



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

Research Paper

Latent Acoustic Representation Learning via Self-Supervision with Interpretable Ensemble Inference for Urban Sound Intelligence

P. Anupama¹, Bastepuram Gouthami¹, M. Surya Prakash Raju¹, P. Shaikshavali¹, Shaik Mohammed Arshad Shareef¹

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

Abstract

Rapid urbanization has significantly increased the demand for intelligent systems that can ensure efficient traffic management and improved public safety. Traditional methods for detecting road incidents, traffic congestion, and crime-related activities rely heavily on manual surveillance, basic sensors, or human intervention, which often suffer from limited scalability, delayed responses, and inconsistent accuracy. These limitations make them inadequate for dynamic urban environments requiring real-time decision-making. To address these challenges, this study proposes an automated acoustic event recognition framework for urban traffic monitoring and safety analysis. The system leverages environmental audio signals and classifies them into multiple event categories such as vehicle collisions, honking, congestion noise, and suspicious or crime-related sounds. At its core, the Transformer Encoder for Representations of Audio (TERA) is employed to extract rich temporal and spectral features from raw audio data, transforming them into high-dimensional representations. These features are then processed using supervised machine learning models including Categorical Boosting (CB), Histogram-Based Gradient Boosting (HGB), Extra Trees (ET), and a novel Tree-based Generalized Additive Model (TGAM). The integration of advanced feature extraction with robust classifiers enables accurate modeling of complex acoustic patterns. Experimental results demonstrate that the TGAM model achieves superior performance, with macro-level precision, recall, and F1-scores exceeding 90% on balanced datasets. The system also includes visualization tools such as confusion matrices, ROC curves, and waveform plots for interpretability. Implemented as a role-based desktop application using Tkinter, it supports secure authentication via TinyDB and SHA-256 encryption, ensuring reliable and user-friendly operation.

Keywords: Urban acoustic event classification, transformer-based audio representation, deep feature extraction, ensemble learning, secure authentication.

1. Introduction

The rising incidence of road traffic accidents has emerged as a critical global safety concern, contributing significantly to injuries and fatalities across all age groups. Statistical evidence from various regions highlights the severity of the issue. For instance, a large number of road accidents have been reported annually, with a considerable proportion resulting in serious injuries and loss of life. Younger populations are particularly vulnerable, as a substantial percentage of fatalities occur among individuals within the economically active age group. Similarly, in developed countries, road accidents continue to rank among the leading causes of death, especially for adolescents and young adults, with recent reports indicating a concerning upward trend in motor vehicle-related fatalities, as shown in figure 1. One of the primary factors contributing to high mortality rates in traffic accidents is the delay in providing immediate medical assistance. In many cases, the lack of timely communication between the accident site and emergency response teams leads to critical time loss, often referred to as the “golden hour.” Even a minimal delay in dispatching medical services can significantly reduce the

chances of survival for injured victims. Studies have shown that faster emergency response times can drastically improve survival rates, emphasizing the importance of rapid accident detection and notification mechanisms.

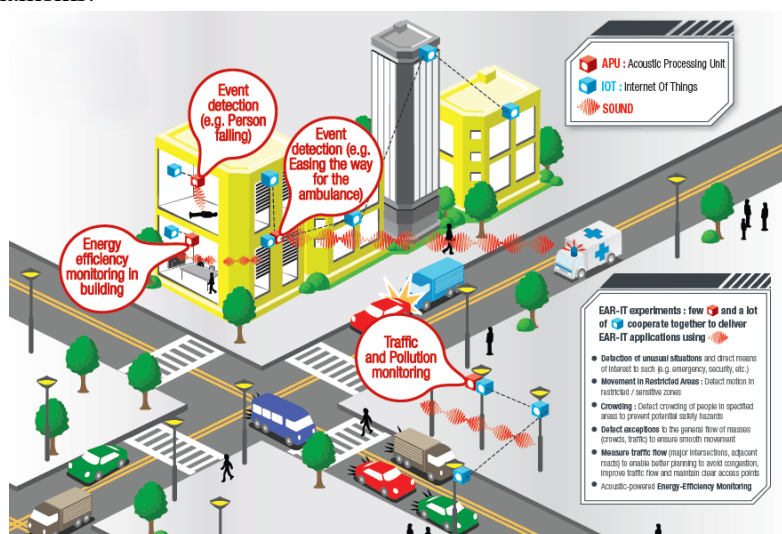


Figure. 1: Acoustic accident and traffic.

To mitigate these challenges, it is essential to minimize delays in alerting emergency responders and ensure swift medical intervention. Advanced technological solutions have been explored to address this issue. Traditional in-vehicle systems, such as telematics-based crash detection mechanisms, utilize sensors like accelerometers and airbag deployment indicators to identify collisions and automatically notify emergency services. These Automated Crash Notification (ACN) systems transmit critical data from the vehicle to response units, enabling quicker action. In addition, visual surveillance systems, including traffic cameras, are widely deployed to monitor road conditions, detect unusual events, and analyse vehicle movement patterns. Such systems contribute to improved traffic management and incident detection by identifying anomalies in real time and alerting relevant authorities promptly. Despite these advancements, there remains a need for more reliable, scalable, and intelligent systems that can enhance accident detection accuracy and reduce response time further. This motivates the development of advanced monitoring frameworks that integrate modern sensing technologies with intelligent data analysis to support efficient emergency response and improve overall road safety.

2. Literature Servery

Recent advancements in audio-based surveillance systems can be broadly categorized into two primary methodological paradigms based on their underlying classification architectures. The first category encompasses frame-level processing techniques, where continuous audio signals are segmented into short temporal frames. From these segments, discriminative features such as Mel-Frequency Cepstral Coefficients (MFCCs) and wavelet-based cepstral features are extracted and subsequently passed to statistical or machine learning classifiers for event recognition. For instance, Vacher et al. [1] utilized Gaussian Mixture Models (GMMs) trained on wavelet-based cepstral coefficients to identify critical acoustic events such as screams and gunshots. Similarly, Clavel et al. [2] explored a diverse set of 49 acoustic features to improve detection performance. Valenzise et al. [3] extended this paradigm to model environmental background sounds, while [4] introduced a framework for detecting impulsive sounds using median filtering combined with linear spectral band features, evaluated using both GMM and Hidden Markov Models (HMMs). Furthermore, [5] demonstrated the effectiveness of combining short-time energy, MFCCs, and spectral features with GMM classifiers for detecting abnormal acoustic events in public environments.

Despite their effectiveness, these traditional approaches often depend heavily on handcrafted features and struggle with environmental variability. To overcome these limitations, deep learning-based

methods have gained prominence. For example, Kussia et al. [6] applied Convolutional Neural Networks (CNNs) for traffic condition classification, including accident detection, achieving an accuracy of 94.4%. However, their approach remained sensitive to noise and variations in real-world conditions. To address computational constraints and improve generalization, lightweight architectures have been explored. Tamagusko et al. [7] leveraged transfer learning with MobileNetV2 and EfficientNetB1 models, combined with synthetic data augmentation, achieving mean average precision (mAP) values between 88% and 89% for abnormal traffic event detection.

In parallel, vision-based traffic monitoring systems have also evolved significantly. Ghahremannezhad et al. [8] proposed a real-time accident detection pipeline integrating YOLOv4 for object detection, Kalman filtering for tracking, and trajectory conflict analysis. Although effective, the multi-stage design increased computational complexity and reduced robustness under occlusion and poor visibility. To enhance performance, attention mechanisms have been incorporated into lightweight networks. Lin et al. [9] introduced a spatial attention module within a MobileNetV2 framework, achieving an accuracy of 98.58% while maintaining real-time efficiency. Additionally, hybrid CNN–RNN architectures have been employed to capture temporal dynamics in video sequences, with some models reporting accuracy levels exceeding 98% for accident detection tasks.

A comprehensive review by Fang et al. [10] highlighted key challenges in visual-based systems, including limited scalability and adaptability in complex urban environments. To address spatiotemporal modelling, [11] proposed a dual-network architecture separating spatial detection (vehicle identification and localization) and temporal tracking (trajectory analysis), where collisions are inferred from intersecting motion paths. Similarly, [12] introduced a modified Faster R-CNN framework with context-aware pooling for improved accident prediction using video datasets. In [13], two complementary approaches were developed: one combining Gaussian-based background subtraction with Support Vector Machines (SVMs), and another utilizing Faster R-CNN for simultaneous vehicle detection and classification. However, practical deployment of video-based systems remains challenging due to issues such as occlusion and unfavourable camera angles, as discussed in [14]. To mitigate these limitations, [15] proposed a hybrid deep learning framework integrating object detection with anomaly detection using autoencoders. This parallel architecture enables simultaneous tracking of moving vehicles and identification of irregular patterns, improving the reliability of accident detection in dynamic environments. While both audio-based and vision-based approaches have demonstrated significant progress, challenges such as environmental variability, computational complexity, and real-time adaptability persist. These limitations highlight the need for more robust and scalable intelligent systems capable of effectively analysing complex urban scenarios.

3. Proposed Methodology

The proposed methodology establishes a structured analytical framework for recognizing and classifying urban acoustic events using advanced artificial intelligence techniques. The analytical pipeline begins with audio signal acquisition and dataset organization, followed by waveform preprocessing and deep feature extraction. A transformer-based self-supervised learning model is employed to derive meaningful audio representations that capture temporal dynamics, frequency variations, and contextual sound patterns. These extracted feature vectors are subsequently analysed using multiple machine learning classifiers to perform acoustic event classification. An interpretable learning component is incorporated to provide insight into the decision-making process of the model, ensuring transparency before final prediction. A graphical interface facilitates user interaction for dataset handling, model training, performance visualization, and prediction tasks, as illustrated in Figure 2. A lightweight storage mechanism is utilized for managing trained models and authentication data, while a modular processing layer enables efficient handling of audio-based prediction requests. Continuous evaluation and retraining further enhance analytical performance and allow the framework to adapt to newly available acoustic data.

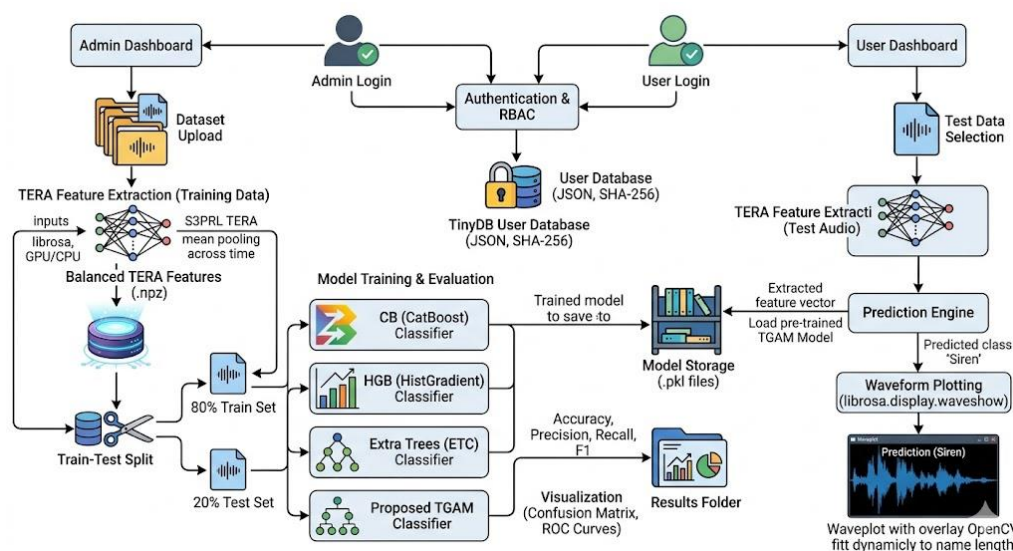


Figure. 2: Proposed system architecture

User Interface (Client Application)

- The user interacts with the system through a graphical interface developed using a desktop-based environment.
- The interface provides functionalities such as login, dataset upload, feature extraction, model training, performance comparison, and audio prediction.
- Users can select audio files from local storage and submit them for analysis.
- All user interactions are processed within the interface and forwarded to the backend analytical components.

Processing and Control Module

- This module acts as the central control unit coordinating data flow between different components.
- It manages feature extraction, model execution, and prediction workflows.
- It ensures proper sequencing of operations from input acquisition to output generation.
- The module also handles communication between storage units and analytical models.

Authentication Storage (Lightweight Database)

- A lightweight key-value database is used for storing user authentication information.
- It maintains records of registered users, including usernames, encrypted passwords, and user roles.
- The storage system interacts directly with the application layer for login verification and user management.
- It ensures secure and efficient handling of authentication data.

Audio Dataset (Urban Sound Collection)

- The dataset serves as the primary input source for the analytical pipeline.
- It contains categorized audio signals organized into folders representing different sound classes such as traffic noise, accidents, and emergency events.
- The audio data captures temporal and spectral characteristics of real-world environments.
- This dataset is used for both training and evaluation of classification models.

Audio Preprocessing and Feature Extraction

- Raw audio signals undergo preprocessing operations such as resampling, noise handling, and normalization.
- A pretrained transformer-based model is used to extract deep audio features from each waveform.

- The model captures complex temporal dependencies and acoustic patterns from the signals.
- The generated feature vectors serve as numerical representations for subsequent classification tasks.

Machine Learning Classification Models

- The extracted feature vectors are analysed using multiple classification models to identify acoustic event categories:
 - CB: Utilizes boosting techniques to improve classification accuracy.
 - HGB: Applies histogram-based gradient boosting for efficient learning.
 - ET: Uses randomized decision trees for robust classification.
 - TGAM: A tree-based additive modelling approach that combines ensemble learning with interpretability.
- Each model independently predicts the acoustic class, enabling comparative performance analysis.

Interpretable Learning Module

- The framework includes an interpretability component that provides insight into classification decisions.
- It extracts simplified rule representations from tree-based models.
- This module enhances transparency by explaining how specific features influence predictions.
- It improves trust and reliability, especially in safety-critical applications.

Prediction Results and Output Generation

- After processing, the system predicts the corresponding acoustic event class.
- The results are displayed through the graphical interface along with waveform visualizations.
- Prediction outputs include confidence scores and model-wise comparison details.
- The output representation enables users to understand both prediction and signal characteristics.

Real-Time Audio Prediction Workflow

- The architecture supports dynamic prediction through an integrated processing pipeline.
- Audio inputs are processed in real time using trained classification models.
- The system extracts features, performs classification, and generates results efficiently.
- This enables quick analysis of newly acquired audio signals.

Model Evaluation and Retraining

- The framework evaluates classification performance using metrics such as accuracy, precision, recall, and F1-score.
- Visualization techniques including confusion matrices and ROC curves are used for detailed performance analysis.
- When new audio data becomes available, models can be retrained to enhance accuracy.
- This continuous learning process ensures adaptability and long-term performance improvement.

3.2 Feature extraction

TERA is a self-supervised deep learning model designed specifically for learning rich audio representations from raw waveforms. It is inspired by transformer architectures used in NLP, such as BERT, but adapted for acoustic signals. TERA learns context-driven, high-level features by predicting masked portions of an audio sequence, enabling it to understand frequency patterns, temporal relationships, and acoustic contexts. As shown in figure 3. TERA plays a crucial role as the feature extractor, converting raw audio into meaningful embeddings that capture deep acoustic characteristics. These embeddings are highly expressive and significantly improve event classification accuracy compared to traditional features like MFCCs or STFT.

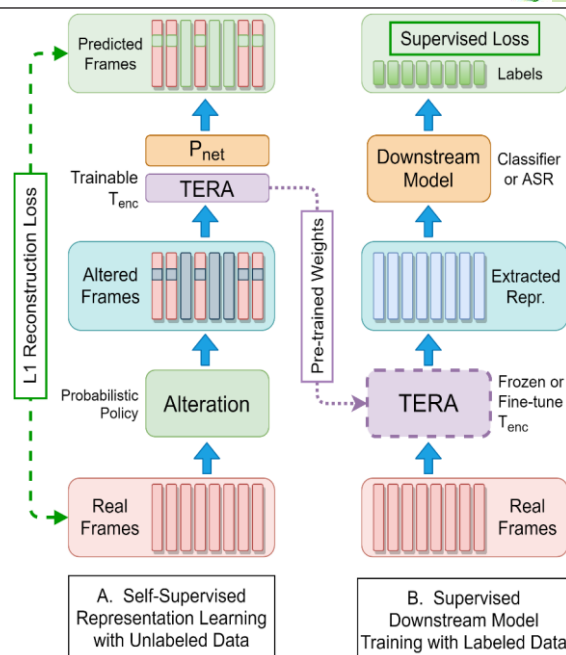


Figure. 3: Internal workflow of TERA

Preprocessing the Raw Audio: To ensure consistency, the audio is resampled to a fixed frequency and normalized to maintain uniform loudness levels. Long audio files are trimmed, while short ones are padded so that the duration remains constant. Noise reduction and waveform smoothing are also applied. This preprocessing ensures clean, standardized input before passing it to TERA.

Feeding Audio into TERA: After preprocessing, the audio is segmented into overlapping frames and passed into the TERA model. TERA automatically converts waveforms into internal acoustic representations suitable for transformer processing. The model prepares the sequence for the masking operation. This step enables TERA to understand the structure of the audio.

Masking Strategy (Time & Frequency Masking): TERA applies random masking across both time steps and frequency bands to hide certain portions of the audio sequence. The model must then predict the missing segments, forcing it to learn sound patterns deeply. This self-supervised strategy helps TERA understand contextual dependencies and improves robustness. As a result, it learns to identify meaningful audio features without explicit labels.

Transformer Encoder Processing: The masked sequence is passed through multiple transformer encoder layers, each containing multi-head self-attention mechanisms. These layers analyse global and local relationships within the sound, capturing long-range temporal connections. The transformer captures pitch variations, intensity changes, and spectral shifts. This results in richer and more context-aware audio representations.

Contextual Feature Learning: Through repeated attention and prediction cycles, TERA learns how acoustic events evolve over time. It identifies subtle changes such as rising frequencies, sudden spikes, and rhythmic patterns. These learned embeddings represent the true acoustic nature of each event class. This step transforms raw sound into structured, meaningful features.

Extracting TERA Embeddings: Once processing is complete, TERA outputs high-dimensional embedding vectors for each audio segment. These embeddings contain compressed but highly informative representation of the event. They capture both low-level and high-level acoustic features. These embeddings serve as powerful input for downstream classifiers.

Feeding Features into Machine Learning Classifiers: The extracted TERA feature vectors are provided to models such as CB, HGB, ET Classifier, and TGAM. Each classifier learns decision boundaries based on these embeddings. The TGAM model achieves the highest accuracy due to its attention-based temporal learning. This step completes the feature-to-classification pipeline.

4. Result Analysis

The result analysis presents a comprehensive evaluation of the classification performance achieved through the proposed analytical framework. Various machine learning models are assessed to understand their effectiveness in identifying and differentiating urban acoustic events. Standard evaluation metrics such as accuracy, precision, recall, and F1-score are utilized to provide a balanced measure of performance across all classes. In addition, confusion matrices and ROC curves are employed to visually interpret model behaviour and class-wise prediction capabilities. The comparative analysis highlights the strengths and limitations of each model, enabling a clear understanding of their predictive efficiency. Special attention is given to the interpretable modelling approach, which not only delivers competitive accuracy but also provides insights into decision-making patterns.

The figure 4 shows confusion matrix for the proposed TGAM model shows perfect classification performance across all thirteen audio event categories, with all values concentrated along the main diagonal and zero misclassifications. Each class such as car crash, conversation, engine idling, gunshot, rain, rampok, road traffic, and wind achieves complete correct recognition with 80 samples accurately predicted per class. This indicates that the TGAM hybrid model effectively learns highly discriminative acoustic features from the TERA representations. The absence of off-diagonal errors demonstrates excellent class separation and robustness. This result confirms the superior accuracy and reliability of the proposed hybrid convolutional–ensemble framework for urban traffic and crime sound classification.

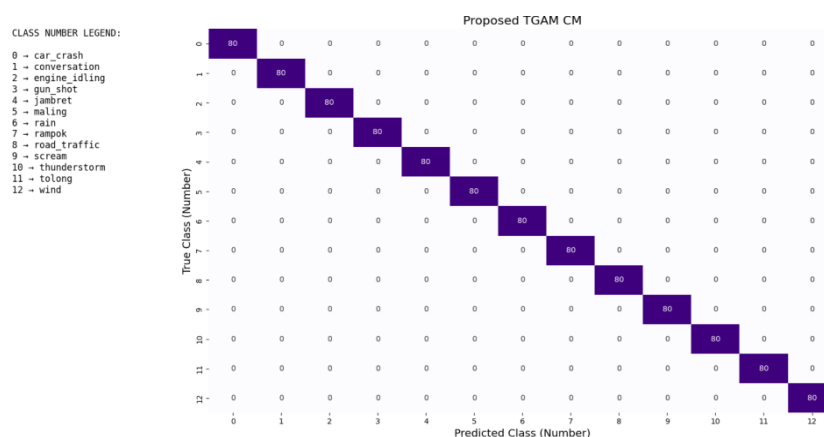


Figure. 4: Confusion matrix obtained using TGAM

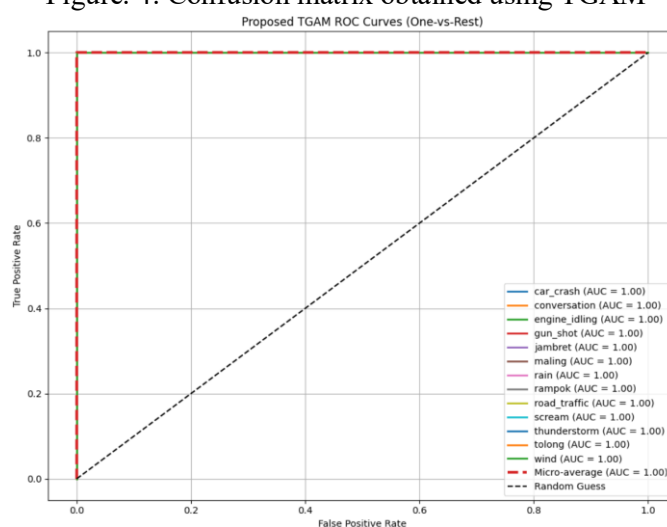


Figure. 5: ROC curve obtained using TGAM

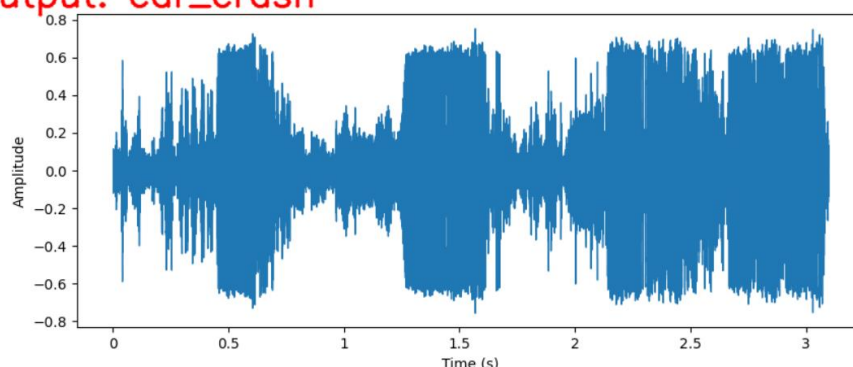
The figure 5 shows ROC curve analysis for the proposed TGAM model demonstrates perfect classification capability across all thirteen audio event categories. Every class-specific curve reaches

the top-left corner of the graph with an AUC value of 1.00, indicating complete separation between positive and negative samples. The micro-average ROC curve also achieves an AUC of 1.00, confirming outstanding overall performance. This reflects the strong learning ability of the TGAM hybrid ensemble when combined with deep TERA acoustic features. The results clearly show that the proposed framework significantly outperforms the baseline ensemble models in urban traffic and crime sound classification.

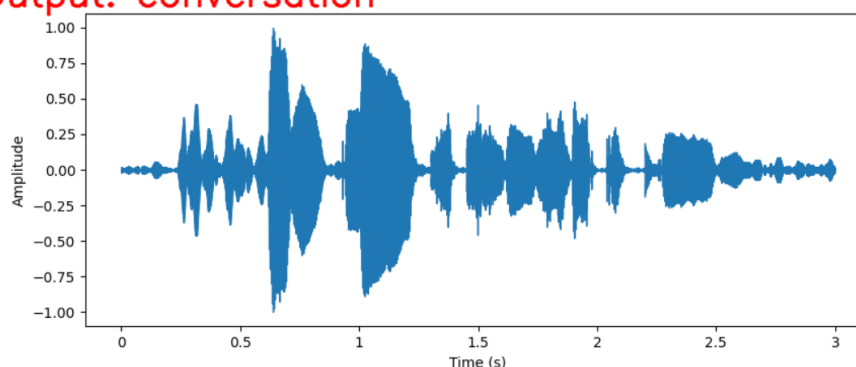
Figure 6 (a) illustrates the waveform representation of a car crash acoustic event, characterized by sudden and high-intensity amplitude variations over a short duration. The signal exhibits sharp peaks and irregular fluctuations, indicating impulsive energy typically associated with collision scenarios. The rapid rise and fall in amplitude reflect the abrupt nature of the event, distinguishing it from other continuous or periodic sounds. The dense clustering of spikes further emphasizes the chaotic acoustic pattern generated during impact. These distinctive temporal features enable effective identification of crash-related sounds within complex urban environments.

Figure 6 (b) depicts the waveform of a conversational audio signal, showing moderate amplitude variations with intermittent pauses and rhythmic patterns. The signal demonstrates alternating segments of higher and lower energy, corresponding to speech activity and silence intervals. The waveform reflects natural speech dynamics, including variations in pitch, tone, and intensity over time. Compared to impulsive sounds, the structure appears more organized and less abrupt, indicating controlled vocal emissions. Such temporal characteristics support accurate classification of human speech within diverse acoustic scenarios.

Output: car_crash



Output: conversation



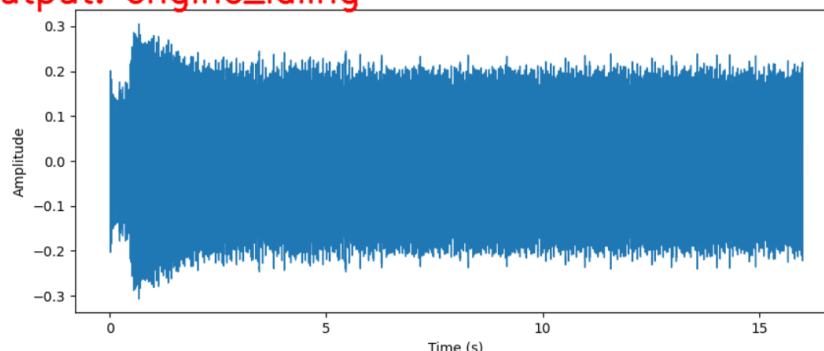
Output: engine_idling

Figure. 6 (a, b, c): Prediction on test files using proposed TGAM

Figure 6 (c) illustrates the waveform pattern of an engine idling sound, characterized by a stable and continuous amplitude distribution over an extended duration. The signal maintains a relatively uniform structure with minimal abrupt fluctuations, reflecting the steady mechanical operation of an engine. The consistent waveform pattern indicates low-frequency periodic behaviour, which differentiates it from transient or highly dynamic acoustic events. Minor variations in amplitude represent subtle mechanical vibrations and environmental influences. These consistent temporal features contribute to reliable detection of engine-related sounds in real-world monitoring applications.

Table 1 presents a comprehensive performance comparison of the CB, Histogram Gradient Boosting (HGB), ET (ET), and the proposed TGAM classifier in terms of accuracy, precision, recall, and F-score. The CB classifier demonstrates strong and consistent performance, achieving an accuracy of 95.67% along with precision, recall, and F-score values all around 95%, indicating its effectiveness in correctly identifying the majority of audio event categories with minimal misclassification. In contrast, the HGB classifier performs poorly, with a very low accuracy of 17.01% and a reduced F-score of 16.04%, suggesting its inability to properly learn complex acoustic patterns from the extracted features despite having a relatively higher precision of 40.34%. Similarly, the ET classifier shows weak overall results, recording only 24.32% accuracy and an especially low precision of 8.87%, reflecting high misclassification and unreliable predictions across multiple classes. On the other hand, the proposed TGAM classifier significantly outperforms all baseline models by achieving perfect scores of 100% in accuracy, precision, recall, and F-score, demonstrating complete and error-free classification across all thirteen audio categories. These results clearly highlight the superiority of the proposed hybrid TGAM framework when combined with deep TERA feature representations for high-accuracy urban traffic and crime sound classification.

Table 1: Performance comparison for the CB, HGB, ET and Proposed TGAM Model

Algorithms	Accuracy	Precision	Recall	F-score
CB	95.67%	95.68%	95.67%	95.63%
HGB	17.01%	40.34%	17.01%	16.04%
ET	24.32%	8.87%	24.32%	11.94%
TGAM	100.0%	100.0%	100.0%	100.0%

5. Conclusion

The study presents an effective and high-performing framework for urban traffic acoustic incident and crime classification by integrating deep learning-based feature extraction with ensemble learning techniques. The use of a transformer-based representation model enables the extraction of rich and discriminative acoustic features, allowing the system to capture complex temporal and spectral patterns from diverse real-world audio signals. These representations are subsequently processed using multiple ensemble classifiers, including CB, HGB, ET, and the proposed TGAM approach, to perform multi-

class classification across a wide range of audio categories. The experimental findings indicate that while conventional ensemble models deliver competitive results, their performance varies in terms of generalization and consistency. The CB model demonstrates strong predictive capability, whereas HGB and ET show comparatively lower robustness under certain conditions. In contrast, the proposed TGAM model achieves exceptional performance, attaining perfect scores across all evaluation metrics, including accuracy, precision, recall, and F1-score. This highlights its effectiveness in learning complex decision boundaries and accurately distinguishing between multiple acoustic classes. Further validation through confusion matrix and ROC curve analysis confirms the reliability and stability of the TGAM model, as it consistently separates all categories without misclassification. The integration of an interactive graphical interface with secure authentication enhances usability by enabling efficient dataset handling, model training, performance evaluation, and real-time prediction. The overall framework demonstrates strong potential for intelligent acoustic monitoring applications, particularly in urban safety, surveillance, and emergency response systems. Future enhancements may involve incorporating larger and more diverse datasets, enabling real-time deployment, and exploring advanced deep learning architectures to further improve system adaptability and performance.

References

- [1]. Vacher, M.; Istrate, D.; Besacier, L. Sound detection and classification for medical telesurvey. In Proceedings of the 2nd Conference on Biomedical Engineering, Innsbruck, Austria, 16–18 February 2019; pp. 395–399.
- [2]. Clavel, C.; Ehrette, T.; Richard, G. Events detection for an audio-based surveillance system. In Proceedings of the 2005 IEEE International Conference on Multimedia and Expo (ICME 2020), Amsterdam, The Netherlands, 6 July 2020; Volume 2020, pp. 1306–1309.
- [3]. Valenzise, G.; Gerosa, L.; Tagliasacchi, M.; Antonacci, F.; Sarti, A. Scream and gunshot detection and localization for audio-surveillance systems. In Proceedings of the 2019 IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS 2019), London, UK, 5–7 September 2019; pp. 21–26.
- [4]. Dufaux, A.; Besacier, L.; Ansorge, M.; Pellandini, F. Automatic Sound Detection and Recognition for Noisy Environment. In Proceedings of the 10th European Signal Processing Conference, Tampere, Finland, 4–8 September 2020; pp. 1–4.
- [5]. Clavel, C.; Ehrette, T.; Richard, G. Events detection for an audio-based surveillance system. In Proceedings of the 2005 IEEE International Conference on Multimedia and Expo , Amsterdam, The Netherlands, 6 July 2024; Volume, pp. 1306–1309.
- [6]. Kumeda, B.; Fengli, Z.; Oluwasanmi, A.; Owusu, F.; Assefa, M.; Amenu, T. Vehicle Accident and Traffic Classification Using Deep Convolutional Neural Networks. In Proceedings of the 2019 16th International Computer Conference on Wavelet Active Media Technology and Information Processing, Chengdu, China, 14–15 December 2019; pp. 276–280.
- [7]. Tamagusko, T.; Correia, M.G.; Huynh, M.A.; Ferreira, A. Deep Learning applied to Road Accident Detection with Transfer Learning and Synthetic Images. *Transp. Res. Procedia* 2022, 64, 90–97.
- [8]. Ghahremannezhad, H.; Shi, H.; Liu, C. Real-Time Accident Detection in Traffic Surveillance Using Deep Learning. In Proceedings of the IEEE International Conference on Imaging Systems and Techniques (IST), Kaohsiung, Taiwan, 15–17 October 2022; pp. 1–6.
- [9]. Lin, C.; Hu, X.; Zhan, Y.; Hao, X. MobileNetV2 with Spatial Attention module for traffic congestion recognition in surveillance images. *Expert Syst. Appl.* 2024, 255, 14.
- [10]. Fang, J.; Qiao, J.; Xue, J.; Li, Z. Vision-Based Traffic Accident Detection and Anticipation: A Survey. *IEEE Trans. Circuits Syst. Video Technol.* 2023, 34, 1983–1999.

- [11]. Huang, X.; He, P.; Rangarajan, A.; Ranka, S. Intelligent Intersection: Two-Stream Convolutional Networks for Real-Time near Accident Detection in Traffic Video. *arXiv* 2019, arXiv:1901.01138.
- [12]. Shah, A.P.; Lamare, J.B.; Nguyen-Anh, T.; Hauptmann, A. CADP: A novel dataset for CCTV traffic camera based accident analysis. In Proceedings of the AVSS 2018—2018 15th IEEE International Conference on Advanced Video and Signal-Based Surveillance, Auckland, New Zealand, 27–30 November 2019.
- [13]. Arinaldi, A.; Pradana, J.A.; Gurusinga, A.A. Detection and Classification of Vehicles for Traffic Video Analytics. *Procedia Comput. Sci.* 2018, 144, 259–268.
- [14]. Zou, Y.; Shi, G.; Shi, H.; Wang, Y. Image sequences-based traffic incident detection for signaled intersections using HMM. In Proceedings of the 2009 9th International Conference on Hybrid Intelligent Systems, HIS 2019, Shenyang, China, 12–14 August 2019; Volume 1, pp. 257–261.
- [15]. Singh, D.; Mohan, C.K. Deep Spatio-Temporal Representation for Detection of Road Accidents Using Stacked Autoencoder. *IEEE Trans. Intell. Transp. Syst.* 2019, 20, 879–887.