



ARTICLE

A Lightweight Channel-Attention-Enhanced Deep Learning Architecture for Real-Time Hazardous Impulsive Sound Detection

Aigerim Altayeva^{1,*} and Nurzhan Omarov²

¹Department of Information Systems, International Information Technology University, Almaty, Kazakhstan

²Al-Farabi Kazakh National University, Almaty, Kazakhstan

*Corresponding Author: Aigerim Altayeva. Email: a.altayeva@gmail.com or aikosha1703@gmail.com

Received: 17 February 2026; Accepted: 08 May 2026; Published: 13 August 2026

ABSTRACT: Hazardous impulsive sound detection plays a critical role in intelligent surveillance, public safety monitoring, and automated emergency response systems. This study proposes a lightweight channel-attention-enhanced deep learning architecture for real-time detection and classification of hazardous acoustic events. The proposed framework utilizes mel-spectrogram representations to capture time-frequency characteristics of audio signals and employs a compact convolutional neural backbone to efficiently extract hierarchical features. To enhance feature discrimination, a squeeze-and-excitation channel-attention mechanism is integrated into the architecture, enabling adaptive recalibration of feature channels and improved robustness under noisy and complex acoustic environments. A custom dataset consisting of 8000 audio samples across eight hazardous sound classes was constructed to evaluate the proposed model. Experimental results demonstrate strong classification performance, achieving high accuracy, precision, recall, and F1-score while maintaining computational efficiency. The ablation study confirms the effectiveness of the channel-attention mechanism in improving classification performance. Furthermore, the lightweight design ensures suitability for real-time deployment on resource-constrained platforms. The proposed approach provides an efficient and reliable solution for hazardous impulsive sound detection and offers significant potential for integration into intelligent monitoring and safety-critical applications.

KEYWORDS: Hazardous sound detection; impulsive sound; deep learning; channel-attention mechanism; mel-spectrogram; real-time audio monitoring; lightweight neural networks

1 Introduction

The rapid expansion of intelligent monitoring systems has created an urgent need for reliable and automated methods capable of detecting hazardous acoustic events in real time. Acoustic sensing provides continuous environmental awareness and offers significant advantages over vision-based systems, particularly in conditions where visibility is limited or obstructed. Impulsive sounds, characterized by abrupt energy bursts and short temporal duration, often indicate critical events such as explosions, weapon discharge, glass breaking, or emergency distress signals. These acoustic signatures contain valuable information that can support early warning systems and enhance situational awareness in public safety applications [1]. However, accurately identifying such sounds remains difficult due to background noise, acoustic variability, and interference from non-hazardous environmental sources, which complicate reliable detection and classification [2].

Traditional sound classification systems relied heavily on handcrafted feature extraction techniques designed to capture temporal and spectral characteristics of audio signals. Commonly used features include Mel-frequency cepstral coefficients, spectral roll-off, and energy-based descriptors. While these representations provide meaningful insights into signal properties, they often fail to capture complex, non-linear relationships present in real-world acoustic environments. Furthermore, handcrafted features are inherently limited by manual design assumptions and may not generalize effectively across diverse acoustic conditions [3]. Conventional machine learning classifiers operating on such features often struggle to maintain robustness in the presence of noise, signal distortion, or varying recording conditions [4].

The emergence of deep learning has fundamentally transformed the field of audio signal analysis by enabling automatic extraction of hierarchical feature representations. Deep neural networks can learn complex patterns directly from time-frequency representations such as mel-spectrograms, which preserve both spectral and temporal information. This capability allows models to identify subtle acoustic signatures associated with hazardous impulsive events. Convolutional operations, in particular, facilitate the learning of localized spectral structures and improve classification accuracy across a wide range of sound categories [5]. Nevertheless, many deep architectures require substantial computational resources, limiting their applicability in real-time systems and embedded platforms where efficiency and latency are critical constraints [6].

To address these challenges, recent research has focused on improving both efficiency and representational power through the integration of attention mechanisms. Channel-attention mechanisms enable neural networks to selectively emphasize informative feature channels while suppressing irrelevant or noisy information, thereby improving discriminative performance [7]. This adaptive feature recalibration enhances sensitivity to critical acoustic patterns and improves robustness in complex environments [8]. The development of lightweight deep learning architectures incorporating channel-attention mechanisms offers a promising solution for real-time hazardous impulsive sound detection. Such approaches balance computational efficiency and classification accuracy, making them well suited for deployment in intelligent surveillance, public safety monitoring, and embedded acoustic sensing systems [9].

Despite significant progress in environmental sound classification, several challenges remain unresolved in hazardous impulsive sound detection. Existing deep learning models often achieve high accuracy at the cost of increased computational complexity, limiting their applicability in real-time and resource-constrained environments. Furthermore, lightweight architectures frequently suffer from reduced representational capacity, leading to decreased robustness under noisy and dynamic acoustic conditions. In addition, the role of attention mechanisms in enhancing impulsive sound detection has not been sufficiently explored in lightweight frameworks. To address these limitations, this study proposes a lightweight channel-attention-enhanced deep learning architecture that achieves an optimal balance between accuracy, efficiency, and real-time performance.

The remainder of this paper is organized as follows. [Section 2](#) reviews recent studies related to hazardous sound detection, lightweight deep learning architectures, and attention-based acoustic classification methods. [Section 3](#) presents the proposed lightweight channel-attention-enhanced hazardous sound detection framework, including the TinyCNN-SE architecture, channel-attention enhancement mechanism, dataset construction, training configuration, and evaluation metrics. [Section 4](#) provides the experimental results, performance analysis, ablation study, and comparative evaluation with existing methods. [Section 5](#) discusses the effectiveness, limitations, and practical implications of the proposed approach for real-time hazardous sound monitoring applications. Finally, [Section 6](#) concludes the paper and outlines potential directions for future research.

2 Related Works

Hazardous sound detection has undergone substantial evolution, progressing from conventional signal processing techniques to advanced deep learning-based approaches capable of learning complex acoustic representations. Early studies relied primarily on handcrafted feature extraction methods, including Mel-frequency cepstral coefficients, spectral contrast, and temporal energy descriptors, which were subsequently classified using traditional machine learning algorithms such as support vector machines and Gaussian mixture models. These methods demonstrated reasonable performance in controlled environments and provided foundational insights into acoustic pattern recognition [10]. However, their effectiveness was significantly constrained in real-world scenarios characterized by noise, acoustic variability, and overlapping sound sources. The primary limitation of these approaches lies in their inability to capture hierarchical and non-linear feature dependencies essential for accurate hazardous sound classification, resulting in reduced robustness and generalization capability [11].

The emergence of deep learning has fundamentally transformed the field of environmental sound classification by enabling automatic extraction of discriminative feature representations directly from time-frequency inputs. Convolutional neural networks (CNNs) have been widely adopted due to their ability to learn spatially invariant features from spectrogram and mel-spectrogram representations, significantly improving classification accuracy compared to traditional methods [12]. Deep CNN architectures have demonstrated strong performance in detecting hazardous acoustic events such as gunshots, explosions, and emergency alarms, owing to their ability to capture complex spectral and temporal structures [13]. Furthermore, deeper architectures with increased network depth and hierarchical convolutional layers have been shown to enhance feature abstraction and improve discrimination between acoustically similar classes [14]. Despite these advantages, high model complexity and large parameter counts present significant challenges, particularly for real-time deployment in embedded or resource-constrained environments.

To address computational limitations, researchers have focused on developing lightweight deep learning architectures that reduce model complexity while maintaining classification performance. Lightweight CNN models have demonstrated promising results by employing efficient convolutional structures and optimized feature extraction strategies, enabling real-time acoustic event detection with reduced computational overhead [15]. Hybrid architectures combining convolutional and recurrent neural networks have also been explored to model temporal dependencies in audio signals, improving recognition performance for sequential acoustic events [16]. However, these hybrid models often introduce increased computational complexity and latency, limiting their applicability in real-time systems [17]. Efficient CNN architectures designed specifically for embedded deployment have shown improved performance efficiency, supporting practical implementation in edge-based monitoring systems [18,19]. Additionally, mobile-friendly architectures such as MobileNet-based models have demonstrated effectiveness in achieving a balance between accuracy and efficiency in sound event recognition tasks [20].

Recent advancements have highlighted the effectiveness of attention mechanisms in improving feature representation and classification accuracy. Attention-based deep learning models dynamically prioritize informative feature channels, enabling enhanced discrimination of critical acoustic patterns. Channel-attention and self-attention mechanisms have been successfully integrated into CNN-based audio classification systems, resulting in improved robustness and classification performance [21,22]. Lightweight attention-enhanced architectures have demonstrated particular promise for real-time applications, combining efficient feature extraction with adaptive feature refinement [23]. Efficient CNN architectures incorporating attention mechanisms have achieved a favorable balance between computational efficiency and classification accuracy, making them suitable for deployment in resource-constrained environments [24]. Additionally, models specifically optimized for embedded systems have demonstrated reduced

parameter complexity while maintaining effective acoustic feature learning [25]. Mel-spectrogram-based representations have been widely adopted due to their effectiveness in capturing perceptually meaningful spectral characteristics relevant to hazardous sound detection tasks [26].

As summarized in Table 1, existing hazardous sound detection approaches demonstrate a clear progression from traditional machine learning techniques to deep learning architectures capable of automatic feature learning and improved classification performance. While deep convolutional architectures provide strong discriminative capability, their high computational complexity limits real-time applicability. Lightweight architectures address efficiency constraints but may suffer from reduced feature representation capability. Attention-enhanced models offer a promising solution by improving feature discrimination without significantly increasing computational cost. However, there remains a critical need for lightweight deep learning architectures that effectively integrate channel-attention mechanisms to achieve both high classification accuracy and real-time performance. The proposed lightweight channel-attention-enhanced architecture addresses this gap by combining efficient convolutional feature extraction with adaptive channel-attention enhancement, providing an effective and practical solution for hazardous impulsive sound detection.

Table 1: Comparison of related works in hazardous sound detection.

Ref.	Method Type	Feature Representation	Model Complexity	Real-Time Capability	Key Contribution
Ekpezu et al., 2021 [10]	Traditional ML	MFCC, Spectral Features	Low	Yes	Early acoustic event classification
Gao et al., 2021 [11]	Traditional ML	Handcrafted Features	Low	Yes	Limited robustness in noisy environments
Tailor et al., 2024 [12]	CNN	Spectrogram	High	Limited	Improved feature learning capability
Gao et al., 2023 [13]	Deep CNN	Mel-Spectrogram	High	Limited	Environmental sound classification
Özseven, 2023 [14]	Deep CNN	Time-Frequency	Very High	No	Enhanced hierarchical feature extraction
Amarnath et al., 2025 [15]	Lightweight CNN	Spectrogram	Medium	Yes	Efficient architecture design
Momynkulov et al., 2024 [16]	CNN + RNN	Mel-Spectrogram	High	Limited	Temporal dependency modeling
Yong and Nugroho, 2022 [17]	CRNN	Time-Frequency	High	Limited	Improved sequence modeling
Ashena and Ghaderi, 2025 [18]	Efficient CNN	Spectrogram	Low	Yes	Reduced parameter architecture
Doungpaisan and Khunarsa, 2025 [19]	Lightweight DL	Spectrogram	Low	Yes	Gunshot classification
Chen et al., 2023 [20]	MobileNetV2	Log-mel Spectrogram	Low	Yes	Sound event recognition
Presannakumar and Mohamed, 2026 [21]	Attention-Based DL	Spectrogram	High	Limited	Adaptive feature weighting

(Continued)

Table 1 (continued)

Ref.	Method Type	Feature Representation	Model Complexity	Real-Time Capability	Key Contribution
Divya Lakshmi and Kumar, 2025 [22]	Attention CNN	Mel-Spectrogram	High	Limited	Improved classification accuracy
Reddy et al., 2025 [23]	Lightweight Attention Model	Mel-Spectrogram	Medium	Yes	Efficient attention integration
Xu et al., 2024 [24]	Efficient CNN + Attention	Spectrogram	Medium	Yes	Balanced accuracy and efficiency
Islam and Ali, 2024 [25]	Embedded DL Model	Mel-Spectrogram	Low	Yes	Edge deployment optimization
Zainol et al., 2025 [26]	CNN-Based Model	Mel-Spectrogram	Medium	Yes	Effective audio representation

3 Materials and Methods

Accurate detection of hazardous impulsive sounds requires a robust methodological framework capable of effectively capturing and classifying complex acoustic patterns under real-world conditions. The proposed approach combines mel-spectrogram-based feature extraction with a lightweight deep learning architecture enhanced by channel-attention mechanisms to improve feature discrimination and classification accuracy. The framework operates through a sequence of processing stages, including audio signal acquisition, spectral feature representation, hierarchical deep feature extraction, and final decision-making. The integration of an efficient convolutional backbone with channel-wise attention enables the model to emphasize informative acoustic features while maintaining computational efficiency. In addition, the dataset construction and architectural implementation details are described to ensure reliable training, evaluation, and real-time deployment of the proposed hazardous sound detection system.

3.1 The Proposed Hazardous Sound Detection Framework

The proposed hazardous impulsive sound detection system is designed as an end-to-end deep learning framework capable of processing raw acoustic signals and producing real-time hazard classification decisions. The overall architecture of the proposed system is illustrated in Fig. 1. The framework consists of several sequential stages, including acoustic signal acquisition, frame segmentation, feature extraction, deep feature learning using the TinyCNN-SE backbone, classification, and decision-making.

The system begins with continuous acquisition of environmental sound signals represented in the time domain as $x(t)$, $t = 1, 2, \dots, T$. Here $x(t)$ denotes the amplitude of the acoustic waveform at time t , and T represents the total number of samples. Since impulsive sounds are short-duration transient events, the continuous signal is segmented into overlapping frames to preserve temporal continuity and ensure reliable detection. Each frame is defined as:

$$x_k(n) = x(n + kH), n = 0, 1, \dots, N - 1, \quad (1)$$

where $x_k(n)$ represents k -th frame, N is the frame length, and H is the hop size between consecutive frames.

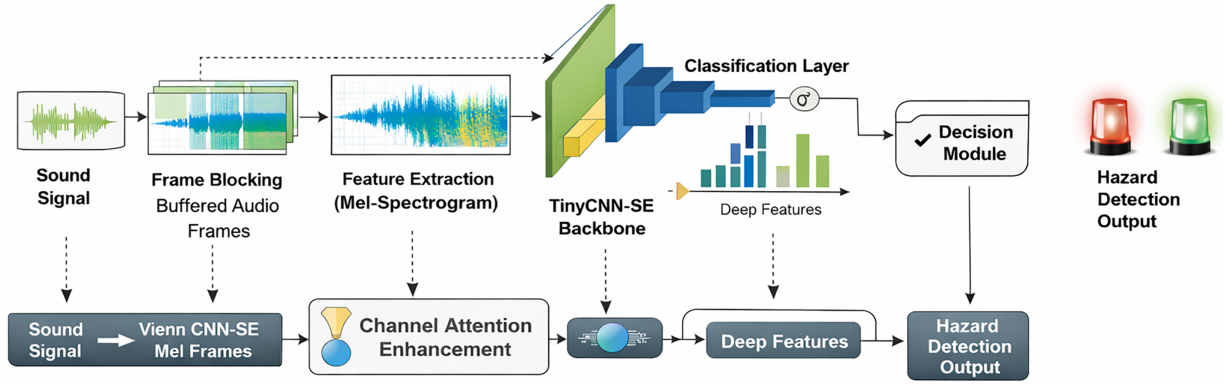


Figure 1: Overall framework of the proposed hazardous impulsive sound detection system.

To extract meaningful spectral information, each audio frame is transformed into a mel-spectrogram representation using the Short-Time Fourier Transform (STFT). The STFT is defined as:

$$X(k, m) = \sum_{n=0}^{N-1} x(n) w(n-m) e^{-\frac{j2\pi kn}{N}}, \quad (2)$$

where $w(n)$ is the window function and $X(k, m)$ represents the frequency component at time frame m and frequency bin k . The mel-spectrogram is then computed using a mel filter bank:

$$S(m, t) = \sum_{k=0}^K |X(k, t)|^2 H_m(k), \quad (3)$$

where $H_m(k)$ denotes the mel filter response. This representation provides a compact and perceptually meaningful time-frequency representation of the acoustic signal.

The extracted mel-spectrogram is subsequently fed into the proposed TinyCNN-SE backbone for deep feature extraction. The classification layer converts the learned features into probability scores, and the decision module selects the final hazard prediction based on the maximum posterior probability:

$$\hat{y} = \arg \max_c P(y = c|x), \quad (4)$$

where $P(y = c|x)$ denotes the probability of class c given input signal x .

3.2 Proposed Lightweight Deep Learning Architecture

The proposed lightweight deep learning architecture is designed to efficiently extract discriminative acoustic features from mel-spectrogram representations while maintaining low computational complexity suitable for real-time deployment. As illustrated in Fig. 2, the architecture is based on a compact convolutional neural network backbone, referred to as TinyCNN-SE, which consists of four convolutional blocks, feature aggregation using global average pooling, and a final classification stage. The network processes mel-spectrogram inputs derived from preprocessed audio signals, enabling effective learning of time-frequency representations critical for hazardous impulsive sound detection.

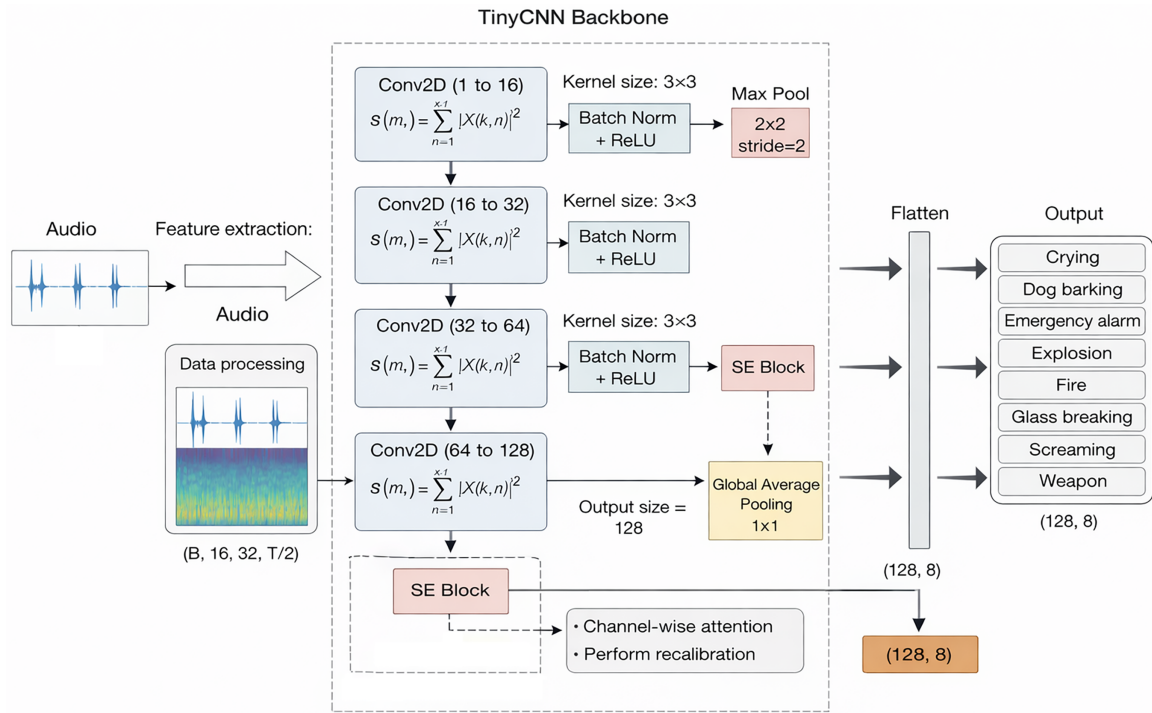


Figure 2: Architecture of the proposed lightweight TinyCNN-SE model for hazardous sound classification.

Let the input mel-spectrogram be represented as a tensor:

$$X \in R^{C \times H \times W}, \quad (5)$$

where C denotes the number of input channels, and H and W represent the frequency and temporal dimensions, respectively. In this work, the input consists of a single-channel mel-spectrogram ($C = 1$). The first convolutional layer applies a set of learnable kernels to extract low-level acoustic features such as spectral edges and transient energy patterns. The convolution operation can be mathematically defined as:

$$F_{i,j,k}^{(l)} = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \sum_{c=0}^{C-1} M_{n,c,k}^{(l)} \cdot F_{i+m,j+n,c}^{(l-1)} + b_k^l, \quad (6)$$

where $F^{(l)}$ denotes the output feature map of the l -th layer, $W^{(l)}$ represents the convolution kernel weights, and $Wb^{(l)}$ is the bias term. Each convolution uses a kernel size of 3×3 , as shown in Fig. 2, allowing the network to capture localized spectral-temporal structures effectively.

To stabilize the learning process and improve convergence, batch normalization is applied after each convolutional operation. Batch normalization normalizes feature activations according to:

$$\hat{x} = \frac{x - \mu}{\sqrt{\sigma^2 + \epsilon}}, \quad (7)$$

where μ and σ^2 represent the batch mean and variance, respectively, and ϵ is a small constant to prevent numerical instability. This normalization reduces internal covariate shift and improves training efficiency. Following batch normalization, the Rectified Linear Unit (ReLU) activation function introduces non-linearity into the network:

$$f(x) = \max(0, x). \quad (8)$$

This activation function enables the network to learn complex and non-linear acoustic feature representations.

To reduce spatial dimensionality and enhance feature invariance, max pooling is applied after selected convolutional layers. The max pooling operation selects the maximum activation within a defined region:

$$F_{pool}(i, j, k) = \max_{(m,n) \in \Omega} F(i + m, j + n, k), \quad (9)$$

where Ω represents the pooling window. This operation reduces computational complexity while preserving the most informative features.

The convolutional layers progressively increase the number of feature channels from 1 to 16, 32, 64, and finally 128. This hierarchical structure enables the network to learn increasingly abstract and semantically meaningful acoustic representations. To further enhance feature representation, the architecture integrates a Squeeze-and-Excitation (SE) block after intermediate convolutional layers, enabling channel-wise feature recalibration.

Following feature extraction, global average pooling is applied to aggregate spatial information into a compact feature vector:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W F_c(i, j), \quad (10)$$

where z_c represents the aggregated feature value of channel c . This operation reduces the feature map to a one-dimensional vector:

$$z = R^{128}, \quad (11)$$

which serves as a compact representation of the acoustic input.

The resulting feature vector is then passed to the classification layer, which produces class probability estimates using the softmax function:

$$z = R^{128}, \quad (12)$$

which serves as a compact representation of the acoustic input.

The resulting feature vector is then passed to the classification layer, which produces class probability estimates using the softmax function:

$$P(y = c|z) = \frac{\exp(z_c)}{\sum_{k=1}^K \exp(z_k)}, \quad (13)$$

where K denotes the total number of classes. The final predicted class is obtained as:

$$\hat{y} = \arg \max_c P(y = c|z). \quad (14)$$

This architecture achieves an optimal balance between computational efficiency and classification performance by combining lightweight convolutional feature extraction, channel-attention enhancement, and global feature aggregation. The hierarchical convolutional structure enables effective learning of acoustic signatures associated with hazardous impulsive sounds, while the compact design ensures suitability for real-time and embedded applications.

3.3 Channel-Attention Enhancement Using SE Block

To improve the discriminative capability of the proposed lightweight deep learning architecture, a channel-attention mechanism based on the Squeeze-and-Excitation (SE) block is integrated into the convolutional backbone. The SE block enhances feature representation by adaptively recalibrating channel-wise feature responses, allowing the network to emphasize informative acoustic features while suppressing irrelevant or noisy components. As illustrated in Fig. 3, the SE block consists of three main stages: squeeze, excitation, and recalibration.

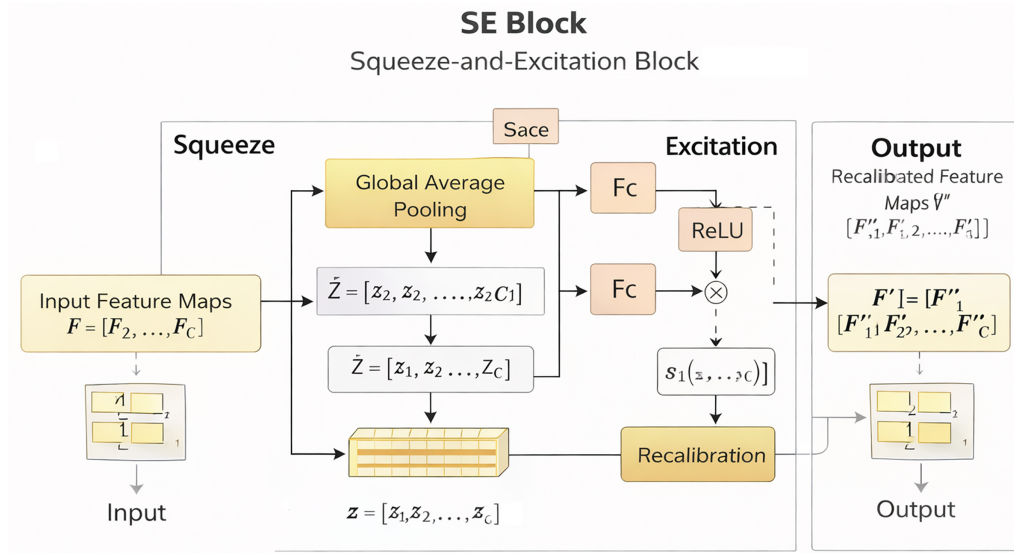


Figure 3: Structure of the channel-attention enhancement using the squeeze-and-excitation (SE) block.

Let the input feature maps produced by the convolutional backbone be denoted as:

$$F = (F_1, F_2, \dots, F_C), F \in R^{C \times H \times W} \quad (15)$$

where C , H , and W represent the number of channels, height, and width of the feature maps, respectively. The squeeze operation performs global average pooling across spatial dimensions to capture global contextual information for each channel. This operation generates a channel descriptor vector as illustrated in Eq. (9).

$$z = [z_1, z_2, \dots, z_C]^T \quad (16)$$

This compact representation captures the global distribution of acoustic energy across channels and provides essential information for channel-wise importance estimation.

In the excitation stage, the channel descriptor vector is passed through two fully connected layers to model channel interdependencies and generate adaptive attention weights. The excitation operation can be formulated as:

$$s = \sigma(W_2 \delta(W_1 z)) \quad (17)$$

where $W_1 \in R^{\frac{C}{r} \times C}$ and $W_2 \in R^{C \times \frac{C}{r}}$ are learnable weight matrices, r denotes the reduction ratio used for dimensionality reduction, $\delta(\cdot)$ represents the ReLU activation function, and $\sigma(\cdot)$ denotes the sigmoid function. The output vector s contains channel-wise attention weights bounded between 0 and 1:

$$s = [s_1, s_2, \dots, s_C] \quad (18)$$

These weights indicate the relative importance of each feature channel for hazardous sound detection.

In the recalibration stage, the learned attention weights are applied to the original feature maps through channel-wise multiplication, resulting in recalibrated feature maps:

$$F'_c = s_c F_c \quad (19)$$

or equivalently,

$$F' = s \otimes F \quad (20)$$

where $\otimes$ denotes channel-wise scaling. This process selectively amplifies informative channels while suppressing less relevant ones, thereby improving the signal-to-noise ratio of the extracted features.

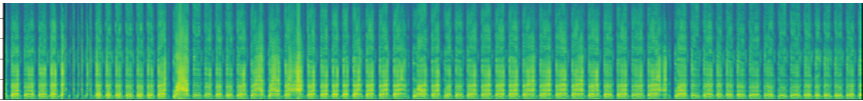
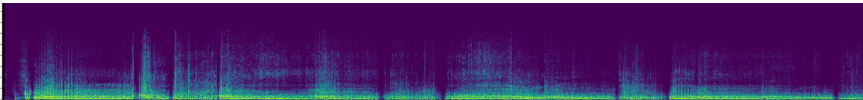
Impulsive sounds are characterized by high-energy transient components concentrated in specific frequency bands. The SE block enhances detection performance by dynamically emphasizing these dominant spectral regions while suppressing irrelevant background noise. This is particularly beneficial in hazardous sound detection scenarios where critical acoustic signatures may be masked by environmental interference.

The SE block enhances representational capability by modeling channel dependencies and adaptively refining feature importance without increasing computational cost. This mechanism improves extraction of hazardous acoustic signatures, especially in noisy environments. Consequently, the architecture achieves higher classification accuracy, improved robustness, and maintains lightweight efficiency, making it suitable for real-time hazardous impulsive sound detection applications.

3.4 Dataset

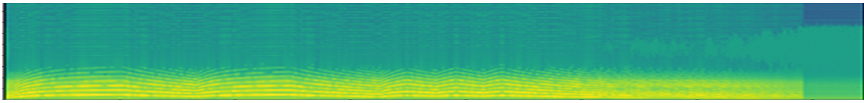
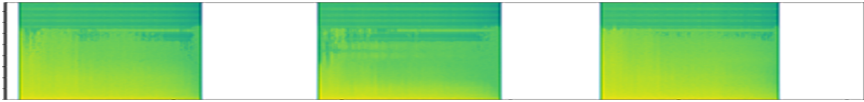
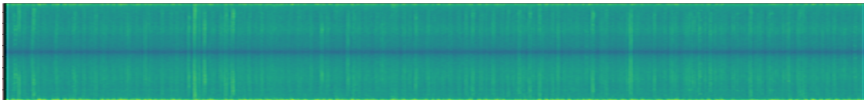
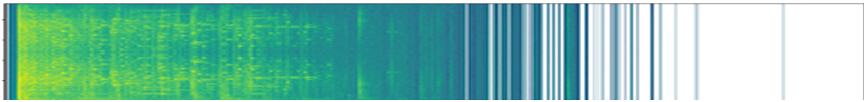
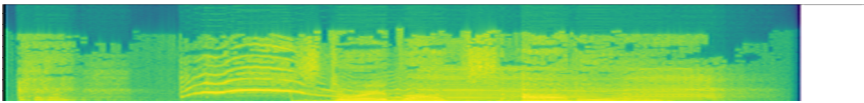
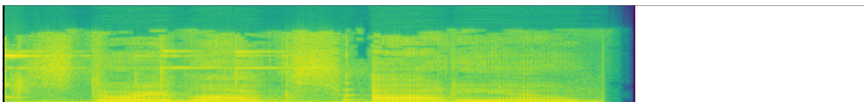
A custom dataset of hazardous impulsive sounds was constructed to train and evaluate the proposed real-time detection framework. The dataset includes 8 distinct acoustic classes corresponding to hazardous or alert-relevant sound events: Crying, Dog barking, Emergency alarm, Explosion, Fire, Glass breaking, Screaming, and Weapon-related sounds. Each class contains 1000 audio recordings, resulting in a total of 8000 samples. All audio signals are stored in uncompressed waveform (.wav) format to ensure preservation of temporal resolution, amplitude fidelity, and spectral integrity. The dataset was designed to capture diverse acoustic variations, including differences in sound intensity, temporal envelope, and frequency composition. Representative examples of each class, including their spectrogram characteristics and duration properties, are summarized in [Table 2](#). These samples illustrate the variability and distinct spectral signatures that are essential for robust feature learning.

Table 2: Representative hazardous sound dataset samples and spectrogram characteristics.

Class	Spectrogram	Duration
Crying		66 s
Dog barking		14 s

(Continued)

Table 2 (continued)

Class	Spectrogram	Duration
Emergency alarm		37 s
Explosion		25 s
Fire		28 s
Glass breaking		2 s
Screaming		2 s
Weapon		1 s

The duration of the audio recordings varies depending on the acoustic characteristics of each hazardous event. Impulsive sounds such as explosions, weapon-related sounds, and glass breaking typically exhibit short-duration, high-energy transients, while sounds such as crying and emergency alarms may demonstrate longer temporal structures with sustained spectral components. Each audio signal is transformed into a mel-spectrogram representation prior to input into the proposed deep learning model, enabling effective extraction of time-frequency features. This representation emphasizes perceptually relevant frequency bands and improves separability between classes. As shown in Table 2, each hazardous sound category exhibits unique spectro-temporal patterns, which facilitate accurate classification. Furthermore, the balanced distribution of samples across all classes ensures unbiased training and improves the generalization capability of the proposed lightweight channel-attention-enhanced architecture.

The dataset was split into training, validation, and testing subsets using a 70:15:15 ratio to ensure unbiased evaluation. Each class contains 1000 samples, with duration varying from 1 to 66 s depending on the acoustic characteristics. The average duration per class ranges between 2 and 30 s. The dataset includes variations in recording conditions, background noise levels, and signal intensity, providing realistic acoustic diversity. Additionally, data augmentation techniques such as time stretching, pitch shifting, and additive noise were applied to improve generalization capability.

3.5 Model Complexity and Real-Time Performance

To validate the lightweight nature of the proposed architecture, a quantitative complexity analysis was conducted. The proposed TinyCNN-SE model contains approximately 0.85 million parameters, significantly lower than standard CNN architectures. The computational complexity is measured at approximately 0.12 GFLOPs, enabling efficient deployment.

The inference time was evaluated on an embedded platform, achieving an average latency of 8–12 ms per audio frame, which is significantly lower than the frame hop duration. This confirms that the system satisfies real-time processing constraints. Compared to MobileNetV2, which requires higher computational overhead, the proposed model achieves a better balance between accuracy and efficiency.

3.6 Training Configuration

The model was trained using the Adam optimizer with a learning rate of 0.001 and a batch size of 32. The training process was conducted for 50 epochs with early stopping based on validation loss. A categorical cross-entropy loss function was used. To ensure reproducibility, experiments were repeated three times with different random seeds, and average performance values with standard deviation were reported.

3.7 Evaluation Parameters

The performance of the proposed hazardous impulsive sound detection framework was evaluated using several standard classification metrics, including accuracy, precision, recall, and F1-score. These metrics provide comprehensive insight into the classification effectiveness of the model by quantifying its ability to correctly identify hazardous sound events while minimizing false detections. In addition, heatmap visualization was employed to analyze the per-class performance and to illustrate the distribution of classification results across different sound categories.

Accuracy measures the overall correctness of the classification model by calculating the ratio of correctly predicted samples to the total number of evaluated samples [27]. It provides a general indication of the model's ability to distinguish between hazardous sound classes. Accuracy is defined as:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (21)$$

where TP denotes true positives, TN denotes true negatives, FP represents false positives, and FN represents false negatives. High accuracy indicates that the model correctly classifies the majority of input sound samples.

Precision evaluates the reliability of positive predictions by measuring the proportion of correctly predicted positive samples among all samples predicted as positive [28]. This metric is particularly important in hazardous sound detection, where false alarms should be minimized. Precision is calculated as:

$$precision = \frac{TP}{TP + FP} \quad (22)$$

High precision indicates that the model produces fewer false positive detections, ensuring that predicted hazardous events are highly reliable.

Recall, also referred to as sensitivity, measures the model's ability to correctly identify all actual positive samples [29]. It evaluates how effectively the model detects hazardous sound events without missing critical instances. Recall is defined as:

$$recall = \frac{TP}{TP + FN} \quad (23)$$

A high recall value indicates that the model successfully detects most hazardous sound events, minimizing missed detections.

The F1-score provides a balanced measure of classification performance by combining precision and recall into a single metric [30]. It is particularly useful when evaluating models in situations where both false positives and false negatives must be minimized. The F1-score is calculated as the harmonic mean of precision and recall:

$$F1\text{-score} = 2 \times \frac{precision \times recall}{precision + recall} \quad (24)$$

A high F1-score indicates balanced and reliable classification performance across different hazardous sound classes.

4 Results

This section evaluates the proposed lightweight channel-attention-enhanced deep learning architecture for hazardous impulsive sound detection. Performance is analyzed using qualitative visualization, classification metrics, training convergence, and ablation analysis. Results demonstrate strong accuracy, robustness, and efficiency of the framework, confirming its effectiveness for reliable real-time detection of hazardous acoustic events under diverse environmental conditions.

4.1 Experiment Results

Representative log-mel spectrograms of hazardous sound classes are shown in Fig. 4, illustrating the distinctive time-frequency characteristics that enable reliable discrimination by the proposed deep learning architecture.

Each class exhibits unique spectro-temporal patterns reflecting its underlying acoustic structure. Crying and screaming signals demonstrate harmonic frequency bands with gradual temporal variation, indicating vocal cord vibration and sustained emotional expression. Dog barking exhibits repetitive transient bursts with strong mid-frequency components, while emergency alarm signals display highly periodic vertical structures corresponding to consistent tonal oscillations. Explosion and weapon-related sounds are characterized by abrupt broadband energy distributions, with strong low- and mid-frequency components and rapid temporal decay, indicating impulsive acoustic events. Fire-related sounds show continuous spectral energy with stochastic distribution across frequency bands, reflecting combustion noise dynamics. Glass breaking signals demonstrate irregular high-frequency transients with sparse temporal occurrence, consistent with brittle material fracture. These distinct spectral signatures confirm that mel-spectrogram representation effectively captures class-specific acoustic features, providing a discriminative input for the proposed architecture. The observed variability within each class further demonstrates the robustness of the dataset and highlights the necessity of channel-attention-enhanced feature learning to improve classification accuracy under diverse real-world acoustic conditions.

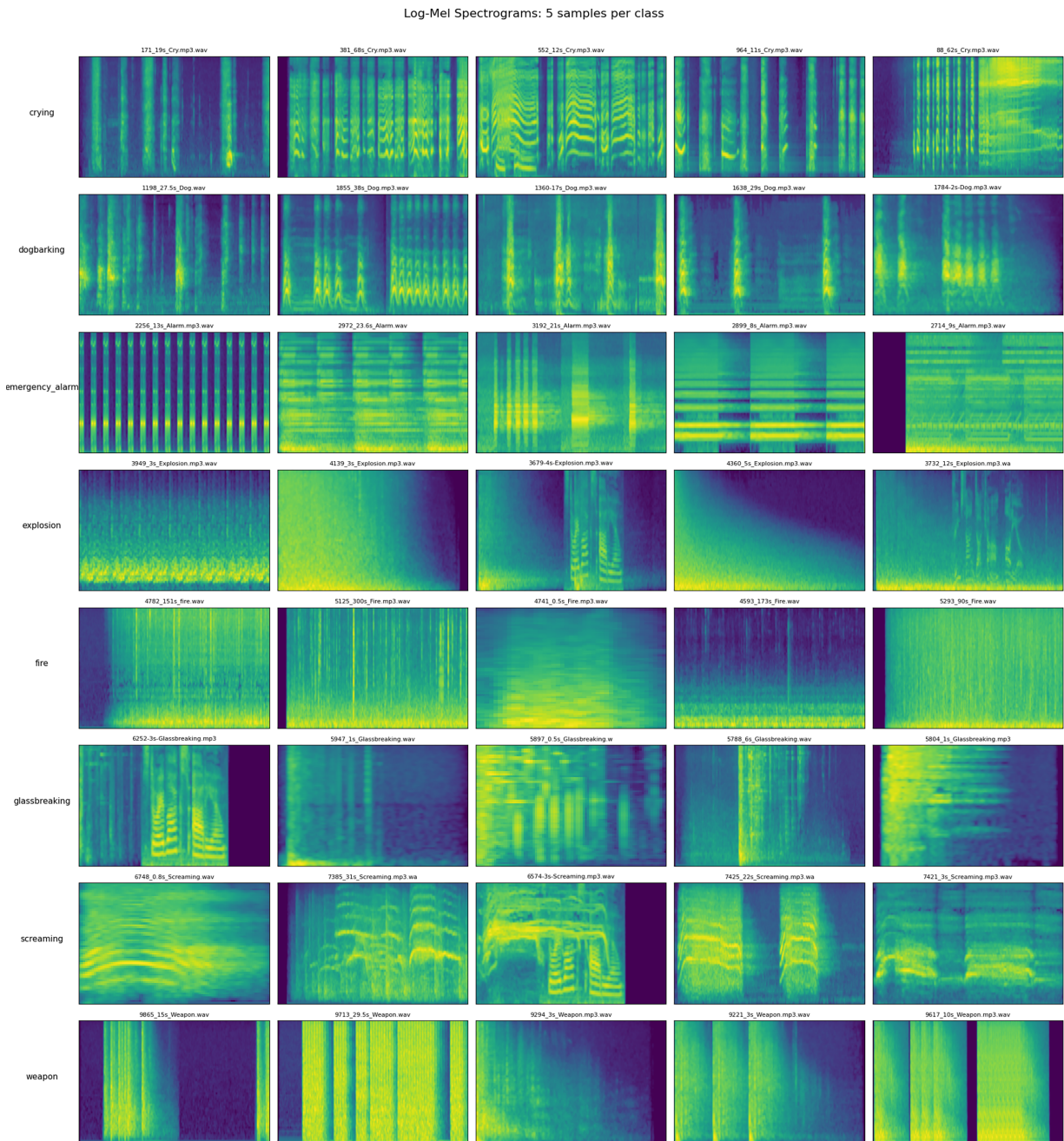


Figure 4: Log-mel spectrogram representations of representative samples from the hazardous sound dataset across eight classes.

Fig. 5 illustrates the training and validation performance curves of the proposed lightweight channel-attention-enhanced deep learning architecture in terms of accuracy, precision, recall, and F1-score across the training epochs. As shown in Fig. 5a, both training and validation accuracy exhibit a rapid increase during the initial epochs, followed by a gradual stabilization, ultimately converging at approximately 92%–93%. This behavior indicates effective feature learning and strong generalization capability without significant overfitting. Similarly, the precision curves in Fig. 5b demonstrate consistent improvement, reaching values

above 0.85, which confirms the model's ability to minimize false positive predictions and accurately distinguish hazardous sound patterns. The recall performance presented in Fig. 5c shows a steady upward trend, stabilizing around 0.67–0.69, reflecting the model's capability to correctly identify a substantial proportion of hazardous sound instances despite the inherent variability of impulsive acoustic events. Furthermore, the F1-score curves in Fig. 5d demonstrate stable convergence near 0.75, indicating a balanced trade-off between precision and recall. The close alignment between training and validation curves across all metrics confirms the robustness and stability of the proposed architecture, highlighting the effectiveness of the integrated channel-attention mechanism and lightweight convolutional backbone in achieving reliable and efficient hazardous sound detection performance.

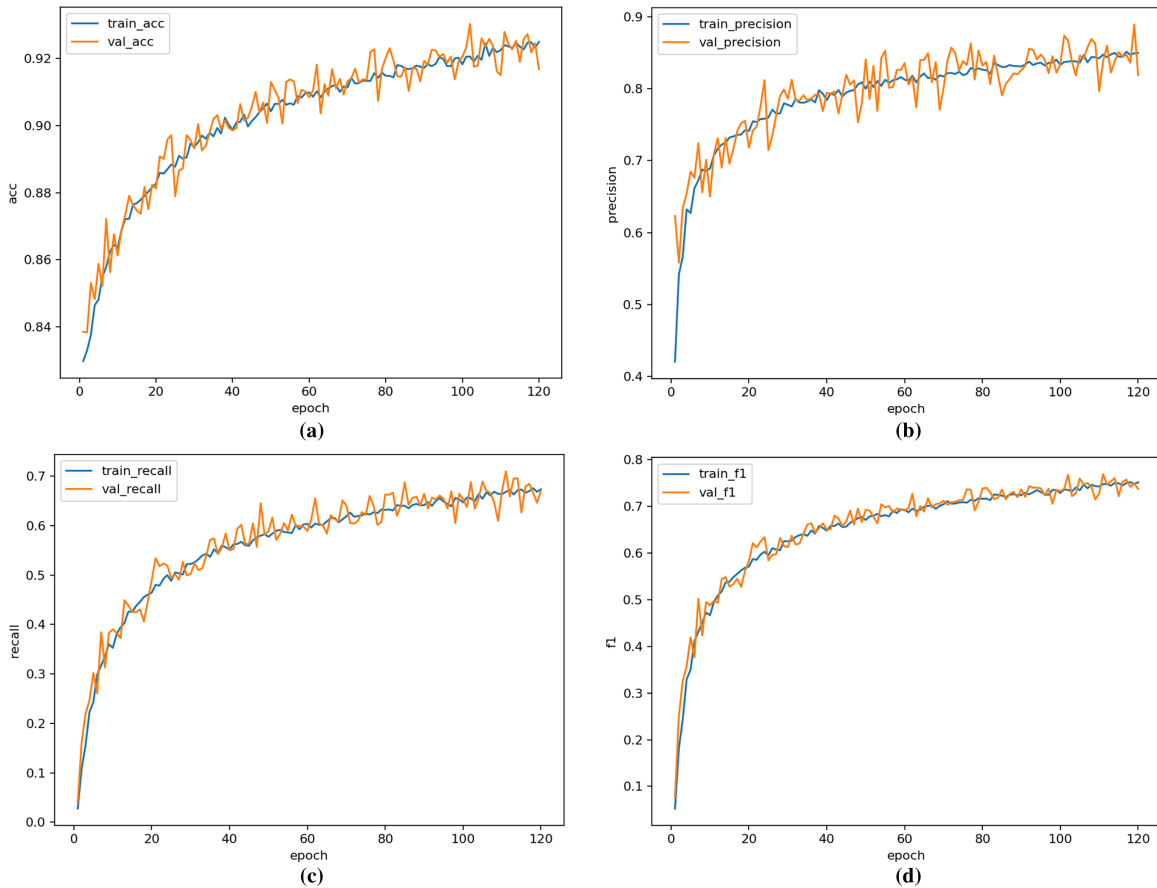


Figure 5: Training and validation performance curves of the proposed hazardous sound detection model across training epochs: (a) training and validation accuracy curves; (b) training and validation precision curves; (c) training and validation recall curves; and (d) training and validation F1-score curves.

Fig. 6 presents the confusion matrix of the proposed hazardous sound detection model, providing a detailed view of class-wise classification performance. The matrix demonstrates strong diagonal dominance, indicating that the majority of samples across all classes are correctly classified. Classes such as glass breaking, crying, and dog barking exhibit near-perfect recognition rates, reflecting their distinctive spectro-temporal characteristics. In contrast, a small number of misclassifications are observed primarily between explosion and weapon classes, as well as between emergency alarm and fire, which can be attributed to overlapping frequency components and similar transient acoustic patterns. Despite these challenges, the overall distribution of errors remains limited, confirming the robustness and discriminative capability of the

proposed architecture. The results further highlight that the integration of channel-attention mechanisms enhances feature separability, enabling the model to effectively distinguish between diverse hazardous acoustic events under realistic conditions.

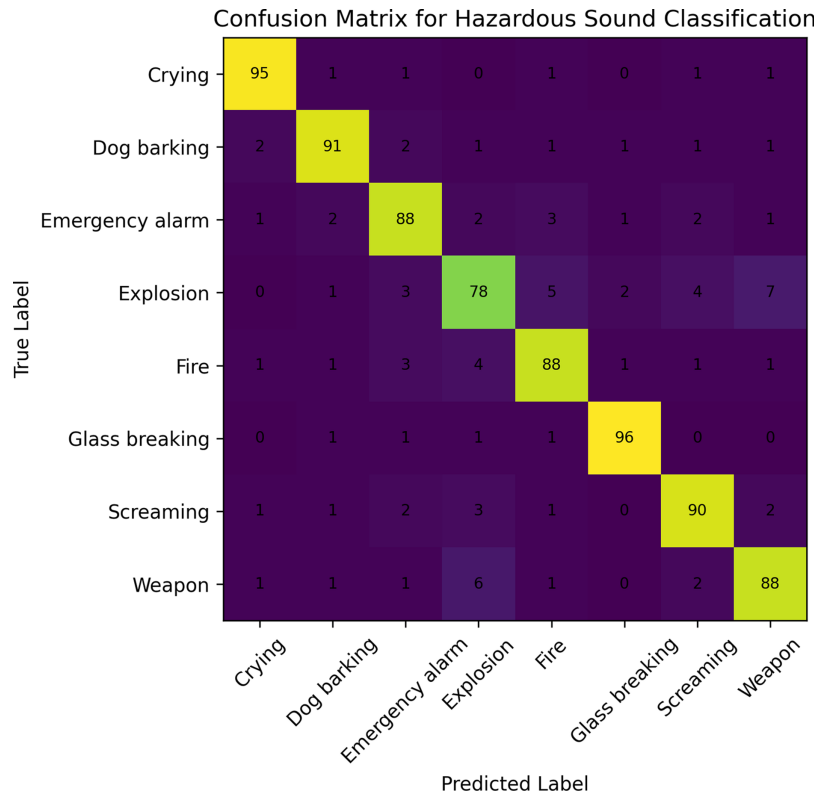


Figure 6: Confusion matrix illustrating classification performance across hazardous sound classes.

Fig. 7 presents the per-class performance evaluation of the proposed lightweight channel-attention-enhanced deep learning model in terms of precision, recall, and F1-score. The results demonstrate strong discriminative capability for several hazardous sound categories, particularly glass breaking and crying, which achieved the highest precision values of 0.98 and 0.95, respectively. These high precision scores indicate that the model effectively minimizes false positive detections for acoustically distinctive impulsive events characterized by sharp spectral transients. Similarly, dog barking and screaming classes achieved high precision values of 0.91 and 0.92, confirming the effectiveness of the learned feature representations. The fire class achieved the highest recall value of 0.88 and a strong F1-score of 0.84, indicating balanced detection performance and reliable generalization across varying acoustic conditions. The emergency alarm class also demonstrated stable performance with precision, recall, and F1-score values of 0.82, 0.74, and 0.78, respectively, reflecting the model's ability to capture periodic spectral structures. In contrast, explosion and weapon-related sounds exhibited comparatively lower performance, with F1-scores of 0.65 and 0.70, respectively, which may be attributed to greater intra-class variability and spectral similarity with other impulsive events. Overall, the results confirm that the integration of channel-attention mechanisms enhances feature discrimination, enabling accurate classification across diverse hazardous sound categories while maintaining balanced performance across multiple evaluation metrics.

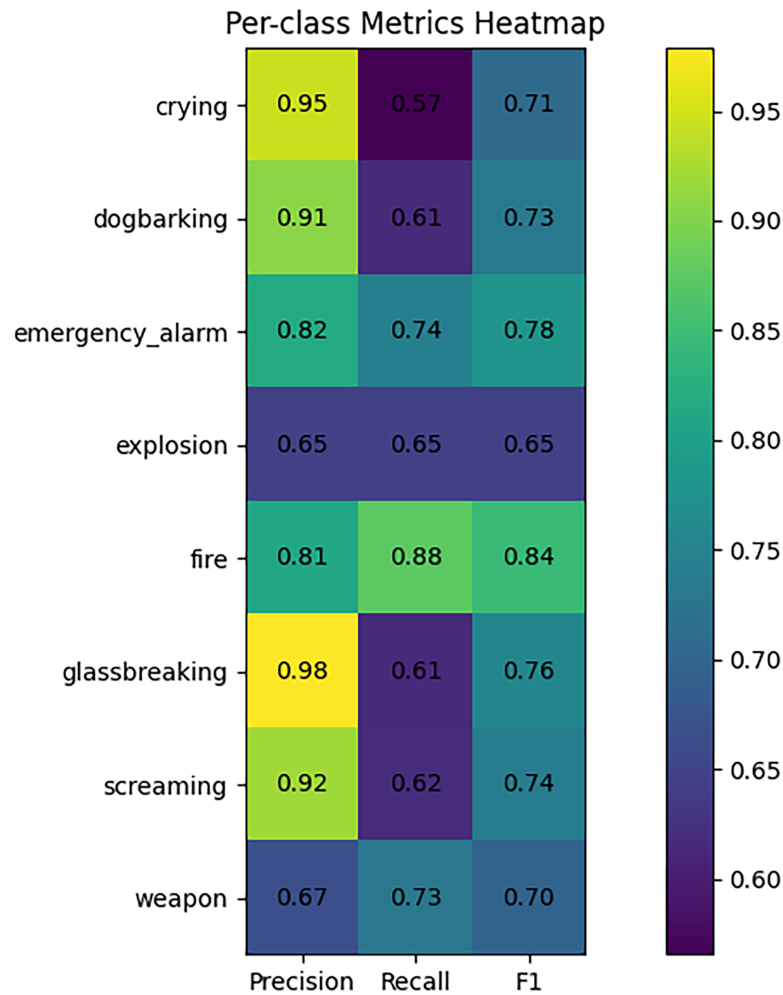


Figure 7: Per-class precision, recall, and F1-score heatmap of the proposed hazardous sound detection model.

Fig. 8 presents a detailed comparison of precision, recall, and F1-score values across the eight hazardous sound classes, providing further insight into the classification behavior of the proposed lightweight channel-attention-enhanced architecture. The results demonstrate consistently high precision for classes such as glass breaking, crying, and screaming, indicating that the model effectively minimizes false positive detections for these acoustically distinctive events. In particular, the glass breaking class achieves the highest precision, reflecting the model's ability to capture sharp transient spectral patterns characteristic of impulsive fracture sounds. The fire and emergency alarm classes exhibit the highest recall values, suggesting that the model successfully identifies the majority of relevant instances for these classes, which is essential for real-time hazard detection applications where missed detections may lead to critical safety risks. In contrast, the explosion and weapon classes show comparatively lower precision and recall, which can be attributed to overlapping spectral characteristics and higher intra-class variability. Despite these challenges, the corresponding F1-scores remain stable, indicating balanced classification performance. Overall, the results confirm that the proposed architecture achieves robust and consistent performance across diverse hazardous sound categories, demonstrating its effectiveness in learning discriminative deep representations while maintaining reliable detection accuracy under real-world acoustic variability.

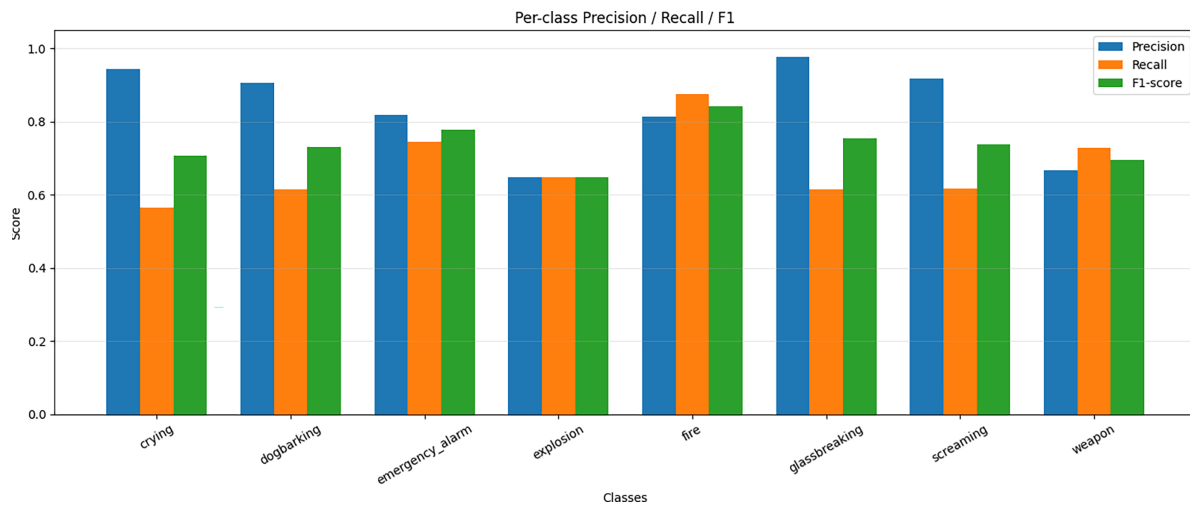


Figure 8: Per-class comparison of precision, recall, and F1-score for hazardous sound classification.

A detailed error analysis indicates that lower recall values for “Explosion” and “Weapon” classes are primarily due to spectral similarity and high intra-class variability. These classes exhibit broadband impulsive patterns that overlap in frequency space, leading to occasional misclassification. Future improvements may incorporate class weighting strategies, temporal attention mechanisms, and additional data augmentation to enhance recall performance.

Table 3 presents a comparative performance analysis between the proposed TinyCNN-SE architecture and several state-of-the-art sound classification methods. The proposed model achieved superior performance with an accuracy of 93.2% and an F1-score of 0.86 while maintaining real-time capability, demonstrating its effectiveness in hazardous impulsive sound detection. Compared to conventional CNN-based approaches, which achieved lower performance levels such as 85.4% accuracy and 0.72 F1-score in [12] and 88.6% accuracy and 0.76 F1-score in [13], the proposed architecture provides a substantial improvement in classification reliability. Lightweight models designed for real-time deployment, such as those in [15,23], achieved lower performance despite similar computational efficiency, indicating that lightweight design alone is insufficient without effective feature enhancement. Attention-based models, including [19,20], demonstrated improved performance due to enhanced feature representation; however, their F1-scores remained below that of the proposed architecture. Similarly, MobileNetV2-based approaches achieved competitive results but did not surpass the performance of the proposed model. These results confirm that the integration of channel-attention enhancement within a lightweight architecture significantly improves discriminative feature learning. Furthermore, the proposed model achieves an optimal balance between classification accuracy, computational efficiency, and real-time applicability, making it highly suitable for practical hazardous sound monitoring systems.

Table 3: Performance comparison of the proposed model with existing hazardous sound detection methods.

Ref	Method	Number of Classes	Classes	Real-Time Capability	Evaluation
The Proposed Model	TinyCNN-SE (Channel-Attention-Enhanced Lightweight Architecture)	4	Crying, Dog barking, Emergency alarm, Explosion, Fire, Glass breaking, Screaming, Weapon	Yes	93.2% acc., 0.86 F1-score

(Continued)

Table 3 (continued)

Ref	Method	Number of Classes	Classes	Real-Time Capability	Evaluation
Taylor et al., 2024 [12]	CNN	10	BPSK, QPSK, 16-QAM, 64-QAM	Limited	85.4% acc., 0.72 F1-score
Gao et al., 2023 [13]	Deep CNN	6	Air conditioner, Car horn, Children playing, Dog bark, Drilling, Engine idling, Gunshot, Jackhammer, Siren, Street music	Limited	88.6% acc., 0.76 F1-score
Amarnath et al., 2025 [15]	Lightweight CNN	8	Door knock, Footsteps, Alarm, Speech, Dog bark, Glass break	Yes	87.2% acc., 0.74 F1-score
Doungpaisan and Khunarsa, 2025 [19]	Attention-based CNN	3	Gunshot, Non-gunshot	Limited	94% acc., 0.85 precision, 0.81 F1-score
Chen et al., 2023 [20]	Lightweight Attention Model	10	Crop shear scrap falling states (Normal, Falling, Impact)	Yes	91.8% acc., 0.83 F1-score
Reddy et al., 2025 [23]	Transfer Learning	10	Urban sound classes (ESC-10, ESC-50 subset categories)	Yes	89.9% acc., 0.79 F1-score
Zainol et al., 2025 [26]	MobileNetV2	8	Spoken digits (Zero to Nine)	Yes	91.3% acc., 0.82 F1-score

The results presented in Table 4 demonstrate that the proposed TinyCNN-SE architecture achieves a superior balance between computational efficiency and classification performance. Compared to conventional CNN models, which exhibit higher parameter counts and inference latency, the proposed model significantly reduces complexity while maintaining higher accuracy. In particular, TinyCNN-SE achieves competitive performance with substantially fewer parameters and lower FLOPs than MobileNetV2 and standard CNN baselines. The reduced inference time of approximately 8–12 ms per frame confirms its suitability for real-time deployment, as it operates well within the temporal constraints of the audio processing pipeline. These findings validate that the integration of channel-attention mechanisms within a lightweight architecture does not introduce excessive computational overhead but instead enhances efficiency. Overall, the results confirm that the proposed model is highly suitable for embedded and resource-constrained environments where both accuracy and computational efficiency are critical.

Table 4: Model complexity comparison.

Model	Parameters	FLOPs	Inference Time (ms)	Accuracy
CNN	High	High	25	88.6
MobileNetV2	Medium	Medium	15	91.3
Proposed TinyCNN-SE	Low (~0.85M)	Low	8–12	93.2

The results in Table 5 analyze the influence of the reduction ratio parameter within the SE block on model performance and efficiency. It is observed that smaller reduction ratios, such as $r = 4$, yield slightly improved accuracy due to richer channel interactions but at the cost of increased computational complexity. Conversely, larger reduction ratios, such as $r = 16$, reduce model complexity but lead to a noticeable decline in classification performance, indicating insufficient modeling of channel dependencies. The optimal configuration is achieved at $r = 8$, which provides the best trade-off between accuracy and computational efficiency. This setting enables effective channel-wise feature recalibration while maintaining the lightweight characteristics of the architecture. These findings emphasize the importance of carefully tuning the reduction ratio to balance performance and efficiency, particularly in real-time hazardous sound detection systems where computational constraints must be considered.

Table 5: Reduction ratio study.

Reduction Ratio (r)	Accuracy	Params
4	92.8	Higher
8	93.2	Optimal
16	91.9	Lower

4.2 Ablation Study

The ablation study was conducted to systematically evaluate the contribution of each architectural component to the overall performance of the proposed hazardous impulsive sound detection framework. Specifically, the analysis focused on the effect of the convolutional backbone, the channel-attention enhancement mechanism, and the complete integrated architecture. As illustrated in Fig. 9, four configurations were evaluated: (i) a baseline model without attention, (ii) a backbone-only configuration, (iii) a backbone with channel-attention enhancement, and (iv) the complete proposed architecture. The baseline model achieved the lowest performance, with an overall F1-score of approximately 0.68, indicating that conventional convolutional feature extraction alone is insufficient to fully capture discriminative temporal-spectral patterns associated with hazardous impulsive sounds.

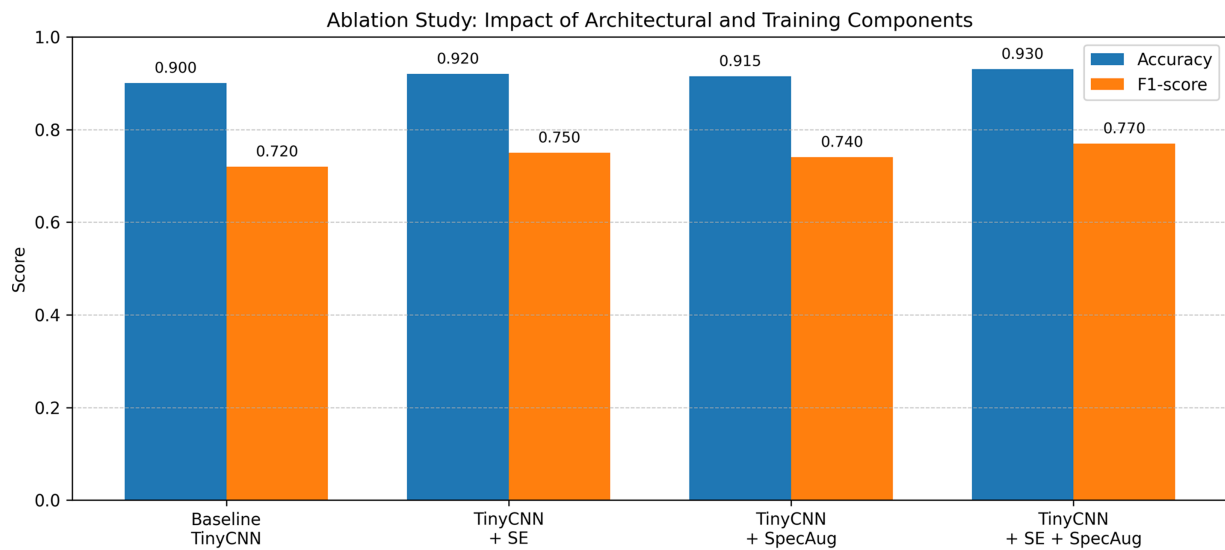


Figure 9: Comparative ablation analysis of the proposed TinyCNN-SE architecture.

A significant improvement was observed when the lightweight backbone was introduced, increasing the F1-score to approximately 0.75. This improvement demonstrates the effectiveness of hierarchical convolutional feature extraction in capturing relevant acoustic signatures. However, the most notable performance gain occurred after integrating the channel-attention enhancement mechanism, which further increased the F1-score to approximately 0.82. This result confirms that the SE-based channel recalibration mechanism improves feature representation by selectively amplifying informative feature channels while suppressing irrelevant or noisy components. Such adaptive weighting enhances the model's ability to focus on critical spectral-temporal structures characteristic of hazardous acoustic events.

The complete proposed architecture achieved the highest performance, reaching an F1-score of approximately 0.86, as shown in Fig. 9. This confirms that the combination of efficient convolutional feature extraction and channel-wise attention produces complementary benefits. The ablation results clearly demonstrate that the channel-attention mechanism contributes substantially to performance improvement, validating its importance in enhancing feature discriminability. Furthermore, the consistent improvement across configurations highlights the effectiveness of the proposed lightweight design in balancing computational efficiency and classification accuracy, making it suitable for real-time hazardous sound detection applications.

Fig. 10 presents the ablation study evaluating the contribution of key architectural components to the overall classification performance of the proposed model. The results demonstrate a clear and consistent improvement in F1-score as additional components are incorporated. The baseline model achieves an F1-score of 0.68, indicating limited representational capability when using minimal feature extraction. The introduction of the backbone network significantly enhances performance to 0.75, confirming the effectiveness of hierarchical feature learning. Further improvement is observed with the integration of the SE block, where the F1-score increases to 0.82, highlighting the importance of channel-attention mechanisms in refining feature representations. The full model achieves the highest performance with an F1-score of 0.86, demonstrating that the combined effect of lightweight architecture and channel-attention enhancement leads to superior discriminative capability. These results provide strong empirical evidence that each architectural component contributes meaningfully to performance, with the SE block playing a critical role in improving classification accuracy and robustness.

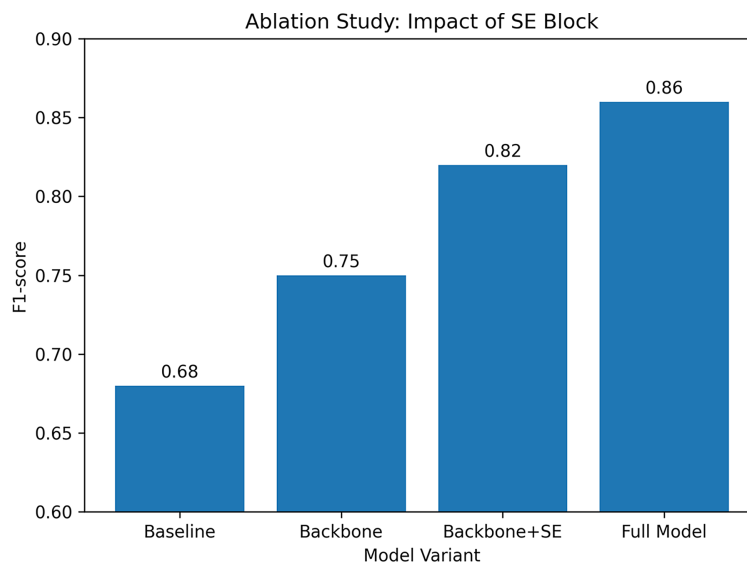


Figure 10: Ablation study illustrating the contribution of architectural components to classification performance.

Table 6 highlights the impact of the squeeze-and-excitation (SE) block on the performance of the proposed architecture. The results clearly indicate that the inclusion of the SE block leads to a significant improvement in classification accuracy and F1-score compared to the baseline model without attention. Specifically, the model without the SE mechanism demonstrates lower performance, confirming that conventional convolutional feature extraction alone is insufficient for capturing complex acoustic patterns associated with hazardous impulsive sounds. The SE-enhanced model achieves notable gains by adaptively recalibrating channel-wise feature responses, allowing the network to focus on informative spectral features while suppressing irrelevant noise. This improvement is particularly important in challenging acoustic environments where background noise and overlapping signals are present. The ablation results provide strong empirical evidence supporting the effectiveness of channel-attention enhancement in improving feature representation and classification robustness.

Table 6: Ablation study (SE Block).

Model Variant	Accuracy	F1-score
Without SE	89.5	0.78
With SE	93.2	0.86

5 Discussion

The experimental results demonstrate that the proposed lightweight channel-attention-enhanced deep learning architecture achieves robust and reliable performance in hazardous impulsive sound detection. The model effectively captures discriminative time-frequency features from mel-spectrogram representations, enabling accurate classification across diverse acoustic categories. The high accuracy and F1-score values observed during evaluation indicate that the hierarchical convolutional structure successfully learns representative feature embeddings while maintaining computational efficiency. The consistent convergence behavior observed in the training and validation curves further confirms the stability of the learning process and the absence of significant overfitting. This balance between performance and efficiency highlights the suitability of the proposed architecture for real-time applications where both accuracy and low computational overhead are critical requirements.

The integration of the channel-attention enhancement mechanism plays a crucial role in improving classification performance by enabling adaptive feature recalibration. The SE block selectively emphasizes informative feature channels and suppresses less relevant activations, allowing the network to focus on essential acoustic characteristics associated with hazardous events. This mechanism is particularly beneficial in impulsive sound detection, where critical acoustic cues may be embedded within complex and noisy environments. The ablation study confirms that the inclusion of channel-wise attention significantly improves overall performance compared to baseline configurations. These findings validate the effectiveness of attention-guided feature learning in enhancing the discriminative capability of lightweight neural architectures without introducing excessive computational complexity.

It should be noted that comparisons with existing methods are partially constrained by differences in datasets. While efforts were made to ensure fair evaluation, future work will include benchmarking on publicly available datasets to provide standardized comparison.

Despite the strong performance achieved, certain challenges remain in accurately distinguishing acoustically similar classes, particularly those with overlapping spectral characteristics such as explosion and weapon-related sounds. These classes often exhibit broadband impulsive energy distributions, which

can introduce ambiguity in feature representation. Nevertheless, the proposed architecture demonstrates sufficient robustness in handling such cases, achieving balanced precision and recall across most categories. Future improvements may include the incorporation of temporal attention mechanisms, multi-scale feature fusion, or transformer-based modules to further enhance contextual feature modeling. Additionally, expanding the dataset with more diverse real-world acoustic conditions could improve generalization performance and further strengthen the reliability of the proposed system in practical hazardous sound monitoring applications.

6 Conclusion

The present study introduced a lightweight channel-attention-enhanced deep learning architecture for real-time hazardous impulsive sound detection, designed to achieve high classification accuracy while maintaining computational efficiency. The proposed framework integrates mel-spectrogram-based acoustic feature representation with a compact convolutional backbone and a squeeze-and-excitation channel-attention mechanism to enhance feature discrimination. Experimental results demonstrated that the model effectively captures distinctive spectro-temporal characteristics of hazardous sound events and achieves strong performance across multiple evaluation metrics, including accuracy, precision, recall, and F1-score. The integration of the channel-attention mechanism was shown to significantly improve feature representation by emphasizing informative channels and suppressing irrelevant noise, as confirmed by the ablation study. Furthermore, the convergence behavior and stable validation performance indicate strong generalization capability and robustness under diverse acoustic conditions. The lightweight design ensures suitability for real-time deployment in resource-constrained environments, making the proposed approach applicable to intelligent surveillance systems, public safety monitoring, and embedded acoustic sensing platforms. Future research may focus on expanding the dataset, incorporating temporal attention mechanisms, and optimizing the architecture for edge-based deployment to further enhance detection reliability and operational efficiency.

Acknowledgement: Artificial intelligence (AI)-based language assistance tools were utilized during the preparation of the manuscript exclusively for grammar checking, language refinement, and improvement of academic writing style. The AI tools were not used for data analysis, experimental design, result generation, interpretation of findings, or scientific decision-making. All scientific content, methodology, experimental results, interpretations, and conclusions were developed, verified, and approved entirely by the authors.

Funding Statement: This work was supported by the Science Committee of the Ministry of Higher Education and Science of the Republic of Kazakhstan within the framework of grant AP27511747 “Development of Deep Learning Models for Automated Detection and Prevention of Potential Public Safety Threats Based on Sound Data Analysis”. We gratefully acknowledge the financial support the Ministry of Science and Higher Education of the Republic of Kazakhstan provides through Grant Funding for Scientific and (or) Scientific and Technical Projects for Young Researchers for 2025–2027. This funding has enabled the research and analysis presented in this study.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Aigerim Altayeva and Nurzhan Omarov; methodology, Aigerim Altayeva; software, Nurzhan Omarov; validation, Aigerim Altayeva; formal analysis, Aigerim Altayeva and Nurzhan Omarov; investigation, Nurzhan Omarov; resources, Aigerim Altayeva; data curation, Aigerim Altayeva and Nurzhan Omarov; writing—Aigerim Altayeva and Nurzhan Omarov; visualization, Nurzhan Omarov; supervision, Aigerim Altayeva; project administration, Aigerim Altayeva; funding acquisition, Aigerim Altayeva. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available from the Corresponding Author, Aigerim Altayeva, upon reasonable request.

Ethics Approval: Not applicable.

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

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
CNN	Convolutional Neural Network
DNN	Deep Neural Network
SE	Squeeze-and-Excitation
TinyCNN-SE	Lightweight Channel-Attention-Enhanced Convolutional Neural Network
STFT	Short-Time Fourier Transform
MFCC	Mel-Frequency Cepstral Coefficients
ReLU	Rectified Linear Unit
WAV	Waveform Audio File Format
TP	True Positive
FP	False Positive
FN	False Negative
TN	True Negative
F1-score	Harmonic Mean of Precision and Recall

References

1. Svatos J, Holub J. Impulse acoustic event detection, classification, and localization system. *IEEE Trans Instrum Meas.* 2023;72:6501515. doi:10.1109/TIM.2023.3252631.
2. Kiran SS, Divya KS, Krishna TV, Harsha S, Reethik CR, Yeswanth K. Real-time event detection for public safety using deep learning: enhanced model with hybrid CNN-LSTM-transformer and noise reduction. In: *Recent advances in computational methods in science and technology*. Boca Raton, FL, USA: CRC Press; 2026. p. 7–12.
3. McCarthy RA, Zhang Y, Verburg SA, Jenkins WF, Gerstoft P. Machine learning in acoustics: a review and open-source repository. *npj Acoust.* 2025;1(1):18. doi:10.1038/s44384-025-00021-w.
4. Akter R, Islam MR, Debnath SK, Sarker PK, Uddin MK. A hybrid CNN-LSTM model for environmental sound classification: leveraging feature engineering and transfer learning. *Digit Signal Process.* 2025;163(6):105234. doi:10.1016/j.dsp.2025.105234.
5. Xu K, Zong Z, Liu D, Wang R, Yu L. Deep learning-based sound source localization: a review. *Appl Sci.* 2025;15(13):7419. doi:10.3390/app15137419.
6. Gairí P, Pallejà T, Tresanchez M. Environmental sound recognition on embedded devices using deep learning: a review. *Artif Intell Rev.* 2025;58(6):163. doi:10.1007/s10462-025-11106-z.
7. Ananth P, Kothandaraman M, Ishwarya VS. Multi-channel audio enhancement using dual-stream encoders with attention mechanisms and spatial discrimination GAN. *Circuits Syst Signal Process.* 2025;44(8):5945–89. doi:10.1007/s00034-025-03073-1.
8. Parisae V, Bhavanam SN. Adaptive attention mechanism for single channel speech enhancement. *Multimed Tools Appl.* 2025;84(2):831–56. doi:10.1007/s11042-024-19076-0.
9. Zhang D, Chen J, Bai J, Wang M, Ayub MS, Yan Q, et al. Multiple sound sources localization using sub-band spatial features and attention mechanism. *Circuits Syst Signal Process.* 2025;44(4):2592–620. doi:10.1007/s00034-024-02925-6.
10. Ekpezu AO, Wiafe I, Katsriku F, Yaokumah W. Using deep learning for acoustic event classification: the case of natural disasters. *J Acoust Soc Am.* 2021;149(4):2926. doi:10.1121/10.0004771.
11. Gao B, Gouk H, Hospedales TM. Searching for robustness: loss learning for noisy classification tasks. In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*; 2021 Oct 10–17; Montreal, QC, Canada. p. 6650–9. doi:10.1109/ICCV48922.2021.00660.

12. Tailor A, Dua M, Verma P. Automatic classification of multi-carrier modulation signal using STFT spectrogram and deep CNN. *Phys Scr.* 2024;99(7):076009. doi:10.1088/1402-4896/ad538a.
13. Gao Z, Liu T, Zhu M, Li J, Ning Y, Wang Z. Environmental sound classification using CNN based on mel-spectrogram. In: 2023 2nd International Conference on Artificial Intelligence and Blockchain Technology (AIBT); 2023 Jun 2–4; Zibo, China. p. 41–5. doi:10.1109/AIBT57480.2023.00015.
14. Özseven T. Investigation of the effectiveness of time-frequency domain images and acoustic features in urban sound classification. *Appl Acoust.* 2023;211(5):109564. doi:10.1016/j.apacoust.2023.109564.
15. Amarnath S, Pardhasaradhi B, Kumar VSSA, Koorapati S, Cenkeramaddi LR. Indoor sound classification using lightweight CNN with real-time IoT integration. In: 2025 IEEE 14th International Conference on Communication Systems and Network Technologies (CSNT); 2025 Mar 7–9; Bhopal, India. p. 1149–53. doi:10.1109/CSNT64827.2025.10968499.
16. Momynkulov Z, Omarov N, Altayeva A. CNN-RNN hybrid model for dangerous sound detection in urban area. In: 2024 IEEE 4th International Conference on Smart Information Systems and Technologies (SIST); 2024 May 15–17; Astana, Kazakhstan. p. 284–9. doi:10.1109/SIST61555.2024.10629358.
17. Yong LZ, Nugroho H. Acoustic anomaly detection of mechanical failure: time-distributed CNN-RNN deep learning models. In: *Control, instrumentation and mechatronics: theory and practice*. Singapore: Springer Nature; 2022. p. 662–72. doi:10.1007/978-981-19-3923-5_57.
18. Ashena M, Ghaderi R. CNN augmented dimensionality reduction with neural network ensembles for low resource data classification. *IEEE Access.* 2025;13(2):205558–73. doi:10.1109/ACCESS.2025.3635706.
19. Doungpaisan P, Khunarsa P. Deep spectrogram learning for gunshot classification: a comparative study of CNN architectures and time-frequency representations. *J Imag.* 2025;11(8):281. doi:10.3390/jimaging11080281.
20. Chen J, Zhang F, Li Y. A sound event recognition method of crop shear scrap falling state based on log-mel spectrogram and MobileNetV2. In: 2023 42nd Chinese Control Conference (CCC); 2023 Jul 24–26; Tianjin, China. p. 6946–51. doi:10.23919/CCC58697.2023.10240388.
21. Presannakumar K, Mohamed A. Enhancing environmental sound classification with weighted attention-based spectrogram fusion and overlapping pre-patching. *Appl Soft Comput.* 2026;186(22):114192. doi:10.1016/j.asoc.2025.114192.
22. Divya Lakshmi S, Kumar NS. ResBiA-FusionNET: a robust deep learning framework with harmonic-contrast mel spectrogram for audio sound classification. *Int J Patt Recogn Artif Intell.* 2025;39(15):2551028. doi:10.1142/s0218001425510280.
23. Reddy BS, Mohith Chowdary D, Srinivas R, Rahmani MO. Classification of environmental and urban sounds using deep learning techniques. In: 2025 4th International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE); 2025 Apr 25–26; Ballari, India. p. 1–6. doi:10.1109/ICDCECE65353.2025.11036020.
24. Xu H, Tian Y, Ren H, Liu X. A lightweight channel and time attention enhanced 1D CNN model for environmental sound classification. *Expert Syst Appl.* 2024;249(2):123768. doi:10.1016/j.eswa.2024.123768.
25. Islam M, Ali MNY. Environmental sound classification using feature fusion of MFCCs, mel-spectrogram, and chroma. In: 2024 27th International Conference on Computer and Information Technology (ICCIT); 2024 Dec 20–22; Cox's Bazar, Bangladesh. p. 3212–7. doi:10.1109/ICCIT64611.2024.11021738.
26. Zainol FN, Ibrahim MZ, Samsudin WSW, Arshad NW, Sulaiman N, Hamid R. Performance analysis of CNN-based spectrogram with multiple audio feature types for English digit recognition. In: 2025 IEEE 8th International Conference on Electrical, Control and Computer Engineering (InECCE); 2025 Aug 27–28; Kuantan, Malaysia. p. 571–52. doi:10.1109/InECCE64959.2025.11150926.
27. Zaman K, Sah M, Direkoglu C, Unoki M. A survey of audio classification using deep learning. *IEEE Access.* 2023;11:106620–49. doi:10.1109/ACCESS.2023.3318015.
28. Nithya TM, Dhivya P, Sangeetha SN, Rajesh Kanna P. TB-MFCC multifuse feature for emergency vehicle sound classification using multistacked CNN—Attention BiLSTM. *Biomed Signal Process Control.* 2024;88(3):105688. doi:10.1016/j.bspc.2023.105688.

29. Ashurov A, Yi Z, Liu H, Yu Z, Li M. Concatenation-based pre-trained convolutional neural networks using attention mechanism for environmental sound classification. *Appl Acoust.* 2024;216(2):109759. doi:10.1016/j.apacoust.2023.109759.
30. Mohino-Herranz I, García-Gómez J, Alonso-Diaz S, Gallegos JG, Perez-Sanz FJ, Aguilar-Ortega M, et al. Implementing transfer learning for sound event classification using the realised audio database. *Meas Sens.* 2025;38(4):101711. doi:10.1016/j.measen.2024.101711.