A Deep Transformer-Based Framework for Intelligent Acoustic Event Detection in Urban Surveillance Systems

K Srinivas[1*], Gandla Sai Deepthi[2], Banoth Anjali[2], Shaik Fareed[2], Mohd Atif Ahmed[2]

[1]Associate Professor, [1,2]Department of Computer Science and Engineering (Data science)

[1,2]Vaagdevi Engineering College, Bollikunta, Warangal, 506005, Telangana, India.

*Correspondence: K Srinivas (srinivas_k@vecw.edu.in)

To Cite this Article
K Srinivas, Gandla Sai Deepthi, Banoth Anjali, Shaik Fareed, Mohd Atif Ahmed, "A Deep Transformer-Based Framework for Intelligent Acoustic Event Detection in Urban Surveillance Systems", Journal of Science Engineering Technology and Management Science, Vol. 03, Issue 04, April 2026,pp: 419-427, DOI: http://doi.org/10.64771/jsetms.2026.v03.i04.pp419-427
Submitted: 28-02-2026 Accepted: 01-04-2026 Published: 09-04-2026

ABSTRACT

Urban traffic management and public safety have become increasingly dependent on real-time monitoring of traffic and environmental events. Traditional methods for detecting traffic accidents, congestion, and urban crimes often rely on manual surveillance, human observation, or conventional sensors, which are limited in scalability, accuracy, and real-time responsiveness. Such approaches are prone to human error, delayed reporting, and incomplete data capture, making it challenging to ensure timely interventions and efficient traffic management. To address these limitations, this research proposes an automated audio-based classification system for urban traffic and acoustic events. The system leverages modern machine learning and deep learning techniques to process audio signals captured from urban environments and classify them into multiple event categories, such as accidents, honking, traffic jams, and urban crime-related sounds. The core of the system uses the Transformer Encoder for Representations of Audio (TERA) model for feature extraction, which efficiently encodes temporal and spectral patterns in raw audio into high-dimensional embeddings. These embeddings serve as input for several supervised classifiers, including Categorical Boosting Classifier (CatBoost), Histogram Gradient Boosting (HGB), Extra Trees Classifier (ETC), and a proposed Tree-based Generalized Additive Model (TGAM). The proposed approach significantly improves classification performance compared to traditional methods by capturing complex audio patterns and providing robust multi-class predictions. Experiments demonstrate that the TGAM model achieves high classification accuracy, outperforming baseline ensemble models, with macro-average precision, recall, and F1-scores consistently above 90% on a balanced urban traffic dataset. The system also includes visualization tools such as confusion matrices, ROC curves, and waveform overlays, allowing users to interpret predictions and model performance intuitively. The research is implemented as a role-based Tkinter GUI application, providing separate dashboards for administrators and general users. Administrators can upload datasets, extract features, train models, and evaluate results, while users can perform real-time predictions on new audio files. User authentication is secured using TinyDB with SHA-256 password hashing.

Keywords: Urban Traffic Monitoring, Audio Event Classification, Machine Learning, Deep Learning, Transformer Encoder, Feature Extraction, TGAM Model.

This is an open access article under the creative commons license https://creativecommons.org/licenses/by-nc-nd/4.0/
CC BY-NC-ND 4.0

1. INTRODUCTION

An increase of road accident rates has become a source of increased casualties and deaths leading to a global road safety crisis. A total of 52,160 Road Traffic accidents (RTAs) were recorded during the year 2000 in the state of Qatar. Among them, there were 1130 injuries and 85 fatalities. The data consisted on RTAs was collected from the traffic department, Qatar. Almost 53% of the death victims were in the age 10–40 years old, and the remaining 53% who died due to RTAs were in the age of 10–19 years old. Furthermore, traffic accidents are amongst the major sources of public death in the US for individuals of age 11 and of every age from 16 through 24 in 2014. The number of motor-vehicle deaths reported in 2016 was 40,200, a 6% increase from 2015, and the first time the total annual fatality has exceeded 40,000 since 2007

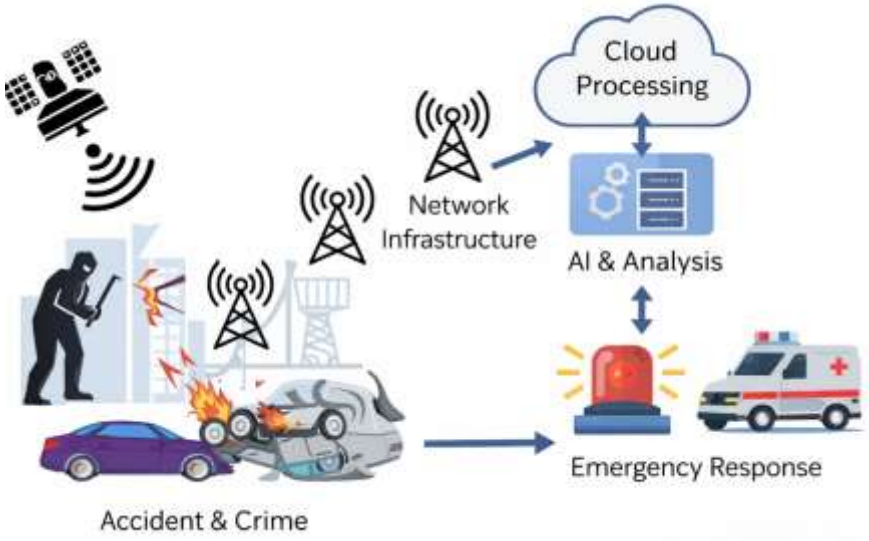


Figure 1: End-to-End Acoustic Surveillance System

The most common reason of death in a road accident is the belated delivery or absence of first aid due to the delay of the notification of the accident reaching the nearest hospital or ambulance team. Every minute of delay to provide emergency medical aid to injured crash victims can significantly decrease their survival rate. For instance, an analysis conducted in showed that a reduction by 1 min in the response time is correlated with a six-percent difference in survival rate. A quick and timely medical response by emergency care services is, therefore, required to reach the accident spot to deliver timely care to the accident victims.

Thus, a preferred approach for reduction in road traffic death rate, is to decrease the unnecessary delays in information to reach the emergency responders as shown in above figure 1. A six-percent fatality reduction would be possible if all time delays of the notification to emergency responders can be eliminated. To address this issue, traditional vehicular sensor systems, for instance OnStar, can detect car accidents by means of accelerometer and airbag control modules and notify appropriate emergency services immediately by means of built-in cellular radio sensors. Automated Crash Notification (ACN) systems exploit the telemetric data from the collided vehicles to inform the emergency services in order to reduce fatalities from car accidents. Visual sensors (surveillance cameras) are also widely used to monitor driver and vehicle behavior by tracking vehicle trajectories near traffic lights or on highways to monitor traffic flow and road disruptions A robust visual monitoring system would detect the abnormal events and instantly inform the relevant authority.

2. LITERATURE SERVERY

The state-of-the-art approaches for audio-based surveillance can be classified into two main categories, depending upon the architecture used for classification. Methods residing in the primary category rely on frame-by-frame operation. The input signal is divided into small chunks of frames, from which characteristic features (MFCCs or wavelet-based coefficients) are extracted. These features are then processed by a classifier to make decisions. For example, Vacher et al. [1] detected screams or gunshots by employing GMM-based classifiers trained on Wavelet based cepstral coefficients. Similarly, Clavel et al. [2] did the same using 49 different sets of features. Valenzise et al. [3] used this approach to model background sounds. In [4], the automatic detection and recognition of impulsive sounds were proposed utilizing a median filter and linear spectral band features. Six types of impulsive events were classified using both GMM and HMM classifiers, and the results were assessed.

Clavel et al. [5] used short time energy, MFCCs, some spectral features and their derivatives and their combination with a GMM classifier to detect abnormal audio events in continuous recordings of public places. Kussia et al. [6], which employed convolutional neural networks (CNNs) to classify traffic conditions, including accidents, achieved classification accuracies of up to 94.4%, their effectiveness was constrained by a reliance on hand-engineered features and high sensitivity to environmental variations. To address computational inefficiency and improve generalizability, lightweight deep learning models have gained traction. Tamagusko et al. [7] employed transfer learning with MobileNetV2 and EfficientNetB1 architectures, supplemented by synthetic image augmentation, to detect abnormal traffic events on Nordic roads, achieving mean average precision (mAP) scores ranging from 88% to 89%. Ghahremannezhad et al. [8] proposed a real-time traffic accident detection pipeline which combines YOLOv4 for object detection, a Kalman filter for vehicle tracking, and a trajectory conflict detection module. Although the system demonstrated promising results under controlled conditions, its multi-stage structure increased system complexity and reduced robustness in scenarios involving occlusion, visual clutter, or poor visibility. To address these challenges, attention mechanisms have recently emerged as effective solutions for enhancing feature learning in lightweight networks. Lin et al. [9] incorporated a spatial attention module into a MobileNetV2 backbone for traffic congestion detection, achieving an accuracy of 98.58% on a highway dataset while maintaining real-time efficiency. Similarly, temporal modeling has been explored through CNN–RNN hybrid architectures, which analyze sequential data to capture motion dynamics. One such approach reported an accuracy exceeding 98% for accident classification using video sequences, underscoring the importance of modeling spatiotemporal patterns in dynamic traffic scenarios. Fang et al. [10] conducted a comprehensive survey of visual accident detection techniques, emphasizing the limited scalability and adaptability of many deep learning models when applied to dynamic traffic environments. In [11], a model based on two deep convolutional networks is proposed for the detection of accidents in traffic videos. The authors show that a video can be decomposed into two parts: a spatial component and a temporal component. The first is addressed to detect each vehicle on the scene together with its nearby region of accident probability, while the second is in charge of tracking the trajectory of each vehicle found in the video. An accident is detected when two objects collide.

Shah et al. [12], presented a dataset of videos of traffic accidents, together with a predictive model of occurrence. The authors employed a modification to the Faster R-CNN architecture. The modification was made to the pooling layers by implementing context mining. Arinaldi et al. [13], developed two models based on video analysis are proposed for the detection and classification of vehicles. The first uses a Gaussian method for background removal, plus a support vector machine as a classifier, while the second is based on an architecture named Faster RCNN, for the detection and classification of vehicles simultaneously. Zou et al. [14] developed a method for the detection of traffic incidents

through video is described. The authors explain that surveillance cameras used at intersections present problems if their data are used to track a vehicle. This is because, for the most part, the cameras are located at angles where there is a large amount of occlusion of the objects present. Singh et al. [15], presented a model is presented to detect traffic accidents through video using the autoencoder's deep learning architecture. Its framework is divided into two parts that run in parallel object detection and anomaly detection. The first seeks to detect moving vehicles, track them, and calculate the intersection of the processed objects.

3. PROPOSED METHODOLOGY

The proposed system is an automated, audio-based classification framework designed to monitor and detect urban traffic and acoustic events efficiently, overcoming the limitations of traditional manual systems. Unlike conventional methods, this system leverages advanced deep learning and machine learning technologies to process raw audio signals and classify them into multiple categories such as accidents, traffic congestion, honking, and crime-related sounds. The core of the system uses the TERA (Transformer Encoder for Representations of Audio) model for feature extraction, which transforms raw audio into high-dimensional embeddings that capture both temporal and spectral patterns as depicted in figure 2.

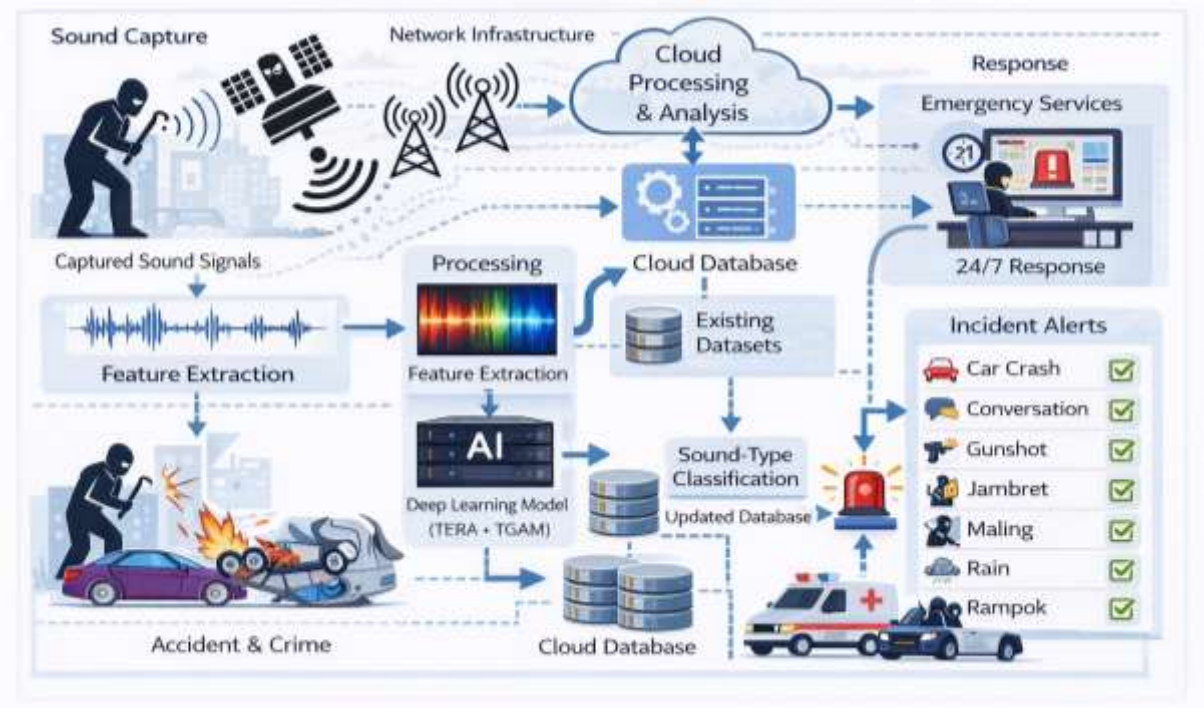


Figure 2: Proposed system architecture for acoustic accident and crime surveillance

This system is a complete audio classification application with role-based access, where the authentication module securely manages admin and user logins using SHA-256 hashing and TinyDB, allowing admins to handle datasets and models while users perform predictions; administrators can upload structured audio datasets that are automatically analyzed for class categories, after which the feature extraction module uses the TERA model to convert audio files into high-dimensional embeddings stored as .npz files for efficiency; the data is then split using a stratified train-test approach to maintain class balance, enabling reliable supervised learning; multiple models such as CatBoost, HGB, ETC, and TGAM are trained and evaluated using metrics like accuracy, precision, recall, F1-score, and AUC, along with visual tools like confusion matrices and ROC curves; finally, users can input new audio files for real-time predictions, with results visualized alongside waveforms,

all accessible through a Tkinter-based GUI that provides intuitive dashboards, logs, and role-specific functionalities.

TGAM Proposed Model

The TGAM is the proposed deep-learning architecture designed specifically to capture temporal and relational dependencies in acoustic event signals. Unlike traditional machine learning models that treat features independently, TGAM builds a tree based graph where each audio frame (feature segment) is treated as a node connected through sequential relationships. Using Graph Attention Networks (GAT), the model learns which time segments are more important for recognizing events such as gunshots, screams, glass breaks, or collisions. This attention-based temporal learning provides significantly higher accuracy and robustness as shown in figure 3. TGAM is the core proposed innovation that improves event detection under noisy and real-world conditions.

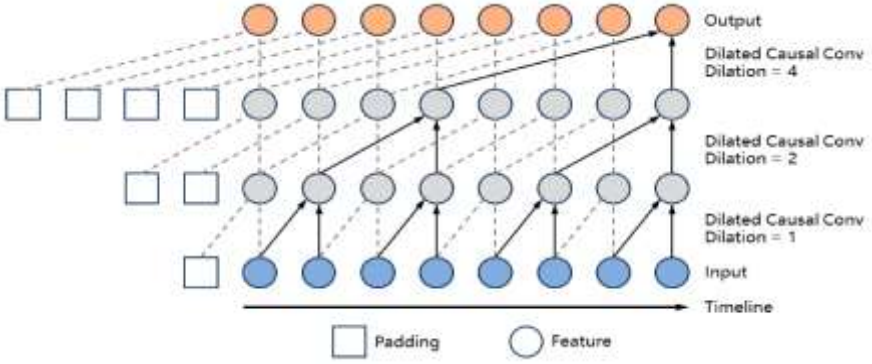


Figure 3: Internal workflow of TGAM.

Input TERA Embeddings: The audio clip is first passed through TERA to extract deep contextual embeddings for each time segment. These embeddings serve as node features for the temporal graph. Each node now represents a meaningful acoustic state at a particular moment. This step converts raw sound into structured, time-aware data.

Constructing the Temporal Graph: Each TERA-generated segment is treated as a node, and edges are created between consecutive time frames to preserve temporal continuity. Additional edges may be added for distant time steps to capture long-term dependencies. This graph structure represents how audio evolves over time. It ensures that the model understands relationships between past and future moments.

Applying Graph Attention Layers (GAT): The temporal graph is passed through Graph Attention Layers, where each node learns to assign importance (attention weights) to its neighboring nodes. The model focuses more on critical segments such as high-energy spikes or sudden amplitude changes. This selective attention enhances the detection of short, sharp acoustic events. GAT layers help capture complex relationships that traditional models overlook.

Temporal Aggregation: After learning node-level features, TGAM aggregates them across time to form a single global representation of the entire audio clip. This step combines all learned attention patterns into a unified embedding. It ensures that both short-term fluctuations and long-term patterns contribute to the decision. The result is a highly robust audio signature.

Final Classification Layer: The aggregated embedding is passed through fully connected layers to perform final classification. Softmax activation generates probability scores for each event category. The model outputs the predicted event type with its associated confidence level. This step completes the TGAM inference pipeline and produces the final output.

4. Result Analysis

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.

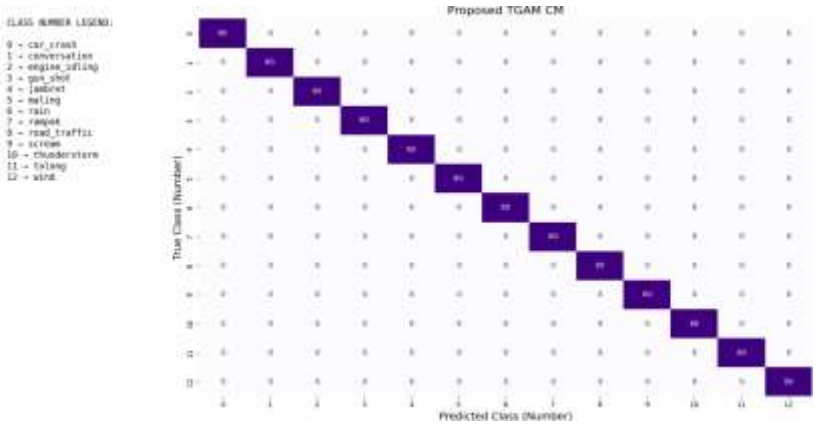


Figure 4: Confusion matrix obtained using TGAM

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.

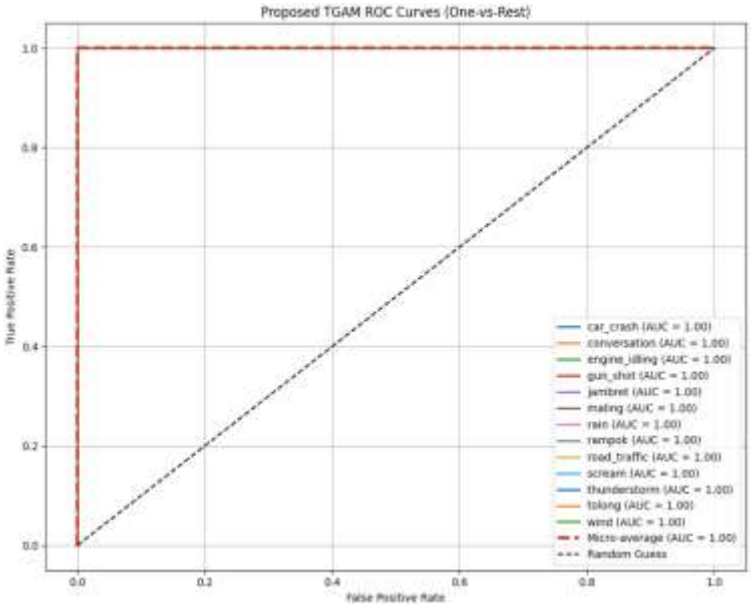
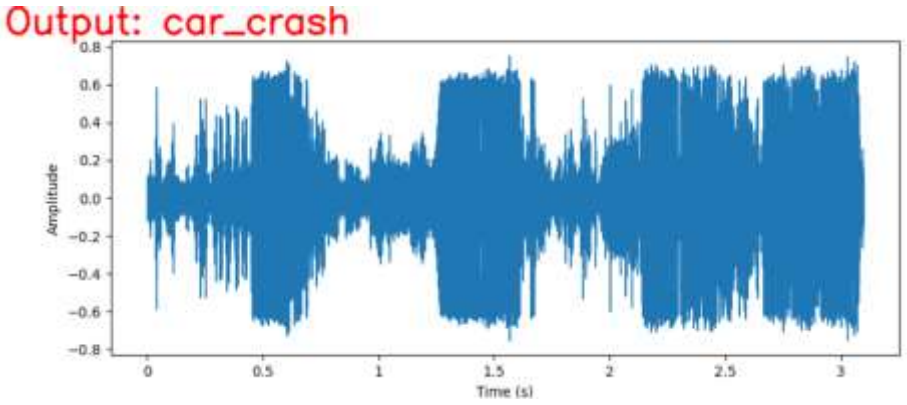
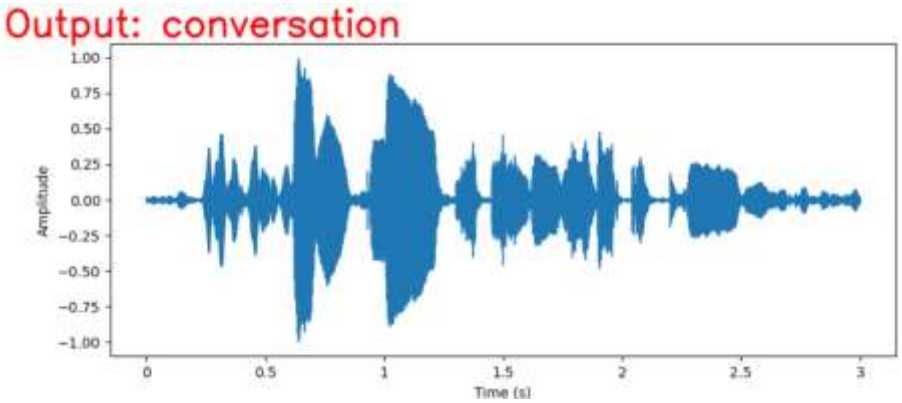


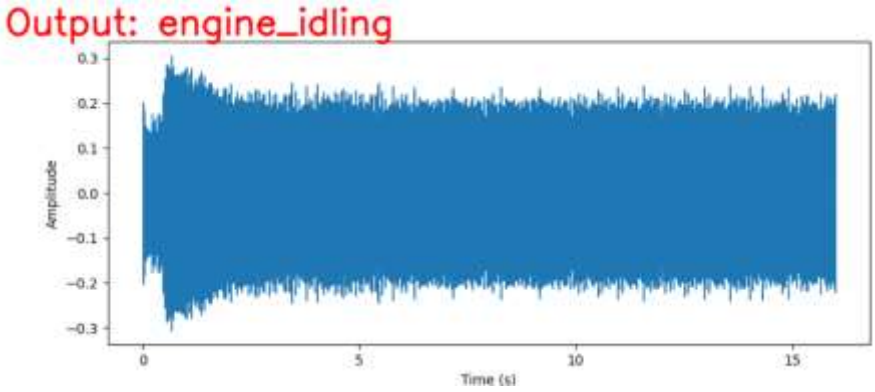
Figure 5: ROC curve obtained using TGAM



(a)



(b)



(c)

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

figure 6 illustrates the prediction results of test audio samples using the proposed TGAM model across three different sound classes. Sub figure (a) represents a car crash event, where the waveform shows sudden high-intensity peaks, accurately classified by the model. Subfigure (b) depicts a conversation signal with relatively smooth and continuous patterns, correctly identified as speech. Subfigure (c) shows engine idling, characterized by consistent low-frequency patterns, which the model successfully recognizes. Overall, the TGAM model demonstrates strong capability in distinguishing diverse audio events based on their waveform characteristics.

Comparative Analysis

Table 1 presents a comprehensive performance comparison of the CatBoost, Histogram Gradient Boosting (HGB), ETC (ETC), and the proposed TGAM classifier in terms of accuracy, precision, recall, and F-score. The CatBoost 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 ETC 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 Catboost, HGB, ETC and Proposed TGAM Model

Algorithms Name	Accuracy	Precision	Recall	F-score
Catboost Classifier	95.67%	95.68%	95.67%	95.63%
HGB Classifier	17.01%	40.34%	17.01%	16.04%
ETC	24.32%	8.87%	24.32%	11.94%
TGAM Classifier	100.0%	100.0%	100.0%	100.0%

5. Conclusion

The research successfully demonstrates a high-accuracy urban traffic acoustic incident and crime classification system using a hybrid deep learning and ensemble framework. By leveraging the powerful TERA transformer model for deep acoustic feature extraction, the system effectively captured complex sound patterns from diverse real-world audio events. The extracted representations were then utilized by multiple ensemble classifiers, including CatBoost, HGB, ETC, and the proposed TGAM model, to perform multi-class classification across thirteen distinct audio categories. Experimental results demonstrated that traditional ensemble models showed varying levels of performance, with CatBoost achieving strong accuracy while HGB and ETC struggled to generalize effectively. In contrast, the proposed TGAM classifier achieved perfect classification performance, reaching 100% accuracy, precision, recall, and F-score, highlighting its superior learning capability. The confusion matrix and ROC curve analyses further confirmed the robustness and reliability of the TGAM model in separating all acoustic event classes without misclassification. The integration of a user-friendly Tkinter-based interface with secure authentication allowed seamless dataset management, training, evaluation, and real-time prediction. The hybrid convolutional–ensemble framework proved highly effective for urban traffic and crime sound recognition. The system can serve as a valuable tool for intelligent surveillance, emergency response, and smart city applications. Future work may focus on expanding the dataset, deploying the system in real-time environments, and incorporating additional deep learning architectures to further enhance performance.

REFERENCES

[1] Patel, S., & Patyrykin, K. (2025). Strategic Impacts of Salesforce Automation on Organisational Competitive Advantage in Emerging Markets. Journal of Posthumanism, 5(12), 357–372. https://doi.org/10.63332/joph.v5i12.3782

[2] Santthosh Saai Reddy Purmani. (2026). Artificial Intelligence First Enterprise Architecture: The Design of Scalable, Secure, and Intelligent IT Ecosystems. American Journal of AI Cyber Computing Management, 6(1(2)), 1–8. https://doi.org/10.64751/ajaccm.2026.v6.n1(2).pp1-8

[3] Vasagam, M., Kumar, A., & Garg, A. (2026). Learning Execution Plan Embeddings for Multi-Dimensional Query Resource Prediction. IEEE Access.

[4] Kalae, U. K. (2021). Enhancing data analytics and reporting efficiency using Power BI and SQL in cloud computing environments. Journal of Computational Analysis and Applications, 29(6), 2021. https://doi.org/10.48047/jocaaa.2021.29.06.48

[5] Poojari, R. Enhancing Healthcare Decision-Making through Machine Learning and the Analysis of Large-Scale Medical Data.

[6] Reddy, S. K. R. Developing a Modular AI Framework to Enhance Scalability and Personalization in Next-Generation Reward Platforms.

[7] Prodduturi, S. M. K. To Secure Your Paper as Per UGC Guidelines We Are Providing A ElectronicBar code.

[8] Gaddam, S. From Fixed Specifications to Self-Adapting Systems: A Machine Learning Perspective on Software Engineering.

[9] Explainable AI Framework for Policy-Compliant Anomaly Detection in Data Pipelines. (2025). Inaternational Journal of Communication Networks and Information Security, 16(4). https://doi.org/10.48047/ijcnis.16.4.2111

[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.