JCTA Journal of Computing Theories and Applications ISSN:3024-9104

Research Article

A Multi-Branch BiLSTM with Multi-Head Self-Attention for Suspicious Sound Recognition

Shehu Mohammed Yusuf *, Hamza Saidu, and Sani Saleh Saminu

Department of Computer Engineering, Ahmadu Bello University, Samaru, Zaria, Kaduna 810211, Nigeria; e-mail : smyusuf@abu.edu.ng; hamzasaidu34@gmail.com; sssaleh@engineering.abu.edu.ng

* Corresponding Author : Shehu Mohammed Yusuf

Abstract: Suspicious urban sound recognition is a critical component of intelligent public safety and urban monitoring systems, enabling the automated identification of anomalous acoustic events such as gunshots, sirens, and other security-sensitive sounds. However, existing deep learning approaches often struggle to simultaneously capture long-range temporal dependencies and global contextual relationships, particularly under noisy and acoustically complex urban conditions. This limitation can reduce reliability in safety-critical scenarios where missed detections carry significant risk. To address these challenges, this study proposes a Multi-Branch Bidirectional Long Short-Term Memory (BiLSTM) framework with Multi-Head Self-Attention (MHSA) for enhanced sequential and contextual feature modeling. Mel-frequency cepstral coefficients (MFCCs) are extracted from a curated subset of the UrbanSound8K dataset, comprising five suspicious sound classes, and used as input to the proposed architecture. The multi-branch design enables complementary temporal representations, while the self-attention mechanism provides lightweight contextual weighting of BiLSTM outputs. Experimental results demonstrate that the proposed model achieves a test accuracy of 95.59%, outperforming conventional Dense and LSTM-based baseline models under identical experimental settings. An ablation study further confirms the contribution of multi-branch integration and attention-based enhancement to overall performance. Class-wise evaluation reveals consistently high recall across all sound categories, particularly for safety-critical classes such as gunshots and sirens. These findings indicate that the proposed framework provides robust and reliable performance, making it suitable for real-time smart city surveillance and public safety applications.

Keywords: Acoustic event detection; BiLSTM; Deep learning; Mel-frequency cepstral coefficients; Multi-head self-attention; Smart city surveillance; Sustainable urban monitoring; Urban sound classification.

Received: February, 3rd 2026
Revised: April, 28th 2026
Accepted: April, 30th 2026
Published: May, 12th 2026

1. Introduction

The rapid advancement of artificial intelligence (AI) and deep learning has significantly transformed intelligent surveillance systems by enabling automated interpretation of complex sensory data for public safety and urban security applications [1], [2]. In modern smart cities, the early detection of hazardous and suspicious activities plays a crucial role in crime prevention, emergency response, and infrastructure protection. Although video-based surveillance systems remain the dominant modality, they are often constrained by occlusion, poor illumination, adverse weather conditions, and increasing privacy concerns [3], [4]. These limitations have motivated growing interest in audio-based surveillance as a complementary sensing modality capable of detecting safety-critical events even when visual information is unreliable or unavailable [5], [6].

Urban sound events such as gunshots, emergency sirens, drilling, jackhammers, and aggressive animal sounds often serve as early acoustic indicators of potentially dangerous situations. Consequently, automatic urban sound classification and suspicious sound recognition have attracted substantial research attention in recent years [7], [8]. Prior studies indicate that sound-based approaches can provide robust situational awareness in noisy and visually

cluttered environments, particularly for public safety monitoring and smart city surveillance [3], [9]. However, real-world urban soundscapes are characterized by high background noise, overlapping sound sources, acoustic reverberation, and significant temporal variability, which pose considerable challenges for reliable and real-time sound event detection [6], [10].

Deep learning techniques have demonstrated strong performance in environmental and urban sound classification by learning discriminative representations from time–frequency features such as Mel-Frequency Cepstral Coefficients (MFCCs) and spectrograms [9], [11]. MFCC features are particularly well suited for this task because they compactly encode perceptually relevant spectral information, capturing formant structure and timbral characteristics essential for distinguishing diverse urban sound events [12], [13]. Early approaches commonly employed fully connected Dense neural networks due to their architectural simplicity; however, these models lack explicit temporal modeling and struggle to capture long-range dependencies inherent in audio signals [4], [14].

To address these limitations, recurrent neural networks (RNNs), particularly Long Short-Term Memory (LSTM) architectures, have been widely adopted for modeling temporal dependencies in sound sequences [4], [15]. Bidirectional LSTM (BiLSTM) further enhances sequential modeling by processing input in both forward and backward directions, providing richer contextual representations that are especially important for time-dependent audio classification tasks [12], [13]. While LSTM-based models improve performance by capturing temporal structure, they remain inherently sequential and may face scalability challenges when handling long audio sequences in noisy environments [9], [14].

In parallel, attention mechanisms have emerged as effective tools for modeling global contextual relationships through self-attention [6], [8]. Multi-head self-attention (MHSA) enables direct modeling of long-range dependencies and supports parallel computation, leading to strong performance across multiple domains [6], [15]. However, when applied independently, attention-based approaches may overlook fine-grained local temporal patterns that are critical for distinguishing acoustically similar or transient urban sound events [16], [17].

Beyond architectural considerations, urban sound datasets introduce additional challenges such as class imbalance, acoustic overlap, and background interference. Many studies emphasize overall classification accuracy while providing limited analysis of class-wise recall, despite the fact that false negatives—such as missed gunshot or siren detections—are particularly critical in public safety applications [18], [19]. Furthermore, high-performing deep learning models often rely on large parameter counts and significant computational resources, which can limit their applicability in real-time and resource-constrained deployment scenarios typical of smart city infrastructures [5], [20].

Recent research trends suggest that hybrid architectures combining recurrent modeling with attention-based mechanisms offer a promising direction for addressing these challenges [16], [17], [21]. By integrating temporal modeling via BiLSTM with contextual weighting through self-attention, such approaches can capture both short-duration impulsive sounds (e.g., gunshots) and long-duration sustained events (e.g., sirens and drilling). However, systematic investigations of multi-branch recurrent–attention architectures for suspicious urban sound recognition, particularly with comprehensive ablation analysis, remain limited [8], [22], [23].

A related recent study, MSA-TCN [24], introduces a multi-scale temporal convolutional network with dual attention mechanisms for robust suspicious sound detection. While MSA-TCN effectively leverages dilated convolutions and attention mechanisms, it relies on convolutional temporal modeling rather than bidirectional recurrent processing, which may limit its ability to capture asymmetric temporal context. In addition, it does not employ a parallel multi-branch architecture with differentiated representational capacities. The present work addresses these limitations by combining bidirectional temporal modeling with multi-branch encoding and lightweight self-attention enhancement.

Despite recent progress, several research gaps remain. First, few frameworks jointly integrate bidirectional temporal modeling and self-attention within a unified architecture tailored for noisy urban environments [17], [21]. Second, comparative evaluations against standard Dense and LSTM baselines under consistent experimental settings are still limited [4], [6]. Third, ablation studies that quantify the individual contributions of architectural components, particularly multi-branch design and attention mechanisms, are often insufficiently explored [7], [9]. Finally, claims of robustness are frequently not supported by detailed analysis of generalization performance [20], [25].

To address these gaps, this study investigates the following research questions. It examines whether a multi-branch BiLSTM architecture augmented with self-attention can outperform conventional Dense and LSTM-based models, how bidirectional temporal modeling contributes to detecting both impulsive and sustained sound events, and how self-attention enhances contextual discrimination under noisy conditions. In addition, the study evaluates the extent to which multi-branch integration improves performance compared to single-branch configurations.

To this end, this paper proposes a Multi-Branch BiLSTM with MHSA for suspicious sound recognition. The framework integrates parallel BiLSTM branches with different representational capacities, followed by self-attention blocks that enhance contextual weighting of temporal features. MFCC features extracted from the UrbanSound8K dataset are used as input, focusing on five security-relevant sound classes: dog barking, drilling, gunshot, jackhammer, and siren. The main contributions of this work are summarized as follows:

- A multi-branch BiLSTM architecture augmented with MHSA for improved contextual weighting of temporal representations.
- Parallel feature learning through capacity-differentiated BiLSTM branches to enhance robustness across diverse urban sound events.
- Improved class-wise discrimination under noisy and overlapping acoustic conditions, particularly for safety-critical sound events.
- A comprehensive experimental evaluation, including baseline comparisons and ablation analysis, demonstrating the effectiveness and generalization capability of the proposed framework.

The remainder of this paper is organized as follows. Section 2 reviews related work on urban sound classification and suspicious audio detection. Section 3 describes the proposed architecture and training methodology. Section 4 presents experimental results and ablation analysis. Finally, Section 5 concludes the paper and outlines directions for future research..

2. Literature Review

2.1. Theoretical Background and Foundational Methods

Early approaches to urban sound analysis predominantly relied on traditional machine learning techniques using handcrafted acoustic features, such as MFCCs, spectral contrast, and chroma features, coupled with classifiers including support vector machines and random forests [9]. While computationally efficient, these methods generally exhibited limited robustness in real-world urban environments due to their inability to effectively model complex temporal dynamics and overlapping sound sources under noisy conditions.

The transition to deep learning-based models marked a significant improvement in environmental sound classification performance. Convolutional neural networks (CNNs) emerged as a foundational architecture for processing time–frequency representations of audio signals, achieving classification accuracies in the range of 80–90% on benchmark datasets such as UrbanSound8K [13], [26]. Despite these advances, CNNs primarily capture local spatial correlations and often fail to model long-range temporal dependencies that are critical for distinguishing acoustically similar or temporally evolving sound events.

To address these limitations, RNNs, particularly LSTM architectures, have been widely adopted for audio sequence modeling [4], [15]. Bidirectional LSTM (BiLSTM) further enhances temporal modeling by processing sequences in both forward and backward directions, enabling richer contextual representations that are especially important for time-dependent audio signals [12]. Prior studies support this design choice: Munirathinam and Vitek [27] demonstrated that bidirectional processing improves detection of time-asymmetric acoustic events such as sirens and emergency sounds, while Bansal and Garg [12] showed that convolutional–recurrent architectures combining spectral feature extraction with bidirectional temporal modeling achieve robust performance across diverse urban noise conditions. These findings provide a strong theoretical and empirical basis for adopting BiLSTM as the primary temporal modeling component in this work.

MFCCs are employed as the feature representation due to their compact and perceptually motivated encoding of the acoustic spectral envelope [12], [13]. Studies by Mukhamadiyev et al. [11] and Padmaja and Banu [9] demonstrate that MFCC-based representations

consistently outperform raw waveform inputs in urban sound classification, owing to their robustness to environmental variability and alignment with human auditory perception. The combination of MFCC features with BiLSTM-based temporal modeling thus forms a well-established and effective foundation for the proposed approach.

2.2. Recent State-of-the-Art in Urban Sound Recognition

Recent studies have explored various deep learning architectures for suspicious and urban sound recognition, focusing on improving classification accuracy, robustness, and deployment efficiency. Shailendra et al. [28] proposed an audio-based suspicious activity detection framework using MFCC features extracted from a subset of the UrbanSound8K dataset, achieving a test accuracy of 90.25% with an LSTM model. Their work demonstrated the effectiveness of temporal modeling for surveillance applications, although generalization performance under diverse urban noise conditions remained a challenge.

In a related domain, study [17] introduced a hybrid architecture combining CNN, BiLSTM, and MHSA for speech emotion recognition. The attention mechanism is formulated as:.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V \tag{1}$$

where Q, K, and V denote the query, key, and value matrices, respectively, and d_k represents the key dimensionality. This formulation enables effective modeling of long-range dependencies and improves contextual representation. The proposed model achieved unweighted accuracies of 95.65% on Emo-DB and 80.19% on RAVDESS, outperforming standalone CNN and LSTM models. Although applied to emotion recognition, this hybrid design provides transferable insights for urban sound classification tasks characterized by temporal and spectral variability.

Recent research has also emphasized computational efficiency and deployability in smart city environments. Research [19] developed a CNN-based framework optimized for IoT deployment, achieving a test accuracy of 92.68% on UrbanSound8K while maintaining low inference latency. The training objective is defined using categorical cross-entropy loss:

$$L = -\sum^{C} y_i \log(\hat{y}_i) \tag{2}$$

where C denotes the number of classes, y_i the ground-truth label, and $\hat{y}_i$ the predicted probability. Their results highlight the increasing importance of lightweight architectures for real-time urban monitoring.

Similarly, study [21] proposed a TinyML-based framework for on-device urban noise classification using Raspberry Pi platforms. Their system achieved precision and recall values between 0.92 and 1.00 across multiple classes, demonstrating effective deployment in resource-constrained environments while reducing latency and communication overhead. Research [29] explored transfer learning using a CNN integrated with the pre-trained YAMNet model. By leveraging embeddings derived from depth-wise separable convolutions, the approach achieved a test accuracy of 98% on UrbanSound8K, significantly outperforming CNN and CNN–LSTM baselines. However, the study reported persistent confusion between acoustically similar stationary sounds, indicating limitations in capturing fine-grained temporal context.

2.3. Research Gaps and Motivation

Collectively, existing studies demonstrate substantial progress in urban sound classification, particularly in improving accuracy and computational efficiency [22]. However, several critical limitations remain that directly motivate this work. First, many approaches treat temporal modeling and contextual attention as separate design components, rather than integrating them within a unified architecture specifically tailored for suspicious sound recognition. While recurrent models capture sequential dependencies and attention mechanisms model global context, their combined effect is often underexplored or loosely integrated [17], [21].

Second, there is a lack of systematic evaluation under consistent experimental settings. Many studies report performance improvements without direct comparison to standard

Dense and LSTM baselines using identical data splits and preprocessing pipelines, making it difficult to isolate the true contribution of architectural design choices [4], [6]. Third, ablation studies are frequently limited or incomplete. In particular, the individual contributions of architectural components—such as multi-branch temporal modeling and attention-based contextual weighting—are rarely analyzed in a structured and reproducible manner [7], [9]. This limits the interpretability and reproducibility of reported improvements.

Fourth, evaluation metrics are often dominated by overall accuracy, with insufficient emphasis on class-wise recall. For safety-critical applications, this is a significant limitation, as false negatives—such as missed gunshot or siren detections—have substantially higher operational risk than false positives [18], [19]. Finally, while recent models achieve high accuracy, many rely on large parameter counts or complex architectures that may limit scalability and real-time deployment in practical smart city environments. There remains a need for architectures that balance performance, interpretability, and computational efficiency.

Motivated by these gaps, this work proposes a Multi-Branch BiLSTM with MHSA framework that integrates parallel bidirectional temporal modeling with lightweight contextual weighting. By combining capacity-differentiated BiLSTM branches with self-attention enhancement, the proposed approach aims to improve both temporal representation and contextual discrimination under noisy urban conditions. In addition, the study emphasizes consistent baseline comparison, comprehensive ablation analysis, and class-wise evaluation to provide a more complete and interpretable assessment of model performance.

3. Proposed Method

This section presents the dataset, preprocessing pipeline, network architecture, and experimental configuration adopted for the proposed Multi-Branch BiLSTM with MHSA for suspicious urban sound recognition. The complete workflow is illustrated in Figure 1, which summarizes the end-to-end process from raw audio input to final classification.

3.1. Overview of the Framework

The proposed methodology begins with preprocessing raw urban audio recordings and transforming them into MFCC representations that capture essential time–frequency characteristics of suspicious sound events. These features are then passed through the proposed architecture, which consists of parallel multi-branch BiLSTM networks designed to extract bidirectional temporal dependencies with different representational capacities. Within each branch, MHSA is applied to enhance contextual weighting of the BiLSTM output representations. In this framework, the self-attention module functions as a lightweight enhancement mechanism rather than a full Transformer sequence backbone. Its role is to selectively emphasize informative temporal regions in the learned sequence representations while preserving the temporal modeling capability of BiLSTM.

The outputs from all branches are subsequently fused through concatenation, followed by global pooling operations to aggregate feature representations. The resulting features are then passed through a dense layer and a softmax classifier to produce the final prediction. Figure 1 illustrates the complete pipeline, including audio preprocessing, MFCC feature extraction, multi-branch BiLSTM-based temporal modeling, self-attention–based contextual enhancement, feature fusion, and supervised classification.

3.2. Experimental Environment

All experiments were conducted using Python-based deep learning frameworks. Audio preprocessing and feature extraction were implemented using Librosa, while model development and training were carried out using TensorFlow/Keras. The proposed Multi-Branch BiLSTM with MHSA model was trained using the Adam optimizer with an initial learning rate of 1×10^{-3}, chosen for its adaptive learning capability and stability in training deep sequence models. Supervised learning employed categorical cross-entropy loss for multi-class classification.

To improve generalization and reduce overfitting, dropout regularization was applied within the attention-enhanced branches and in the final classification layers. A ReduceLROnPlateau scheduler was used to decrease the learning rate by a factor of 0.25 when validation loss did not improve for five consecutive epochs, with a minimum learning rate of 1×10^{-6}. Training was conducted for up to 100 epochs with a batch size of 32, and early

stopping with a patience of 10 epochs was applied based on validation loss. The hyperparameter configuration is summarized in Table 1, while branch-specific architectural parameters and fusion settings are presented in Table 2 and Table 3, respectively.

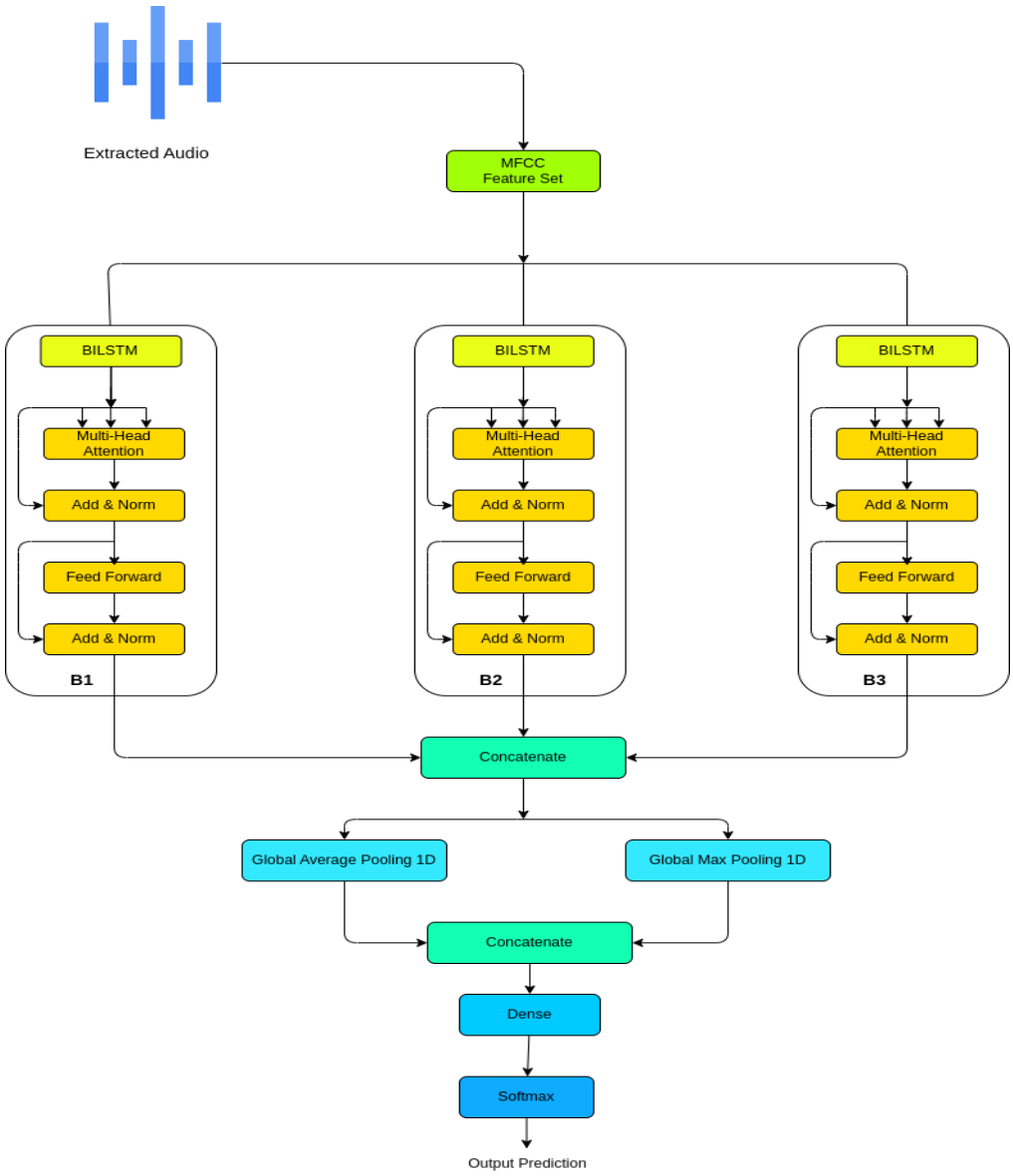

Figure 1. Overall workflow of the proposed Multi-Branch BiLSTM with MHSA framework for suspicious urban sound recognition.

Table 1. Global training configuration.

Hyperparameter	Value
MFCC Coefficients (N_MFCC)	64
Maximum Sequence Length	50
Optimizer	Adam
Initial Learning Rate	1e-3
Loss Function	Categorical Cross-Entropy
Early Stopping Patience	10
LR Reduction Patience	5
LR Reduction Factor	0.25
Minimum Learning Rate	1e-6
Batch Size	32

Table 2. Branch-specific parameters (each branch is parameterized independently).

Parameter	Branch 1	Branch 2	Branch 3
BiLSTM Units	32	64	128
Key Dimension (MHSA)	32	64	128
Feed-Forward Size	64 → 32	128 → 64	256 → 128
Number of Attention Heads	3	3	3
Dropout Rate	0.3	0.3	0.3

Table 3. Fusion and classification settings

Parameter	Value
Fusion Method	Concatenation of branch outputs
Pooling Strategy	Global Average Pooling + Global Max Pooling (concatenated)
Dense Layer Units	64 (ReLU)
Dropout Rate (Classifier)	0.2
Output Layer	5-class Softmax

3.3. Dataset Description and Feature Extraction

The experiments in this study are conducted using the UrbanSound8K dataset, a widely used benchmark for environmental and urban sound classification [3]. The dataset consists of 8,732 labeled audio clips, each with a maximum duration of 4 seconds, organized into 10 predefined folds to support cross-validation. For suspicious sound recognition, a curated subset is constructed by selecting five security-relevant classes from the original ten-class UrbanSound8K dataset: dog bark (class 0), drilling (class 1), gunshot (class 2), jackhammer (class 3), and siren (class 4). The resulting subset contains 4,303 audio samples.

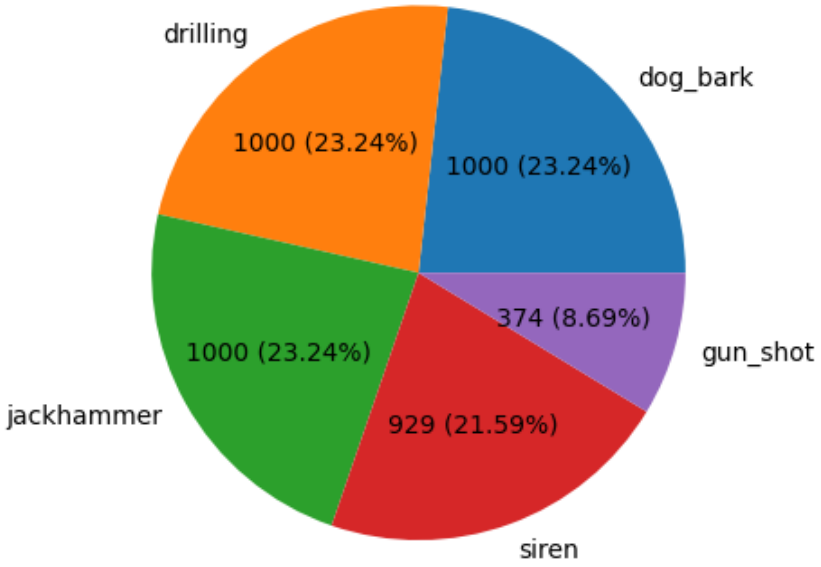

Figure 2. Class distribution of the selected subset (five security-relevant classes) from the UrbanSound8K dataset.

The selection of these classes is motivated by their direct relevance to urban surveillance applications, where early detection of anomalous acoustic events supports incident prevention and emergency response. The diversity of recording environments in UrbanSound8K ensures that the curated subset reflects realistic urban acoustic variability. The class distribution is approximately balanced, as shown in Figure 2, which supports unbiased model training. All audio files are loaded using the Librosa library while preserving their native sampling rates. Feature extraction is performed using MFCCs, with 64 coefficients computed per frame, resulting in two-dimensional feature matrices of size (time steps × 64). MFCC features are

selected due to their compact and perceptually motivated representation of the spectral envelope, which is effective for modeling diverse urban sound characteristics.

The MFCC extraction process includes pre-emphasis, framing, windowing, Fast Fourier Transform (FFT), Mel filter bank processing, and Discrete Cosine Transform (DCT), expressed as:

$$c_k = \sum_{m=1}^{M} \log(E_m) \cos\left[\frac{\pi k}{M}(m - 0.5)\right] \tag{3}$$

where E_m denotes the Mel filter bank energy, M is the number of filters, and k is the cepstral coefficient index.

To ensure consistent input dimensions for batch processing, MFCC sequences are padded or truncated to a fixed length of 50 frames, resulting in input tensors of size (50 × 64). Feature standardization is applied using zero-mean and unit-variance normalization (with $\epsilon = 1 \times 10^{-6}$) to improve numerical stability and convergence. The dataset is partitioned using stratified sampling into 80% training, 10% validation, and 10% testing subsets to preserve class distribution across all splits.

3.4. Architectural Design

The architectural design of the proposed model builds upon the framework overview in Section 3.1 by providing a more detailed description of the multi-branch BiLSTM and self-attention components. The network processes MFCC inputs of size (50 × 64) through three parallel branches, each consisting of a Bidirectional LSTM layer followed by a MHSA block. The branches differ in representational capacity, allowing the model to capture complementary temporal features across varying levels of complexity. Detailed hyperparameter settings for each branch are provided in Table 2.

The term multi-branch refers to parallel branches with different representational capacities rather than different temporal resolutions. Each branch is parameterized independently, enabling the model to jointly learn diverse temporal representations from the same input feature sequence. Within each branch, the self-attention mechanism refines the contextual weighting of BiLSTM outputs by allowing each temporal position to attend to all others in the sequence. The formulation follows the standard MHSA mechanism described in Section 2.2. In this architecture, it functions as a lightweight enhancement layer rather than a standalone sequence modeling backbone.

The outputs from all branches are concatenated and passed through global average pooling and global max pooling to capture both dominant and extreme temporal responses. The pooled features are then concatenated and fed into a dense classifier consisting of a Dense layer with 64 units (ReLU activation), followed by dropout (0.2) and a softmax output layer with five neurons corresponding to the target classes.

3.5. Baseline Models for Performance Comparison

To evaluate the effectiveness of the proposed architecture, two baseline models are implemented:

- Dense Neural Network: Multiple fully connected layers with ReLU activations and dropout
- Bidirectional LSTM Model: A single BiLSTM layer followed by a dense classifier

All baseline models are trained using identical dataset splits, optimization settings, and evaluation protocols to ensure fair and unbiased comparison.

3.6. Evaluation Metrics and Visualization

Model performance is evaluated using four standard classification metrics: accuracy, precision, recall, and F1-score, as described in [30]–[32]. These metrics provide a comprehensive assessment of classification performance across suspicious sound categories.

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$\text{precision} = \frac{TP}{TP + FP} \tag{5}$$

$$\text{recall} = \frac{TP}{TP + FN} \tag{6}$$

$$\text{f1} - \text{score} = 2 \times \frac{(precision \times recall)}{(precision + recall)} \tag{7}$$

where TP, FP, TN, and FN denote true positives, false positives, true negatives, and false negatives, respectively.

Class-wise precision, recall, and F1-scores are computed to evaluate performance across individual sound categories, with particular attention to safety-critical classes such as gunshot and siren. Macro-averaged and weighted F1-scores are also reported to account for class distribution. Performance visualization includes confusion matrices and training curves, which are used to analyze classification behavior, convergence, and generalization. Detailed interpretation of these results is provided in Section 4.

4. Results and Discussion

This section presents a comprehensive evaluation of the proposed Multi-Branch BiLSTM with MHSA for suspicious urban sound recognition. All experiments follow the methodology described in Section 3, using identical data splits, feature representations, and evaluation metrics to ensure fair comparison and reproducibility. The primary emphasis is placed on test-set performance as the main indicator of model generalization. The results are organized as follows: (i) baseline model comparison; (ii) per-class performance evaluation on the test set; (iii) ablation study and architectural contribution analysis; (iv) confusion matrix analysis; and (v) training dynamics as supporting evidence.

4.1. Feature Visualization: Waveform, Spectrogram, and MFCC Analysis

Figure 3 shows representative waveform and Mel-spectrogram examples. The waveform illustrates amplitude variations over time, where impulsive events such as gunshots exhibit sharp, short-duration spikes, while sustained events such as sirens display more regular oscillatory patterns. The Mel-spectrogram further reveals distinct frequency–time energy distributions: impulsive sounds tend to produce broadband bursts, whereas sustained sounds exhibit structured and periodic frequency patterns. These temporal and spectral differences support the use of MFCC features, which provide a compact representation of the spectral envelope. They also motivate the multi-branch BiLSTM design: lower-capacity branches capture broader temporal structures, while higher-capacity branches are better suited to modeling rapid temporal variations.

Figure 4 presents MFCC heatmaps for jackhammer and siren samples. The jackhammer signal exhibits regular, high-energy cepstral patterns consistent with repetitive mechanical activity, while the siren shows smoothly varying cepstral structures corresponding to frequency modulation. These differences indicate that MFCC features capture discriminative temporal–cepstral patterns across sound classes. The self-attention mechanism further complements this representation by emphasizing informative temporal regions within the BiLSTM outputs, improving contextual discrimination between acoustically distinct yet temporally varying events.

4.2. Training Dynamics and Convergence Behavior

To provide context for model learning behavior, the training and validation dynamics of the proposed Multi-Branch BiLSTM with MHSA model are analyzed. Figure 5 illustrates the training and validation accuracy and loss curves across epochs. The accuracy curves show rapid improvement during the initial epochs, followed by gradual stabilization as training progresses. The validation accuracy closely follows the training accuracy, with only minor fluctuations, indicating stable learning behavior and effective generalization during training.

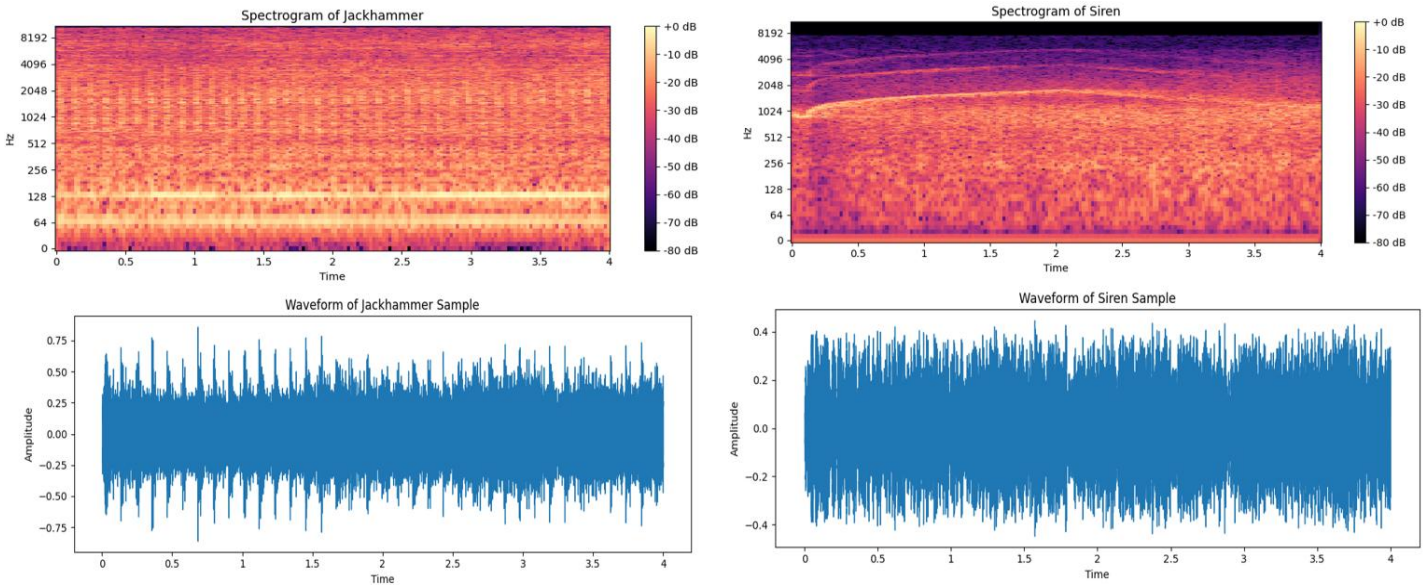

Figure 3. Waveform (top) and Mel-spectrogram (bottom) of representative audio samples from the curated UrbanSound8K subset, illustrating distinct temporal and spectral patterns across sound types.

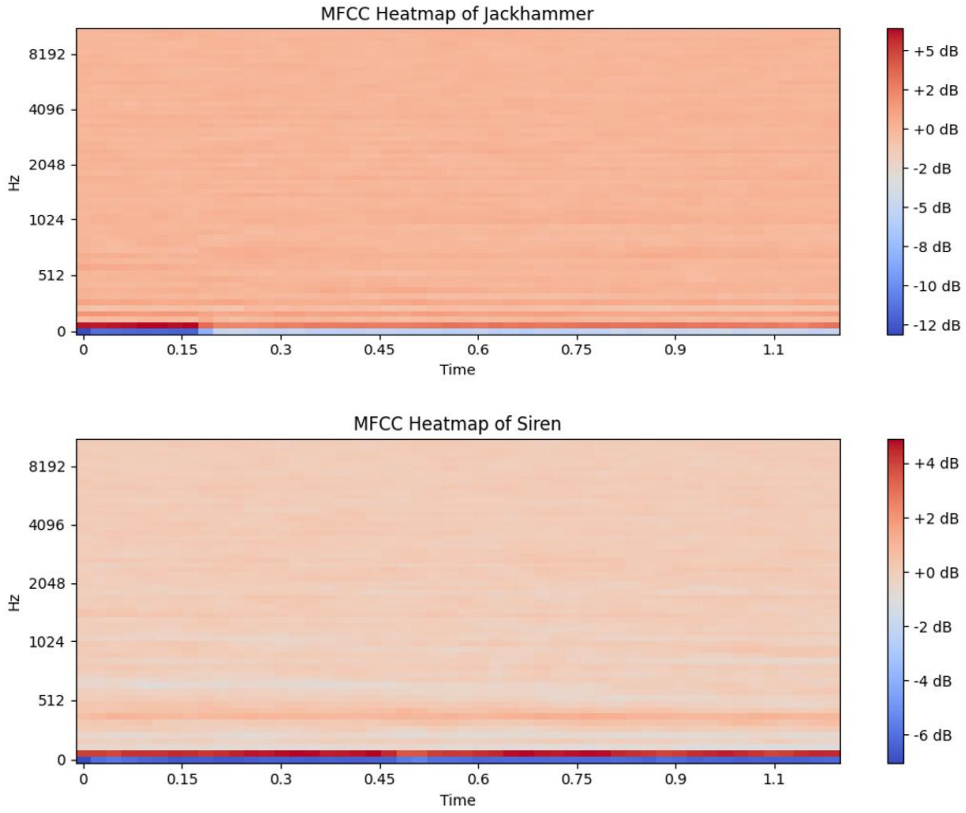

Figure 4. MFCC heatmaps for jackhammer (left) and siren (right) samples, demonstrating distinct temporal–cepstral patterns captured by the MFCC representation.

Similarly, the loss curves exhibit a consistent downward trend in training loss, while validation loss decreases initially and stabilizes after several epochs. The absence of large divergence between training and validation loss suggests that overfitting is effectively controlled through dropout regularization and early stopping. A summary of training and validation performance is presented in Table 4 for reference.

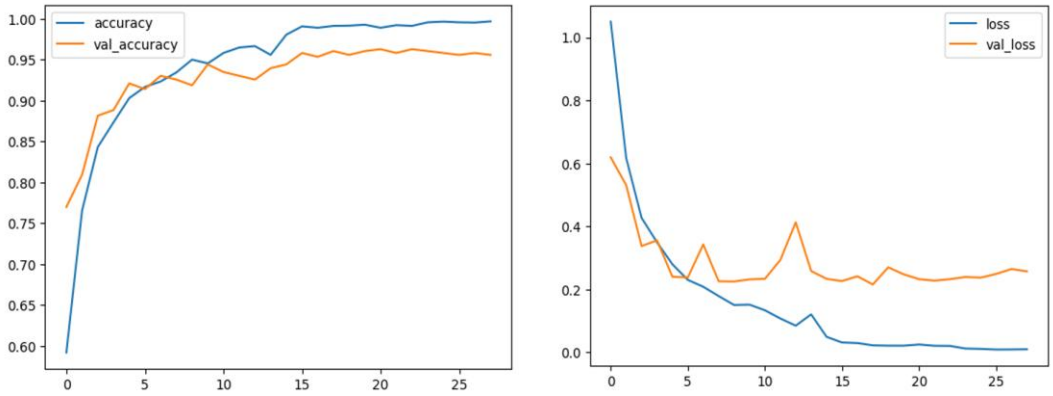

Figure 5. (a) Training and validation accuracy curves; (b) training and validation loss curves.

Table 4. Summary of training and validation performance.

Model	Train Acc.	Val. Acc.	Train Loss	Val. Loss
Dense [28]	85.96%	88.63%	0.3889	0.3779
LSTM [28]	96.02%	90.49%	0.1036	0.3184
Proposed Model	99.78%	96.28%	0.0085	0.2325

The proposed model achieves the highest training and validation accuracy among all compared models. While the training loss is relatively low, its interpretation should be considered together with validation performance. The consistency between training and validation metrics indicates that the model learns effectively without exhibiting significant overfitting. These training dynamics are presented as supporting evidence of stable optimization, while the primary evaluation of model performance is discussed using test-set results in subsequent sections.

4.3. Baseline Models and Overall Test Performance

To benchmark the effectiveness of the proposed architecture, two baseline models from Shailendra et al. [28] were implemented under identical experimental conditions. The first baseline is a Dense neural network without explicit temporal modeling, while the second is a standard LSTM-based model that captures sequential dependencies but does not incorporate attention mechanisms or multi-branch processing. All models were trained, validated, and tested using identical 80/10/10 stratified splits, MFCC-based feature representations, and optimization settings. No data augmentation was applied in order to isolate architectural effects.

Table 4. Test-set performance comparison (primary evaluation)

Model	Test Accuracy	Test Precision	Test Recall	Test F1-Score	Test Loss
Dense Model [28]	88.86%	88.4%†	88.2%†	88.3%†	0.3488
LSTM Model [28]	90.25%	90.1%†	89.8%†	89.9%†	0.3471
Proposed Model	**95.59%**	**95.4%**	**95.3%**	**95.3%**	**0.2308**

† Precision, recall, and F1-score values for baseline models are approximate macro-averaged estimates derived from reported results in [28] and are provided for reference only.

The proposed model consistently outperforms both baselines across all metrics. It achieves improvements of +4.73 percentage points over the Dense model and +5.34 percentage points over the LSTM model in test accuracy. Similar gains are observed in precision, recall, and F1-score, indicating improved classification robustness across sound categories.

4.4. Per-Class Performance on the Test Set

To provide a more detailed evaluation beyond overall metrics, class-wise performance on the test set is analyzed. This analysis is particularly important in the context of public

safety, where detection reliability varies across sound categories and certain classes carry higher operational risk.

Table 5. Per-class test-set performance of the proposed model

Class	Precision	Recall	F1-Score
Dog Bark	0.96	0.95	0.96
Drilling	0.93	0.94	0.93
Gunshot	0.97	0.97	0.97
Jackhammer	0.94	0.94	0.94
Siren	0.97	0.96	0.97
Macro Average	0.95	0.95	0.95

As shown in Table 5, the proposed model achieves consistently high performance across all classes. In particular, recall is emphasized for safety-critical events such as gunshot and siren, where false negatives can lead to delayed or missed emergency responses. Both classes achieve high recall values (0.97 and 0.96, respectively), indicating reliable detection capability.

The lowest recall is observed for drilling (0.94), which can be attributed to acoustic similarity with jackhammer sounds. Such confusion between mechanically repetitive sound sources has been widely reported in urban sound classification studies [29], where overlapping temporal and spectral patterns make fine-grained discrimination challenging. The implications of this behavior are further examined in the confusion matrix analysis (Section 4.6).

4.5. Ablation Study and Architectural Contribution Analysis

To further examine the contribution of each architectural component, an ablation study was conducted using single-branch (B1, B2, B3), dual-branch, and full multi-branch configurations. The evaluation focuses on test-set performance as the primary indicator of generalization.

Table 6. Ablation study – test-set performance comparison

Model Variant	Test Acc.	Test Prec.	Test Recall	Test F1	Test Loss
B1 only (32 units)	93.74%	93.5%	93.4%	93.4%	0.2661
B2 only (64 units)	93.97%	93.8%	93.6%	93.7%	0.2361
B3 only (128 units)	94.43%	94.2%	94.1%	94.2%	0.2192
B1 + B2	93.97%	93.7%	93.5%	93.6%	0.2390
B1 + B3	93.97%	93.8%	93.7%	93.7%	0.2128
B2 + B3	94.43%	94.2%	94.0%	94.1%	0.2752
B1 + B2 + B3 (Proposed)	95.59%	95.4%	95.3%	95.3%	0.2308

As summarized in Table 6, the full multi-branch configuration consistently achieves the best performance across all evaluation metrics. This indicates that combining branches with different representational capacities provides complementary information that improves overall model robustness. Among single-branch variants, B3 (128 units) achieves the highest performance, suggesting that higher-capacity representations contribute significantly to capturing complex temporal patterns. However, the integration of lower-capacity branches (B1 and B2) further improves performance, highlighting the importance of multi-scale representational diversity.

The contribution of the self-attention component is observed indirectly through the overall performance improvements. While an explicit comparison between BiLSTM-only and BiLSTM with attention is not included, the results align with prior findings that attention mechanisms improve contextual weighting in sequence modeling tasks [17].

4.6. Confusion Matrix Analysis

To further analyze class-wise prediction behavior, a normalized confusion matrix is presented in Figure 5.

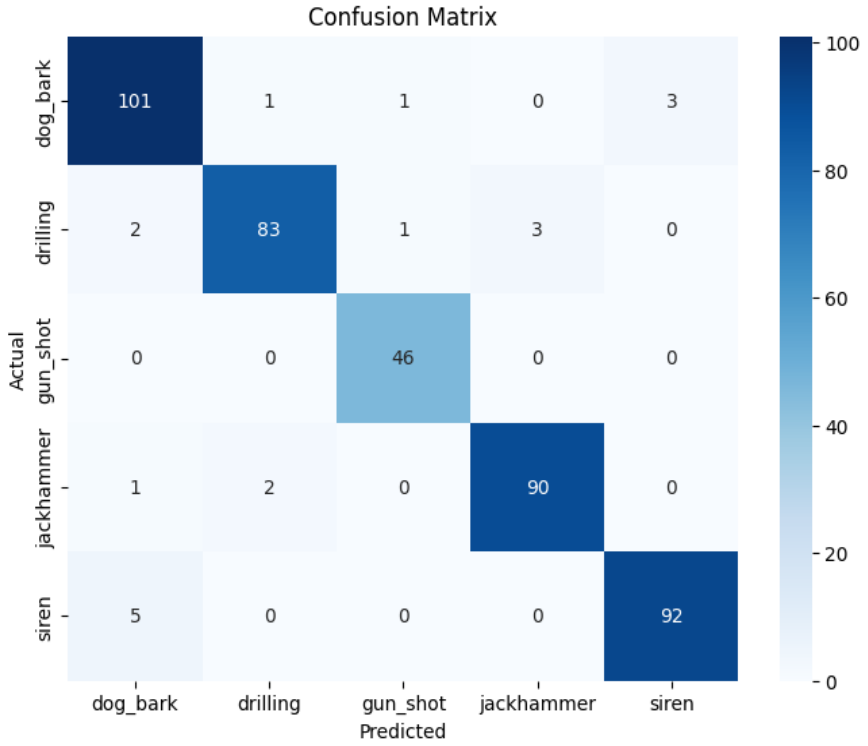

Figure 5. Confusion matrix of the proposed Multi-Branch BiLSTM with MHSA model on the test set.

The diagonal elements in Figure 5 indicate high recall across all classes, consistent with the results reported in Table 5. Misclassifications are limited and primarily occur between drilling and jackhammer, reflecting their similar repetitive mechanical characteristics. This observation is consistent with prior studies on urban sound datasets, where such classes exhibit overlapping temporal–spectral signatures [29]. Importantly, gunshot detection shows minimal confusion with other classes, and siren detection also maintains high class separability. The high recall values for these safety-critical classes indicate low false negative rates, which is essential for surveillance applications where missed detections may lead to delayed response. These findings suggest that the proposed architecture effectively balances class discrimination and robustness, particularly for events with distinct acoustic signatures, while remaining competitive for acoustically similar classes.

4.7. Discussion of Robustness and Limitations

The experimental results demonstrate that the proposed Multi-Branch BiLSTM with MHSA provides a robust framework for suspicious urban sound recognition. The integration of bidirectional temporal modeling and attention-based contextual weighting enables the model to capture both short-duration acoustic transients (e.g., gunshots) and longer, structured temporal patterns (e.g., sirens and drilling). Compared to MSA-TCN [24], the proposed approach introduces bidirectional recurrent modeling and capacity-differentiated multi-branch encoding. These design choices offer a complementary perspective to convolution-based temporal modeling, particularly for time-asymmetric acoustic events where forward and backward temporal context can be informative.

Despite the strong performance, some limitations remain. Misclassification is primarily observed between acoustically similar classes, especially drilling and jackhammer, which share repetitive mechanical patterns. This limitation is consistent with observations reported in prior studies on urban sound classification [29], suggesting that it is partly attributable to dataset characteristics rather than model design alone. In addition, the evaluation is conducted on a curated five-class subset of UrbanSound8K, which, while relevant for surveillance scenarios, does not fully reflect the complexity of broader environmental sound classification tasks. Furthermore, the specific contribution of the MHSA component has not been explicitly isolated through a dedicated ablation study. Future work may address these limitations by

incorporating data augmentation strategies, exploring complementary spectral representations, and extending the framework to multi-modal or larger-scale datasets. Investigating lightweight deployment strategies for edge and IoT environments also remains an important direction for practical adoption.

5. Conclusion

This study presents a Multi-Branch BiLSTM with MHSA framework for suspicious urban sound recognition. The proposed approach combines parallel BiLSTM branches with different representational capacities and applies attention-based contextual weighting to enhance temporal feature representations. The results indicate that the model achieves strong generalization performance on the test set, outperforming conventional Dense and LSTM baselines. In particular, high recall values for safety-critical classes such as gunshot and siren demonstrate the model's effectiveness in minimizing false negatives, which is essential for surveillance applications.

The ablation analysis highlights the importance of multi-branch integration, showing that combining branches of different capacities provides complementary information that improves overall performance. The MHSA component functions as a lightweight enhancement to refine BiLSTM outputs, contributing to improved contextual discrimination without acting as a standalone sequence modeling backbone. Several limitations remain, including the use of a five-class subset, the absence of an isolated attention ablation, and the lack of deployment-oriented evaluation. Future work will focus on expanding the evaluation to larger and more diverse datasets, further analyzing the contribution of attention mechanisms, and optimizing the model for real-time deployment in resource-constrained environments.

Author Contributions: Conceptualization: M.S.B. and S.M.Y.; Methodology: M.S.B.; Software: M.S.B.; Validation: M.S.B., S.M.Y. and S.S.S.; Formal analysis: M.S.B.; Investigation: M.S.B.; Resources: S.M.Y.; Data curation: M.S.B.; Writing—original draft preparation: M.S.B.; Writing—review and editing: S.M.Y. and S.S.S.; Visualization: M.S.B.; Supervision: S.M.Y.; Project administration: S.M.Y.; Funding acquisition: S.S.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by TETFund Nigeria, grant number NRF-2023-SETI-AFS-00158.

Data Availability Statement: The dataset used in this study is the UrbanSound8K dataset, which is publicly available at https://urbansounddataset.weebly.com/urbansound8k.html. The processed data, implementation code, and trained model checkpoints generated during this study are available from the corresponding author upon reasonable request.

Acknowledgments: The authors acknowledge the support of TETFund Nigeria for funding this research under grant NRF-2023-SETI-AFS-00158.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] B. Kim, J. Kim, H. Chae, D. Yoon, and J. W. Choi, "Deep neural network-based automatic modulation classification technique," in *2016 International Conference on Information and Communication Technology Convergence (ICTC)*, Oct. 2016, pp. 579–582. doi: 10.1109/ICTC.2016.7763537.

[2] S. Abdoli, P. Cardinal, and A. Lameiras Koerich, "End-to-end environmental sound classification using a 1D convolutional neural network," *Expert Syst. Appl.*, vol. 136, pp. 252–263, Dec. 2019, doi: 10.1016/j.eswa.2019.06.040.

[3] J. Salamon, C. Jacoby, and J. P. Bello, "A Dataset and Taxonomy for Urban Sound Research," in *Proceedings of the 22nd ACM international conference on Multimedia*, Nov. 2014, pp. 1041–1044. doi: 10.1145/2647868.2655045.

[4] Y. Tokozume and T. Harada, "Learning environmental sounds with end-to-end convolutional neural network," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2017, pp. 2721–2725. doi: 10.1109/ICASSP.2017.7952651.

[5] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, Sep. 2015, pp. 1–6. doi: 10.1109/MLSP.2015.7324337.

[6] A. Vaswani *et al.*, "Attention Is All You Need," *arXiv*, vol. 30, Aug. 2023, [Online]. Available: http://arxiv.org/abs/1706.03762

[7] S. Hershey *et al.*, "CNN architectures for large-scale audio classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2017, pp. 131–135. doi: 10.1109/ICASSP.2017.7952132.

[8] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," in *arXiv*, May 2016, pp. 1–15. [Online]. Available: http://arxiv.org/abs/1409.0473

[9] S. Padmaja and N. Sharmila Banu, "A Systematic Literature Review on Sound Event Detection and Classification," in *2025 5th International Conference on Trends in Material Science and Inventive Materials (ICTMIM)*, Apr. 2025, pp. 1580–1587. doi: 10.1109/ICTMIM65579.2025.10988199.

[10] A. S. Roman, I. R. Roman, and J. P. Bello, "Robust DoA Estimation from Deep Acoustic Imaging," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2024, pp. 1321–1325. doi: 10.1109/ICASSP48485.2024.10447883.

[11] A. Mukhamadiyev, I. Khujayarov, D. Nabieva, and J. Cho, "An Ensemble of Convolutional Neural Networks for Sound Event Detection," *Mathematics*, vol. 13, no. 9, p. 1502, May 2025, doi: 10.3390/math13091502.

[12] A. Bansal and N. K. Garg, "Robust technique for environmental sound classification using convolutional recurrent neural network," *Multimed. Tools Appl.*, vol. 83, no. 18, pp. 54755–54772, Dec. 2023, doi: 10.1007/s11042-023-17066-2.

[13] N. N. Wijaya, D. R. I. M. Setiadi, and A. R. Muslikh, "Music-Genre Classification using Bidirectional Long Short-Term Memory and Mel-Frequency Cepstral Coefficients," *J. Comput. Theor. Appl.*, vol. 1, no. 3, pp. 243–256, Jan. 2024, doi: 10.62411/jcta.9655.

[14] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[15] J. Devlin, M.-W. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North*, Oct. 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

[16] K. Zaman, K. Li, M. Sah, C. Direkoglu, S. Okada, and M. Unoki, "Transformers and audio detection tasks: An overview," *Digit. Signal Process.*, vol. 158, p. 104956, Mar. 2025, doi: 10.1016/j.dsp.2024.104956.

[17] S. Kim and S.-P. Lee, "A BiLSTM–Transformer and 2D CNN Architecture for Emotion Recognition from Speech," *Electronics*, vol. 12, no. 19, p. 4034, Sep. 2023, doi: 10.3390/electronics12194034.

[18] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for Polyphonic Sound Event Detection," *Appl. Sci.*, vol. 6, no. 6, p. 162, May 2016, doi: 10.3390/app6060162.

[19] S. Domazetovska Markovska, V. Gavriloski, D. Pecioski, M. Anachkova, D. Shishkovski, and A. Angjusheva Ignjatovska, "Urban Sound Classification for IoT Devices in Smart City Infrastructures," *Urban Sci.*, vol. 9, no. 12, p. 517, Dec. 2025, doi: 10.3390/urbansci9120517.

[20] B. Koçak, R. Cuocolo, D. P. dos Santos, A. Stanzione, and L. Ugga, "Must-have Qualities of Clinical Research on Artificial Intelligence and Machine Learning," *Balkan Med. J.*, vol. 40, no. 1, pp. 3–12, Jan. 2023, doi: 10.4274/balkanmedj.galenos.2022.2022-11-51.

[21] M. S. Remolina Soto, B. Amaya Guzmán, P. A. Aya-Parra, O. J. Perdomo, M. Becerra-Fernandez, and J. Sarmiento-Rojas, "Intelligent Classification of Urban Noise Sources Using TinyML: Towards Efficient Noise Management in Smart Cities," *Sensors*, vol. 25, no. 20, p. 6361, Oct. 2025, doi: 10.3390/s25206361.

[22] Z. Huang, C. Liu, H. Fei, W. Li, J. Yu, and Y. Cao, "Urban sound classification based on 2-order dense convolutional network using dual features," *Appl. Acoust.*, vol. 164, p. 107243, Jul. 2020, doi: 10.1016/j.apacoust.2020.107243.

[23] K. Zaman, M. Sah, C. Direkoglu, and M. Unoki, "A Survey of Audio Classification Using Deep Learning," *IEEE Access*, vol. 11, pp. 106620–106649, 2023, doi: 10.1109/ACCESS.2023.3318015.

[24] M. S. Bawa, S. M. Yusuf, and S. S. Saminu, "MSA-TCN: Robust Urban Suspicious Sound Detection Using Multi-Scale Temporal Convolutions and Dual Attention," *J. Futur. Artif. Intell. Technol.*, vol. 3, no. 1, pp. 36–52, Apr. 2026, doi: 10.62411/faith.3048-3719-313.

[25] M. Cantarini, L. Gabrielli, A. Mancini, S. Squartini, and R. Longo, "A3CarScene: An audio-visual dataset for driving scene understanding," *Data Br.*, vol. 48, p. 109146, Jun. 2023, doi: 10.1016/j.dib.2023.109146.

[26] S. Suzić *et al.*, "UNS Exterior Spatial Sound Events Dataset for Urban Monitoring," in *2024 32nd European Signal Processing Conference (EUSIPCO)*, Aug. 2024, pp. 176–180. doi: 10.23919/EUSIPCO63174.2024.10715448.

[27] R. Munirathinam and S. Vitek, "Sound Source Localization and Classification for Emergency Vehicle Siren Detection Using Resource Constrained Systems," in *2024 34th International Conference Radioelektronika (RADIOELEKTRONIKA)*, Apr. 2024, pp. 1–5. doi: 10.1109/RADIOELEKTRONIKA61599.2024.10524053.

[28] A. Shailendra, C. Bengani, K. S. Kumari, P. Senthilraja, A. Prithivi, and S. Ramesh, "Suspicious Activity Detection based on Audio Detecting Methodology using Deep Learning," in *Recent Trends in Data Science and its Applications*, 2023, pp. 683–687. doi: 10.13052/rp-9788770040723.131.

[29] D. Trivedi, R. Sarmukaddam, and V. C. Gandhi, "Deep Learning for Urban Sound Classification: Using CNN and YAMNet Model Integration," in *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, 2026, pp. 332–349. doi: 10.1007/978-3-031-94283-9_29.

[30] M. Abubakar, Y. Ibrahim, O.-O. Ajayi, and S. S. Saminu, "A Lightweight Maize Leaf Disease Recognition Using PCA-Compressed MobileNetV2 Features and RBF-SVM," *J. Comput. Theor. Appl.*, vol. 3, no. 3, pp. 334–348, Jan. 2026, doi: 10.62411/jcta.15675.

[31] S. M. Yusuf, E. A. Adedokun, M. B. Mu'azu, I. J. Umoh, and A. A. Ibrahim, "A Novel Multi-Window Spectrogram Augmentation Approach for Speech Emotion Recognition Using Deep Learning," in *2021 1st International Conference on Multidisciplinary Engineering and Applied Science (ICMEAS)*, Jul. 2021, pp. 1–6. doi: 10.1109/ICMEAS52683.2021.9692411.

[32] S. M. Yusuf, E. A. Adedokun, M. B. Muazu, I. J. Umoh, and A. A. Ibrahim, "RMWSaug: Robust Multi-window Spectrogram Augmentation Approach for Deep Learning based Speech Emotion Recognition," in *2021 Innovations in Intelligent Systems and Applications Conference (ASYU)*, Oct. 2021, pp. 1–6. doi: 10.1109/ASYU52992.2021.9598956.