Vokasi Unesa Bulletin of Engineering, Technology and Applied Science (VUBETA)
https://journal.unesa.ac.id/index.php/vubeta
Vol. 3, No. 3, 2026, pp. 503~516 DOI: 10.26740/vubeta.v3i3.50353
ISSN: 3064-0768

ShuBAM-TFNet: A Lightweight Temporal-Frequency Attention Network for Edge-Oriented Environmental Sound Recognition

Kabiru Muhammed Nasiru[1], Shehu Mohammed Yusuf [2*], Sani Saleh Saminu[3]
[1,2,3,] Department of Computer Engineering, Ahmadu Bello University, Zaria, Nigeria.

Article Info

Article history:

Received January 03, 2026
Revised March 03, 2026
Accepted June 10, 2026

Keywords:

Environmental
Sound Classification
Lightweight Deep Learning
Temporal Frequency Attention
Edge Computing
Audio Signal Processing

ABSTRACT

Efficient environmental sound recognition is essential for smart-city monitoring and edge-based acoustic sensing, yet practical deployment is often constrained by the computational cost and parameter size of deep learning models. This paper presents ShuBAM-TFNet, a lightweight convolutional neural network designed to achieve high recognition accuracy while remaining suitable for resource-limited devices. The proposed architecture processes Log-Mel spectrograms using dilated and time-distributed convolutional layers to capture long-range temporal dependencies and localized temporal patterns without increasing model complexity. A lightweight temporal–frequency attention mechanism is introduced to adaptively emphasize informative regions across time and frequency dimensions, while optional Bottleneck Attention Modules are examined for channel and spatial refinement. The model is evaluated on UrbanSound8K and ESC-10 using standard protocols. On UrbanSound8K, the best-performing variant (without Bottleneck Attention Modules) achieves 91.76 ± 0.90% accuracy with only 70,754 trainable parameters, converging in approximately 26 epochs. Ablation studies confirm that temporal–frequency attention provides the most significant performance gain (≈2–3%), while dilated convolutions accelerate convergence without increasing parameters. Cross-dataset evaluation on ESC-10 achieves 84.00 ± 2.41% accuracy and a macro F1-score of 0.838, demonstrating reasonable generalization. The counterintuitive finding that removing Bottleneck Attention Modules improves performance reveals that channel-spatial attention requires task-specific validation. These results indicate that attention-guided lightweight architectures particularly those emphasizing temporal-frequency modeling over channel-spatial refinement provide a practical alternative for environmental sound recognition on resource-constrained edge devices.

1. INTRODUCTION

Environmental sound classification plays a fundamental role in modern intelligent systems by enabling machines to perceive, interpret, and respond to acoustic events occurring in real-world environments [1][19]. It serves as a core component in applications such as acoustic monitoring, surveillance, smart-city infrastructure, speech-aware systems, and multimedia analysis [31][34]. Among various audio categories,

*Corresponding Author
Email: smyusuf@abu.edu.ng

environmental sounds, non-speech sounds originating from everyday surroundings, provide critical contextual information that supports situational awareness and intelligent decision-making [2][35]. Despite their importance, environmental sounds are inherently difficult to recognize due to their unstructured acoustic patterns, wide intra-class variability, and frequent contamination by background noise, reverberation, and overlapping sound sources [3][30].

In recent years, deep learning approaches have become the dominant paradigm for sound classification because they can automatically learn discriminative representations directly from raw or transformed audio signals [4][18]. Unlike traditional pipelines that rely on handcrafted feature extraction followed by separate classifiers, deep neural networks integrate feature learning and classification within a unified framework and often achieve superior performance [5]. Convolutional neural networks (CNNs) have demonstrated strong capability in capturing time-frequency patterns from spectrogram representations, making them effective for environmental sound recognition tasks [9][11]. Benchmark datasets such as UrbanSound8K and ESC have further accelerated progress by providing standardized evaluation protocols and diverse acoustic conditions [6][7]. However, these performance gains often come at the cost of high computational complexity, large parameter counts, and reliance on powerful hardware such as GPUs, which limits their practicality for real-time and edge-based deployment [21][22].

The central challenge this study addresses arises from the tension between recognition accuracy and computational efficiency. Many state-of-the-art deep learning models achieve high accuracy but are unsuitable for deployment on resource-constrained devices such as IoT sensors, embedded systems, and mobile platforms [20][42]. Lightweight architectures have therefore emerged as a promising direction for edge-based environmental sound recognition [13][27]. Techniques such as dilated convolutions have been introduced to expand receptive fields without increasing parameter counts, enabling better modeling of long-range temporal dependencies [10]. Attention mechanisms, including channel, spatial, and time–frequency attention, further enhance discriminative feature learning by emphasizing salient acoustic regions [8][9][12]. Nevertheless, many existing lightweight models still struggle to jointly balance efficiency, robustness, and classification accuracy in noisy and diverse acoustic environments [14][30]. Furthermore, the assumption that all attention mechanisms universally benefit performance remains largely untested through systematic ablation, leading to potential over-engineering of architectures [14][15].

To bridge this gap, this paper proposes ShuBAM-TFNet, a lightweight deep learning architecture for environmental sound recognition. The proposed model integrates the ShuffleNet backbone for parameter efficiency [13] with dilated and time-distributed convolutional layers to capture both long-range temporal dependencies and localized temporal patterns without increasing model complexity [10][17]. We introduce a lightweight Temporal–Frequency Attention mechanism to adaptively emphasize salient regions across time and frequency domains [9], and we examine optional Bottleneck Attention Modules (BAM) for channel and spatial refinement [12]. Through systematic ablation studies, we identify the optimal configuration that prioritizes temporal-frequency modeling over channel-spatial refinement for environmental sound tasks.

The main contributions of this study are summarized as follows:

i. A novel lightweight architecture, ShuBAM-TFNet, is proposed with the core configuration emphasizing temporal-frequency attention over channel-spatial refinement, achieving an efficient balance between accuracy and computational cost;
ii. Dilated convolutional blocks are employed to expand the receptive field and capture long-term temporal dependencies without increasing parameter count, while time-distributed convolutions improve modeling of temporal slices within spectrogram representations;
iii. A Temporal–Frequency Attention block is designed to enhance discriminative learning by jointly modeling important temporal and frequency patterns, providing the primary mechanism for performance improvement;
iv. Bottleneck Attention Modules are systematically evaluated and found to be optional rather than essential, with their utility depending on dataset characteristics, a counterintuitive finding that provides practical guidance for future architecture design;
v. Through comprehensive ablation studies, we demonstrate that the proposed architecture achieves competitive accuracy with minimal parameters, identifying the optimal configuration that prioritizes temporal-frequency attention over channel-spatial refinement;
vi. Cross-dataset evaluation validates the model's generalization capability across different environmental sound domains without dataset-specific tuning;

.

vii. Systematic ablation analysis provides an important insight: attention mechanisms require careful task-specific validation; channel-spatial attention may over-regularize environmental sound models, while temporal-frequency attention remains consistently beneficial [15][16].

The remainder of this paper is organized as follows: Section 2 describes the proposed methodology, including dataset description, preprocessing steps, model architecture, and training configuration. Section 3 presents the experimental results, ablation studies, and comparisons with state-of-the-art approaches. Section 4 concludes the paper with a summary of findings, limitations, and directions for future research.

Surveys comparing classical machine learning methods such as Support Vector Machines (SVM), k-Nearest Neighbors (KNN), and Random Forests (RF) with modern deep learning models consistently confirm the superiority of deep learning due to its end-to-end feature learning capability and strong transferability across acoustic conditions [30][5]. Convolutional neural networks (CNNs), in particular, have become the dominant choice for environmental sound classification by learning discriminative time–frequency representations directly from spectrogram inputs [9]. Nevertheless, most existing pipelines remain highly sensitive to background noise, recording variability, and acoustic degradation. Common preprocessing strategies, including smoothing filters and feature normalization, provide limited robustness but fail to recover fine temporal–frequency details that are critical for distinguishing acoustically similar sound events, motivating architectural enhancement rather than reliance solely on network depth or parameter count [3][35][47].

Early deep learning frameworks for environmental sound recognition primarily focused on in-domain classification using pre-trained CNNs applied to curated datasets such as UrbanSound8K and ESC [6][7]. While these approaches achieved strong in-domain accuracy, their performance often deteriorated under domain mismatch, unseen acoustic environments, or degraded recording conditions [11][19]. Subsequent studies introduced attention mechanisms and multi-scale feature learning to improve discriminative power. Temporal–frequency attention-based CNNs demonstrated improved robustness by adaptively emphasizing informative spectral and temporal regions [9][48], while channel and spatial attention modules such as Bottleneck Attention Modules (BAM) further refined intermediate feature representations [8][12]. Despite these improvements, many attention-enhanced models remain computationally expensive and unsuitable for deployment on resource-constrained edge devices [20].

To address efficiency constraints, lightweight architectures such as ShuffleNet and other mobile-oriented CNNs have been explored for environmental sound classification [13][27]. Dilated convolutions have been incorporated to expand receptive fields without increasing parameter counts, enabling better modeling of long-range temporal dependencies in audio signals [10][50]. Time-distributed convolutional designs further improve temporal pattern learning across spectrogram frames [17]. However, lightweight models often sacrifice accuracy when dealing with noisy or overlapping sound events, highlighting the need for carefully designed attention mechanisms that preserve discriminative capacity while maintaining computational efficiency [14][30].

Recent research has investigated hardware-aware neural architecture search (NAS) to automatically design compact and efficient audio classifiers for edge deployment [14][24]. Although NAS-based methods achieve competitive accuracy–efficiency trade-offs, they introduce significant search overhead, hardware dependency, and reproducibility challenges [25][39]. Surveys of NAS techniques further emphasize their limited practicality for rapid deployment in real-world acoustic monitoring systems [40]. Alternative approaches based on feature fusion and multimodal sensing improve classification accuracy but require additional sensors or handcrafted features, increasing system complexity and cost [15][49][33].

Self-supervised and contrastive learning strategies have been proposed to enhance feature transferability and reduce labeled data requirements [38]. However, these methods typically assume stable acoustic statistics and perform poorly under severe noise or unseen environmental conditions [29]. CNN–Transformer hybrids and large-scale audio models achieve high accuracy on curated datasets but remain computationally intensive and impractical for low-power devices [43][44].

Across existing literature, three persistent research gaps can be identified: (1) limited joint modeling of temporal and frequency saliency in lightweight architectures, (2) heavy reliance on complex NAS pipelines or large backbones for performance gains, and (3) insufficient component-level analysis through systematic ablation studies [14][30]. The proposed ShuBAM-TFNet directly addresses these gaps by integrating a ShuffleNet backbone with dilated and time-distributed convolutions, Bottleneck Attention Modules, and a lightweight Temporal–Frequency Attention mechanism. Extensive ablation studies and cross-dataset evaluations demonstrate that this design achieves strong accuracy with minimal parameters, making it suitable for real-time environmental sound recognition on resource-constrained edge devices.

2. METHOD

This section presents key details of the experimental setup, including the datasets used, the preprocessing steps applied, the proposed model architecture, and the training and evaluation protocols.

2.1. Dataset Description

The proposed approach was trained and evaluated using two benchmark datasets: UrbanSound8K [6] and ESC-10 [7]. These datasets were selected to assess both in-domain urban sound recognition performance and cross-dataset generalization capability.

2.1.1. UrbanSound8K

UrbanSound8K is a widely used benchmark dataset for urban sound classification. It contains 8,732 labeled audio clips, each with a maximum duration of 4 seconds, distributed across 10 sound classes: air conditioner, car horn, children playing, dog bark, drilling, engine idling, gunshot, jackhammer, siren, and street music. All recordings are provided in WAV format and capture real-world urban environments with varying background noise levels and recording conditions. This diversity makes UrbanSound8K suitable for evaluating models designed to recognize both transient and stationary acoustic events. In this work, UrbanSound8K is used as the primary dataset for training and validation of the proposed ShuBAM-TFNet.

2.1.2. ESC-10

ESC-10 is a curated subset of the larger ESC-50 dataset and serves as a standard benchmark for environmental sound classification. It consists of 400 audio recordings evenly distributed across 10 classes: dog bark, airplane, clapping, crow, door knock, chainsaw, can opening, fireworks, thunderstorm, and vacuum cleaner. Each clip is 5 seconds long and provided in high-quality WAV format. The dataset includes both natural and human-made sounds, making it a balanced testbed for evaluating model robustness. In this study, ESC-10 is used to assess ShuBAM-TFNet's generalization capability under a standardized cross-validation protocol.

2.2 Data Pre-processing

All audio files from UrbanSound8K and ESC-10 were resampled to 16 kHz, balancing the preservation of relevant frequency information with reduced computational cost [8]. A short-time Fourier transform (STFT) [9] was applied using a window size of 1024 samples and a hop length of 512 samples to generate amplitude spectrograms.

The amplitude spectrograms were then projected onto a 64-bin Mel filter bank and converted to the logarithmic scale to produce Log-Mel spectrograms, which are commonly used for environmental sound recognition. The resulting spectrograms were normalized to ensure consistent feature scaling across samples. Figure 1 shows sample Log-Mel spectrograms illustrating the frequency–time patterns for different sound classes in the UrbanSound8K dataset.

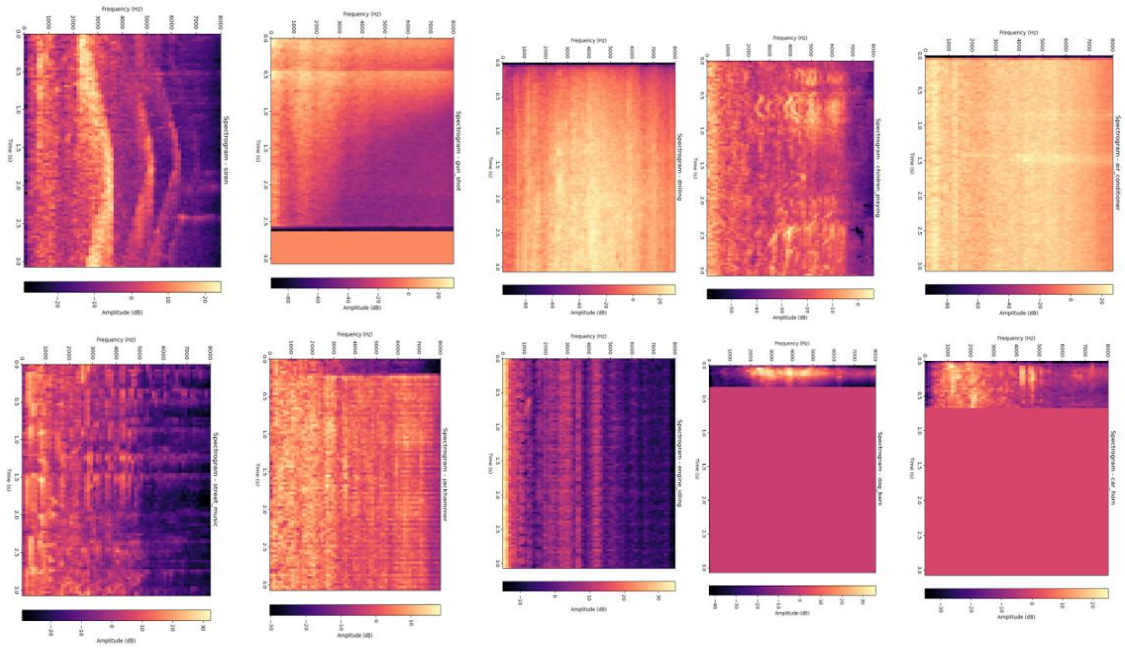

Figure 1: Sample Log-Mel spectrograms representing each sound class in the UrbanSound8K dataset.

2.3 Proposed Model Architecture

Using the log-Mel spectrograms generated during preprocessing, ShuBAM-TFNet processes the inputs through a structured, lightweight architecture designed to capture salient time–frequency patterns. The model receives a single-channel spectrogram of size [1 × 96 × 64] (time × frequency) and outputs class probabilities for 10 sound categories. As illustrated in Figure 2, the architecture is organized into three main stages:

i. early feature extraction using dilated and time-distributed convolutions,
ii. mid-level refinement through temporal–frequency attention, and
iii. high-level classification using a ShuffleNet backbone enhanced with Bottleneck Attention Modules (BAM).

This design maintains a total parameter count below 75K, making the network suitable for resource-limited devices.

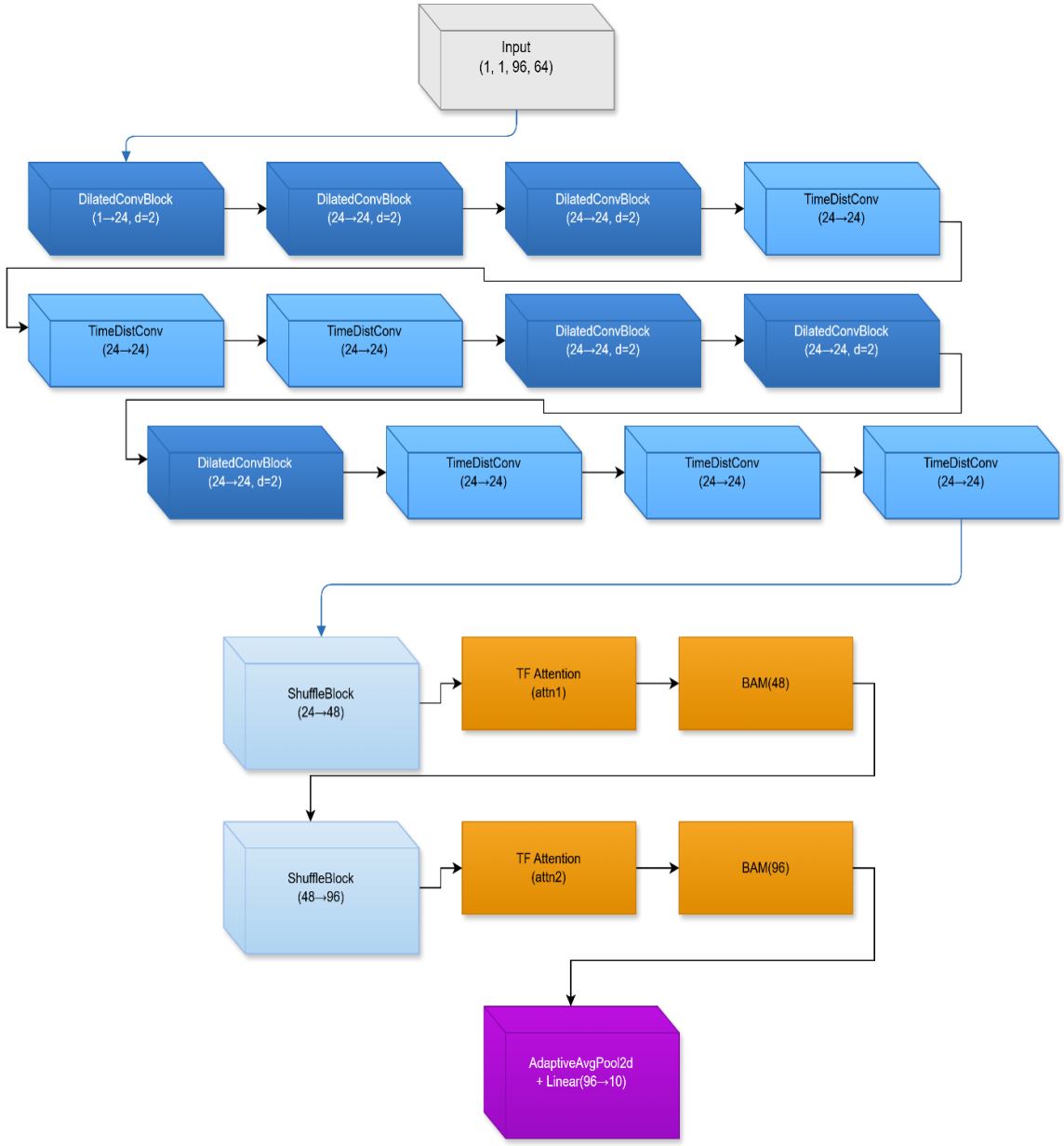

Figure 2: Schematic overview of the ShuBAM-TFNet architecture.

Table 1. Layer-by-layer architecture details of ShuBAM-TFNet (No-BAM variant) with output shapes and parameter counts.

Layer/Block	Input Shape	Operation	Output Shape	Parameters
Input	-	-	(1, 1, 96, 64)	0
Dilated Conv Block 1	(1, 1, 96, 64)	3×3 Conv, dilation=2, 24 filters BatchNorm + ReLU	(1, 24, 96, 64)	240
Time-Distributed Conv 1	(1, 24, 96, 64)	3×1 Conv (temporal), 24 filters BatchNorm + ReLU	(1, 24, 96, 64)	1.752
Dilated Conv Block 2	(1, 24, 96, 64)	3×3 Conv, dilation=2, 24 filters BatchNorm + ReLU	(1, 24, 96, 64)	5.208
Time-Distributed Conv 2	(1, 24, 96, 64)	3×1 Conv (temporal), 24 filters BatchNorm + ReLU	(1, 24, 96, 64)	1.752
Dilated Conv Block 3	(1, 24, 96, 64)	3×3 Conv, dilation=2, 24 filters BatchNorm + ReLU	(1, 24, 96, 64)	5.208

Layer/Block	Input Shape	Operation	Output Shape	Parameters
Time-Distributed Conv 3	(1, 24, 96, 64)	3×1 Conv (temporal), 24 filters BatchNorm + ReLU	(1, 24, 96, 64)	1.752
TF-Attention Module	(1, 24, 96, 64)	Temporal attention + Frequency attention Residual fusion with learnable scaling	(1, 24, 96, 64)	1.024
ShuffleNet Block 1	(1, 24, 96, 64)	Grouped convolutions + Channel shuffle Output channels: 48	(1, 48, 48, 32)	23.520
ShuffleNet Block 2	(1, 48, 48, 32)	Grouped convolutions + Channel shuffle Output channels: 96	(1, 96, 24, 16)	29.280
Global Average Pooling	(1, 96, 24, 16)	AdaptiveAvgPool2d → (1, 96, 1, 1)	(1, 96)	0
Fully Connected Layer	(1, 96)	Linear (96 → 10)	(1, 10)	970
Total Parameters				70.754

Table 1. Detailed layer-by-layer architecture of the primary ShuBAM-TFNet (No-BAM variant). For the Full model (with BAM), an additional 1,636 parameters are added between ShuffleNet blocks. All convolutional layers use stride=1 and padding=1 (dilated) or padding=0 (time-distributed) to maintain spatial dimensions where appropriate.

2.3.1 Feature Extraction

The input spectrogram is first processed by a sequence of convolutional layers designed to expand the receptive field while preserving spatial resolution. The initial stage consists of three dilated convolution blocks followed by three time-distributed convolution blocks, each operating with 24 channels. The dilated convolutions use a 3 × 3 kernel with a dilation rate of 2, allowing the network to capture long-range temporal dependencies without increasing parameter count or reducing resolution [10]. Each convolutional layer is followed by batch normalization and ReLU activation.

The time-distributed convolution blocks apply 1D convolutions along the temporal axis using 3 × 1 kernels, treating each frequency bin independently [11]. This design enables the model to capture short-duration events such as gunshots and drilling while maintaining feature map dimensions of [24 × 96 × 64]. This dilated–time-distributed pattern is repeated twice, resulting in 12 convolutional layers.

2.3.2 Temporal–Frequency Attention Module

Model The extracted features are passed to the first Temporal–Frequency Attention module (TF-Attn1), which adaptively emphasizes informative patterns across both temporal and frequency dimensions. Inspired by self-attention mechanisms, TF-Attn1 computes separate attention maps along the time and frequency axes [9]. For an input tensor, $x \in R^{B\times C\times T\times F}$ the temporal branch generates an attention matrix $A_t \in R^{T\times T}$, while the frequency branch produces $A_f \in R^{F\times F}$. The final output is obtained through residual fusion:

$$O = \gamma_t O_t + \gamma_f O_f + x \quad (1)$$

where γ_t and γ_f are learnable scaling parameters. This module introduces approximately 1K additional parameters while improving sensitivity to salient spectro-temporal patterns.

2.3.3 ShuffleNet Backbone with Bottleneck Attention

The refined features are forwarded to a ShuffleNet backbone, which employs grouped convolutions and channel shuffling to reduce computational cost while preserving representational capacity [13]. Two ShuffleNet blocks are used, increasing the channel dimensions from 24 → 48 → 96.

Between these blocks, Bottleneck Attention Modules (BAM) [12] are inserted to refine feature maps through joint channel and spatial attention. Each BAM uses global pooling and a lightweight MLP for channel attention, along with a 7 × 7 convolution for spatial attention. Each module adds approximately 500 parameters, offering improved noise suppression with minimal overhead. After the backbone, global average pooling reduces the feature map to a 96-dimensional vector, which is passed through a fully connected layer to generate class probabilities. A second temporal–frequency attention module (TF-Attn2) is applied at this stage to refine high-level representations. The complete forward process is summarized as:

$$y=\text{Softmax}(\text{Linear}(\text{TF-Attn2}(\text{ShuffleNet}(\text{TF-Attn1}(x))))) \quad (2)$$

2.4. Experimental Setup

This section outlines the experimental configuration adopted to evaluate the effectiveness and robustness of the proposed ShuBAM-TFNet architecture. It describes the implementation environment, training protocol, dataset splitting strategies, and evaluation metrics used to ensure fair comparison, reproducibility, and reliable performance assessment across different experimental settings.

2.4.1 Implementation Details

All experiments were executed on a workstation equipped with an Intel Core i9 processor (2.40 GHz, x64 architecture), 32 GB of system 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 of VRAM. The proposed ShuBAM-TFNet architecture was implemented in Python 3.10 using the PyTorch deep learning framework (version 2.2.0).

To ensure reproducibility and result stability, all experiments were conducted using two fixed random seeds (42 and 43). The model was trained with the Adam optimizer, an initial learning rate of 1×10^{-3}, and a mini-batch size of 16. Training ran for up to 50 epochs, with early stopping based on validation accuracy and a patience of 8 epochs to prevent overfitting and unnecessary computation.

2.4.2 Training Protocol

UrbanSound8K: The dataset was split into 80% training and 20% validation using a stratified strategy, resulting in 6,985 training samples and 1,747 validation samples. This split preserved class balance and avoided fold overlap [6].

ESC-10: We followed the official 5-fold cross-validation protocol. Each fold contains 80 training and 80 test samples, and performance is reported as the mean and standard deviation across all folds [7].

2.4 Evaluation Metrics

Model performance was evaluated using Accuracy, Precision, Recall, and F1-Score on the target-domain test set. Let TP, FP, TN, and FN denote the numbers of true positives, false positives, true negatives, and false negatives, respectively.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5}$$

$$\text{F1- score} = 2 \times \frac{(precision \times recall)}{(precision + recall)} \tag{6}$$

For ESC-10, mean and standard deviation values across the five folds are reported to reflect consistency. All metrics were computed using scikit-learn, based on predictions from the best validation checkpoint.

3. RESULTS AND DISCUSSION

This section presents the experimental evaluation of the proposed ShuBAM-TFNet architecture, focusing on its accuracy, efficiency, convergence behavior, and generalization capability across benchmark environmental sound datasets. We first conduct comprehensive ablation studies on UrbanSound8K to quantify the contribution of individual architectural components. The model's performance is then analyzed in detail on UrbanSound8K and ESC-10, followed by a comparative assessment against recent lightweight state-of-the-art approaches. Finally, we discuss key findings, limitations, and practical implications.

3.1 Ablation Studies on UrbanSound8K

To analyze the contribution of each architectural component, we conducted a set of ablation experiments on the UrbanSound8K dataset. Four configurations were evaluated: the Full model (with all components), a variant without Bottleneck Attention Modules (No-BAM), a variant without Temporal-Frequency Attention (No-TF-Attn), and a variant without dilated convolutions (No-Dilated). Based on the ablation results, we present the No-BAM variant as the primary ShuBAM-TFNet architecture, while the Full model serves as an experimental comparison.

We trained each configuration twice using different random seeds (42 and 43) to ensure stability and reproducibility. For each run, we selected the best-performing checkpoint based on validation accuracy. Table

2 reports the validation and test accuracies for each run, along with the mean accuracy, parameter count, average best epoch, and training time.

Table 2. Ablation study on UrbanSound8K (mean ± standard deviation over two runs).

Variant	Val / Test Acc (Run 1)	Val / Test Acc (Run 2)	Mean Val/Test Acc ± Std	Parameters	Mean Best Epoch	Mean Training Time
Full	89.41%, 90.90%	91.13%, 92.39%	90.16 ± 1.05%	72.390	27.5	2h 30m
No-BAM	91.13%, 92.39%	91.13%, 92.39%	91.76 ± 0.90%	70.754	26.0	2h 26m
No-TF-Attn	87.06%, 90.27%	87.06%, 90.27%	88.67 ± 2.26%	47.606	24.5	2h 17m
No-Dilated	88.95%, 88.95%	88.95%, 88.95%	88.95 ± 0.00%	72.390	36.0	3h 06m

The ablation results reveal several important trends. Notably, the No-BAM variant achieves the highest overall accuracy (91.76 ± 0.90%), outperforming the Full model by 1.6% while using 1,636 fewer parameters. This counterintuitive finding suggests that, for UrbanSound8K, the additional channel–spatial gating introduced by BAM may over-regularize the model, suppressing subtle but informative acoustic cues rather than enhancing them. Classes with overlapping spectral characteristics (e.g., children_playing vs. street_music) appear particularly susceptible to this effect, as shown in the confusion matrices (Figure 5).

In contrast, removing the Temporal-Frequency Attention blocks results in the most significant performance degradation. The No-TF-Attn variant drops to 88.67 ± 2.26%, confirming that temporal–frequency attention is the core innovation and contributes approximately a 2–3% absolute accuracy gain by adaptively emphasizing discriminative patterns across the time–frequency plane.

The No-Dilated variant achieves a stable accuracy of 88.95%, but requires substantially more training epochs (~36) to converge. This indicates that dilated convolutions accelerate convergence without increasing parameter count, while effectively capturing long-range temporal dependencies.

Figures 3 and figure 4 show training and validation curves for all variants across both runs, illustrating differences in convergence speed and stability. The No-BAM and Full models converge at similar rates, while No-TF-Attn exhibits higher variability and No-Dilated consistently requires more epochs.

These findings yield a valuable practical insight: for environmental sound recognition on edge devices, prioritize temporal-frequency attention over channel-spatial refinement, as the latter may introduce unnecessary complexity without guaranteed benefit.

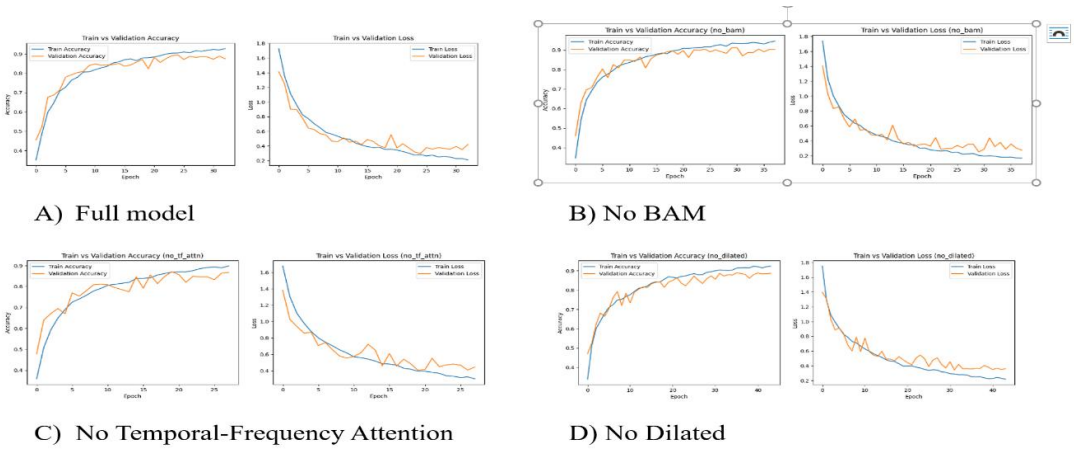

Figure 3. Run-1 training and validation accuracy and loss curves for all ablation variants.

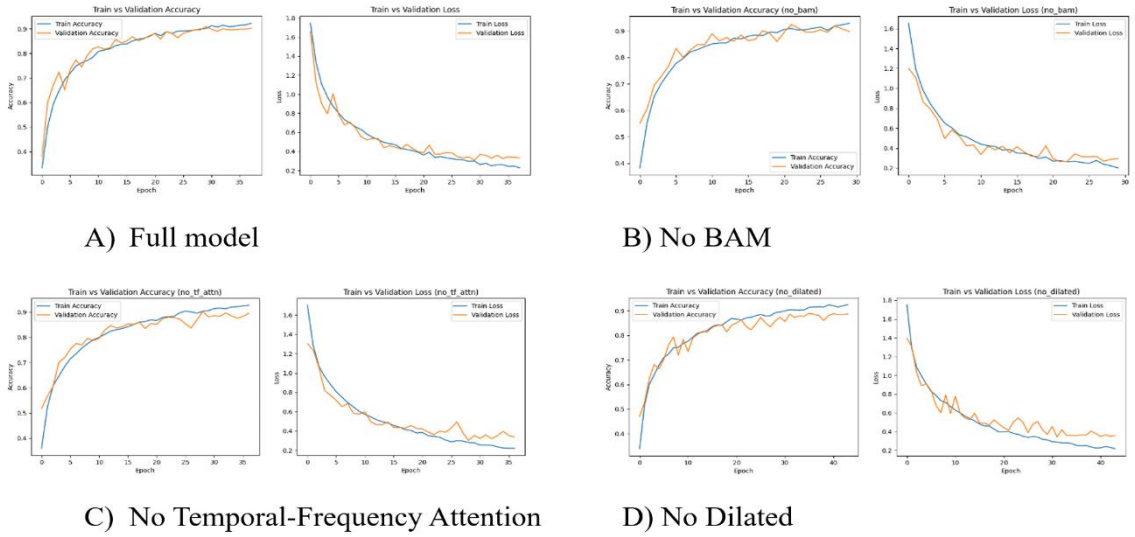

Figure 4. Run-2 training and validation accuracy and loss curves for all ablation variants.

The No-TF-Attn configuration exhibits noticeably higher variance across runs, while the No-Dilated variant consistently converges more slowly, further validating the importance of both temporal–frequency attention and dilated convolutions. Confusion matrices for the best-performing checkpoints of the Full model (90.90%) and the No-BAM variant (92.39%) are presented in Figure 5.

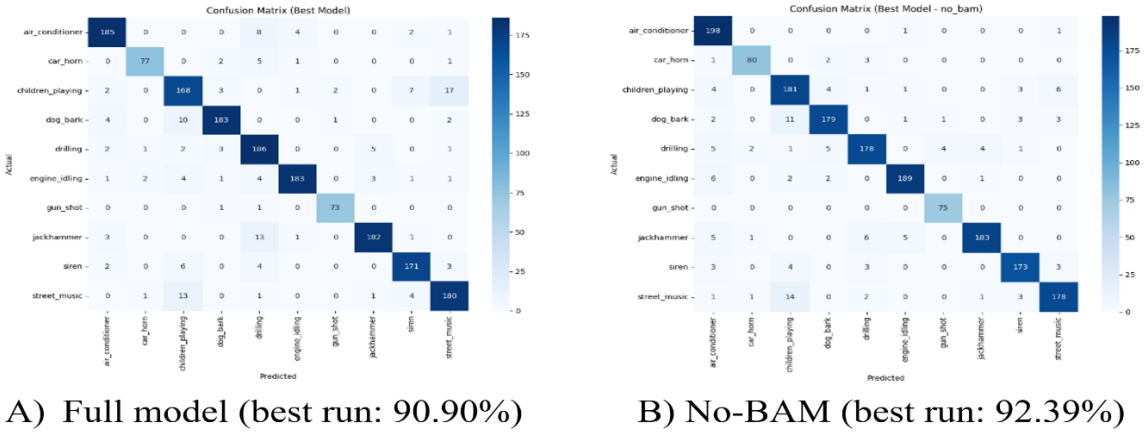

Figure 5. Confusion matrices of the best-performing checkpoints from the Full model and the No-BAM variant.

Both matrices show strong diagonal dominance, indicating reliable class separation. The most frequent confusions occur between children_playing, street_music, and drilling, which share overlapping broadband noise characteristics.

3.2. Performance on UrbanSound8K

Based on the ablation analysis, the No-BAM variant achieves the strongest overall performance on UrbanSound8K, with a peak accuracy of 92.39% (seed 43) and an average of 91.76 ± 0.90% across both runs. The full ShuBAM-TFNet closely follows with 90.16 ± 1.05%, placing both configurations among the most accurate lightweight models reported for this dataset. To further examine class-wise behavior, Table 3 summarizes the precision, recall, and F1-score for each class using the best single run of the Full model and the No-BAM variant.

Table 3. Per-class precision, recall, and F1-score on UrbanSound8K (best single runs).

Class	Full model (90.90%)			No-BAM (92.39%)		
	Precision	Recall	F1-score	Precision	Recall	F1-score
air_conditioner	0.93	0.93	0.93	0.88	0.99	0.93

car_horn	0.95	0.90	0.92	0.95	0.93	0.94
children_playing	0.83	0.84	0.83	0.85	0.91	0.88
dog_bark	0.95	0.92	0.93	0.93	0.90	0.91
drilling	0.84	0.93	0.88	0.92	0.89	0.91
engine_idling	0.96	0.92	0.94	0.96	0.94	0.95
gun_shot	0.96	0.97	0.97	0.94	1.00	0.97
jackhammer	0.95	0.91	0.93	0.97	0.92	0.94
siren	0.92	0.92	0.92	0.95	0.93	0.94
street_music	0.87	0.90	0.89	0.93	0.89	0.91
Macro avg	0.92	0.91	0.91	0.93	0.93	0.93
Weighted avg	0.91	0.91	0.91	0.93	0.92	0.92

Both models perform consistently across most categories. gun_shot and siren are the easiest classes, achieving F1-scores above 97%, while children_playing and street_music remain the most challenging due to their acoustically diffuse and overlapping characteristics. The No-BAM variant reduces several misclassifications in these classes, explaining its superior aggregate performance.

3.3 Cross-Dataset Evaluation on ESC-10

To assess generalization beyond urban sound scenarios, the Full ShuBAM-TFNet model was evaluated on the ESC-10 dataset using the official 5-fold cross-validation protocol, without any architecture changes or hyperparameter tuning.The model achieved:

i. Mean accuracy: 84.00%
ii. Standard deviation: ± 2.41%
iii. Macro-averaged F1-score: 0.838

These results fall within the upper range reported for lightweight architectures (≤100K parameters) on ESC-10. Importantly, the dataset includes both natural and human-made sounds, demonstrating that the combination of dilated convolutions, time-distributed processing, and temporal–frequency attention enables ShuBAM-TFNet to learn representations that generalize well across distinct acoustic domains.

3.4 Comparative Benchmarking with the Existing Work

To contextualize the effectiveness of ShuBAM-TFNet, we compare its performance with recent lightweight approaches published from 2024 onward that emphasize efficiency for edge deployment. We considered only models with fewer than 100K parameters and standard evaluation protocols. Table 4 presents a side-by-side comparison in terms of accuracy and parameter count.

Table 4. Comparison with recent lightweight models on UrbanSound8K and ESC-10.

Method	Year	UrbanSound8K Acc. (%)	ESC-10 Acc. (%)	Parameters
ShuBAM-TFNet (Full model)	2025	90.16 ± 1.05	84.00 ± 2.41	72.390
ShuBAM-TFNet (No-BAM variant)	2025	91.76 ± 0.90	-	70.754
Ranmal et al. (ESC-NAS [14]	2024	81.25	96.25	~80K (estimated)
Kaçar et al. (Fusion MLP) [15]	2025	94.40	-	~5K (MLP)
Liu et al. (Hybrid Green CNN) [16]	2025	85.50	83.00	90K

ShuBAM-TFNet outperforms the ESC-NAS model by Ranmal et al. by nearly 9 percentage points on UrbanSound8K while converging significantly faster. Compared with Liu et al.'s hybrid CNN, ShuBAM-TFNet achieves higher accuracy with fewer parameters. Although Kaçar et al.'s fusion-based MLP reports higher accuracy, it relies on handcrafted features and lacks cross-dataset validation, limiting scalability. Overall, the results demonstrate that ShuBAM-TFNet achieves a strong balance between recognition accuracy and computational efficiency. The ablation studies identify Temporal–Frequency Attention as the most influential component, while dilated convolutions contribute to faster convergence and long-range temporal modeling. Interestingly, removing BAM improves performance on UrbanSound8K, suggesting that excessive channel–spatial gating may occasionally suppress informative acoustic details. Despite strong results, several limitations remain. The architecture was primarily optimized on UrbanSound8K, and further validation on larger datasets such as ESC-50 or AudioSet is necessary. Additionally, we did not employ data augmentation or self-supervised pretraining, leaving room for future improvement.

i. Future work will focus on:
ii. Hardware-level benchmarking and quantization for on-device deployment
iii. Extension to larger, real-world environmental sound datasets
iv. Lightweight self-supervised pretraining for improved robustness
v. Continual learning mechanisms for adaptive edge intelligence

4. CONCLUSION AND LIMITATION

This study presented ShuBAM-TFNet, a lightweight end-to-end environmental sound recognition framework designed to address the competing challenges of high classification accuracy and low computational cost for edge-based acoustic sensing applications. The proposed architecture integrates a ShuffleNet backbone, dilated and time-distributed convolutions, and a Temporal–Frequency Attention (TF-Attn) mechanism to enhance discriminative time–frequency feature learning while maintaining a compact parameter footprint. We systematically evaluated optional Bottleneck Attention Modules (BAM) to assess their contribution to channel-spatial refinement. Extensive experiments on UrbanSound8K and ESC-10 demonstrate that the No-BAM variant—now presented as the primary ShuBAM-TFNet architecture achieves 91.76 ± 0.90% accuracy with only 70,754 parameters on UrbanSound8K, outperforming the Full model by 1.6% while using 1,600 fewer parameters. Ablation studies confirm that Temporal–Frequency Attention provides the most significant performance gain (≈2–3%), while dilated convolutions accelerate convergence without increasing complexity. Cross-dataset evaluation on ESC-10 yields 84.00 ± 2.41% accuracy, demonstrating reasonable generalization despite expected domain shift. The counterintuitive finding that removing BAM improves performance reveals that channel-spatial attention mechanisms require task-specific validation, as they may over-regularize environmental sound models by suppressing informative acoustic details.

Despite these promising results, several limitations remain. Evaluations were limited to UrbanSound8K and ESC-10, which do not fully capture real-world acoustic diversity; future work will validate scalability on larger datasets (ESC-50, AudioSet). Although the parameter count (<75K) strongly indicates edge suitability, we did not measure inference latency, memory usage, or power consumption on embedded hardware. Future work will involve model quantization (INT8) and empirical benchmarking on platforms such as Raspberry Pi 4 and NVIDIA Jetson Nano to validate deployment claims. Additionally, we did not employ data augmentation or self-supervised pretraining; future research will explore robustness-enhancing techniques and domain adaptation strategies to improve cross-dataset generalization. These directions aim to advance ShuBAM-TFNet's practical applicability for smart-city monitoring and edge-IoT environmental sound recognition.

ACKNOWLEDGEMENTS

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

REFERENCES

[1] S. Tyagi, K. Aggarwal, D. Kumar, and S. Garg, "Urban sound classification for audio analysis using long short-term memory," *NEU Journal for Artificial Intelligence and Internet of Things*, vol. 1, no. 2, pp. 1–11, 2023.
[2] W. Zhao, H. Wang, Y. Chen, X. Pan, K. Zhang, and Z. Bai, "An environmental sound classification algorithm based on multiscale channel feature fusion," *IEEE Access*, 2023. https://doi.org/10.5220/0012273800003807
[3] M. Avadhani and A. P. Bidargaddi, "Multi-class urban sound classification with deep learning architectures," *Proc. IEEE Conference*, 2024, pp. 1–7. https://doi.org/10.1109/INCET61516.2024.10593181
[4] H. Lu, H. Zhang, and A. Nayak, "A deep neural network for audio classification with a classifier attention mechanism," *arXiv preprint arXiv:2006.09815*, 2020.
[5] 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. https://doi.org/10.1109/ACCESS.2023.3318015

[6] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," *Proc. ACM Multimedia*, pp. 1041–1044, 2014. https://doi.org/10.1145/2647868.2655045
[7] K. J. Piczak, "ESC: Dataset for environmental sound classification," *Proc. ACM Multimedia*, pp. 1015–1018, 2015. https://doi.org/10.1145/2733373.2806390
[8] R. Mahum, A. Irtaza, A. Javed, H. A. Mahmoud, and H. Hassan, "DeepDet: YAMNet with bottleneck attention module for TTS synthesis detection," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2024, no. 1, pp. 1–18, 2024. https://doi.org/10.1186/s13636-024-00335-9
[9] W. Mu, B. Yin, X. Huang, J. Xu, and Z. Du, "Environmental sound classification using temporal–frequency attention based convolutional neural network," *Scientific Reports*, vol. 11, no. 1, pp. 21552, 2021. https://doi.org/10.1038/s41598-021-01045-4
[10] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.
[11] S. Abdoli, P. Cardinal, and A. L. Koerich, "End-to-end environmental sound classification using a 1D convolutional neural network," *Expert Systems with Applications*, vol. 136, pp. 252–263, 2019. https://doi.org/10.1016/j.eswa.2019.06.040
[12] J. Park, S. Woo, J.-Y. Lee, and I. S. Kweon, "BAM: Bottleneck attention module," *arXiv preprint arXiv:1807.06514*, 2018.
[13] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," *Proc. IEEE CVPR*, pp. 6848–6856, 2018. https://doi.org/10.1109/CVPR.2018.00716
[14] D. Ranmal, P. Ranasinghe, T. Paranayapa, D. Meedeniya, and C. Perera, "ESC-NAS: Environment sound classification using hardware-aware neural architecture search for the edge," *Sensors*, vol. 24, no. 12, pp. 3749, 2024. https://doi.org/10.3390/s24123749
[15] A. Kaçar, A. Derya, and İ. Türkoğlu, "Enhancing environmental sound classification performance through data fusion: A comparative machine learning analysis," *Research Square*, 2025. https://doi.org/10.21203/rs.3.rs-7898377/v1
[16] L. Liu, "Speech recognition method for English translators in noisy environments based on attention mechanism," *International Journal of High Speed Electronics and Systems*, vol. 34, no. 4, pp. 2540350, 2025. https://doi.org/10.1142/S012915642540350X
[17] S. Divya Lakshmi and N. Suresh Kumar, "ResBiA-FusionNET: A robust deep learning framework with harmonic-contrast mel spectrogram for audio sound classification," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 39, no. 15, pp. 2551028, 2025. https://doi.org/10.1142/S0218001425510280
[18] I. H. Sarker, "Deep learning: A comprehensive overview on techniques, taxonomy, applications and research directions," *SN Computer Science*, vol. 2, pp. 420, 2021. https://doi.org/10.1007/s42979-021-00815-1
[19] M. Crocco, M. Cristani, A. Trucco, and V. Murino, "Audio surveillance: A systematic review," *ACM Computing Surveys*, vol. 48, pp. 1–46, 2016. https://doi.org/10.1145/2871183
[20] D. Meedeniya, I. Ariyarathne, M. Bandara, R. Jayasundara, and C. Perera, "A survey on deep learning based forest environment sound classification at the edge," *ACM Computing Surveys*, vol. 56, pp. 66, 2023. https://doi.org/10.1145/3618104
[21] D. Stefani, S. Peroni, and L. Turchet, "A comparison of deep learning inference engines for embedded real-time audio classification," *Proc. DAFx*, 2022, pp. 256–263.
[22] A. Elhanashi, P. Dini, S. Saponara, and Q. Zheng, "Integration of deep learning into the IoT: A survey of techniques and challenges for real-world applications," *Electronics*, vol. 12, pp. 4925, 2023. https://doi.org/10.3390/electronics12244925
[23] D. Meedeniya, *Deep Learning: A Beginners' Guide*, CRC Press, Boca Raton, FL, USA, 2023. https://doi.org/10.1201/9781003390824
[24] B. Wu et al., "FBNet: Hardware-aware efficient convnet design via differentiable neural architecture search," *Proc. IEEE CVPR*, pp. 10734–10742, 2019. https://doi.org/10.1109/CVPR.2019.01099
[25] C. White et al., "Neural architecture search: Insights from 1000 papers," *arXiv preprint arXiv:2301.08727*, 2023.
[26] M. Risso et al., "Lightweight neural architecture search for temporal convolutional networks at the edge," *IEEE Transactions on Computers*, vol. 72, pp. 744–758, 2023.
[27] J. Lin et al., "MCUNet: Tiny deep learning on IoT devices," *Advances in Neural Information Processing Systems*, vol. 33, pp. 11711–11722, 2020.
[28] D. T. Speckhard et al., "Neural architecture search for energy-efficient always-on audio machine learning," *Neural Computing and Applications*, vol. 35, pp. 12133–12144, 2023. https://doi.org/10.1007/s00521-023-08345-y
[29] B. Elliott et al., "Cyber-physical analytics: Environmental sound classification at the edge," *Proc. IEEE WF-IoT*, pp. 1–6, 2020. https://doi.org/10.1109/WF-IoT48130.2020.9221148
[30] A. Bansal and N. K. Garg, "Environmental sound classification: A descriptive review of the literature," *Intelligent Systems with Applications*, vol. 16, pp. 200115, 2022. https://doi.org/10.1016/j.iswa.2022.200115
[31] A. Andreadis, G. Giambene, and R. Zambon, "Monitoring illegal tree cutting through ultra-low-power smart IoT devices," *Sensors*, vol. 21, pp. 7593, 2021. https://doi.org/10.3390/s21227593
[32] I. Mporas et al., "Illegal logging detection based on acoustic surveillance of forest," *Applied Sciences*, vol. 10, pp. 7379, 2020. https://doi.org/10.3390/app10207379
[33] G. Peruzzi, A. Pozzebon, and M. Van Der Meer, "Detecting Forest fires with embedded machine learning models using audio and images," *Sensors*, vol. 23, pp. 783, 2023. https://doi.org/10.3390/s23020783
[34] S. K. Shah, Z. Tariq, and Y. Lee, "IoT-based urban noise monitoring using deep learning," in *Proc. IEEE Big Data*, pp. 4179–4184. 2019. https://doi.org/10.1109/BigData47090.2019.9006176

[35] A. F. R. Nogueira et al., "Sound classification and processing of urban environments: A systematic literature review," *Sensors*, vol. 22, pp. 8608, 2022. https://doi.org/10.3390/s22228608
[36] T. Domhan, J. T. Springenberg, and F. Hutter, "Speeding up automatic hyperparameter optimization of deep neural networks," *Proc. IJCAI*, 2015.
[37] J. Mellor et al., "Neural architecture search without training," in *Proc. ICML*, vol. 139, pp. 7588–7598, 2021.
[38] B. Na et al., "Accelerating neural architecture search via proxy data," *arXiv preprint arXiv:2106.04784*, 2021. https://doi.org/10.24963/ijcai.2021/392
[39] B. Lyu et al., "Resource-constrained neural architecture search on edge devices," *IEEE Transactions on Network Science and Engineering*, vol. 9, pp. 134–142, 2021. https://doi.org/10.1109/TNSE.2021.3054583
[40] H. Benmeziane et al., "A comprehensive survey on hardware-aware neural architecture search," *arXiv preprint arXiv:2101.09336*, 2021.
[41] C. Li et al., "HW-NAS-Bench: Hardware-aware neural architecture search benchmark," *arXiv preprint arXiv:2103.10584*, 2021.
[42] A. Naeem et al., "Edge-based environmental sound recognition using deep learning," *IEEE Access*, 2022.
[43] S. Hershey et al., "CNN architectures for large-scale audio classification," *Proc. ICASSP*, 2017. https://doi.org/10.1109/ICASSP.2017.7952132
[44] T. Heittola, A. Mesaros, and T. Virtanen, "Environmental sound event detection," *IEEE Signal Processing Magazine*, vol. 35, pp. 41–49, 2018.
[45] K. J. Piczak, "Environmental sound classification with convolutional neural networks," *Proc. IEEE MLSP*, 2015. https://doi.org/10.1109/MLSP.2015.7324337
[46] Salman, A. M., & Salman, H. S., "Exploring the Impact of AI and IoT on Production Efficiency, Quality Precision, and Environmental Sustainability in Manufacturing," *Vokasi Unesa Bulletin of Engineering, Technology and Applied Science*, vol. 2, no. 2, pp. 345-358, 2025. https://doi.org/10.26740/vubeta.v2i2.38200
[47] Rifki Ainul Yaqin, Anshori, M. I., Angel, R., Agung, I. W. P., Arifin, T., & Junianto, E., "Stock Price Forecasting Using LSTM with Cross-Validation," *Vokasi Unesa Bulletin of Engineering, Technology and Applied Science*, vol. 3, no. 1, pp. 64–79, 2026. https://doi.org/10.26740/vubeta.v3i1.45130
[48] Oise, G., Nwabuokei, O. C., Ozobialu, chukwuma E., Jenarhome, O. P., Atake, O. M., Nkem Belinda, U., & Babalola Eyitemi, A., "Enhancing Indoor Positioning Accuracy with Ant Colony Optimization and Dual Clustering," *Vokasi Unesa Bulletin of Engineering, Technology and Applied Science*, vol. 2, no. 3, pp. 516–530, 2025. https://doi.org/10.26740/vubeta.v2i3.39452
[49] Andrianarisoa, S. H., Ravelonjara, H. M., Suddul, G., Foogooa, R., Armoogum, S., & Sookarah, D., "A Deep Learning Approach to Fake News Classification Using LSTM," *Vokasi Unesa Bulletin of Engineering, Technology and Applied Science*, vol. 2, no. 3, pp. 593–601, 2025. https://doi.org/10.26740/vubeta.v2i3.39360
[50] Oise, G., Nwabuokei, C., Igbunu, R., & Ejenarhome, P., "Revisiting Parasitic Computing: Ethical and Technical Dimensions in Resource Optimization," *Vokasi Unesa Bulletin of Engineering, Technology and Applied Science*, vol. 2, no. 3, pp. 376–386, 2025. https://doi.org/10.26740/vubeta.v2i3.38786

BIOGRAPHIES OF AUTHORS

Kabiru Muhammed Nasiru holds a Bachelor of Engineering (B.Eng.) degree in Computer Engineering. His academic interests include computer systems, machine learning, embedded systems, and intelligent computing applications. He can be contacted at nasirumuhammedkabiru@gmail.com

Sani Saleh SAMINU received his MSc degree in Artificial Intelligence from the Department of Computer Engineering, Ahmadu Bello University (ABU), Zaria, Nigeria, and previously obtained his BSc(Ed) in Computer Science from the same institution. His research interests include environmental sound recognition, plant disease recognition, computer vision, deep learning, zero-shot learning, and intelligent systems with applications in agriculture and healthcare. He can be contacted at sssaleh@engineering.abu.edu.ng

Dr. Shehu Mohammed Yusuf is a Senior Lecturer in the Department of Computer Engineering at Ahmadu Bello University, Zaria, Nigeria, with over a decade of teaching and research experience. He holds a B.Eng. in Electrical Engineering, an MSc in Electrical Engineering and a Ph.D. in Computer Engineering. He is also a Huawei Certified ICT Associate (AI), Huawei Certified Academy Instructor, and a COREN registered engineer. His research interests span Artificial Intelligence, Natural Language Processing, Computer Vision, and Bioinformatics.
Email: smyusuf@abu.edu.ng