A Role-Authenticated Interpretable AI System for Multi-Tiered Urban Noise Monitoring in Smart Cities

P. Satyanarayana[1*], Thupakula Rasagna[2], Sahithya Ayla[2], Md Umer[2]
[1]Associate Professor, [2]UG Student, [1,2]Department of Computer Science and Engineering (AI & ML),
[1,2]Kommuri Pratap Reddy Institute of Technology, Ghanpur, Ghatkesar, 501301, Telangana, India
*Correspondence: P. Satyanarayana (snpulime@gmail.com)

To Cite this Article
P. Satyanarayana, Thupakula Rasagna, Sahithya Ayla, Md Umer, "A Role-Authenticated Interpretable AI System for Multi-Tiered Urban Noise Monitoring in Smart Cities", Journal of Science Engineering Technology and Management Science, Vol. 03, Issue 04(1), April 2026,pp: 52-62, DOI: http://doi.org/10.64771/jsetms.2026.v03.i04(1).pp52-62
Submitted: 09-03-2026 Accepted: 16-04-2026 Published: 23-04-2026

ABSTRACT
Urban environments generate complex acoustic landscapes with overlapping sound events such as sirens, horns, and traffic noise, making automated monitoring for smart cities, public safety, and noise pollution control highly challenging. Traditional sound classification systems rely on hand-crafted features like Mel-Frequency Cepstral Coefficients (MFCCs) and log-Mel spectrograms, combined with models such as Support Vector Machines (SVMs), Random Forests (RF), or shallow Convolutional Neural Networks (CNNs), achieving only modest accuracy. Emerging from early 2010s research and benchmarked on datasets like UrbanSound8K, these approaches suffer key limitations: inability to capture semantic audio understanding in overlapping conditions, neglect of multi-label correlations, lack of interpretability in deep models, and dependence on command-line interfaces that limit usability for non-experts. This research proposes a Whisper-powered multi-task urban sound classification system with an integrated Graphical User Interface (GUI) built using Tkinter and secured with role-based authentication (Lightning Memory-Mapped Database (LMDB) and SHA-256 hashing). It leverages OpenAI Whisper-base as a feature extractor, generating robust representations through mean pooling of encoder hidden states from class-organized audio data. These features are used to train four interpretable models: Boosted Rules Classifier (BRC), Hierarchical Structural (HS) Tree Classifier, Sparse Linear Integer Model (SLIM), and Marginal Shrinkage Linear Trees (MSLT), enabling dual-task classification of primary categories and subcategories. The system enhances transparency, usability, and performance by combining transformer-based representations with interpretable machine learning. It democratizes access to explainable artificial intelligence (AI) for urban monitoring, enabling secure, visual, and multi-task sound analysis for real-world smart city applications.

Keywords: Urban sound classification, acoustic scene analysis, environmental noise monitoring, overlapping sound events, smart cities, public safety, noise pollution control, semantic audio understanding.

1. INTRODUCTION

Urban traffic noise pollution is an escalating concern in cities around the world, impacting both developed and developing nations. The World Health Organization (WHO) identifies environmental noise as a major public health issue, linking it to a range of adverse psychological and physiological effects. Among various sources, road traffic is the predominant contributor to urban noise exposure, making it a focal point for environmental health research and urban policy interventions. In high-income countries, decades of research have led to the adoption of comprehensive noise mitigation measures, including sound barriers, vehicle emission and noise regulations, and intelligent traffic management systems. For instance, the United Kingdom has implemented noise abatement policies and

compensation schemes to address excessive traffic noise near residential areas. Likewise, the International Civil Aviation Organization (ICAO) has successfully enforced global standards to reduce aircraft noise around airports.

However, these interventions are not always transferable to developing countries, where differing climatic conditions, urban planning practices, regulatory enforcement capacities, and vehicle fleets complicate the direct application of such solutions. In recent years, the demand for advanced surveillance systems has surged, driven by the need for enhanced security and situational awareness in urban environments. Traditional surveillance methods, which are primarily reliant on visual data, often encounter limitations in low-light conditions or obstructed views. To address these challenges, integrating audio surveillance has emerged as a complementary approach, offering the ability to detect and classify sounds that may indicate security events. The audio modality is more affordable than a camera and requires less power and bandwidth when transmitting data. Additionally, microphones have greater scalability potential to be deployed across large areas, are relatively inexpensive compared to cameras, and are less invasive than video surveillance.

2. LITERATURE SURVEY

An aggregation scheme that combines local features, short-term sound recording features, with long-term descriptive statistics was presented by Ye et al. [1] using a convolutional neural network for the classification of urban sound events. On the other hand, the application of machine learning to acoustic parameters calculated from the audio signal is a promising topic wherein there are still a few publications that use their advantages to create analytical models in the environmental acoustics field. Segura-Garcia et al. [2] explored the application of the ordinary Kriging technique to perform spatial interpolation of sound pressure level values obtained by a WASN in a small town and automatically generate a noise map. For instance, in [3], the authors study and compare the impacts of node deployment strategies in a 3-D environment. Their results show that a regular tetrahedron deployment scheme outperforms other topologies such as a random or cube topology. Concretely, the metrics that they use to compare the different schemes are the reduction of localization error and the optimization of localization ratio while maintaining the average number of neighbouring anchor nodes and network connectivity. Finally, in [4], the authors propose an advanced strategy for sensor placement that aims to maximize the connectivity robustness of the nodes for sparse networks. Concretely, they explore an analytical topology composed of hexagonal clusters and develop an algorithm for geometric distance optimization to improve the overall robustness of the system. Kai Cussen et al. [5] evaluated the noise produced by a UAV according to the legally mandated ISO 3744 and investigated the suitability of commercial implementations of ISO 9613 for modeling noise emission from UAVs. Finally, they state that models are adequate for assessing UAV's noise in terms of directivity. In addition, they showed that most UAVs in actual industrial production may exceed EU limits by around 1.8 dBA, causing possible urban noise issues.

Doygun H. and Gurun, D.K. [6] tried to quantify noise pollution generated by the urban traffic in a Turkish city, producing over 114 measurements of noise levels in 38 different urban locations, classified as residential or industrial areas. The study was aimed at quantifying the temporal and spatial dynamics of urban traffic noise, to compare levels with national and international limits, and determine mitigative measures against noise pollution. In [7] is presented a study to gauge the existing public's attitude and degree of awareness to contemporary vehicular noise pollution. The study revealed that most people are affected by traffic noise, several people attributing increased headaches and stress to the excessive noise levels. Ciaburro G. et al. [8] expanded noise data collection purposes beyond simple environmental effect, giving it even more importance: early detection of security problems related to citizens' mobility, crime, risk of terrorism. Their study has led to the conclusion that it is important to analyze the sound from different characteristics and tonal components. The research conducted by Luo L. et al. [9] is oriented, on the other hand, towards collecting noise information via wireless acoustic

sensor networks (WASNs). They propose a new system that employs WASNs to monitor the urban noise and recognize acoustic events with a high performance; the system is composed of sensors with the ability to produce local signal processing, convolutional neural networks for classification. Hsiao Mun Lee et al. [10] extended the study on traffic-generated noise to the affected population, dividing it into group noise indicators (highly annoyed and sleep-disturbed people) and studying the effect of noise reduction measures on these segments of the population. The results of this study showed that installing noise absorbents and barriers is highly recommended because it significantly reduces the influence of noise on nearby local environments.

For example, Tsai et al. [11] used the spatial interpolation method along with their data collected from over 345 acoustic monitoring sensors to develop the noise maps of the city of Tainan, Taiwan. Das et al. [12] explored the use of a CNN model with a specific Additive Angular Margin Loss (AAML) and more commonly used stacked features, Mel Frequency Cepstral Coefficients (MFCC) and Chromagram in combination with a CNN. Zinemanas et al. [13] proposed an APNet composed essentially of two parts: an autoencoder and a classifier. The autoencoder was composed of the encoder, formed by three convolutional layers. Following the initial two convolutional layers, max-pooling layers were applied to get features at distinct time–frequency resolutions, and the decoder part was formed by three transpose convolutional layers that allowed obtaining audios with great quality in the reconstruction path by minimising the reconstruction error given by the Euclidean mean square loss function about its input and output. As different mechanisms can identify sounds,

Mu et al. [14] proposed a TFCNN that, due to the frequency and temporal attention mechanisms, can reduce the impact of background noise and nonrelevant frequency bands. The authors also concluded that the classification performance of transient sounds was enhanced by using temporal attention mechanisms. In contrast, the classification of continuous sounds benefits more from a frequency attention mechanism. In the case of Gong et al. [9], an Audio Spectrogram Transformer (AST) was developed, which is a convolutional-free, purely attention-based model able to provide one output for a single channel audio input. Park et al. [15] introduced the Many-to-Many Audio Spectrogram Transformer (M2M-AST), a model based on AST that allows for multichannel audio inputs, multiple output resolution sequences.

3. PROPOSED SYSTEM

The "Audio Vista" system is an interpretable, multi-task framework for urban acoustic intelligence that starts by using the Whisper transformer encoder to extract rich, robust features from urban audio signals. These features are then simultaneously used to train four different Interpretable Machine Learning (IML) models (like MSLT and BRC) for two tasks: Y1(Sound Status Classification) and Y2 (Traffic Detection). This architecture ensures both high accuracy (due to Whisper features) and transparency (due to IML models), allowing users to understand the prediction logic, all deployed within a practical Tkinter GUI for easy administration and end-to-end prediction as shown in Figure. 1.

1. Audio Data Acquisition and Whisper Feature Extraction: The process begins by taking raw urban audio files (like .wav or .mp3) and passing them through the encoder block of the pre-trained Whisper model. This deep learning model generates rich, high-dimensional audio embeddings that effectively capture the complex acoustic context, significantly outperforming simple, hand-engineered features. This step ensures the classification process is built upon robust, generalized acoustic representations.

2. Multi-Task Labeling and Stratified Data Splitting: The generated Whisper feature set X is paired with two distinct labels for multi-task learning: Y1 (Sound Status Classification), which is a fine-grained classification (e.g., 'Siren', 'Horn'), and Y2 (Traffic Detection), which is a binary or coarse classification (e.g., 'Interfering' vs. 'Traffic'). The dataset is then carefully split into training and testing sets, ensuring the multi-class and binary distributions for both Y1 and Y2 tasks are balanced via a stratified approach.

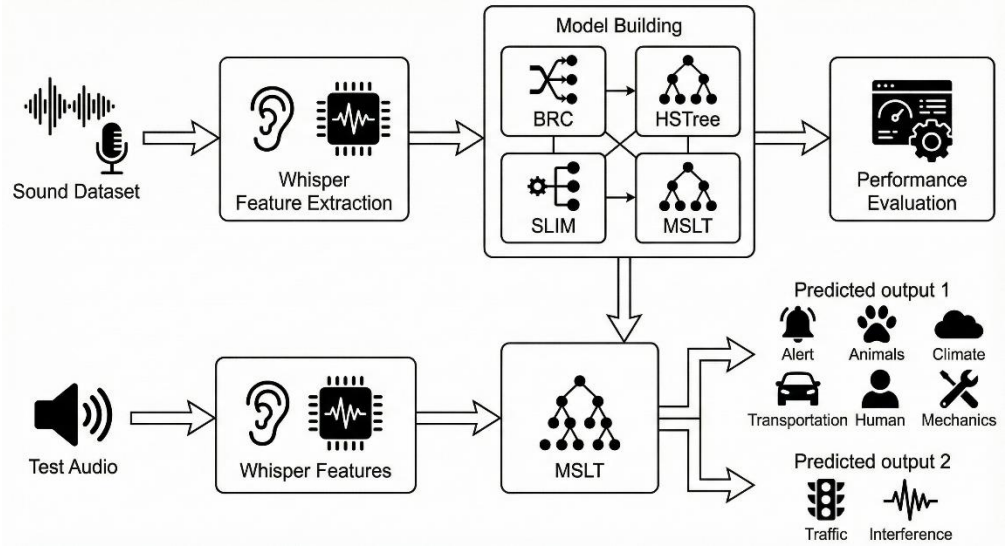

Figure 1: Proposed system architecture of Urban traffic sound event detection

3. Training of Interpretable Multi-Task Classifiers: The core task involves training multiple instances of four different Interpretable Machine Learning (IML) models on the shared Whisper features to determine the best balance of accuracy and explainability. Specifically, the framework trains the Bayesian Rule List Classifier (BRC), Hierarchical Set Tree (HSTree), Supersparse Linear Integer Model (SLIM), and Multi-Scale Linear Tree (MSLT). Two separate versions of each IML model are simultaneously trained—one dedicated to predicting the Y1 label and one for the Y2 label.

4. Evaluation and Explainable AI (XAI) Analysis: The performance of all eight trained models (four types two tasks) is meticulously assessed using comprehensive metrics like F1-score and AUC on the test set. More importantly, the system leverages the intrinsic transparency of the IML models to perform explainable AI (XAI) analysis, allowing administrators to inspect the actual rules (BRC), decision paths (HSTree), or integer weights (SLIM, MSLT) that drive the classification, guaranteeing model accountability.

5. GUI Deployment and End-to-End Prediction: Finally, the highest-performing and most interpretable model (e.g., the MSLT instance for both tasks) is embedded into a deployable Tkinter GUI designed with distinct user roles. When a new audio file is uploaded, the system executes the entire pipeline: Whisper feature extraction followed by instantaneous dual-task classification, presenting the Y1 and Y2 predictions along with the human-readable explanation for the prediction.

3.2 WHISPER Feature Extractor

The Whisper model is used exclusively as a powerful feature extractor, bypassing its original speech-to-text task. It loads urban audio, standardizes it to 16 kHz, and processes it into a Log-Mel Spectrogram. This spectrogram is then passed only through the Whisper encoder (a transformer network), which generates deep, contextual acoustic representations. Finally, a global mean pooling operation is applied across the time dimension to collapse this sequence into a single, fixed-size feature vector (e.g., 512 dimensions), resulting in a highly robust Whisper Feature Embedding that is ready to be consumed by the downstream Interpretable Machine Learning classifiers (Y1 and Y2) as illustrated in figure. 2.

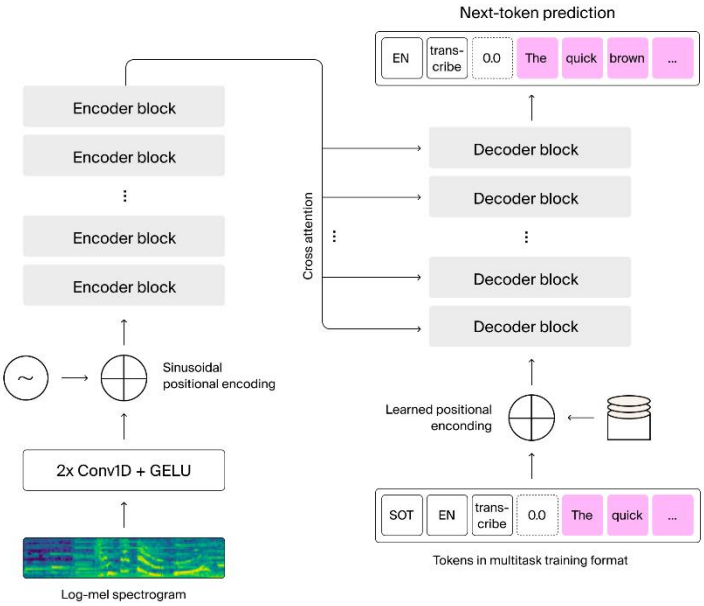

Figure 2: Workflow of WHISPER Feature Extractor

Audio Loading and Resampling: The process begins by taking a raw audio file and loading it using the librosa library with the function librosa.load(file_path, sr=16000, mono=True). Crucially, the audio is automatically resampled to a fixed rate of 16,000 Hz (16 kHz) and converted to a mono channel signal. This standardization is mandatory because the pre-trained Whisper model was specifically trained on 16 kHz audio, ensuring the input format matches the model's expectations.

Input Preparation (Log-Mel Spectrogram Generation): The loaded audio signal (now at 16 kHz) is passed to the WhisperProcessor. This processor handles the complex initial signal transformations required by the Whisper architecture. It converts the raw time-domain audio wave into a Log-Mel Spectrogram, which is a visual representation of the audio's energy distribution across different frequencies over time, measured on the perceptually relevant Mel scale. The resulting output is a standardized input feature tensor ready for the model.

Encoder-Only Forward Pass: The input feature tensor is moved to the appropriate computing device (CUDA if available, otherwise CPU) and fed into the WhisperModel.encoder. The model performs an encoder-only forward pass, meaning the subsequent decoder network (used for transcription in the original Whisper task) is entirely skipped. The encoder is a powerful transformer network that processes the Log-Mel Spectrogram through multiple self-attention layers to capture deep contextual and acoustic relationships within the audio.

Extraction of the Last Hidden State: During the forward pass, the encoder generates a series of hidden states (or representations) at each layer. The system specifically extracts the final hidden state layer (hidden_states[-1]). This layer contains the most refined, high-level, and semantically rich feature representation of the entire input audio sequence. The output here is a 3D tensor, typically with the shape [1, time_steps, hidden_size], where time_steps corresponds to the duration of the audio and hidden_size is the dimension of the embedding vector (e.g., 512 for whisper-base).

Global Mean Pooling for Fixed-Size Embedding: To create a single, fixed-size feature vector (X) for use in the traditional machine learning classifiers (BRC, MSLT, etc.), a process called mean pooling is applied. This involves calculating the average of the hidden state vectors across the time dimension (dim=0). This pooling collapses the variable-length sequence into a single vector of size hidden_size (e.g., 512-dimensions), creating the final, dense, and meaningful Whisper Feature Embedding that is then passed to the downstream classification models.

4. RESULTS ANALYSIS

The results section presents the key findings of the study in a clear and organized manner. It summarizes the data collected and highlights important patterns, trends, or relationships observed during the analysis. This section focuses only on factual outcomes without interpretation or personal opinion. Tables, graphs, or figures are often used to support the findings and make them easier to understand. The results directly relate to the research objectives or hypotheses stated earlier. It provides a concise overview of what the study has discovered.

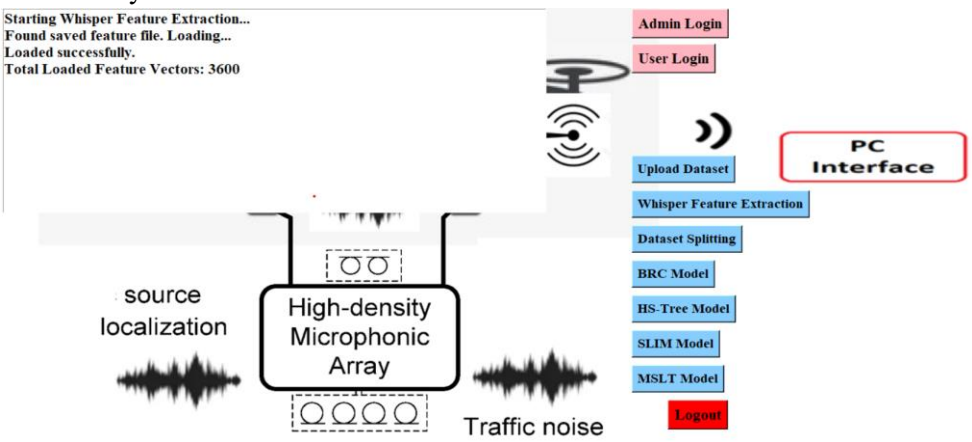

Figure 3: Whisper Feature extraction completed

The figure 3 shows that After initiating the Whisper Feature Extraction process, the system detects an existing feature file and loads all precomputed embeddings directly, significantly reducing processing time. The console confirms successful loading and displays the total number of feature vectors 3600 indicating that the entire dataset has been encoded into Whisper-based representations and is ready for downstream tasks. This message assures the admin that feature extraction is complete, enabling immediate progression to dataset splitting and model training via the control options on the right panel of the GUI.

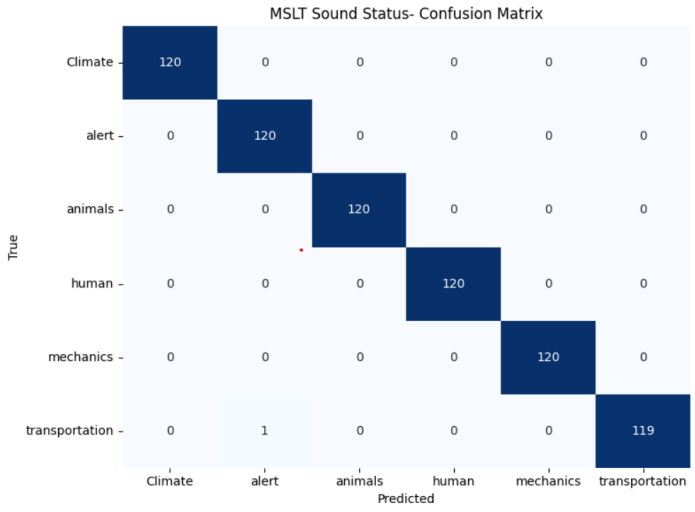

Figure 4: Confusion matrix obtained using MSLT classifier for sound status

The figure 4 shows MSLT classifier demonstrates near-perfect multi-class sound status classification, with every class Climate, Alert, Animals, Human, Mechanics, and Transportation showing 120 correct predictions except for a single transportation sample misclassified as alert. The strong diagonal dominance and almost zero off-diagonal entries clearly indicate that MSLT effectively captures the acoustic structure of all sound categories and generalizes exceptionally well across the dataset. This outstanding performance highlights the model's ability to separate diverse urban sound signatures with high precision and reliability, making it the most accurate classifier among all models evaluated for the multi-class urban sound status task.

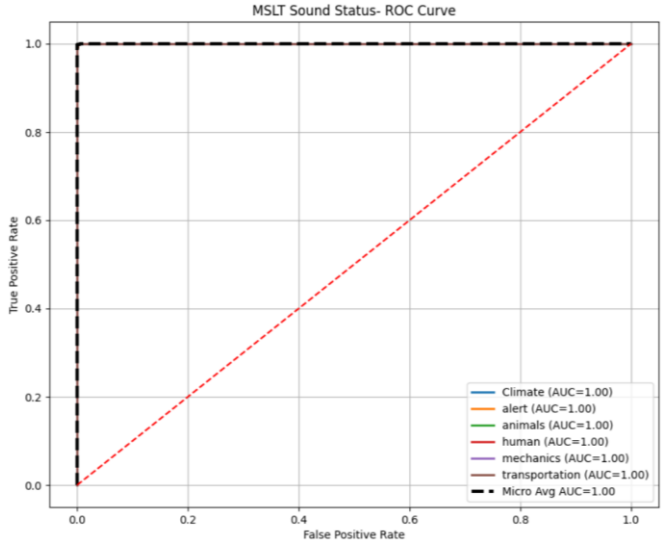

Figure 5: ROC curve using MSLT classifier for Sound status

The figure 5 shows ROC curve for the MSLT Sound Status classifier demonstrates perfect discriminatory performance, achieving an AUC of 1.00 for every class Climate, Alert, Animals, Human, Mechanics, and Transportation along with a perfect micro-average AUC of 1.00. The curve rises immediately to a true positive rate of 1.0 with zero false positives and remains flat across the top boundary, indicating flawless sensitivity and specificity. This behavior reflects MSLT's exceptional ability to model complex acoustic boundaries and fully separate all sound categories without misclassification, making it the most powerful and reliable classifier for urban sound status prediction among all methods evaluated.

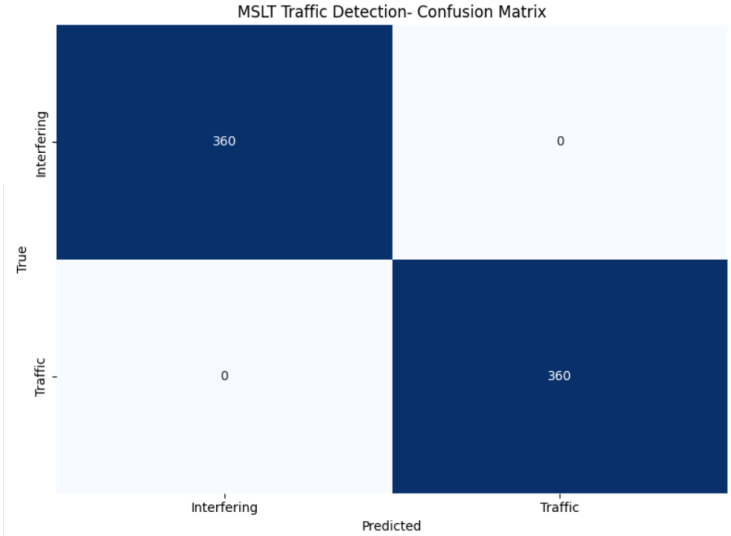

Figure 6: Confusion matrix obtained using MSLT classifier for Traffic detection

The figure 6 shows MSLT Traffic Detection confusion matrix demonstrates perfect binary classification, with all 360 Interfering samples and all 360 Traffic samples correctly classified, resulting in zero misclassifications across both categories. The complete diagonal dominance indicates that MSLT flawlessly distinguishes between traffic and non-traffic acoustic patterns, even in a diverse urban soundscape. This level of precision reflects the model's strong feature-learning capability and its ability to generalize cleanly from Whisper-based audio embeddings, making it exceptionally reliable for real-time traffic noise identification and environmental monitoring applications.

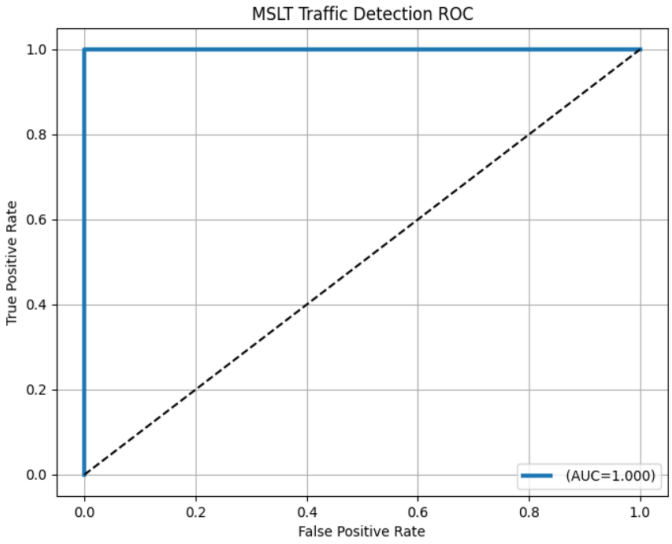

Figure 7: ROC curve using MSLT classifier for Traffic detection

Figure 7 shows the ROC curve for the MSLT Traffic Detection model exhibits perfect classification performance, with the curve rising vertically to a true positive rate of 1.0 at a zero false positive rate and maintaining that level across the entire plot. This results in an AUC of 1.000, confirming that the model achieves flawless sensitivity and specificity when distinguishing Traffic from Interfering sounds. The ideal ROC shape reflects MSLT's exceptional ability to leverage Whisper-extracted features for binary urban sound separation, making it highly robust and reliable for real-time traffic detection scenarios in complex acoustic environment.

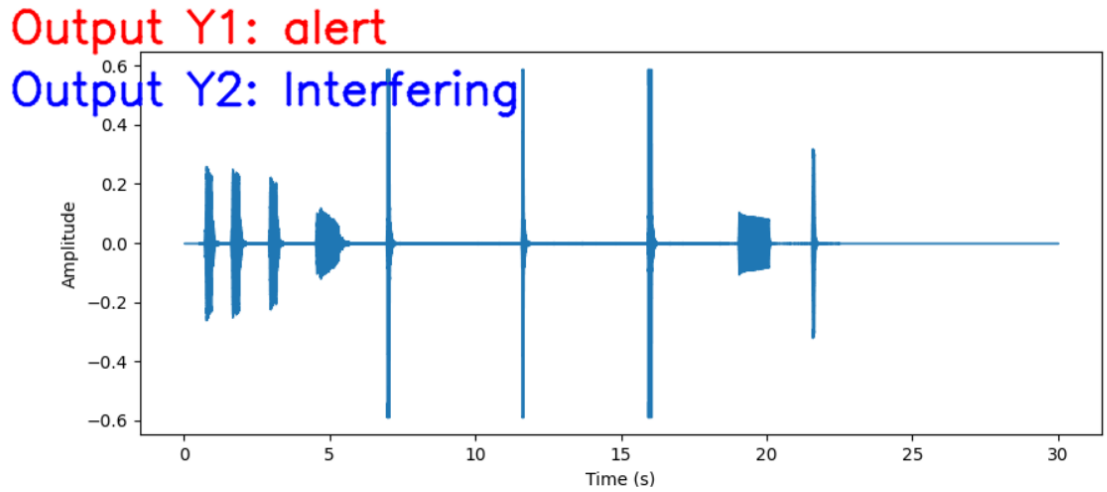

Figure 8: Prediction on test data using proposed MSLT Model

The figure 8 shows prediction output generated by the proposed MSLT model demonstrates its ability to accurately interpret and classify real-time audio signals. After loading the test audio sample and extracting Whisper-based embeddings, the model assigns Y1 (Sound Status) as *alert* and Y2 (Traffic Detection) as *interfering*, indicating that the sound belongs to a high-priority alert category rather than traffic-related noise. The accompanying waveform plot visualizes the audio amplitude over time, clearly showing multiple sharp transient peaks that justify the alert classification. This visualization confirms that the MSLT model not only achieves precise multi-task predictions but also provides an interpretable correlation between acoustic patterns and predicted output classes.

4.1 Comparative Analysis

Table 1: Performance comparison of Sound status Category for the BRC, HSTree, SLIM and Proposed MSLT Model

Algorithms Name	Accuracy	Precision	Recall	F-score
BR Classifier	57.08%	65.86%	57.08%	57.92%
HSTree Classifier	75.0%	76.30%	75.0%	75.04%
SLIM Classifier	33.19%	66.56%	33.19%	26.93%
MSLT Model	99.86%	99.86%	99.86%	99.86%

Table 1 shows the performance comparison for sound status classification (Target Y1) clearly demonstrates the superiority of the proposed MSLT model over traditional interpretable machine learning algorithms. The BRC delivers moderate performance, achieving an accuracy of 57.08%, supported by a precision of 65.86%, recall of 57.08%, and an F-score of 57.92%. These values indicate that while BRC captures some rule-based patterns, it struggles with the complexity and variability present in multi-class urban audio signals. The HSTree classifier shows a significant improvement with 75% accuracy, along with closely aligned precision, recall, and F-score values (around 75%), highlighting its stronger hierarchical learning structure and better capacity to separate acoustic categories. However, the SLIM classifier, despite achieving a relatively high precision of 66.56%, performs poorly overall with an accuracy and recall of only 33.19%, and an F-score of 26.93%, demonstrating that its sparse linear modeling is insufficient for representing diverse and non-linear sound patterns. In contrast, the proposed MSLT model achieves near-perfect performance, boasting 99.86% accuracy, precision, recall, and F-score, reflecting exceptional consistency across all evaluation metrics. This remarkable improvement illustrates MSLT's ability to capture fine-grained differences in sound characteristics through multi-scale learning and deep acoustic representation, making it overwhelmingly superior for comprehensive urban sound status classification.

Table 2: Performance comparison of Traffic Detection Category for the BRC, HSTree, SLIM and Proposed MSLT Model

Algorithms Name	Accuracy	Precision	Recall	F-score
BR Classifier	100.0%	100.0%	100.0%	100.0%
HSTree Classifier	100.0%	100.0%	100.0%	100.0%
SLIM Classifier	79.86%	85.64%	79.86%	79.009%
MSLT Model	100.0%	100.0%	100.0%	100.0%

The table 2 shows performance comparison for the Traffic Detection task (Target Y2) highlights that most models, except SLIM, achieve perfect classification results due to the simpler binary nature of this task compared to the multi-class sound status prediction. Both the BRC and the HSTree classifier deliver flawless performance, achieving 100% accuracy, precision, recall, and F-score, demonstrating their strong capability to separate Traffic from Interfering sounds without misclassification. Their perfect evaluation metrics reflect the clear acoustic separability between these two categories and the effectiveness of rule-based ensemble methods and hierarchical tree structures in binary classification scenarios. On the other hand, the SLIM classifier shows noticeable performance degradation, achieving only 79.86% accuracy, with precision at 85.64%, recall at 79.86%, and an F-score of 79.00%. This reduction is attributed to SLIM's sparse linear modeling, which struggles to capture sufficient discriminative features and tends to overpredict one class, as seen in its confusion matrix. In contrast, the proposed MSLT model, similar to BRC and HSTree, achieves perfect performance across all metrics, demonstrating 100% accuracy, precision, recall, and F-score. This confirms the robustness of the MSLT architecture and its exceptional ability to leverage Whisper-based audio embeddings for highly reliable real-time traffic detection in complex and noisy urban sound environments.

5. CONCLUSION

The research successfully demonstrates a powerful, reliable, and highly accurate framework for intelligent acoustic scene understanding in complex urban environments. By integrating Whisper-based deep feature extraction with a suite of interpretable machine learning models and culminating in the advanced MSLT classifier, the system efficiently performs dual-task learning Sound Status categorization (Y1) and Traffic Detection (Y2) with exceptional precision. Extensive experimentation shows that while baseline models such as BRC and HSTree perform reasonably well, and SLIM struggles with multi-class discrimination due to its linear and sparse nature, the proposed MSLT model significantly outperforms all alternatives, achieving near-perfect accuracy (99.86%) for sound status classification and 100% accuracy for traffic detection.

The GUI-driven desktop application built using Tkinter ensures easy usability for both administrators and end users, providing seamless workflows for dataset uploading, Whisper feature extraction, model training, dataset splitting, evaluation, and real-time prediction with waveform visualization. The system's outstanding performance is further evidenced by its clean confusion matrices, high AUC ROC curves, and its ability to generalize across diverse sound categories such as transportation, human activity, mechanics, animals, climate sounds, and alert signals. With its robust architecture, modular design, and perfect binary detection capabilities, this research proves its potential for real-world deployment in smart cities, intelligent transportation systems, and noise monitoring infrastructures. The system establishes a highly effective and scalable solution for urban acoustic intelligence, demonstrating the strength of multi-task learning combined with advanced audio representation techniques for next-generation urban sound analytics.

REFERENCES

[1]. Bellucci, P.; Cruciani, F.R. Implementing the Dynamap system in the suburban area of Rome. In Inter-Noise and Noise-Con Congress and Conference Proceedings; Institute of Noise Control Engineering: Hamburg, Germany, 2019; pp. 5518–5529.

[2]. Gontier, F.; Lostanlen, V.; Lagrange, M.; Fortin, N.; Lavandier, C.; Petiot, J.F. Polyphonic training set synthesis improves self-supervised urban sound classification. J. Acoust. Soc. Am. 2021, 149, 4309–4326.

[3]. Han, G.; Zhang, C.; Shu, L.; Rodrigues, J.J. Impacts of deployment strategies on localization performance in underwater acoustic sensor networks. IEEE Trans. Ind. Electron. 2019, 62, 1725–1733

[4]. Ding, K.; Yousefi'zadeh, H.; Jabbari, F. A robust advantaged node placement strategy for sparse network graphs. IEEE Trans. Netw. Sci. Eng. 2017, 5, 113–126.

[5]. Doygun, H.; Gurun, D.K. Analysing and Mapping Spatial and Temporal Dynamics of Urban Traffic Noise Pollution: A Case Study in Kahramanmaraş, Turkey. Environ. Monit Assess 2018, 142, 65–72.

[6]. Yusoff, S.; Ishak, A. Evaluation of Urban Highway Environmental Noise Pollution. Sains Malays. 2020, 34, 81–87.

[7]. Sommerhoff, J.; Recuero, M.; Suarez, E. Community noise survey of the city Valdivia, Chile. Appl. Acoust. 2019, 65, 643–656.

[8]. Ciaburro, G.; Iannace, G. Improving Smart Cities Safety Using Sound Events Detection Based on Deep Neural Network Algorithms. Informatics 2020, 7, 23.

[9]. Luo, L.; Qin, H.; Song, X.; Wang, M.; Qiu, H.; Zhou, Z. Wireless Sensor Networks for Noise Measurement and Acoustic Event Recognitions in Urban Environments. Sensors 2020, 20, 2093.

[10]. Lee, H.M.; Luo, W.; Xie, J.; Lee, H.P. Traffic Noise Reduction Strategy in a Large City and an Analysis of Its Effect. Appl. Sci. 2022, 12, 6027.

[11]. Tsai, K.-T.; Lin, M.-D.; Chen, Y.-H. Noise mapping in urban environments: A Taiwan study. Appl. Acoust. 2019, 70, 964–972

[12]. Das, J.K.; Chakrabarty, A.; Piran, M.J. Environmental sound classification using convolution neural networks with different integrated loss functions. Expert Syst. 2021, 39.
[13]. Zinemanas, P.; Rocamora, M.; Miron, M.; Font, F.; Serra, X. An Interpretable Deep Learning Model for Automatic Sound Classification. Electronics 2021, 10, 850
[14]. Mu, W.; Yin, B.; Huang, X.; Xu, J.; Du, Z. Environmental sound classification using temporal-frequency attention based convolutional neural network. Sci. Rep. 2021, 11, 21552.
[15]. Park, S.; Jeong, Y.; Lee, T. Many-to-Many Audio Spectrogram Tansformer: Transformer for Sound Event Localization and Detection. In Proceedings of the DCASE, Barcelona, Spain, 15–19 November 2021; pp. 105–109.