Journal of Future Artificial Intelligence and Technologies

E-ISSN: 3048-3719

Research Article

MSA-TCN: Robust Urban Suspicious Sound Detection Using Multi-Scale Temporal Convolutions and Dual Attention

Micah Shallom Bawa, Shehu Mohammed Yusuf *, and Sani Saleh Saminu

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

* Corresponding Author : Shehu Mohammed Yusuf

Abstract: Efficient suspicious sound event detection is critical for urban surveillance and public safety applications; however, many existing deep learning approaches face challenges related to high computational cost, unstable optimization, and limited suitability for resource-constrained environments. This paper proposes a Multi-Scale Attention-based Temporal Convolutional Network (MSA-TCN) for urban suspicious sound classification. The model operates on MFCC representations and integrates dilated multi-scale temporal convolutions to capture both short-duration impulsive events and long-duration acoustic patterns. Spectral normalization is employed to improve training stability and generalization, while channel and spatial attention mechanisms are incorporated to enhance feature representation. The proposed model is evaluated on the UrbanSound8K dataset using a stratified 80/10/10 train–validation–test split and compared with Dense and LSTM baseline models under the same experimental setup. MSA-TCN achieves a test accuracy of 92.34% and a macro-averaged F1-score of 0.92, outperforming both baselines while using only 128K parameters. Ablation results indicate that multi-scale temporal modeling and spectral normalization are the primary contributors to model performance and stability, whereas attention mechanisms mainly improve convergence behavior and training efficiency rather than final accuracy. Per-class evaluation shows strong recall for safety-critical events such as gunshot (0.92) and siren (0.88), with misclassifications largely occurring between acoustically similar classes. Overall, the results demonstrate that MSA-TCN provides an effective balance between performance, efficiency, and training stability for urban sound classification, while highlighting areas for further improvement in handling acoustically similar events and real-world deployment conditions.

Keywords: Acoustic signal processing; Audio event detection; Deep learning for audio; Edge AI for sound analysis; Multi-scale temporal convolutional networks; Sustainable Cities and Communities; Suspicious sound detection; Urban sound classification.

Received: January, 5th 2026
Revised: March, 29th 2026
Accepted: March, 30th 2026
Published: April, 15th 2026
Curr. Ver.: April, 15th 2026

Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY SA) license (https://creativecommons.org/licenses/by-sa/4.0/)

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 critical role in crime prevention, emergency response, and infrastructure protection. While video-based surveillance remains the dominant modality, it is often constrained by occlusion, poor illumination, adverse weather conditions, and privacy concerns [3], [4]. These limitations have motivated increasing interest in complementary sensing modalities capable of detecting safety-critical events even when visual cues are unreliable or unavailable, including vibration signals, thermal gradients, and acoustic emissions captured through industrial telemetry sensors [5]–[7].

Urban sound events such as gunshots, sirens, drilling, jackhammers, and aggressive animal sounds serve as early acoustic indicators of potentially dangerous situations. Consequently, automatic urban sound classification and detection have attracted significant research

attention [8], [9]. However, real-world urban soundscapes are inherently challenging due to high background noise, overlapping sound sources, and significant temporal variability, which hinder reliable and real-time sound event detection [6], [10]. Furthermore, many studies emphasize overall classification accuracy while under-analyzing class-wise recall, despite the fact that false negatives—such as missed gunshot or siren detections—have critical implications in public safety applications [11], [12].

Deep learning techniques have demonstrated strong performance in environmental sound classification by learning discriminative representations from time–frequency features such as Mel-Frequency Cepstral Coefficients (MFCCs) and spectrograms [13], [14]. Early approaches based on fully connected Dense networks are simple and effective in controlled settings but lack explicit temporal modeling, limiting their ability to capture long-range dependencies in audio signals [2], [4]. Recurrent architectures, particularly Long Short-Term Memory (LSTM) networks, improve temporal modeling but introduce sequential computation, higher computational cost, and susceptibility to overfitting in noisy environments [4], [6], [14]. These limitations highlight the need for architectures that can efficiently capture temporal dependencies while maintaining computational efficiency.

In addition to architectural challenges, urban audio datasets present inherent issues such as class imbalance, acoustic overlap, and background interference. Many deep learning-based approaches rely on high-capacity models that require substantial memory and processing resources, limiting their applicability in resource-constrained or edge-based surveillance systems [5], [15]. At the same time, Temporal Convolutional Networks (TCNs) have emerged as a promising alternative, offering parallel computation, stable gradient propagation, and efficient temporal modeling [16]. However, most existing TCN-based approaches operate at a single temporal scale, limiting their ability to simultaneously capture short-duration impulsive events (e.g., gunshots) and long-duration patterns (e.g., sirens and drilling) [11], [17], [18].

Attention mechanisms have been introduced to enhance feature discrimination by focusing on salient time–frequency regions while suppressing background noise [19]. Channel attention emphasizes informative feature maps, whereas spatial attention highlights important temporal or spectral regions [16], [17]. Although attention-based models have shown strong performance, their contribution is often not systematically analyzed, particularly in lightweight architectures, and their role in improving performance versus training dynamics remains unclear [20].

Another critical yet underexplored aspect is training stability. Noisy data and small batch sizes frequently lead to unstable optimization and degraded generalization [6], [14]. Spectral normalization has been shown to stabilize training by constraining Lipschitz continuity and improving convergence behavior [19], [21]. Despite these advantages, its application in lightweight audio-based surveillance models remains limited, and its contribution is rarely evaluated systematically.

Based on these limitations, this study addresses the following problem: how to design a lightweight and robust architecture that can simultaneously capture multi-scale temporal dynamics, maintain training stability, and operate efficiently in noisy urban environments. The central hypothesis is that integrating multi-scale temporal modeling with stabilization mechanisms can achieve robust performance under resource constraints, while attention mechanisms primarily contribute to improved training behavior and feature refinement rather than significant gains in final accuracy.

To address this problem, this paper proposes MSA-TCN, a Multi-Scale Attention-based Temporal Convolutional Network for urban suspicious sound detection. The proposed architecture integrates dilated multi-scale temporal convolutions to capture diverse temporal patterns, dual channel–spatial attention mechanisms to enhance discriminative time–frequency representations, and spectral normalization to stabilize training and improve generalization. Each component is incorporated based on a targeted problem–solution mapping, ensuring that the architecture is not an arbitrary combination of techniques but a structured design addressing specific limitations in prior work. The main contributions of this study are summarized as follows:

- Multi-Scale Temporal Modeling: A temporal convolutional architecture with dilated receptive fields is designed to capture both short-duration impulsive sounds and long-duration acoustic patterns.
- Stability-Oriented Training Design: Spectral normalization is applied to improve optimization stability and generalization in lightweight audio models.

- Attention-Based Feature Refinement: Channel and spatial attention mechanisms are incorporated to enhance feature representation, with analysis showing their primary impact on convergence behavior rather than final accuracy.
- Parameter-Efficient Architecture: The proposed model achieves competitive performance with only 128K parameters, supporting potential deployment in resource-constrained environments.

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

2. Related Work

The integration of AI and deep learning has significantly advanced surveillance systems, particularly in audio-based suspicious activity detection, where acoustic analysis improves detection accuracy in complex urban environments [2]. In modern smart cities, public safety is increasingly challenged by hazardous acoustic events such as gunshots, sirens, drilling, jackhammers, and aggressive animal sounds, which often serve as early indicators of criminal activity or emergency situations [3], [6]. Unlike visual cues, these sound events can be detected independently of lighting conditions and line-of-sight constraints. Consequently, audio-based sound event detection systems have become essential components of intelligent surveillance infrastructures for early warning and urban monitoring [14].

Earlier approaches to suspicious sound detection relied on handcrafted acoustic descriptors, including spectral energy distributions, zero-crossing rates, and temporal statistics, combined with classical machine-learning classifiers. Although computationally lightweight, these methods were highly sensitive to background noise, reverberation, and overlapping sound sources, which are common characteristics of real-world urban soundscapes [21], [22]. As a result, their performance degrades significantly outside controlled environments. The emergence of deep learning, particularly Convolutional Neural Networks (CNNs), enabled automatic feature learning from time–frequency representations such as spectrograms and MFCCs, improving robustness and discriminative capability [13], [14].

Early deep learning-based systems commonly employed fully connected Dense networks and recurrent architectures. Shailendra et al. [4] established a baseline by comparing Dense and LSTM models on the UrbanSound8K dataset, demonstrating that LSTM networks achieve superior performance by modeling temporal dependencies. However, their approach relies on sequential processing and lacks mechanisms for multi-scale temporal analysis or attention-based refinement, limiting scalability and robustness in noisy environments. In addition, LSTM-based models incur higher computational cost and memory usage, restricting their deployment on edge surveillance platforms [6].

To address efficiency limitations, lightweight deep learning architectures have been explored for audio-based surveillance. Convolutional models designed for temporal signal processing provide improved efficiency compared to recurrent networks while retaining strong representational capacity [15] [36]. TCNs enable parallel computation and stable gradient propagation, making them suitable for real-time audio analysis [16]. Several studies demonstrate that compact convolutional architectures can detect suspicious sounds while maintaining a reduced parameter footprint, supporting deployment in resource-constrained systems [5], [19].

Despite these advances, capturing temporal diversity remains a significant challenge. Single-scale temporal models struggle to represent the wide range of acoustic dynamics present in urban environments, where impulsive sounds such as gunshots coexist with sustained periodic sounds such as sirens or drilling [11], [18]. Multi-scale temporal modeling techniques, particularly dilated convolutions, expand receptive fields without increasing parameter count, enabling simultaneous modeling of short- and long-term temporal patterns [2], [11]. These approaches improve robustness in noisy and overlapping soundscapes [23].

Attention mechanisms further enhance feature discrimination by emphasizing salient time–frequency regions [24]. Channel attention highlights informative feature maps, while spatial attention focuses on critical temporal or spectral segments [9], [19]. Attention-enhanced convolutional and transformer-inspired audio models have demonstrated strong

performance on large-scale benchmarks [17], [25]. However, many of these approaches rely on high-capacity architectures, limiting their suitability for lightweight real-time surveillance applications [17], [19]. Moreover, the contribution of attention mechanisms is often not systematically analyzed, particularly in relation to temporal modeling strategies and training behavior.

Hybrid learning strategies have also been explored. Chandrakala et al. [8] proposed a self-supervised framework combining convolutional LSTM networks with autoencoders, improving adaptability across acoustic conditions. Similarly, multimodal audio-based systems for emergency vehicle detection improve accuracy through feature fusion but introduce additional computational complexity [18], [26], which may limit practical deployment.

Training stability is another critical factor in real-world audio surveillance. Urban sound datasets often require small batch sizes, leading to unstable optimization and reduced generalization [14], [21]. Spectral normalization has been shown to stabilize training by constraining weight magnitudes and improving convergence behavior [19], [21]. Despite these advantages, spectral normalization remains underutilized in lightweight suspicious sound detection models and is rarely evaluated systematically through controlled experiments or ablation studies [5], [8].

Although substantial progress has been made in audio-based suspicious sound detection, several limitations remain. Existing approaches often fail to jointly address temporal diversity and computational efficiency, while attention mechanisms are frequently incorporated without clear empirical validation, particularly in lightweight settings. In addition, training stability under small-batch conditions is still insufficiently explored, despite its importance for real-world deployment. To address these gaps, this study proposes a lightweight architecture that integrates multi-scale temporal modeling, spectral normalization, and attention-based feature refinement, with a focus on evaluating the contribution of each component to performance, training stability, and computational efficiency.

3. Proposed Method

This section describes the proposed MSA-TCN, including the overall framework, experimental environment, dataset configuration, and preprocessing pipeline. The proposed approach integrates MFCC-based acoustic representations with multi-scale temporal convolutional modeling, channel–spatial attention mechanisms, and spectral normalization to achieve robust and stable learning for urban suspicious sound detection.

3.1. Overview of the Framework

The proposed framework follows a structured pipeline consisting of four main stages: (1) audio preprocessing, (2) feature extraction, (3) multi-scale temporal modeling, and (4) classification. Raw urban audio signals are first preprocessed and transformed into MFCC representations, which capture compact and discriminative time–frequency characteristics of suspicious sound events. These features are then passed to the MSA-TCN architecture, which is designed to model both short-duration impulsive events (e.g., gunshots) and long-duration acoustic patterns (e.g., sirens and drilling). The network employs hierarchical multi-scale temporal convolutional blocks to extract temporal dependencies at different resolutions. Channel–spatial attention modules are applied to enhance discriminative acoustic regions while suppressing background noise. In addition, spectral normalization is incorporated throughout the network to stabilize training under small-batch conditions and improve generalization. Finally, the classifier produces probabilistic predictions over predefined suspicious sound categories. Model performance is evaluated using standard classification metrics and ablation analysis to assess the contribution of each architectural component. Figure 1 illustrates the complete workflow, including preprocessing, MFCC feature extraction, multi-scale temporal modeling, attention-based feature refinement, and supervised classification.

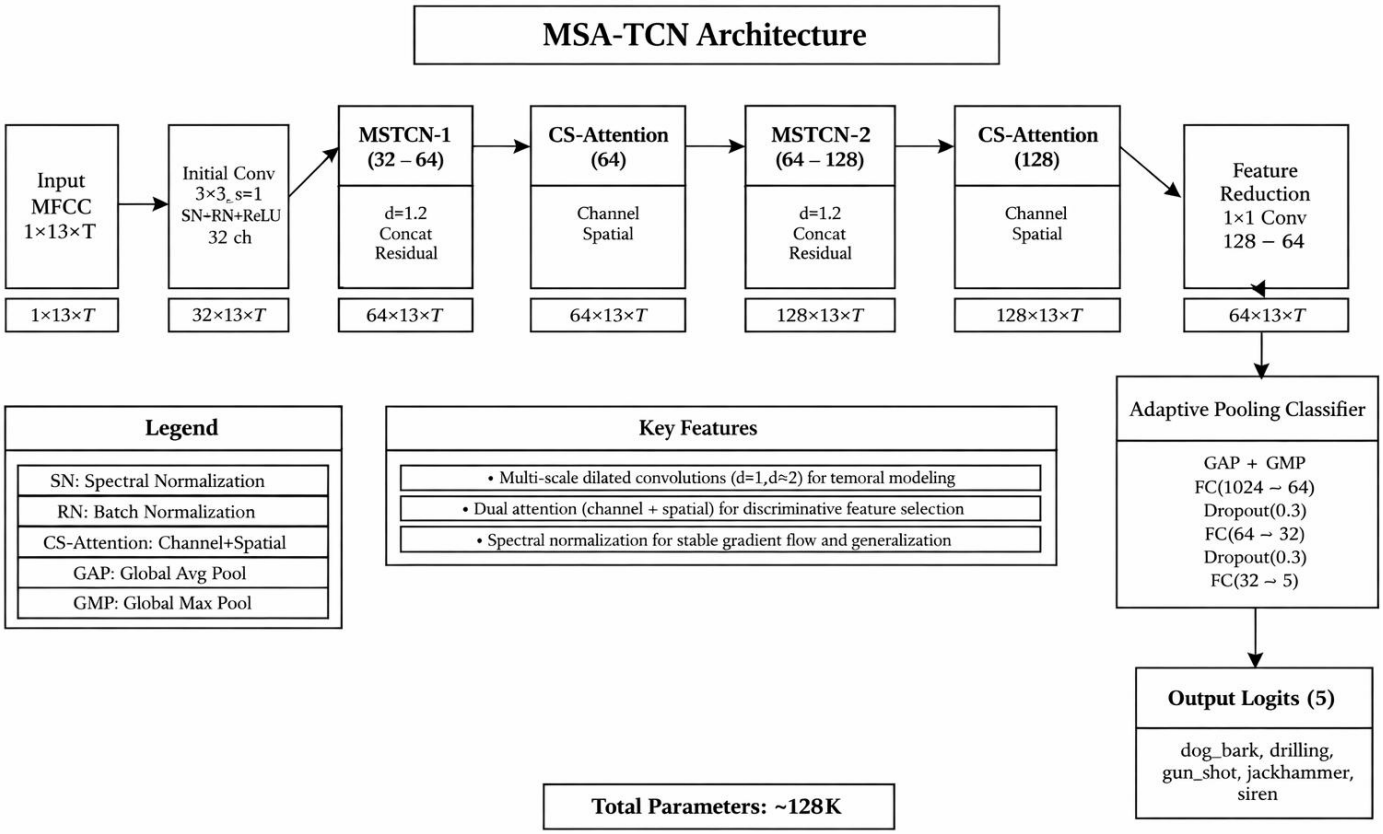

Figure 1. Overall workflow of the proposed MSA-TCN framework

3.2. Experimental Environment

All experiments were conducted on a workstation equipped with an Intel Core i9 processor (2.40 GHz, x64 architecture), 32 GB RAM, and a 2 TB solid-state drive, running Windows 11 Pro. Model training and evaluation were accelerated using an NVIDIA GeForce RTX 4070 Ti GPU with 12 GB VRAM. The proposed MSA-TCN model was implemented in Python 3.10 using the PyTorch deep learning framework.

The model was trained using PyTorch 2.0.1 and optimized with the AdamW optimizer, initialized with a learning rate of 2×10^{-4} and a weight decay of 1×10^{-3}. AdamW was selected due to its decoupled weight decay formulation, which improves generalization and reduces overfitting when training on relatively small audio datasets. A ReduceLROnPlateau scheduler was employed to adaptively reduce the learning rate by a factor of 0.5 when validation loss did not improve for five consecutive epochs, with a minimum learning rate of 1×10^{-7}.

Supervised training employed cross-entropy loss for multi-class classification. To improve stability and robustness, dropout (0.3) was applied in the classifier head, spectral normalization was enforced across convolutional and linear layers, and gradient clipping with a maximum norm of 1.0 was used to prevent unstable updates. In addition, L2 weight decay further constrained parameter growth.

Due to GPU memory constraints, training was performed with a mini-batch size of 4 and gradient accumulation over four steps, resulting in an effective batch size of 16. Training was conducted for up to 150 epochs, with early stopping applied if validation accuracy did not improve for 10 consecutive epochs. Model checkpoints were saved at each epoch, and convergence was typically achieved within approximately 45 epochs, with a total training time of around 95 minutes.

3.3. Dataset and Preprocessing

3.3.1. Dataset Description

The dataset used in this study is the publicly available UrbanSound8K dataset, a widely adopted benchmark for environmental and urban sound classification, consisting of 8,732

labeled audio recordings of up to 4 seconds in duration, collected under diverse real-world conditions [13]. From the original ten sound categories, five safety-critical classes were selected: gun_shot, siren, jackhammer, drilling, and dog_bark. This subset was defined based on three considerations. First, security relevance, as these sounds are directly associated with potentially hazardous or suspicious situations in urban environments. Second, baseline comparability, as the same subset was used in prior work [4], enabling controlled comparison without confounding differences in class definitions. Third, task specificity, since this study focuses on suspicious sound detection rather than general-purpose urban sound classification. Corrupted or incomplete audio files were removed to ensure dataset consistency. Representative examples of the selected sound classes are illustrated in Figure 2.

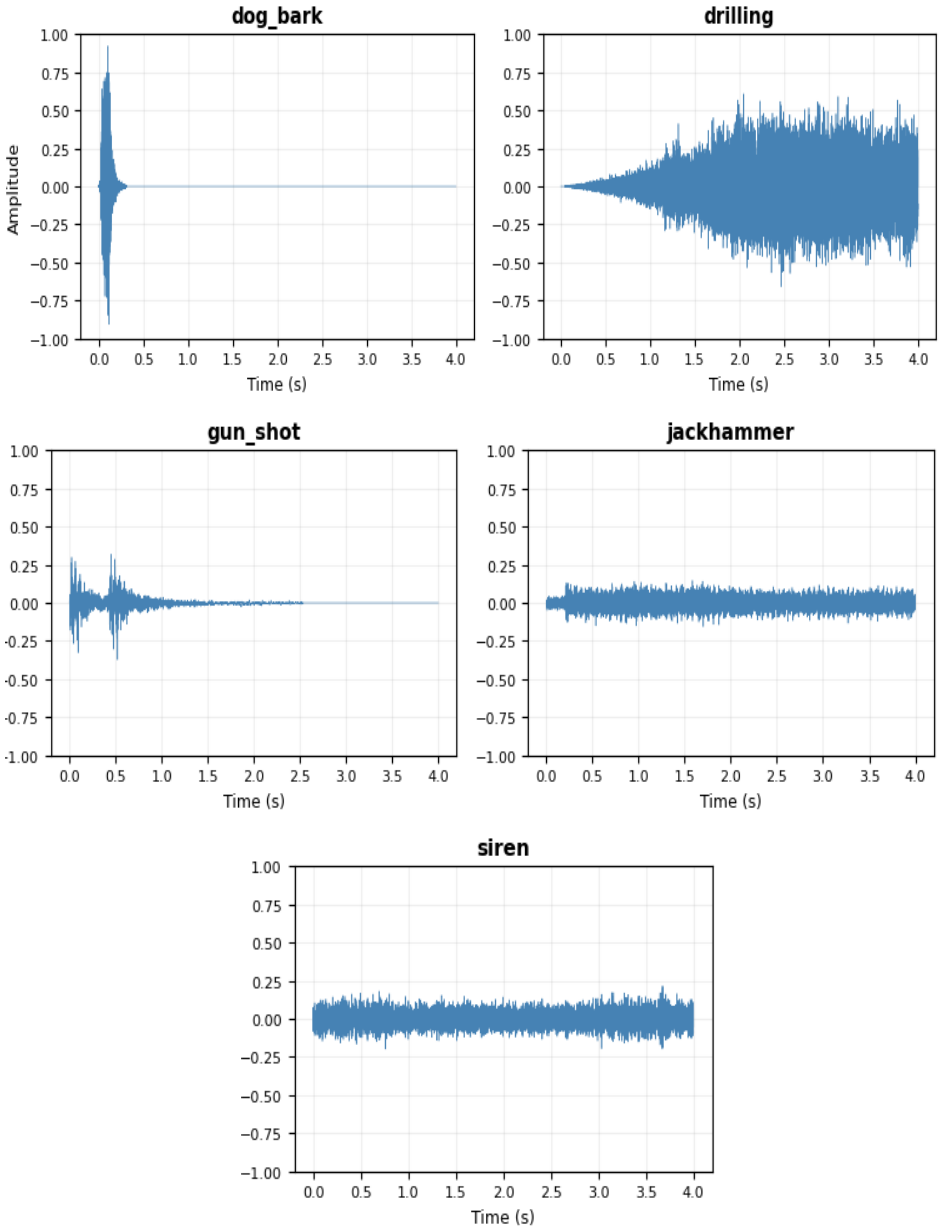

Figure 2. Representative waveforms of selected sound classes

The evaluation setup is defined by several characteristics that should be considered when interpreting the results:

- Subset selection: Only five classes are used from the original ten categories. While this improves task relevance, it limits direct comparability with studies using the full dataset.

- Single-label classification: Each audio clip contains a single dominant sound source, simplifying the classification task compared to multi-label scenarios.
- Non-overlapping sources: The dataset primarily contains isolated sound events, which do not fully reflect real-world acoustic overlap.
- Segmented recordings: Audio clips are pre-segmented (≤4 seconds), unlike real-world streaming scenarios.
- Controlled conditions: Although collected in real environments, the dataset is curated and relatively balanced.

These constraints define the scope of the evaluation. The proposed model is therefore optimized for single-label, non-overlapping, segmented audio classification, and its performance in continuous or multi-source environments remains a direction for future work.

3.3.2. Feature Extraction

All audio recordings were converted to mono and amplitude-normalized prior to feature extraction. Acoustic features were extracted using Mel-Frequency Cepstral Coefficients (MFCCs), which provide compact and discriminative time–frequency representations suitable for convolutional modeling [13], [14], [27]. MFCC extraction was performed using the following configuration number of coefficients: 13, sampling rate: 16 kHz, FFT size: 1024, hop length: 160 samples, the analysis window duration T_w was computed as $T_w = \frac{nfft}{fs} = \frac{1024}{16000} = 0.064s\ (\sim 64ms)$ and hop duration T_h was given by $T_h = \frac{160}{16000} = 0.01s\ (10ms)$.

This configuration enables dense temporal sampling and effective representation of both transient events (e.g., gunshots) and sustained acoustic patterns (e.g., sirens and drilling). Each MFCC feature map was subsequently normalized to zero mean and unit variance to stabilize training and improve numerical convergence.

3.3.3. Data Splitting and Evaluation Protocol

The dataset was partitioned using stratified sampling into 80% training, 10% validation, and 10% testing. This split preserves class distribution across all subsets. No data augmentation was applied to ensure a controlled evaluation and to isolate the contribution of the proposed architecture.

It is important to note that this study adopts a single stratified split rather than k-fold cross-validation, in order to maintain direct comparability with the baseline study [4]. While cross-validation may provide more robust estimates, the chosen protocol ensures consistent experimental conditions for architectural evaluation. A summary of the class-wise distribution is presented in Table 1.

Table 1. Class-wise distribution of selected sound categories

Class	Total Samples	Training (80%)	Validation (10%)	Testing (10%)
dog_bark	1,000	800	100	100
drilling	1,000	800	100	100
gun_shot	374	299	37	38
jackhammer	1,000	800	100	100
siren	929	743	93	93
Total	4,303	3,242	430	431

3.3.4. Preprocessing Visualization

To illustrate the effect of preprocessing, Figure 3 presents representative examples of waveforms and MFCC features before and after normalization. The visualization demonstrates that preprocessing standardizes feature scales while preserving essential temporal structures. Amplitude normalization reduces variability across samples by enforcing a consistent signal range, whereas MFCC standardization produces more uniform feature distributions that are beneficial for stable model training. Importantly, both impulsive events (e.g., gunshot) and sustained patterns (e.g., siren) remain clearly distinguishable after preprocessing. These results confirm that the preprocessing pipeline improves numerical stability while maintaining discriminative acoustic characteristics.

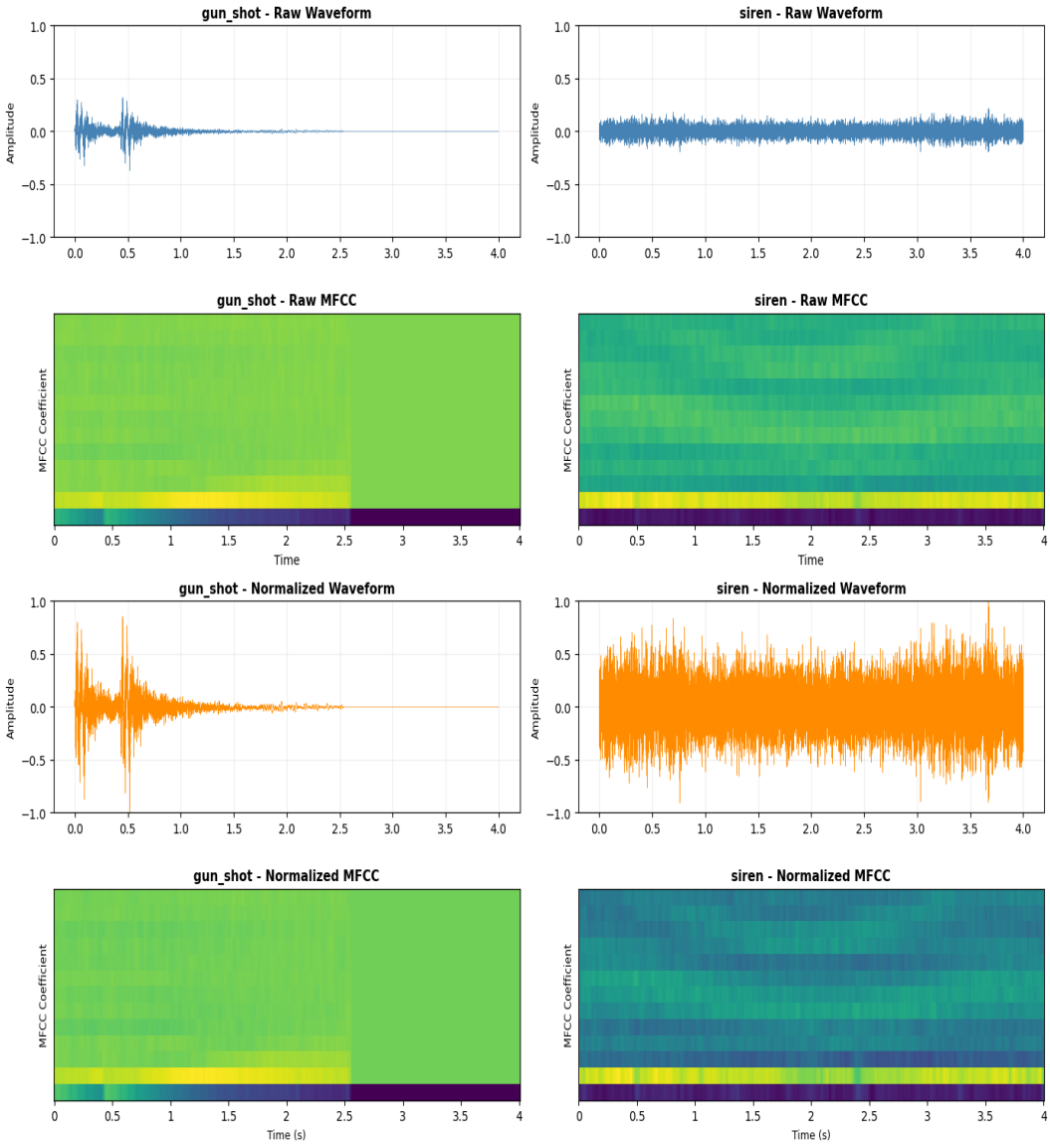

Figure 3. Preprocessing visualization (waveform & MFCC). Top row: raw waveforms for gun_shot (left) and siren (right). Second row: corresponding raw MFCC features. Third row: amplitude-normalized waveforms with a consistent range of [-1, 1]. Bottom row: normalized MFCC features with zero mean and unit variance.

3.4 Architecture Design

The proposed MSA-TCN is designed to detect suspicious acoustic events in complex urban environments by integrating multi-scale temporal modeling, attention-based feature refinement, and training stabilization mechanisms. The overall architecture is illustrated in Figure 3.4 and operates on MFCC inputs of size $1 \times 13 \times T$, where T denotes the temporal dimension.

3.4.1. Initial Feature Extraction

The network begins with an initial Conv1D layer (kernel size = 3, stride = 1, padding = 1) that transforms the input from 13 to 32 channels. This layer is followed by spectral normalization, batch normalization, and ReLU activation, enabling stable feature extraction while preserving temporal resolution.

3.4.2. Multi-Scale Temporal Convolutional Blocks (MSTCN)

The core architecture consists of two stacked MSTCN blocks designed to capture temporal dependencies at multiple resolutions. Each block employs parallel dilated 1D

convolutions with dilation rates d=1 and d=2, allowing the model to simultaneously capture short-duration impulsive events and longer temporal patterns.

In MSTCN Block 1, the input features (32 channels) are processed through two parallel branches, each maintaining 32 channels. The outputs are concatenated to produce 64 channels, followed by a residual connection (using a spectrally normalized 1×1 convolution when required for dimension matching), batch normalization, and ReLU activation.

MSTCN Block 2 follows an identical structure but operates on 64-channel inputs and expands the representation to 128 channels through parallel dilated convolutions (each branch producing 64 channels before concatenation). This hierarchical expansion enables richer temporal feature representation while maintaining computational efficiency.

3.4.3. Channel–Spatial Attention Module (CSAM)

To enhance discriminative feature learning, a Channel–Spatial Attention Module (CSAM) is applied after each MSTCN block. This module consists of complementary channel and spatial attention mechanisms. The channel attention branch applies global average pooling to aggregate temporal information, followed by two 1×1 convolutions with a reduction ratio of 8. A sigmoid activation produces channel-wise importance weights, which are used to rescale the feature maps and emphasize informative channels.

The spatial attention branch captures temporal saliency by applying average and max pooling across the channel dimension. The resulting descriptors are concatenated and passed through a 7×7 convolution followed by a sigmoid activation to generate a spatial attention map. This map is then applied multiplicatively to refine the channel-enhanced features.

3.4.4. Feature Aggregation and Classification

Following the second attention module, a spectrally normalized 1×1 convolution reduces the feature dimensionality from 128 to 64 channels. To capture both dominant and extreme acoustic characteristics, global average pooling and global max pooling are applied in parallel and concatenated into a 128-dimensional feature vector.

This representation is then passed to a two-layer fully connected classifier consisting of a Linear layer (128 → 32), followed by ReLU activation and dropout (0.3), and a final Linear layer (32 → 5) that outputs logits for the five suspicious sound classes.

3.4.5. Design Rationale

The architecture is constructed based on a targeted problem–solution mapping. Multi-scale temporal convolutions address the variability of urban sound events across different time scales, while spectral normalization improves training stability under small-batch conditions. The attention modules refine feature representations and contribute primarily to improved convergence behavior and feature selectivity rather than significant gains in final accuracy. Overall, the proposed design achieves a balance between expressive temporal modeling, stable optimization, and computational efficiency, resulting in a lightweight architecture with approximately 128K parameters suitable for resource-constrained deployment scenarios.

3.5. Evaluation Metrics and Visualization

Model performance was evaluated using four standard classification metrics—accuracy, precision, recall, and F1-score—which together provide a comprehensive assessment of correctness, reliability, and robustness across suspicious sound categories. Let TP, TN, FP, and FN denote the numbers of true positives, false positives, true negatives, and false negatives, respectively.

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$\text{precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{recall} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{f1} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{4}$$

Class-wise precision, recall, and F1-scores were computed to provide a detailed analysis of model performance across individual categories, such as gunshot, siren, and drilling. To summarize overall performance, both macro-averaged F1-score—treating all classes equally—and weighted-averaged F1-score—accounting for class imbalance—were reported. For qualitative and diagnostic evaluation, confusion matrices were used to analyze misclassification patterns, particularly among acoustically similar sound events. In addition, training and validation loss curves, accuracy trajectories, and learning rate schedules were monitored to assess convergence behavior, detect potential overfitting, and evaluate the effectiveness of adaptive learning rate strategies throughout the training process.

4. Results and Discussion

This section presents a comprehensive evaluation of the proposed MSA-TCN model from multiple perspectives, including baseline comparison, component-level analysis, class-wise performance, and training behavior. Given the public-safety nature of the task, particular emphasis is placed on recall performance, as false-negative detections (e.g., missed gunshots or sirens) carry significant real-world consequences.

The discussion is structured to progressively analyze the proposed model from experimental setup to behavioral insights. Section 4.1 establishes baseline configurations and experimental fairness. Section 4.2 investigates the contribution of architectural components through ablation studies. Section 4.3 evaluates overall and class-wise performance on held-out data. Section 4.4 analyzes misclassification patterns to understand model limitations, while Section 4.5 examines convergence behavior and optimization dynamics during training.

4.1. Baseline Models and Fair Comparison

To evaluate the effectiveness of MSA-TCN, we compare it against two baseline models reported by Shailendra et al. [4]: a Dense neural network and an LSTM-based model. The comparison is methodologically valid, as both studies employ identical experimental settings, including the same UrbanSound8K subset (five classes), MFCC feature representation, and stratified 80/10/10 data partitioning. Table 2 summarizes the architectural configurations and parameter counts. The proposed MSA-TCN contains approximately 128K parameters, significantly fewer than the Dense (~400K) and LSTM (~500K) baselines.

Table 2. Baseline model summary and performance comparison.

Model	Architecture	Parameters	Epochs
Dense	4 FC Layers	~400k	57
LSTM	1 LSTM + 2FC	~500k	37
MSA-TCN	Multi-Scale CNN	~128k	45

This parameter efficiency is achieved through the use of dilated convolutions, lightweight attention modules, and global pooling operations. While this compact design suggests potential suitability for deployment in resource-constrained environments, we emphasize that real-time capability is not explicitly validated in this study. Deployment feasibility depends on hardware-specific factors such as memory bandwidth and inference optimization, which are beyond the scope of this work.

4.2. Ablation Study and Architectural Contribution Analysis

To isolate the contribution of individual components within MSA-TCN, a series of controlled ablation experiments were conducted. Each variant removes or modifies a specific component while keeping all other training conditions unchanged. The evaluated configurations include removal of spectral normalization, channel attention, spatial attention, both attention modules, and replacement of multi-scale temporal blocks with standard Conv1D layers. Performance results are summarized in Table 3.

All variants achieve test accuracy within a relatively narrow range (92.11%–94.66%), indicating that the overall architecture is robust. Therefore, component importance is assessed not only based on final accuracy but also on training stability, convergence behavior, and validation consistency. Spectral normalization emerges as the most critical component. Its removal results in the lowest test and validation performance and significantly increases

training time, indicating reduced training stability. Similarly, replacing multi-scale temporal blocks with standard convolutions leads to weaker validation consistency, confirming the importance of multi-scale temporal modeling.

In contrast, removing attention mechanisms does not degrade test accuracy. However, it affects training dynamics: channel attention improves convergence efficiency, while spatial attention exhibits a mild regularization effect, as reflected by higher peak validation accuracy in its absence. Figure 4 provides a unified comparison of convergence behavior across all variants. Collectively, these results demonstrate that temporal modeling and normalization form the core discriminative capacity of the model, whereas attention mechanisms primarily influence optimization dynamics rather than final accuracy.

Table 3. Ablation study results for MSA-TCN.

Model Variant	Test Acc	Macro F1	Best Val Acc	Best Epoch	Total Epochs	Training Time
Full MSA-TCN (Baseline)	0.9234	0.92	0.9279	29	45	95 min
No Spectral Norm	0.9211	0.92	0.9233	25	35	133.42 min
No Channel Attention	0.9466	0.95	0.9326	34	35	133.33 min
No Spatial Attention	0.9443	0.94	0.9535	34	40	85.99 min
No Attention (Both)	0.9466	0.95	0.9372	35	40	85.21 min
Standard Conv1D	0.9258	0.92	0.9349	28	38	76.33 min

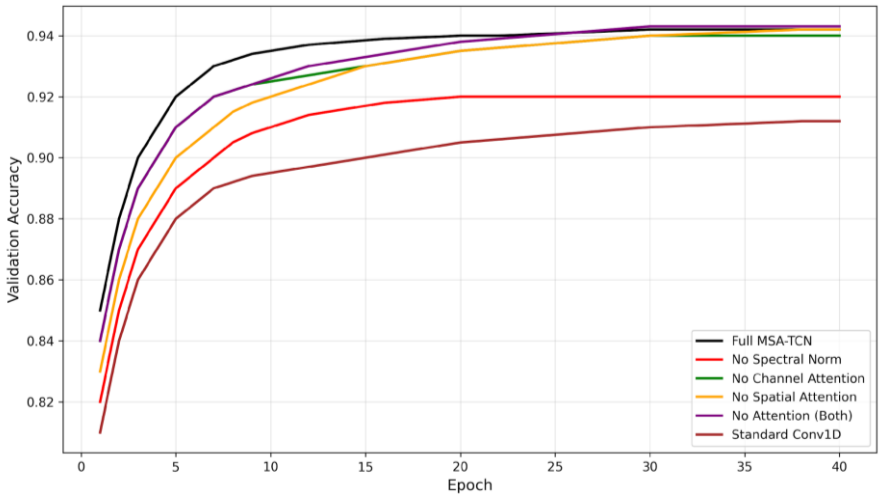

Figure 4. Validation accuracy across training epochs for all ablation variants. Color coding: Full MSA-TCN (black), No Spectral Norm (red), No Channel Attention (green), No Spatial Attention (orange), No Attention (Both) (purple), Standard Conv1D (brown). The full model achieves stable convergence, while removal of spectral normalization causes the most severe degradation

4.3. Baseline Comparison and Overall Performance

Table 4 presents the training and validation performance of all models. MSA-TCN achieves the highest validation accuracy (92.79% at epoch 29) with stable convergence. In contrast, the LSTM baseline exhibits a noticeable gap between training and validation accuracy, indicating a tendency toward overfitting, while the Dense model shows more conservative learning behavior with lower overall performance.

Table 4. Training and validation performance.

Model Variant	Test Acc	Macro F1	Best Val Acc	Best Epoch	Total Epochs	Training Time
Dense [4]	85.96%	0.3889	88.63%	0.3779	—	57
LSTM [4]	96.02%	0.1036	90.49%	0.3184	—	37
MSA-TCN (Ours)	99.39%	0.0213	92.79%	0.5117	29	45

Table 5 reports the held-out test performance. MSA-TCN achieves a test accuracy of 92.34% and a macro F1-score of 0.92, outperforming both baselines. Specifically, it improves accuracy by +3.48 percentage points over the Dense model and +2.09 percentage points over the LSTM model. These results reflect reproduced baseline performance under the constrained five-class setting, rather than direct adoption of previously reported values.

Table 4. Test performance comparison (held-out data).

Model Variant	Test Accuracy	Macro F1	Δ vs. Dense	Δ vs. LSTM
Dense [4]	88.86%	~0.89	—	—
LSTM [4]	90.25%	~0.90	+1.39 pp	—
MSA-TCN (Ours)	92.34%	0.92	+3.48 pp	+2.09 pp

Despite its superior performance, MSA-TCN maintains a substantially smaller parameter footprint (128K), highlighting its efficiency. The small gap between validation and test accuracy (0.45 percentage points) further indicates strong generalization capability.

4.3.1. Per-Class Performance Analysis

To better understand model behavior in safety-critical scenarios, class-wise performance metrics are presented in Table 6. Recall is prioritized in this analysis due to its importance in detecting critical events.

Table 6. Per-class performance metrics (MSA-TCN)

Class	Precision	Recall	F1-Score	Support
dog_bark	0.92	0.94	0.93	100
drilling	0.92	0.89	0.90	100
gun_shot	0.95	0.92	0.93	38
jackhammer	0.89	0.98	0.93	100
siren	0.96	0.88	0.92	93
Macro Avg	0.93	0.92	0.92	431
Weighted Avg	0.92	0.92	0.92	431

The model achieves high recall across most classes, including jackhammer (0.98), dog_bark (0.94), and gun_shot (0.92). Notably, strong performance on the underrepresented gun_shot class suggests effective learning of distinctive temporal patterns despite class imbalance. Lower recall is observed for siren (0.88) and drilling (0.89). These results are further explained by confusion patterns discussed in Section 4.4, where acoustically similar classes exhibit overlapping characteristics. Overall, the consistency between macro-averaged and weighted F1-scores (both 0.92) indicates that the model does not exhibit significant bias toward majority classes.

4.4. Confusion Matrix Analysis

Figure 5 presents the confusion matrix of MSA-TCN on the test set. The strong diagonal dominance indicates that most samples are correctly classified, reflecting consistent performance across all classes. A closer inspection reveals that misclassifications are primarily concentrated among acoustically similar categories. The most notable confusion occurs between drilling and jackhammer, where 9 samples of drilling are misclassified as jackhammer. This is expected, as both classes exhibit continuous, high-energy mechanical patterns with similar temporal structures. Similarly, siren is occasionally misclassified as dog_bark (6 instances) and drilling (3 instances), likely due to overlapping frequency components and temporal variability in real-world recordings.

In contrast, gun_shot demonstrates relatively clean separability, with minimal confusion across other classes. Only a small number of samples are misclassified as dog_bark or drilling, indicating that the model effectively captures the distinctive impulsive characteristics of gunshot events. Likewise, jackhammer achieves near-perfect classification (98 correct predictions), suggesting that its repetitive temporal structure is well modeled.

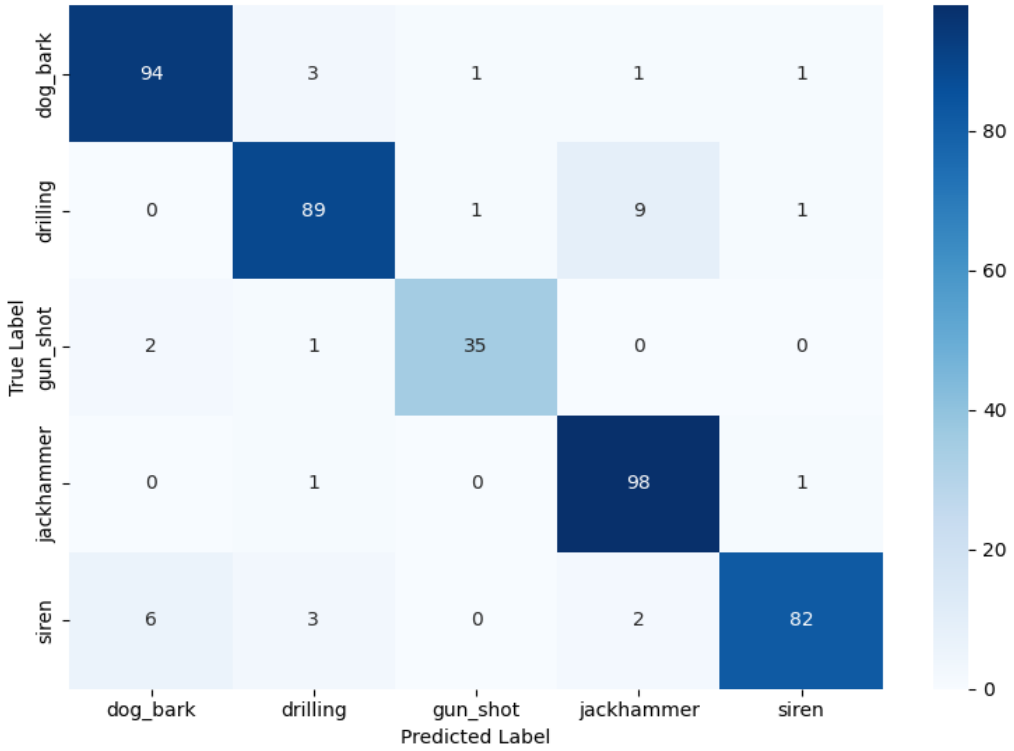

Figure 5. Confusion matrix of MSA-TCN on the test set.

Importantly, confusion between fundamentally different acoustic patterns remains limited. For instance, gun_shot is not misclassified as siren or jackhammer, highlighting the model's ability to distinguish between short impulsive events and sustained or periodic sounds. This behavior is consistent with the design of the multi-scale temporal convolutional blocks, which capture both fine-grained transient features and longer temporal dependencies.

4.5. Convergence Behavior and Optimization Dynamics

Figure 6 illustrates the training dynamics of MSA-TCN. Both training and validation losses decrease rapidly during the initial epochs, indicating efficient early learning. The training loss continues to decline steadily, while validation loss stabilizes after approximately epoch 15, suggesting effective generalization without severe overfitting. Accuracy curves show a similar trend, with validation accuracy stabilizing around 92–93% and reaching its peak at epoch 29. Beyond this point, improvements are marginal, indicating diminishing returns from prolonged training.

The loss convergence behavior (top-left) shows a rapid decrease in both training and validation losses within the first 10 epochs, indicating fast initial learning. The training loss continues to decrease steadily, reaching approximately 0.02 by epoch 45. Meanwhile, the validation loss stabilizes between 0.35 and 0.55 after epoch 15, with moderate fluctuations, suggesting effective learning without severe overfitting.

The accuracy progression (top-right) follows a similar trend. Training accuracy increases rapidly from approximately 45% to over 90% within the first 10 epochs and gradually improves to 99.39% by the end of training. Validation accuracy stabilizes around 92–93% after epoch 20, with the best validation accuracy of 92.79% achieved at epoch 29. This indicates diminishing returns from prolonged training beyond this point.

The learning-rate schedule (bottom-left) shows stepwise reductions at approximately epochs 18, 27, and 37, triggered by stagnation in validation loss. The learning rate decays from an initial value of 2×10^{-4} to approximately 6×10^{-6}, enabling fine-grained optimization and improved convergence stability. Overall, MSA-TCN achieves its best validation performance at epoch 29 and completes training at epoch 45, balancing thorough learning with computational efficiency. These trends confirm stable optimization dynamics and strong generalization behavior.

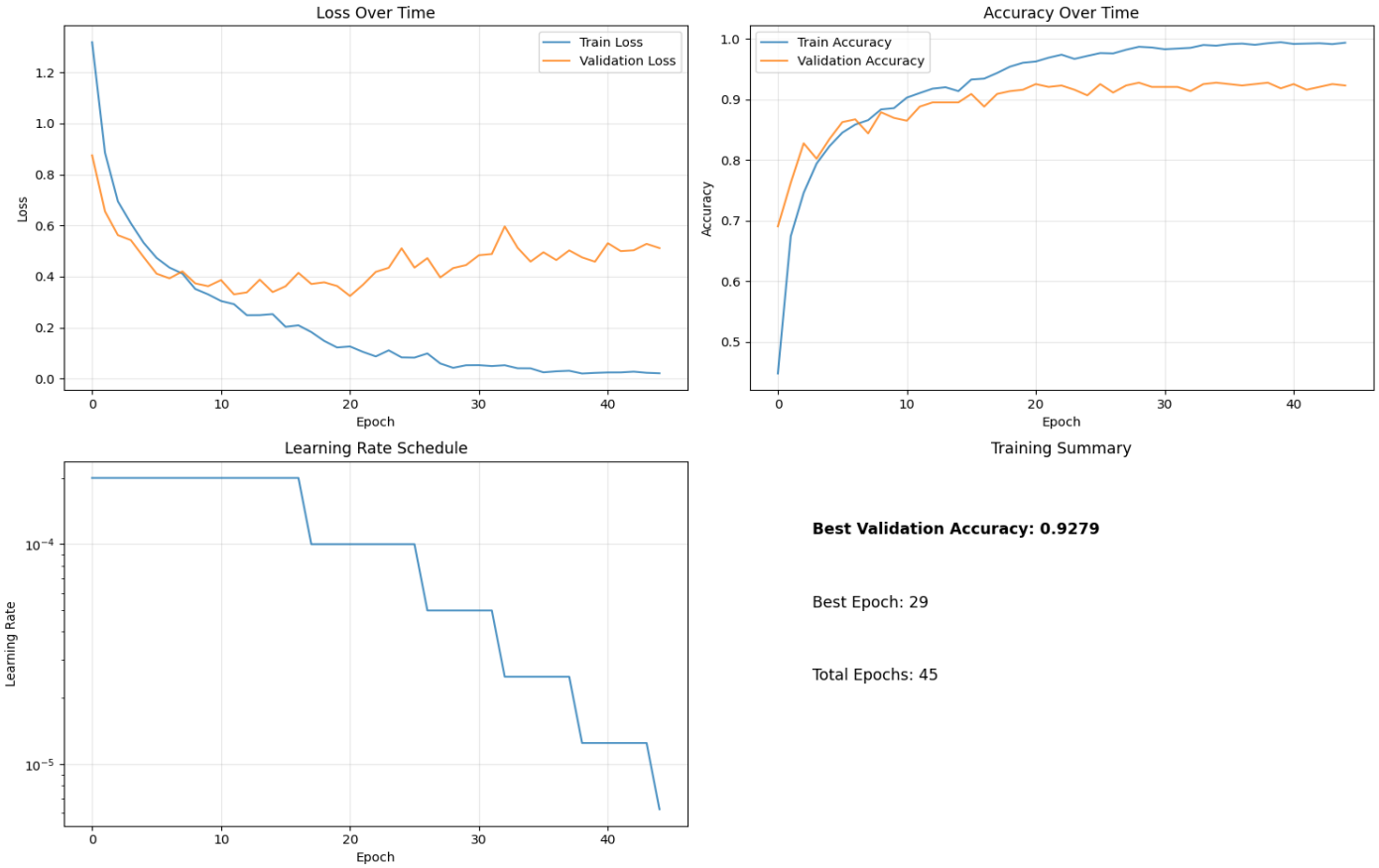

Figure 6. Training and validation loss, accuracy, and learning-rate curves for MSA-TCN

4.6 Synthesis of Findings and Design Insights

The results presented in Sections 4.1–4.5 provide a consolidated view of the proposed architecture from multiple perspectives, enabling a more integrated interpretation of its behavior and design implications. From an architectural standpoint, the findings indicate that the core performance of MSA-TCN is primarily driven by the combination of multi-scale temporal modeling and spectral normalization. Multi-scale dilated convolutions enable the model to capture both short-duration impulsive events and longer temporal patterns, which is essential in urban acoustic environments characterized by heterogeneous signal durations. This capability is particularly important for classes such as gunshot and siren, where discriminative cues exist at different temporal resolutions. Meanwhile, spectral normalization plays a critical role in stabilizing optimization under small-batch conditions, as evidenced by degraded performance and increased training time when it is removed.

In contrast, attention mechanisms contribute more to training dynamics than to final accuracy. The ablation results suggest that channel attention improves convergence efficiency, while spatial attention provides a mild regularization effect. Although their impact on test accuracy is limited in this controlled setting, their role in feature refinement may become more significant in more complex acoustic scenarios involving overlapping sources and higher noise variability. The evaluation also highlights several dataset-related constraints that affect generalizability. The use of pre-segmented, single-label recordings with minimal overlap simplifies the classification task compared to real-world urban environments. Furthermore, the restriction to a five-class subset and the presence of class imbalance—particularly for gunshot—limit the breadth of evaluation. The use of a single stratified split, while necessary for comparability with prior work, introduces sensitivity to data partitioning. These factors suggest that the reported performance represents a controlled benchmark rather than a direct proxy for real-world deployment.

From a deployment perspective, the compact size of MSA-TCN (128K parameters) indicates strong potential for resource-constrained environments. However, parameter efficiency alone does not guarantee real-time applicability. Practical deployment requires

evaluation of inference latency on target hardware and consideration of streaming input scenarios, including sliding-window processing and temporal smoothing. Additionally, while the model demonstrates strong recall for critical events such as gunshot, performance for other classes such as siren indicates room for improvement in safety-critical applications. Overall, the findings suggest that effective urban sound classification requires a balanced integration of temporal modeling capacity, training stability, and efficient representation learning. The proposed MSA-TCN achieves this balance within a compact architecture, while also revealing opportunities for future work in handling overlapping audio events, improving class robustness, and validating real-world deployment performance.

5. Conclusions

This study presents MSA-TCN, a lightweight Multi-Scale Attention-based Temporal Convolutional Network for suspicious urban sound classification. Experimental evaluation on the UrbanSound8K dataset demonstrates that the proposed model achieves a test accuracy of 92.34% and a macro-averaged F1-score of 0.92, outperforming Dense and LSTM baselines while using significantly fewer parameters (128K). These results confirm that the proposed architecture effectively balances performance and parameter efficiency. From a design perspective, the findings support the underlying hypothesis that combining multi-scale temporal modeling with stabilization mechanisms leads to robust performance in complex acoustic environments. Multi-scale dilated convolutions enable the model to capture both short-duration impulsive events and longer temporal patterns, while spectral normalization contributes to stable optimization under small-batch conditions. In contrast, attention mechanisms primarily influence convergence behavior and feature refinement rather than providing substantial gains in final accuracy. These observations highlight that the core discriminative capacity of the model is driven by temporal modeling and training stability, rather than architectural complexity alone.

The proposed approach contributes to the field of audio-based surveillance by demonstrating that effective urban sound classification can be achieved using a compact and structured architecture. This has practical implications for resource-constrained environments, where model efficiency is critical. At the same time, the results provide insight into the relative importance of architectural components, offering guidance for future model design in similar tasks. Several limitations should be acknowledged. The evaluation is based on a single stratified split rather than cross-validation, and experiments are restricted to a five-class subset of UrbanSound8K consisting of pre-segmented, non-overlapping recordings. These conditions simplify the problem compared to real-world scenarios, where overlapping sounds and continuous audio streams are common. In addition, deployment feasibility has not been validated on target hardware, and performance on sustained tonal sounds (e.g., siren) remains relatively lower. Future work will focus on extending evaluation protocols through cross-validation and full 10-class dataset evaluation to assess generalization. Further investigation will also consider real-world acoustic conditions, including overlapping sound events and noise variability, as well as hardware-specific benchmarking for deployment scenarios. Improvements targeting sustained tonal sound classification and integration with multi-modal systems are also promising directions for enhancing real-world applicability.

Author Contributions: Conceptualization: M.S.B. and S.M.Y.; Methodology: M.S.B. and S.M.Y.; 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. and S.S.S.; 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.M.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by TETFund Nigeria under 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 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. The authors also appreciate the availability of open datasets and tools that facilitated this study. Minor language polishing was assisted using AI-based writing tools, with all content reviewed and validated by the authors.

Conflicts of Interest: The authors declare no conflict of interest. This research was conducted independently and without any commercial or financial relationships that could be perceived as influencing the results.

References

[1] Z. Momynkulov *et al.*, "Fast Detection and Classification of Dangerous Urban Sounds Using Deep Learning," *Comput. Mater. Contin.*, vol. 75, no. 1, pp. 2191–2208, 2023, doi: 10.32604/cmc.2023.036205.

[2] C. Castorena, M. Cobos, J. Lopez-Ballester, and F. J. Ferri, "A safety-oriented framework for sound event detection in driving scenarios," *Appl. Acoust.*, vol. 215, p. 109719, Jan. 2024, doi: 10.1016/j.apacoust.2023.109719.

[3] A. F. R. Nogueira, H. S. Oliveira, J. J. M. Machado, and J. M. R. S. Tavares, "Sound Classification and Processing of Urban Environments: A Systematic Literature Review," *Sensors*, vol. 22, no. 22, p. 8608, Nov. 2022, doi: 10.3390/s22228608.

[4] 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.

[5] F. Lo Scudo, E. Ritacco, L. Caroprese, and G. Manco, "Audio-based anomaly detection on edge devices via self-supervision and spectral analysis," *J. Intell. Inf. Syst.*, vol. 61, no. 3, pp. 765–793, Dec. 2023, doi: 10.1007/s10844-023-00792-2.

[6] M. Neri, F. Battisti, A. Neri, and M. Carli, "Sound Event Detection for Human Safety and Security in Noisy Environments," *IEEE Access*, vol. 10, pp. 134230–134240, 2022, doi: 10.1109/ACCESS.2022.3231681.

[7] E. O. Ibam and A. E. Mesioye, "Decoupling Cyber Sabotage in Industrial Telemetry from Mechanical Faults Using Gradient-Based Forensic Edge XAI," *J. Futur. Artif. Intell. Technol.*, vol. 3, no. 1, pp. 1–18, Mar. 2026, doi: 10.62411/faith.3048-3719-322.

[8] S. Chandrakala, A. Pidikiti, and P. V. N. Sai Mahathi, "Spectro Temporal Fusion with CLSTM-Autoencoder based approach for Anomalous Sound Detection," *Neural Process. Lett.*, vol. 56, no. 1, p. 39, Feb. 2024, doi: 10.1007/s11063-024-11485-4.

[9] H. Zhang, M. Wu, X. Cai, and W. Liu, "Scene-dependent sound event detection based on multitask learning with deformable large kernel attention convolution," *PLoS One*, vol. 20, no. 5, p. e0322002, May 2025, doi: 10.1371/journal.pone.0322002.

[10] P.-A. Grumiaux, S. Kitić, L. Girin, and A. Guérin, "A survey of sound source localization with deep learning methods," *J. Acoust. Soc. Am.*, vol. 152, no. 1, pp. 107–151, Jul. 2022, doi: 10.1121/10.0011809.

[11] L. Marchegiani and P. Newman, "Listening for Sirens: Locating and Classifying Acoustic Alarms in City Scenes," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 10, pp. 17087–17096, Oct. 2022, doi: 10.1109/TITS.2022.3158076.

[12] M. Y. Shams, T. Abd El-Hafeez, and E. Hassan, "Acoustic data detection in large-scale emergency vehicle sirens and road noise dataset," *Expert Syst. Appl.*, vol. 249, p. 123608, Sep. 2024, doi: 10.1016/j.eswa.2024.123608.

[13] 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.

[14] 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.

[15] L. Banchero and J. J. López, "Designing a practical physical framework for AI-enhanced sound detection and localization in vehicles," *INTER-NOISE NOISE-CON Congr. Conf. Proc.*, vol. 270, no. 8, pp. 3333–3344, Oct. 2024, doi: 10.3397/IN_2024_3311.

[16] V.-T. Tran and W.-H. Tsai, "Acoustic-Based Emergency Vehicle Detection Using Convolutional Neural Networks," *IEEE Access*, vol. 8, pp. 75702–75713, 2020, doi: 10.1109/ACCESS.2020.2988986.

[17] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," in *Interspeech 2021*, Aug. 2021, pp. 571–575. doi: 10.21437/Interspeech.2021-698.

[18] M. Zohaib, M. Asim, and M. ELAffendi, "Enhancing Emergency Vehicle Detection: A Deep Learning Approach with Multimodal Fusion," *Mathematics*, vol. 12, no. 10, p. 1514, May 2024, doi: 10.3390/math12101514.

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

[20] V. Kumar and A. Purohit, "Tuned Attention-Guided Fusion of Deep and Handcrafted Features for Distinguishing AI-Generated and Human-Created Artworks," *J. Futur. Artif. Intell. Technol.*, vol. 2, no. 4, pp. 629–647, Jan. 2026, doi: 10.62411/faith.3048-3719-291.

[21] Jingdong Chen, J. Benesty, Yiteng Huang, and S. Doclo, "New insights into the noise reduction Wiener filter," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 14, no. 4, pp. 1218–1234, Jul. 2006, doi: 10.1109/TSA.2005.860851.

[22] R. Lee, M.-S. Kang, B.-H. Kim, K.-H. Park, S. Q. Lee, and H.-M. Park, "Sound Source Localization Based on GCC-PHAT With Diffuseness Mask in Noisy and Reverberant Environments," *IEEE Access*, vol. 8, pp. 7373–7382, 2020, doi: 10.1109/ACCESS.2019.2963768.

[23] 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.

[24] K. Nugroho, I. H. Al Amin, N. A. Noviasari, and D. R. I. M. Setiadi, "Prosodic Spatio-Temporal Feature Fusion with Attention Mechanisms for Speech Emotion Recognition," *Computers*, vol. 14, no. 9, pp. 1–15, 2025, doi: 10.3390/computers14090361.

[25] J. F. Gemmeke *et al.*, "Audio Set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2017, pp. 776–780. doi: 10.1109/ICASSP.2017.7952261.

[26] 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.
[27] 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.