International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

# Audio Vista: Whisper-Embedded Interpretable Multi-Task Framework for Urban Acoustic Intelligence with GUI Deployment

Akula Varsha[1], R. Swathi[2], Rekha Gangula[3*], K Uday Kiran[1], Kurapati Rahul[1], Pakanati Shivakumari[1]

[1]UG Student, [2]Assistant Professor, [3]Associate Professor and Head, [1,2,3]Department of Computer Science and Engineering (AI&ML)

[1,2,3]Vaagdevi Engineering College, Bollikunta, Warangal, 506005, Telangana, India

*Correspondence: Rekha Gangula (gangularekha@gmail.com)

## ABSTRACT

Urban environments produce complex acoustic landscapes with overlapping sound events like sirens, horns, and traffic noise, complicating automated monitoring for smart cities, public safety, and noise pollution control. Traditional sound classification systems rely on hand-crafted features such as MFCCs and log-Mel spectrograms fed into SVMs, Random Forests, or shallow CNNs, achieving modest accuracies. These approaches emerged from early 2010s research addressing environmental audio challenges, evolving through competitions that established datasets like UrbanSound8K as standards. However, traditional systems face critical limitations: hand-engineered features fail to capture semantic audio understanding amid co-occurring sounds and abnormal noise conditions; single-task models ignore label correlations; black-box deep networks lack interpretability for regulatory use; and command-line interfaces exclude non-experts, hindering real-world deployment. No integrated GUI exists for multi-task urban sound classification combining transformer representations with interpretable models, creating a gap in accessible, transparent AI for urban monitoring. This research addresses these needs through a Whisper-powered multi-task urban sound classifier GUI, an end-to-end Tkinter application with role-based authentication (LMDB + SHA-256). It leverages OpenAI's Whisper-base encoder for state-of-the-art feature extraction like mean-pooling hidden states from audio files organized by class folders yielding fixed-length vectors. These features train four interpretable classifiers such as Boosted Rules Classifier (BRC), Hierarchical Structural (HS) Tree Classifier, Sparse Linear Integer Model (SLIM) Classifier, Marginal Shrinkage Linear Trees (MSLT) for dual tasks like primary sound categories and internal subcategories. The system's significance lies in democratizing interpretable AI: Whisper provides superior representations without fine-tuning, interpretable models ensure trust via rule-based decisions, and the GUI enables non-technical deployment. It advances smart city applications by enabling secure, visual multi-task classification of urban sounds, bridging transformer power with human-understandable analytics for environmental intelligence.

**Key words:** UrbanSound, Role based authentication, Whisper-base encoder, Rotor noise compensation, Multi-task urban sound classifier

## 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.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

Urban environments generate complex, overlapping sound events that require simultaneous classification of both primary sound categories (sirens, horns, voices) and contextual attributes (traffic-related vs interfering noise). Existing audio classification systems struggle with multi-task learning, lack interpretability for real-world deployment, and demand command-line expertise, creating barriers for non-technical users in smart city monitoring and noise pollution management

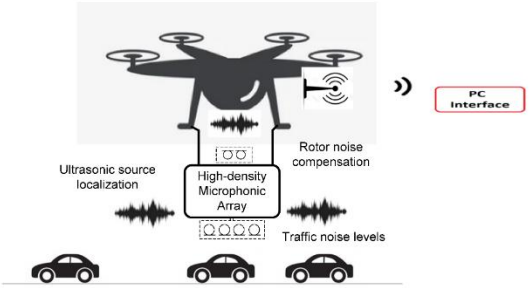


Figure 1. Urban traffic noise analysis

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.

Rapid urbanization increases urban noise pollution, necessitating automated sound event detection for traffic management, public safety, and environmental monitoring. While deep learning transformers like Whisper excel at audio representation learning, their black-box nature limits trust in critical applications. Interpretable ML models integrated with transformer features provide transparent decision-making while leveraging state-of-the-art representations, but no accessible GUI framework exists for end-to-end multi-task urban sound classification workflows.

## 2. LITERATURE SURVEY

### 2.1 Machine Learning and Deep Learning for Acoustic Event Classification

Machine learning and deep learning techniques have been widely applied for urban sound classification and acoustic event recognition. Ye et al. [1] proposed a CNN-based framework combining local features, short-term recordings, and long-term statistical descriptors for urban sound classification. Das et al. [12] utilized CNN models with MFCC and chromagram features enhanced by Additive Angular Margin Loss (AAML), improving classification performance. Zinemanas et al. [13] introduced an APNet architecture integrating an autoencoder and classifier for high-quality audio reconstruction and classification. Furthermore, Mu et al. [14]

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

proposed a Temporal-Frequency CNN (TFCNN) incorporating attention mechanisms to improve classification of transient and continuous sounds. Advanced transformer-based models such as Gong et al. [9] and Park et al. [15] further enhanced audio classification by leveraging attention mechanisms for single-channel and multi-channel audio inputs.

### 2.2 Wireless Acoustic Sensor Networks and Smart Monitoring

Wireless Acoustic Sensor Networks (WASNs) have been increasingly used for large-scale environmental noise monitoring. Segura-Garcia et al. [2] applied the ordinary Kriging technique to generate spatial noise maps from sensor data. Luo L. et al. [9] proposed a WASN-based system integrating local signal processing with CNN models for real-time acoustic event recognition. Additionally, studies such as Tsai et al. [11] utilized large-scale sensor deployments to develop urban noise maps, demonstrating the effectiveness of distributed acoustic sensing.

### 2.3 Sensor Deployment and Network Optimization Strategies

Efficient sensor deployment and network design play a crucial role in acoustic monitoring systems. 3D deployment topology study authors [3] analyzed various node deployment strategies in 3D environments, showing that tetrahedron-based configurations improve localization accuracy and network connectivity. sensor placement robustness study authors [4] proposed a hexagonal cluster-based sensor placement strategy to enhance connectivity and robustness in sparse networks.

### 2.4 Environmental Noise Analysis and Impact Assessment

Several studies focus on analyzing environmental noise and its effects on urban populations. Kai Cussen et al. [5] evaluated UAV noise emissions using ISO standards, highlighting potential urban noise concerns. Doygun H. and Gurun D.K. [6] conducted extensive measurements of traffic noise levels, analyzing spatial and temporal variations. public perception noise study authors [7] investigated public awareness of traffic noise, revealing its impact on stress and health. Similarly, Hsiao Mun Lee et al. [10] examined the effects of noise on populations and recommended mitigation strategies such as noise barriers.

### 2.5 Advanced Acoustic Analysis for Security and Urban Applications

Beyond environmental monitoring, acoustic analysis has been extended to security and smart city applications. Ciaburro G. et al. [8] emphasized the importance of analyzing sound characteristics for early detection of security threats such as crime and terrorism. Their work highlights the potential of acoustic data in enhancing urban safety and intelligent surveillance systems.

## 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. 2.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

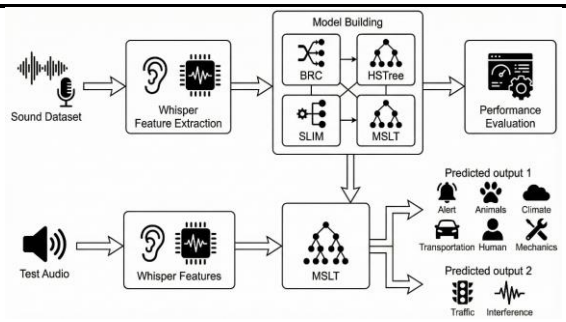


Figure 2. Proposed system architecture of Urban traffic sound event detection

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 and 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.

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 and 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. 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.

**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. 3.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

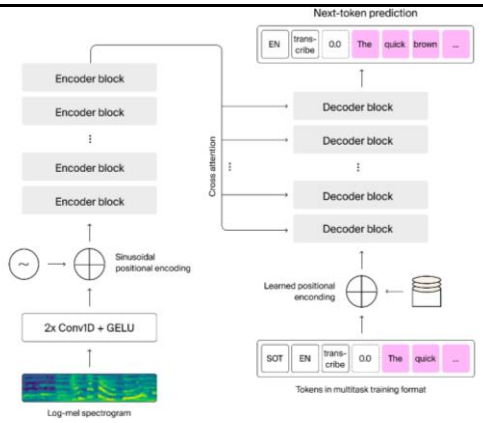


Figure 3. Workflow of WHISPER Feature Extractor

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 and 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.

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. During the forward pass, the encoder generates a series of hidden states (or representations) at each layer. The project 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). 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. Result Analysis

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

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

evaluated for the multi-class urban sound status task.

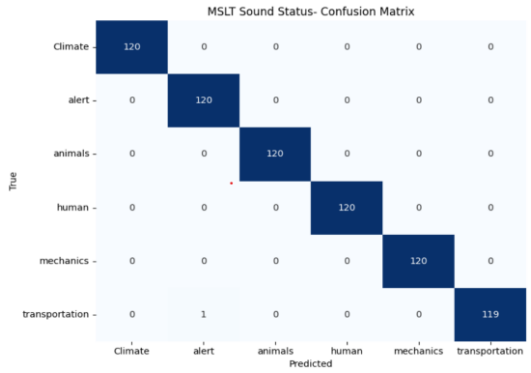


Figure 4. Confusion matrix obtained using MSLT classifier for sound status

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.

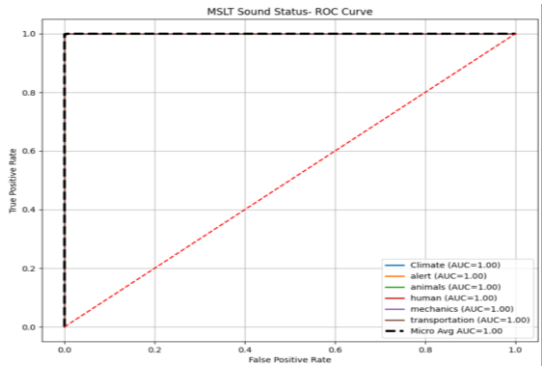


Figure 5. ROC curve using MSLT classifier for Sound status

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.

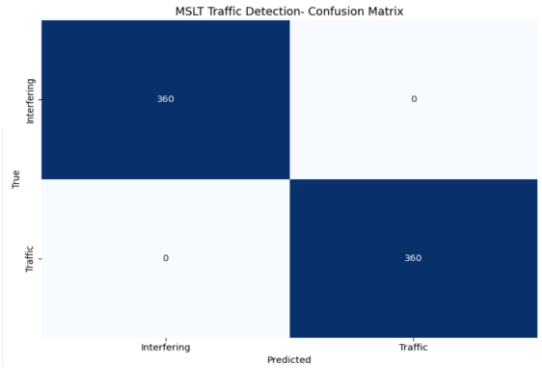


Figure 6. Confusion matrix obtained using MSLT classifier for Traffic detection

The Figure 7 shows 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.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

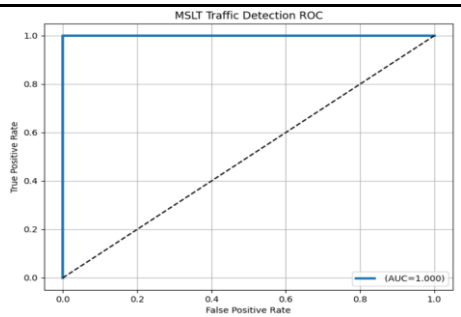


Figure 7. ROC curve using MSLT classifier for Traffic detection

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.

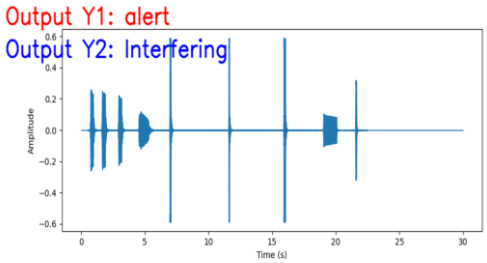


Figure 8. Prediction on test data using porposed MSLT Model

**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% |

The 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

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

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

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

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 project proves its potential for real-world deployment in smart cities, intelligent transportation systems, and noise monitoring infrastructures. The project 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

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

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.