monitoring applications, thereby contributing to safer and more efficient transportation systems.

Keywords: Intelligent Transportation Systems, Vehicle Audio Analysis, Acoustic Event Detection, Hierarchical Classification, Machine Learning, Vehicle Fault Detection, In-Vehicle Monitoring Systems

1. INTRODUCTION

Everyday acoustic cues such as vehicle horns, emergency sirens, engine vibrations, braking noises, and tire–road interactions provide rich contextual information that enables drivers to make rapid and informed decisions, thereby enhancing road safety. The in-vehicle acoustic environment is inherently dynamic and information-dense, offering complementary insights that extend beyond visual perception. In intelligent and autonomous driving systems, incorporating auditory perception has been increasingly recognized as essential for achieving robust situational awareness. However, most existing systems rely predominantly on visual sensing modalities such as cameras, along with RADAR and LiDAR for environmental understanding. While these technologies provide accurate spatial and depth information, they are fundamentally limited to line-of-sight perception and cannot capture critical auditory events such as approaching emergency vehicles or abnormal engine sounds. Furthermore, visual sensors often suffer performance degradation under adverse conditions such as low illumination, fog, rain, and occlusions, whereas acoustic signals remain comparatively resilient and omnidirectional, making them a reliable complementary modality for perception tasks [1,2]. This motivates the integration of Acoustic Event Detection (AED) systems to enhance redundancy, improve robustness, and support safer navigation in complex driving scenarios.

AED, also referred to as Sound Event Detection (SED), focuses on identifying and temporally localizing distinct sound events within continuous audio streams [3]. Based on the nature of sound occurrences, AED systems are broadly categorized into monophonic and polyphonic frameworks. Monophonic AED assumes that only one sound event is present at a given time, whereas polyphonic AED is capable of detecting multiple overlapping events simultaneously [4]. Real-world acoustic environments, particularly inside vehicles, are inherently polyphonic, consisting of concurrent sound sources with varying characteristics such as frequency, amplitude, duration, and timbre. This overlap introduces significant challenges in feature extraction, representation learning, and accurate classification. In this work, a polyphonic AED framework is developed to effectively model such complex auditory scenes by analyzing short audio segments and predicting the presence of multiple sound classes simultaneously, thereby formulating the problem as a multi-label classification task [5]. When predictions are generated for entire audio clips, the task is referred to as audio tagging, whereas generating time-resolved predictions that indicate the onset and offset of events corresponds to event detection. This dual capability enables detailed auditory scene understanding, facilitating improved perception, decision-making, and safety in next-generation intelligent transportation systems.

2. LITERATURE SURVEY

Visconti, et al. [6] presented a comprehensive review of intelligent transportation safety systems by analyzing key components such as driver monitoring, vehicle condition assessment, and traffic management. Their study discussed various Machine Learning (ML) models used for detecting driver fatigue, abnormal driving behavior, and vehicle instability under different road conditions. In addition, they explored sensor-based architectures integrating cameras, inertial sensors, and environmental sensors for real-time monitoring. The work also highlighted applications such as driving style classification, tire condition monitoring, and emission analysis, providing a complete overview of modern vehicle safety frameworks. Yang, et al. [7] conducted an extensive survey on Artificial Intelligence (AI)-based tools used for driver monitoring and safety-critical event detection. Their work examined different deep learning models, sensor technologies, and datasets used in both academic research and industry systems. They identified key challenges such as lack of generalization, dataset limitations, and difficulties in real-time deployment. The study also emphasized the importance of multimodal approaches and scalable AI models for improving system reliability and performance. Khan, et al. [8] provided a detailed survey on driver assistance systems by focusing on the interaction between driver behavior, vehicle dynamics, and environmental conditions. Their framework highlighted the importance of continuous monitoring using multiple sensors, including physiological signals, vehicle parameters, and external environmental data. The study emphasized that integrating these heterogeneous data sources improves early detection of unsafe driving conditions and enhances overall driving safety and comfort. Krstačić, et al. [9] reviewed the evolution of In-Vehicle Infotainment Systems and their impact on driver safety by analyzing multiple studies over time. Their work identified major issues such as driver distraction, increased cognitive workload, and reduced situational awareness caused by complex human–machine interactions. The study compared different interface types and concluded that touchscreen-based systems negatively affect driving performance, while speech-based and hands-free interfaces offer safer alternatives.

Li, et al. [10] explored Electroencephalogram-based driver drowsiness detection techniques by categorizing existing approaches into open-loop detection systems and closed-loop management systems. Their survey examined signal acquisition, preprocessing, feature extraction, and classification methods used to detect fatigue from brain signals. The study also highlighted challenges such as noise sensitivity, variability across individuals, and difficulties in real-time implementation, emphasizing the need for robust and practical solutions. Fu, et al. [11] analyzed recent advancements in machine vision-based driver monitoring systems, focusing on facial feature detection and behavioral analysis. Their work covered techniques such as eye tracking, blink detection, head pose estimation, and facial expression recognition using deep learning models. The study also discussed safety mechanisms that generate alerts based on driver state and highlighted future directions such as multimodal integration and adaptive human–machine interaction systems. Almaadeed, et al. [12] investigated the role of audio-based surveillance systems for detecting hazardous road events such as vehicle collisions and tire skidding. Their study highlighted the limitations of vision-only systems, particularly under adverse weather conditions like rain, fog, and snow, where visual clarity is reduced. By incorporating audio signal processing and acoustic event detection techniques, they demonstrated improved reliability and accuracy in identifying critical road incidents.

Zohaib, et al. [13] proposed a multimodal emergency vehicle detection framework that integrates both acoustic and visual information. Their approach utilized an attention-based temporal spectrum network for detecting siren sounds and a multi-level spatial fusion YOLO architecture for visual detection. By combining both modalities using ensemble learning techniques, the system achieved high detection accuracy and reduced misclassification, demonstrating the effectiveness of sensor fusion in real-world scenarios. Fei, et al. [14] developed an interpretable framework for evaluating driving behavior using statistical and pattern mining techniques. Their study employed Principal Component Analysis (PCA) to determine the importance of different driving indicators and used the Frequent Pattern Growth (FP-growth) algorithm to analyze relationships between these indicators. This approach enabled the identification of unsafe driving patterns and provided a structured method for behavior-based safety assessment. Florez, et al. [15] proposed a real-time driver drowsiness detection system based on Convolutional Neural Networks (CNN) by analyzing facial features such as eye movements and yawning patterns. Their system utilized near-infrared imaging and edge computing devices for real-time performance. The proposed approach achieved high accuracy and demonstrated robustness in both controlled and real-world driving environments.

3. PROPOSED SYSTEM

The proposed system architecture is designed as a comprehensive framework for vehicle audio event classification by integrating advanced feature extraction and machine learning models. Initially, the system accepts structured audio datasets organized into hierarchical categories representing main (Y1) and sub (Y2) classes.

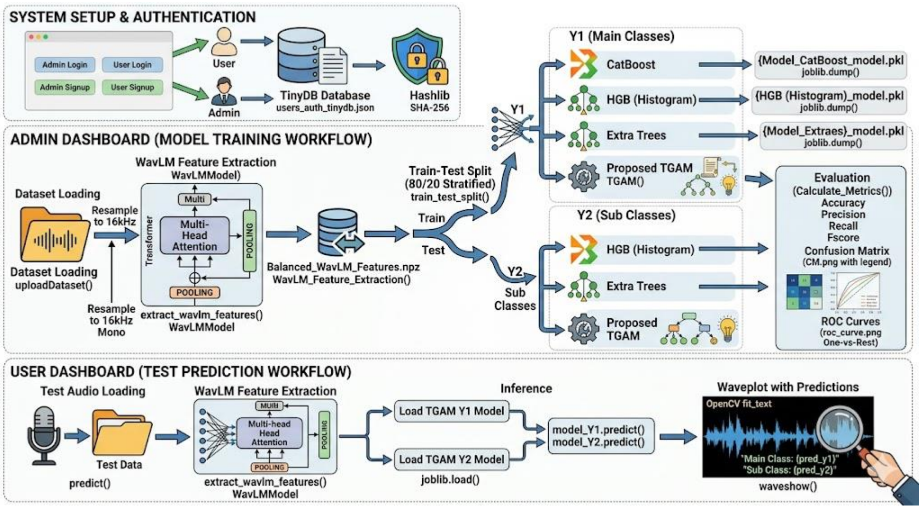


Fig. 1: Proposed system architecture

The audio signals are preprocessed through resampling to 16 kHz and conversion to mono format to ensure consistency across inputs. These standardized signals are then passed through the WavLM model, a transformer-based architecture that extracts rich contextual embeddings capturing both temporal and spectral characteristics of vehicle sounds. The extracted features are stored in a dedicated feature repository to avoid redundant computations and improve efficiency. Subsequently, the dataset is split into training and testing sets using stratified sampling to preserve class distribution. Multiple classification models, including CBC, HGBC, ETC, and the proposed TGAM, are trained separately for hierarchical classification tasks. The system performs dual-level prediction, where Y1 identifies the primary vehicle state and Y2 determines the specific fault condition. Among the models, TGAM demonstrates superior performance due to its ability to capture complex feature interactions and provide interpretable decision rules. The trained models are evaluated using metrics such as accuracy, precision, recall, F1-score, confusion matrix, and ROC curves to ensure robustness. During inference, new audio inputs undergo the same feature extraction process and are passed through trained TGAM models for prediction. The final output includes predicted class labels along with waveform visualization, enabling real-time monitoring and interpretation of vehicle acoustic events as illustrated in Fig. 1.

3.1 TGAM

As depicted in the given fig. 2. 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.

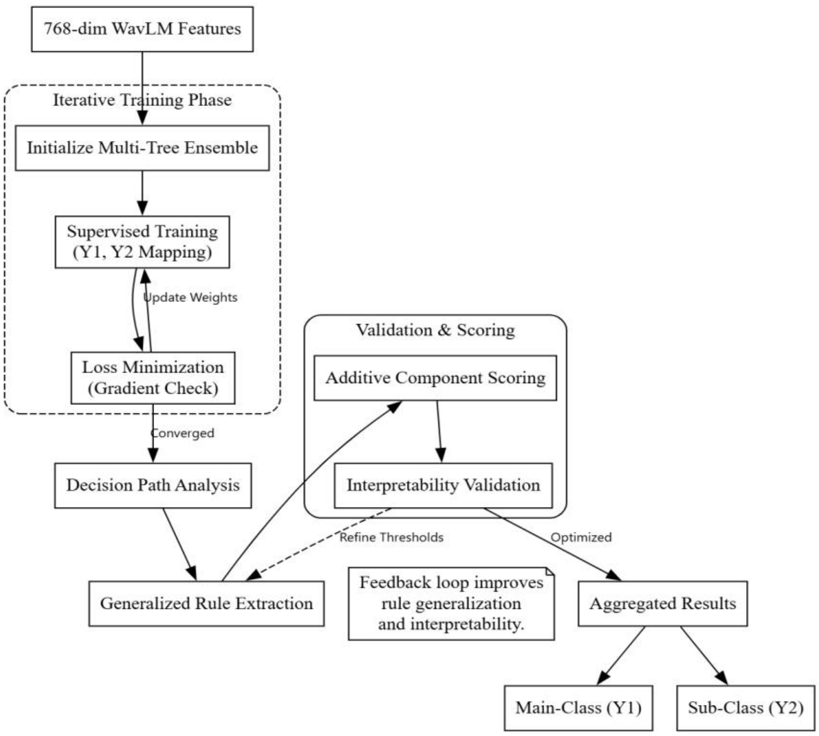


Fig. 2: Operational flow of Proposed TGAM.

Input Feature Initialization: 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.

Feature Pattern Learning: 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.

Tree Structure Construction: 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.

Rule Extraction Process: 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.

Ensemble Decision Formation: 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.

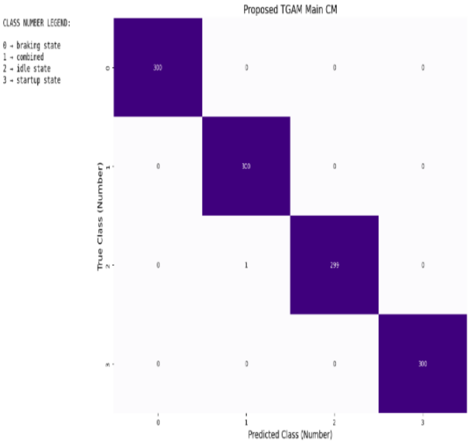
Prediction Generation: Once the model is trained, it can analyze new audio feature vectors and classify them into the corresponding audio event categories. The trained TGAM evaluates the input features using the learned rules and tree structures. This process enables the system to accurately identify vehicle audio events from previously unseen data.

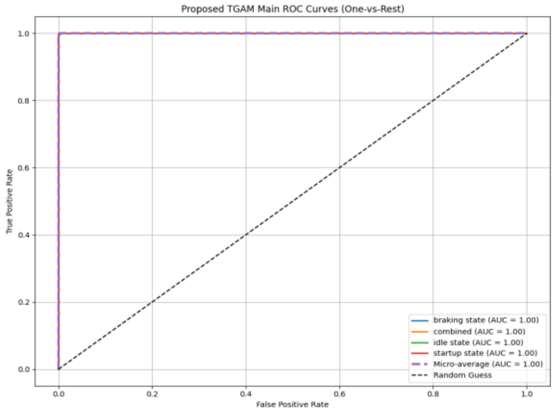
4. RESULTS ANALYSIS

The results obtained from the research demonstrate the effectiveness of machine learning techniques in analysing and classifying vehicle audio signals. After extracting audio features using the WavLM feature extraction method, multiple classification models including CBC, HGBC, ETC, and TGAMC were trained and evaluated using the prepared dataset. The performance of these models was measured using standard evaluation metrics such as accuracy, precision, recall, and F-score. Visualization techniques such as confusion matrices and ROC curves were used to better understand the classification performance of each model. The comparative analysis of different models helps in identifying the most effective algorithm for recognizing vehicle audio events. The results also demonstrate how deep feature extraction combined with machine learning classification can successfully interpret complex acoustic patterns present in vehicle environments.

Fig. 3 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.

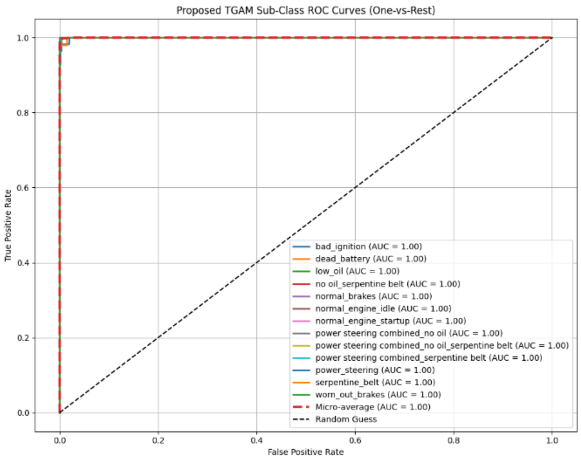
Fig. 4 illustrates the TGAMC model's performance on sub-class classification, covering multiple detailed engine and system conditions. Subfigure (b) shows a confusion matrix with strong diagonal values, where most classes are predicted with exceptionally high accuracy, reflecting the model's ability to separate subtle differences in audio signatures. Misclassifications are minimal, indicating robust feature extraction and representation. Subfigure (a) presents ROC curves for all sub-classes, each achieving an AUC of 1.00, confirming excellent discriminative capability across highly diverse fault categories.

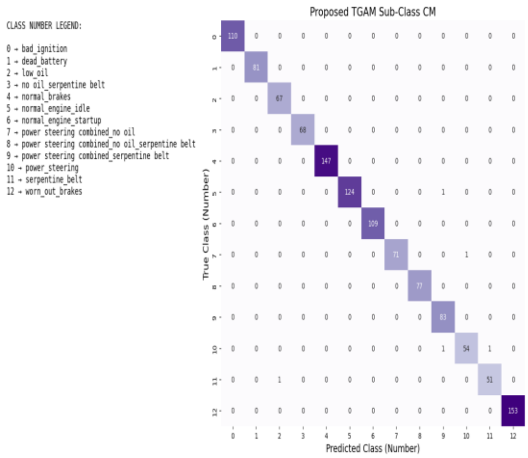




(a) (b)

Fig. 3: TGAMC Main-Class (a) Confusion Matrix (b) ROC Curve





(a)

(b)

Fig. 4: TGAMC Sub-Class (a) Confusion Matrix (b) ROC Curve.

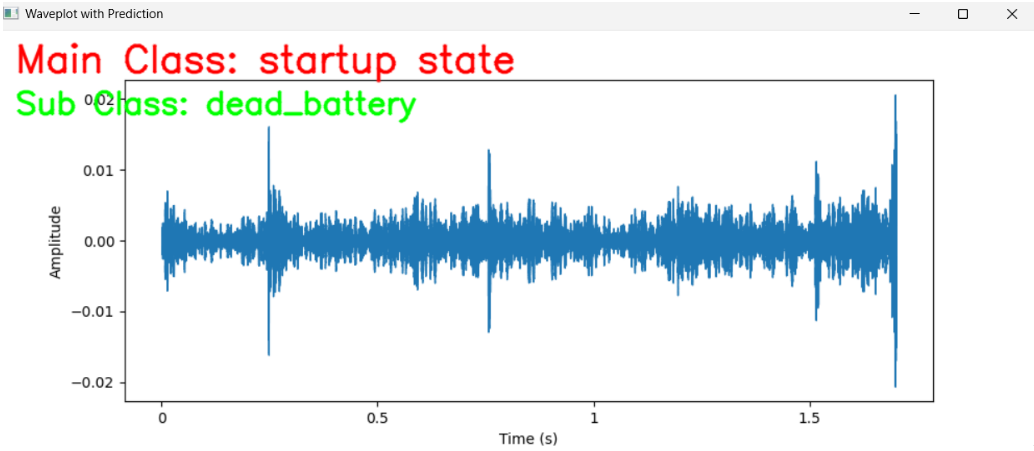


Fig. 5: Prediction output screen for dead battery

Fig. 5 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.

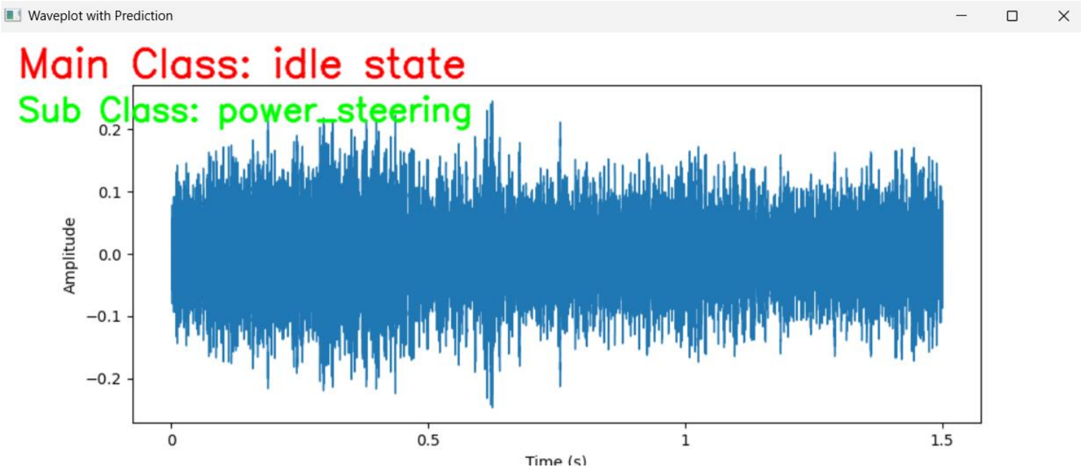


Fig. 6: Prediction output screen for power steering

Fig. 6 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.

The comparative analysis evaluates the performance of different machine learning classification models used for vehicle audio event classification. In this research, models such as CBC, HGBC, ETC, and TGAMC are analysed to understand their ability to learn and classify complex acoustic patterns extracted from vehicle audio signals. Each model is trained using the same feature dataset obtained from the WavLM feature extraction process to ensure a fair comparison. The performance of these models is assessed using evaluation metrics such as accuracy, precision, recall, and F-score. Visualization techniques like confusion matrices and ROC curves are used to interpret the classification results and understand model behaviour. By comparing the results of different algorithms, the analysis helps identify the most effective model for accurately classifying vehicle audio events.

Table. 1: Performance comparison of existing and proposed algorithms for Main Class Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Micro-AUC
CBC	91.67	92.24	91.67	91.45	0.9897
HGBC	69.17	69.10	69.17	68.96	0.7984
ETC	42.25	33.14	42.25	35.59	0.6836
TGAMC	99.92	99.91	99.92	99.91	1.0000

The comparative results show clear performance differences among the machine learning models used for vehicle audio event classification. The CBC model achieved a strong accuracy of 91.67%, demonstrating reliable classification performance with high precision and recall values. The HGBC model obtained an accuracy of 69.17%, indicating moderate performance in recognizing audio event categories compared to other models. The ETC model recorded the lowest accuracy of 42.25%, showing limited ability to capture complex acoustic patterns in the dataset. In contrast, the TGAMC model achieved the highest accuracy of 99.92%, along with very high precision, recall, and F1-score values of 99.91%. Additionally, TGAMC obtained a Micro-AUC score of 1.0000, indicating excellent classification capability and near-perfect discrimination between different audio classes. These results demonstrate that TGAMC significantly outperforms the other models in accurately classifying vehicle audio events.

Table. 2: Performance comparison of existing and proposed algorithms for Sub Class Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
CBC	88.58	89.27	84.56	85.72
HGBC	23.25	30.68	19.62	18.30
ETC	24.42	9.40	17.48	12.00
TGAMC	99.58	99.41	99.41	99.41

The performance comparison of the sub class models shows noticeable differences among the implemented machine learning algorithms. The CBC model achieved an accuracy of 88.58%, demonstrating good classification capability with strong precision and recall values. In contrast, the HGBC model produced a lower accuracy of 23.25%, indicating limited effectiveness in identifying sub class audio categories. Similarly, the ETC model obtained an accuracy of 24.42%, showing weaker performance in distinguishing between detailed audio subclasses. The TGAMC model achieved the highest accuracy of 99.58%, along with very high precision, recall, and F1-score values of 99.41%. These results indicate that TGAMC provides significantly better performance compared to the other models when classifying detailed vehicle audio sub categories.

5. CONCLUSION

The study validates the effectiveness of machine learning approaches for vehicle audio event classification using advanced feature extraction and modeling techniques. Audio signals are processed through WavLM, transforming raw waveforms into high-quality feature representations suitable for classification. Multiple models, including CBC, HGBC, ETC, and TGAM, were trained and evaluated for both main and sub class prediction tasks. Among these, TGAM demonstrated superior performance, achieving 99.92% accuracy for main classes and 99.58% for sub classes, along with consistently high precision, recall, and F1-score. In comparison, CBC showed moderate results, while HGBC and ETC delivered relatively lower performance. The improved results of TGAM are attributed to its capability to model complex relationships within the feature space effectively. Additionally, WavLM-based feature extraction enhances the discriminative power of the system by capturing rich acoustic patterns

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] 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

[7] 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

[8] Khan, M.Q.; Lee, S. A Comprehensive Survey of Driving Monitoring and Assistance Systems. Sensors 2019, 19, 2574. https://doi.org/10.3390/s19112574

[9] 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

[10] Li, G.; Chung, W.-Y. Electroencephalogram-Based Approaches for Driver Drowsiness Detection and Management: A Review. Sensors 2022, 22, 1100. https://doi.org/10.3390/s22031100

[11] Fu, S.; Yang, Z.; Ma, Y.; Li, Z.; Xu, L.; Zhou, H. Advancements in the Intelligent Detection of Driver Fatigue and Distraction: A Comprehensive Review. Applied Sciences 2024, 14, 3016. https://doi.org/10.3390/app14073016

[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] Zohaib, M.; Asim, M.; ELAffendi, M. Enhancing Emergency Vehicle Detection: A Deep Learning Approach with Multimodal Fusion. Mathematics 2024, 12, 1514. https://doi.org/10.3390/math12101514

[14] Fei, M.; Zhou, W.; Zhao, H.; Pan, C.; Shi, D.; An, X. Enhancing Driving Safety Evaluation Through Correlation Analysis of Driver Behavior. Sustainability 2025, 17, 4067. https://doi.org/10.3390/su17094067

[15] Florez, R.; Palomino-Quispe, F.; Alvarez, A.B.; Coaquira-Castillo, R.J.; Herrera-Levano, J.C. A Real-Time Embedded System for Driver Drowsiness Detection Based on Visual Analysis of the Eyes and Mouth Using Convolutional Neural Network and Mouth Aspect Ratio. Sensors 2024, 24, 6261. https://doi.org/10.3390/s24196261