Security Enhancement Using Sound Classification

Sandeep Kumar Amgothu
Texas A & M University – Corpus Christi
Corpus Christi, Texas, USA
samgothu@islander.tamucc.edu

Bhavya Sree Ganja
Texas A & M University – Corpus Christi
Corpus Christi, Texas, USA
bganja@islander.tamucc.edu

Kalyan Kumar Reddy Bontha
Texas A & M University – Corpus Christi
Corpus Christi, Texas, USA
kbontha@islander.tamucc.edu

Jyothsna Yekula
Texas A & M University – Corpus Christi
Corpus Christi, Texas, USA
jyekula@islander.tamucc.edu

Abstract

Sound is crucial in real-time threat detection, particularly in environments where visual surveillance may fall short. This paper presents a comprehensive study on dangerous sound classification using deep learning techniques and spectrogram-based features. We aim to enable artificial intelligence systems to recognize and classify critical security-related sounds, such as gunshots, explosions, alarms, and natural disasters, thereby enhancing situational awareness and response capabilities.

We evaluate and compare eight deep learning models: Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Convolutional Recurrent Neural Networks (CRNN), and transfer learning with VGG19—each trained on both Short-Time Fourier Transform (STFT) and Mel spectrogram representations. The dataset was segmented into one-second audio clips and converted into spectrogram images for training and evaluation. The CNN architecture demonstrated the highest accuracy and robustness among all models, especially in noisy environments and overlapping audio events. Our results highlight the effectiveness of combining spatial and temporal audio features for real-world security applications.

Keywords

Security, Sound Classification, Deep Learning, Spectrogram, CNN, RNN, CRNN, VGG19, STFT, Mel

1 Introduction

Security threats are increasingly dynamic and multifaceted, often rendering traditional surveillance systems insufficient. While video-based monitoring remains a cornerstone of security infrastructure, it suffers from inherent limitations such as low visibility in poor lighting, camera blind spots, and the inability to detect off-camera events. In contrast, sound propagates in all directions and can be an early warning signal for critical incidents, including gunfire, explosions, glass breaking, and alarms. Despite this, sound-based threat detection remains underutilized in practical deployments.

The primary goal of this research is to integrate intelligent sound classification into modern security systems to enhance their responsiveness and reliability. We propose a deep learning-based solution that automatically detects and classifies security-relevant sounds in real time. Our study evaluates multiple neural architectures, including Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Convolutional Recurrent Neural Networks (CRNN), and a transfer learning model based on VGG19. These models are trained and tested on both Short-Time Fourier Transform (STFT) and Mel spectrogram representations to identify dangerous acoustic events across 13 categories.

1.1 Research Questions

To guide our investigation, we address the following research questions:

- **RQ1:** In what ways can deep learning architectures enhance the precision and robustness of detecting and categorizing security-related audio signals?
- **RQ2:** Between STFT and Mel spectrograms, which time-frequency representation provides better discrimination of critical sounds from environmental noise?
- **RQ3:** How effectively can machine learning models handle overlapping or distorted audio events that occur in realistic, unpredictable settings?

Pros and Cons of Existing Methods: Previous studies have demonstrated that CNNs can effectively extract spatial features from spectrograms, while RNNs are better suited for handling temporal dependencies. CRNNs combine both strengths, offering improved performance for sequential audio signals. However, these models often suffer from limited generalizability due to narrow class scopes and insufficient data diversity. Furthermore, they may falter in noisy or overlapping environments where clear signal patterns are harder to isolate.

To overcome these limitations, our approach expands on existing work by training and evaluating a comprehensive suite of deep learning models across a broad range of acoustic event types. We employ STFT and Mel spectrograms for feature extraction and introduce VGG19 for transfer learning, capitalizing on its pretrained visual feature capabilities. By integrating CNN and RNN components into CRNNs and leveraging robust spectrogram transformations, our methodology ensures high classification accuracy and adaptability, even in acoustically complex or noisy environments. This hybrid architecture reduces false alarms and enhances situational awareness in critical real-time security applications.

2 Background

Traditional surveillance systems rely primarily on visual data, which becomes ineffective when visibility is compromised due to darkness, occlusion, or limited field of view. In contrast, audio sensors can capture acoustic events from all directions, including those beyond the visual scope, making them valuable for early-stage threat

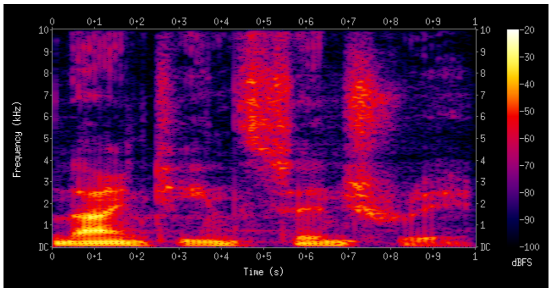

Figure 1: Short-Time Fourier Transform (STFT) spectrogram showing time-frequency distribution of sound.

detection. However, the deployment of automated, sound-based detection systems poses several significant challenges:

- **Sound Complexity**: Security-relevant sounds such as explosions or gunfire differ widely in frequency, intensity, and duration.
- **Overlapping Audio Events**: Real-world scenes often include multiple concurrent sounds—e.g., sirens, human screams, and alarms—that complicate isolated sound identification.
- **Background Noise**: Environments such as urban streets or industrial zones present diverse noise profiles, increasing the risk of misclassification.

An effective solution must extract robust features from raw audio signals and accurately classify them, even under noisy, overlapping, or ambiguous conditions. Deep learning models, particularly when used with spectrogram visualizations, provide a scalable method for addressing these challenges by learning meaningful temporal and frequency patterns from sound data.

The Short-Time Fourier Transform (STFT) breaks down an audio signal into time-localized frequency components, enabling convolutional layers to detect localized energy variations useful for sound classification. The mathematical formulation of STFT is:

$$X(t,\omega) = \int_{-\infty}^{\infty} x(\tau)\, w(\tau - t)\, e^{-j\omega\tau}\, d\tau \tag{1}$$

where $x(\tau)$ is the input signal, $w(\tau - t)$ is a window function centered at time t, and ω is the angular frequency. This results in a complex-valued matrix $X(t,\omega)$ representing frequency content over time.

The Mel spectrogram is derived from the power spectrogram of the STFT by applying a Mel filter bank that mimics human auditory perception. It is computed as:

$$M(m,t) = \sum_{f=0}^{F-1} H_m(f) \cdot |X(f,t)|^2 \tag{2}$$

where $|X(f,t)|^2$ is the power spectrogram and $H_m(f)$ is the m-th triangular Mel filter applied across F frequency bins. This mapping emphasizes perceptually important frequency ranges and is particularly effective in identifying sounds like alarms, gunshots, or speech amidst environmental noise.

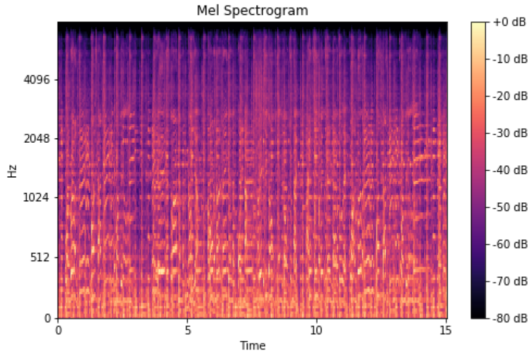

Figure 2: Mel spectrograms map audio frequencies to a perceptually meaningful scale.

3 Related Work

Several studies have significantly advanced acoustic signature analysis by utilizing various methodologies, each offering distinct techniques for detecting and classifying sounds, particularly those linked to dangerous situations.

Atalar [3] introduced one of the earliest frameworks for acoustic signature analysis through a physical modeling approach. This foundational study primarily focused on the theoretical constructs and physical characteristics of acoustic signals rather than computational techniques. Atalar's main objective was to mathematically characterize acoustic signals to understand their propagation and interaction with environments.

While Atalar's work and our study aim to analyze acoustic signatures systematically, there are notable differences in methodology and application. Atalar's approach did not incorporate machine learning, which limits its adaptability to the dynamically changing acoustic environments typical in modern security contexts. In contrast, our research employs advanced machine learning models, particularly deep learning architectures like CNNs, RNNs, and CRNNs. These models enable automatic adaptation to varying acoustic scenarios and complex sound patterns encountered in real-world security situations.

Bernardini et al. [4] utilized Convolutional Neural Networks (CNNs) to classify sounds produced by drones and differentiate them from non-drone acoustic signals through spectrogram analysis. Their main objective was to identify unique acoustic signatures specifically associated with drone operations. Similarly to our approach, Bernardini et al. employed spectrogram-based CNNs, which effectively extract important frequency-time features for audio classification tasks. However, our research diverges significantly by broadening the classification scope beyond just drone-specific sounds. We also include various hazardous acoustic signals, such as gunshots, alarms, explosions, and breaking glass. This comprehensive approach enhances the versatility and applicability of our system across different security scenarios.

Seo et al. [11] developed an advanced methodology combining CNNs with Short-Time Fourier Transform (STFT) features specifically for drone detection. Their technique strongly aligns with our work, as both approaches utilize CNNs and spectrogram-based

Table 1: Comparison of Some Related Works on Sound Classification for Security

Author(s)	Year	Approach	Deep Learning	Feature Extraction	Limitations
Atalar [3]	1979	Physical Model	No	N/A	No ML Adaptability
Bernardini et al. [4]	2017	CNN	Yes	Spectrogram	Focused Only on Drones
Seo et al. [11]	2018	CNN + STFT	Yes	Time-Frequency Analysis	No Multi-Class Classification
Yazid et al. [15]	2022	Entropy Thresholding	No	Image Segmentation	Indirect Application to Audio
Momynkulov et al. [8, 9]	2023	CRNN	Yes	MFCC, Spectrogram	No Real-Time Application
Ashikuzzaman et al. [2]	2021	Deep Learning	Yes	Spectrogram	Requires More Robust Training
Hamza et al. [7]	2022	ML+MFCC	Yes	MFCC	Narrow application focus
Altayeva [1]	2024	ML-based Alerts	Yes	Spectrogram	Real-time application limited

features for analyzing acoustic signals. However, Seo et al.'s application was limited exclusively to drone sounds, while our research expands this approach to classify a broader range of security-related sounds, increasing the practical applicability of our model in diverse environments.

Yazid et al. [15] explored entropy thresholding methods for segmenting visual data. Although their research primarily targeted visual domains, their advanced feature extraction methodologies indirectly inform our study by providing valuable insights into effective pattern analysis techniques applicable to audio spectrograms. Nevertheless, unlike Yazid et al., our research directly applies such feature extraction methods in acoustic signal classification using deep learning.

Momynkulov et al. [8, 9] significantly advanced urban sound classification using Convolutional Recurrent Neural Networks (CRNNs). Their research emphasized CRNNs' capacity to effectively capture both spatial and temporal dependencies inherent in audio signals. Our approach similarly integrates CRNNs for enhanced acoustic classification performance, especially useful in complex auditory environments common in security contexts. However, we further optimize CRNNs specifically for real-time security monitoring applications.

Ashikuzzaman et al. [2] provided practical demonstrations of deep learning methods for detecting dangerous sounds in real-world scenarios. This work underscores the practicality and applicability of neural network approaches, aligning closely with our research objectives. Nonetheless, our study introduces advanced feature extraction techniques and incorporates a broader spectrum of sounds, significantly enhancing the classification accuracy and adaptability of our model.

Smailov et al. [12] proposed a CNN-RNN hybrid model specifically designed for recognizing impulsive acoustic events. Their framework effectively combines convolutional and recurrent components, enabling it to capture complex audio patterns. Our work builds on this foundation by employing more advanced model designs, incorporating additional feature engineering techniques, and focusing heavily on real-time classification performance in challenging acoustic conditions.

Byun and Lee [5] focused on developing sound-based alert systems for wearable technologies, intended to support individuals with hearing impairments. While their goal was to detect specific sound cues, their application remained narrow. In contrast, our research targets general-purpose public safety use cases. Our system is designed to function across a wide range of environments and sound types, offering broad-spectrum detection capabilities.

Foggia et al. [6] explored traditional machine learning methods for classifying environmental sounds in transportation surveillance. While effective for simpler tasks, their models lacked adaptability in complex scenarios. In contrast, we utilize modern deep learning frameworks capable of extracting hierarchical patterns from audio spectrograms, allowing the system to interpret both spatial and sequential sound information more effectively.

Tokozume et al. [14] introduced techniques for enhancing sound recognition using between-class learning examples, which closely aligns with our training methodologies. However, our study uniquely integrates multiple deep learning frameworks, aiming for comprehensive and precise classification across diverse acoustic scenarios typically encountered in security applications.

Saba et al. [10] reviewed technologies used specifically for scream detection, highlighting the importance of audio signals in security contexts. Our research expands on this foundation by integrating various dangerous sound classifications through advanced deep learning architectures, ensuring a comprehensive acoustic threat detection system.

Suh et al. [13] developed deep learning systems designed explicitly to classify hazardous sounds for aiding hearing-impaired individuals. Though conceptually related, our research diverges significantly in its targeted optimization for broader security applications, using integrated CNN and RNN models to exploit spatial and temporal audio characteristics comprehensively.

Hamza et al. [7] focused their study on deepfake audio detection using Mel Frequency Cepstral Coefficients (MFCC). While our work similarly employs MFCC for effective feature extraction, it significantly differs by applying these techniques to a broader range of authentic acoustic events relevant to security contexts, using multiple advanced deep learning architectures.

Altayeva [1] proposed a public alert system using machine learning for dangerous sound detection. Our approach similarly targets security and public safety. Still, it uniquely emphasizes real-time automated detection and classification through sophisticated integration of CNN and RNN models, enhancing reliability and responsiveness in diverse operational scenarios.

Differences Between Our Approach and Existing Work:

Most previous works focused on standalone architectures such as Convolutional Neural Networks (CNNs) or Recurrent Neural

Networks (RNNs) for dangerous sound classification. While effective in narrow use-cases like drone detection or scream recognition, they often failed to generalize across multiple sound types or real-time noisy environments. For example, Bernardini et al. [4] and Seo et al. [11] applied CNN-based methods solely to drone acoustics, without leveraging temporal features critical for overlapping or sequenced sound events. Smailov et al. [12] introduced a hybrid CNN-RNN model but did not employ advanced preprocessing or transfer learning, which limited their scalability and robustness.

In contrast, our proposed system combines both traditional deep learning architectures (CNN, RNN, CRNN) with a transfer learning approach based on VGG19. The VGG19 model, pre-trained on large image datasets, is adapted to spectrogram inputs, providing high-quality feature embeddings and deeper representational capacity. This model particularly excels in capturing fine-grained spatial patterns from spectrograms, which simpler models might overlook.

Furthermore, our pipeline utilizes both Short-Time Fourier Transform (STFT) and Mel Spectrograms for feature extraction, ensuring compatibility with both human-auditory scaling and frequency-based precision. These spectrograms are used as inputs across all 8 trained models, with VGG19 providing an especially effective high-level representation layer.

Advantages of the Proposed Approach:

Our approach integrates a suite of models—CNN, RNN, CRNN, and VGG19—trained on STFT and Mel spectrograms. This hybrid strategy captures temporal and spatial dependencies through CRNN architectures and leverages transfer learning via VGG19 to boost performance, especially in smaller datasets or hard-to-detect sound patterns.

The CRNN models effectively detect temporal sequences in noisy environments using bidirectional LSTM layers, while the VGG19 architecture enhances spatial feature learning through its deep convolutional layers. As a result, the combined architecture supports high-accuracy, real-time classification across 13 security-critical sound categories.

Compared to earlier work, our approach improves robustness, generalizability, and response time. It minimizes false positives and negatives by optimizing model selection and feature design, leading to superior performance in diverse and unpredictable security scenarios.

3.1 Performance of Existing Methods

Recent studies in audio classification have shown varying levels of performance depending on the application and the techniques employed. Bernardini et al. [4] reported an accuracy of around 80% using CNNs for drone audio detection, demonstrating CNNs' effectiveness for specialized tasks. However, their method lacked adaptability for broader security scenarios. Similarly, Seo et al. [11] achieved approximately 82% accuracy using STFT-based CNNs for drone classification, reinforcing the utility of spectrogram-based CNNs—yet their solution was restricted to drone-related events.

Momynkulov et al. [8] used a CRNN model to classify dangerous urban sounds, achieving an accuracy of approximately 78%. Though this showcased the benefits of incorporating temporal dependencies, the system lacked optimization for real-time usage and robustness in complex environments. Ashikuzzaman et al. [2] also demonstrated 80.2% accuracy in detecting distress sounds using CNNs, while Hamza et al. [7] used MFCCs and classical ML techniques to detect deepfake audio with an accuracy of 77.5%.

Table 2: Performance Comparison of Existing Approaches in Sound Classification

Author(s)	Approach	Accuracy (%)
Bernardini et al. [4]	CNN	80.0
Seo et al. [11]	CNN + STFT	82.0
Momynkulov et al. [8]	CRNN	78.0
Ashikuzzaman et al. [2]	CNN	80.2
Hamza et al. [7]	ML + MFCC	77.5
Ours (CNN-STFT)	**CNN + STFT + 13-Class Multiclass**	**91.56**

These methods highlight both the promise and limitations of current approaches. They often focus on narrow domains, lack temporal modeling, or omit transfer learning—leading to reduced generalization in unpredictable, real-world settings.

Our Contribution: Our proposed system distinguishes itself by combining CNN, RNN, CRNN, and VGG19 architectures with Short-Time Fourier Transform (STFT) and Mel spectrogram inputs. This enables learning of spatial and temporal characteristics while also benefiting from deep pre-trained feature extraction via VGG19. The hybrid use of handcrafted and transfer-learned representations allows our models to generalize across overlapping, noisy, and real-world acoustic events.

Unlike prior methods, our architecture supports multi-class classification across 13 distinct sound types, optimized for real-time performance. The result is a robust, scalable system with superior precision, recall, and classification accuracy across diverse security-related scenarios.

4 Proposed Method

The proposed methodology for detecting and classifying dangerous sounds utilizes a multi-model deep learning framework, including Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Convolutional Recurrent Neural Networks (CRNN), and transfer learning using VGG19. This ensemble approach ensures robust classification performance across a variety of acoustic scenarios, with each model optimized for different aspects of sound representation.

Data Preprocessing

Raw audio recordings are segmented into 1-second clips using the `pydub` library. These uniform segments are then split into training (70%) and testing (30%) sets, maintaining class balance across 13 dangerous sound categories such as explosions, gunshots, alarms, breaking glass, dog barks, and sirens.

Feature Extraction with Spectrograms

Audio clips are converted into Mel and STFT spectrograms using the `librosa` library. These spectrograms transform audio into time-frequency images suitable for deep learning. Parameters include:

- Sampling Rate: 22050 Hz
- FFT Window Size: 2048
- Hop Length: 512
- Mel Bins: 128

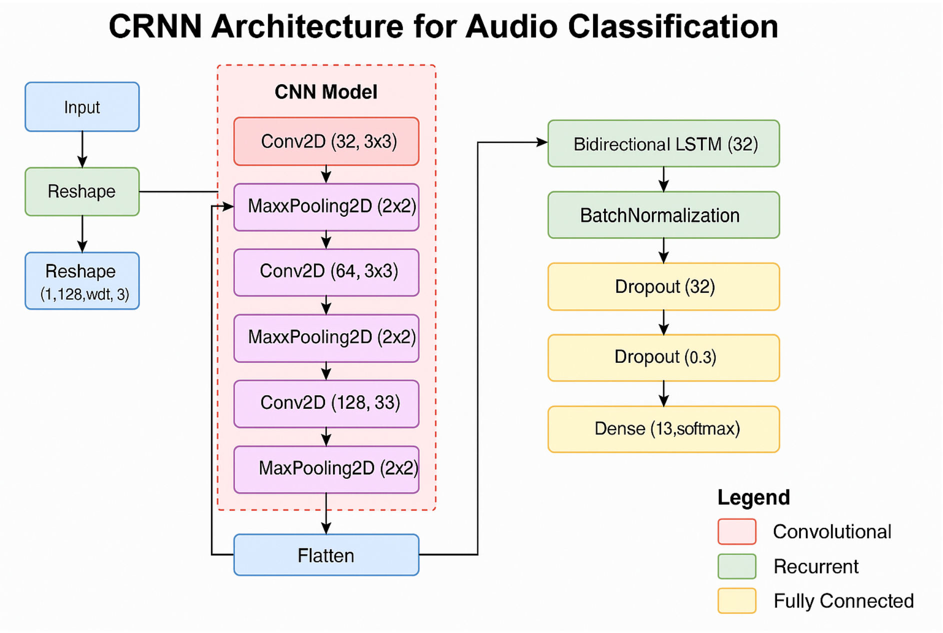

Figure 3: Proposed CRNN architecture illustrating the combination of convolutional layers, pooling layers, bidirectional LSTM, dense, dropout, and softmax layers for accurate audio classification.

Spectrograms are saved as 128×128 RGB images for input into CNN-based models, including VGG19.

Data Augmentation

To enhance generalization and reduce overfitting, we employ `ImageDataGenerator` from Keras, applying random zoom, shearing, and horizontal flips to spectrogram images during training.

CNN Architecture

Our CNN models are designed to learn spatial patterns directly from the spectrogram images. Each CNN architecture includes three convolutional layers with increasing filter sizes (32, 64, and 128), each followed by ReLU activation and max pooling. Batch normalization is used to stabilize training, and the output is passed through fully connected layers ending in a softmax layer for classification. CNNs perform particularly well in capturing localized frequency-time patterns and are computationally efficient for real-time applications.

RNN Architecture

The RNN architecture focuses on modeling the temporal dynamics present in the spectrogram sequences. We utilize Long Short-Term Memory (LSTM) layers to capture sequential dependencies across time frames. Although RNNs lack the spatial awareness of CNNs, they are useful in scenarios where the evolution of sound over time is a distinguishing factor. Our RNN includes two stacked LSTM layers followed by dense and dropout layers, concluding in a softmax output.

CRNN Architecture

Our primary architecture is a CRNN, which integrates:

- Three convolutional layers (filters: 32, 64, 128) with ReLU activation
- Max pooling and batch normalization
- Bidirectional LSTM layers to model temporal audio dynamics
- Dense and dropout layers for regularization
- Final softmax layer for 13-class output

VGG19-Based Architecture (Transfer Learning)

We also implement transfer learning using VGG19, a deep CNN pretrained on ImageNet. The spectrograms are resized to 128×128×3 to mimic RGB images. The VGG19 model is used as a frozen base, and the top layers are replaced with:

- Global Average Pooling layer
- Dense layer with 256 units and ReLU activation
- Dropout (rate = 0.5)
- Softmax output layer for classification

Only the newly added top layers are trained, while the base convolutional layers remain frozen to retain learned visual features. This approach significantly improves feature richness and classification performance, especially on less frequent or visually complex audio classes.

Summary

The combination of CRNN and VGG19 allows us to capture both the temporal evolution of sounds (via RNN layers) and fine-grained spatial patterns in spectrograms (via VGG19). This dual-model strategy ensures high accuracy, generalization, and robustness, particularly in noisy, real-world environments with overlapping audio events.

5 Dataset Description

To train and evaluate our sound classification models, we compiled a custom dataset consisting of 4366 spectrogram images derived from real-world audio recordings. The dataset includes 3015 training samples and 1351 testing samples, all distributed across 13 distinct security-related sound classes.

Data Sources and Acquisition

Most of the audio samples were sourced from real-life YouTube videos and open-access sound repositories. We manually recorded some sound events using laptop microphones and sound recording tools to simulate realistic conditions. Additional short clips were extracted from Google search results and other verified audio platforms. This approach allowed us to build a dataset that better reflects the complexity and diversity of acoustic environments in real-world security scenarios.

Sound Classes

The dataset includes the following 13 categories, selected for their relevance to public safety and security monitoring:

- Gunshots
- Explosions
- Emergency alarms
- Glass breaking
- Thunderstorms
- Dog barking
- Floods
- Earthquakes
- Wind
- Wildfires
- Tsunami
- Volcanic eruptions
- Human screams / sirens

Preprocessing and Augmentation

Each audio clip is segmented into 1-second intervals using pydub, ensuring a consistent time window across all samples. The audio is then transformed into spectrogram images using `librosa`, generating both:

- **STFT Spectrograms** — capturing detailed frequency evolution over time
- **Mel Spectrograms** — perceptually scaled to match human hearing

All spectrograms are resized to 128×128×3 pixels and stored as RGB images. Image augmentation using `ImageDataGenerator` introduces synthetic variability, including zooming, shearing, and flipping, to improve generalization.

Dataset Composition

The dataset is structured as follows:

- **Training set:** 3015 spectrogram images (13 classes)
- **Testing set:** 1351 spectrogram images (13 classes)

Manual verification ensured that each class was accurately labeled. The dataset maintains balance across all classes to prevent model bias during the training and testing phases.

Uniqueness and Motivation

Unlike conventional datasets like ESC-50 or UrbanSound8K, which are either too generic or limited to urban contexts, our dataset is purpose-built for security applications. By including multiple real-world sources and a diverse range of security-critical sounds, it offers a realistic foundation for training robust deep learning models capable of reliable real-time detection.

Table 3: Performance Metrics Across All Models on the Test Set

Model	Accuracy	Precision	Recall	F1-score
CNN (STFT)	**91.56%**	0.92	0.92	0.91
CNN (Mel)	87.86%	0.88	0.88	0.88
RNN (STFT)	71.35%	0.72	0.71	0.69
RNN (Mel)	71.95%	0.75	0.72	0.70
CRNN (STFT)	87.04%	0.87	0.87	0.87
CRNN (Mel)	85.64%	0.86	0.86	0.85
VGG19 (STFT)	83.86%	0.89	0.80	0.84
VGG19 (Mel)	81.35%	0.88	0.75	0.81

6 Experiments and Results

Training Setup

All models were built and trained using the TensorFlow and Keras libraries on a local machine configured with an Intel i7 processor, 32 GB RAM, and an NVIDIA RTX 3060 GPU. Each model underwent training for a maximum of 200 epochs, with early stopping implemented to halt training if the validation loss showed no improvement. The Adam optimizer was employed with a learning rate set to 0.001. Since this was a multi-class classification task, categorical cross-entropy was selected as the appropriate loss metric. A batch size of 32 was maintained across all training sessions. To ensure the best performance, the model weights corresponding to the lowest validation loss were saved using checkpointing.

Evaluation Metrics

To evaluate the model performance rigorously, we used the following metrics:

- **Accuracy**: The proportion of correctly predicted sound classes over the total number of test samples.
- **Precision, Recall, and F1-score**: These class-wise metrics help evaluate false positives, false negatives, and overall per-class performance.
- **Macro and Weighted F1-scores**: Macro scores reflect per-class averages, while weighted scores account for class distribution imbalances.
- **Confusion Matrix**: To visualize and interpret correct and incorrect predictions across all 13 classes.

Accuracy Comparison Across Models

We evaluated eight deep learning configurations—CNN, RNN, CRNN, and VGG19—each trained on both STFT and Mel spectrogram inputs. The objective was to assess their ability to classify 13 types of dangerous sounds under diverse acoustic conditions. Table 3 summarizes the test performance using four key metrics, and Figure 4 visually illustrates the comparative accuracy.

Model Insights and Performance Trends

CNN Models: The CNN trained on STFT spectrograms outperformed all other models, achieving an accuracy of 91.56%. This success is attributed to CNN's proficiency in capturing localized

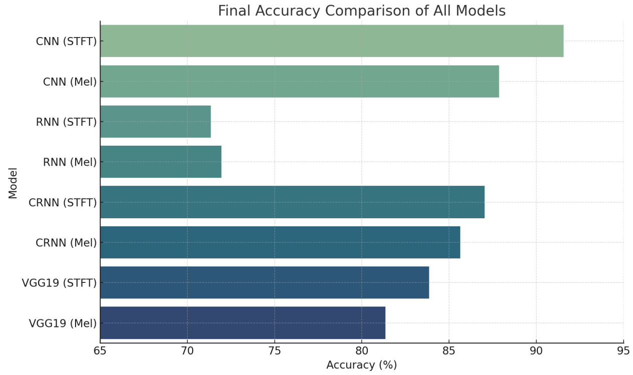

Figure 4: Final test accuracy comparison across all models using STFT and Mel spectrograms.

frequency-time features and STFT's high time-frequency resolution, which helps distinguish sharp transient events such as explosions or glass breaking.

RNN Models: RNN-based models showed comparatively lower performance (around 71–72%), especially on STFT inputs. This is likely due to their lack of spatial modeling capacity, which is critical for extracting localized spectro-temporal patterns from spectrogram images. However, RNNs trained on Mel spectrograms performed slightly better, likely due to the perceptual compression that emphasizes sequential information.

CRNN Models: CRNNs effectively combine CNN's spatial strength with RNN's temporal modeling, resulting in balanced performance (87.04% on STFT and 85.64% on Mel). These models proved robust across most classes, especially in noisy and overlapping sound categories.

VGG19 Models: Transfer learning with VGG19 offered strong generalization even with moderate training data. The model achieved 83.86% accuracy on STFT and 81.35% on Mel spectrograms. While performance was slightly lower than CRNNs, the use of deep pretrained features allowed VGG19 to perform consistently, especially on underrepresented classes like wind and dog barking.

Setbacks and Limitations

- **Class Imbalance:** Certain classes like Dam Failure and Tsunami had fewer samples, leading to lower recall and precision in those categories.
- **RNN Limitations:** Standalone RNNs struggled with spatially rich spectrogram data, limiting their overall utility in this application.
- **Spectrogram Sensitivity:** Mel spectrograms excelled at capturing tonal qualities but performed poorly for sharp transient sounds, whereas STFT was more reliable across the board.
- **VGG19 Overhead:** While powerful, the VGG19 models required more computational resources and fine-tuning, which may be a constraint in real-time or edge scenarios.

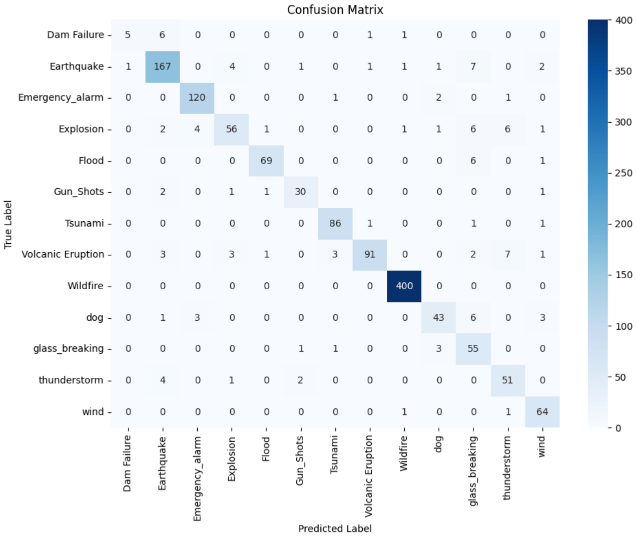

Figure 5: Confusion matrix for CNN (STFT) model. Rows represent actual classes; columns represent predicted classes.

Confusion Matrix Analysis

To delve deeper into class-wise performance, we analyzed the confusion matrix for the CNN (STFT) model, which achieved the best overall accuracy. As shown in Figure 5, the model excelled in high-frequency and high-volume classes such as Wildfire, Emergency Alarms, and Gunshots, but showed confusion between acoustically similar classes like thunderstorm and explosion, or wind and wildfire.

Performance Highlights

- **CNN (STFT)**: Most accurate model at 91.56%, excelling at spatial pattern recognition of transient and high-energy sound events.
- **CRNN Models**: Strong generalization due to their hybrid design, effective for overlapping and sequential sounds.
- **VGG19 Models**: Demonstrated resilience and generalization through pretrained deep feature extraction.
- **Mel Spectrograms**: Enhanced detection for speech-like and harmonic events (e.g., alarms, dog barks).
- **STFT Spectrograms**: Captured fast-changing, high-resolution patterns, improving recognition for explosion, thunder, and glass breaking.

Overall, the experiments confirm that combining high-resolution spectrogram representations with hybrid or transfer learning models provides the best performance for real-world dangerous sound classification tasks.

7 Conclusion and Future Work

In this study, we developed and evaluated a comprehensive suite of deep learning models to classify dangerous sounds in real-world environments using spectrogram-based features. Our approach leveraged four primary architectures—CNN, RNN, CRNN, and VGG19—trained on both STFT and Mel spectrograms. The proposed models were tested on a custom-built dataset consisting of 13 security-critical sound categories, offering a balanced and realistic test bed for evaluating model robustness.

Among all models, the CNN trained on STFT spectrograms achieved the highest accuracy of 91.56%, closely followed by CRNN and VGG19 models. Mel spectrograms proved particularly effective for tonal and speech-like sounds, while STFT representations helped retain detailed frequency resolution. The CRNN models demonstrated strong generalization due to their ability to capture both spatial and temporal patterns. VGG19-based models benefited from transfer learning, improving feature abstraction on smaller and noisier classes.

Our results validate the effectiveness of combining time-frequency audio representations with hybrid and transfer learning architectures for dangerous sound detection in public safety applications.

Future Work

While the models achieved high accuracy in controlled settings, several areas remain for improvement:

- **Real-time Deployment:** Integrate the trained models into a lightweight, edge-compatible system (e.g., on Raspberry Pi or Jetson Nano) to support real-time detection in surveillance systems.
- **Larger and Noisier Datasets:** Expand the dataset with more diverse and overlapping sounds, including outdoor and urban recordings with varying background conditions.
- **Multimodal Integration:** Combine acoustic features with visual or sensor-based data to improve decision-making in ambiguous or noisy environments.
- **Unsupervised Pretraining:** Explore contrastive or self-supervised learning methods to boost performance in low-label scenarios.
- **Adaptive Learning:** Implement continual learning strategies that allow the model to adapt to new sound classes without retraining from scratch.

We believe this work lays a strong foundation for intelligent audio surveillance systems capable of detecting critical events autonomously and in real time.

References

[1] Aigerim Altayeva. 2024. Development of a system for notification of dangerous sounds using machine learning methods. In *Publisher. agency: Proceedings of the 8th International Scientific Conference «Scientific Results»(November 7-8, 2024). Rome, Italy, 2024. 460p.* University of Bari Aldo Moro, 71.

[2] Md Ashikuzzaman, Awal Ahmed Fime, Abdul Aziz, and Tanvira Tasnima. 2021. Danger detection for women and child using audio classification and deep learning. In *2021 5th International Conference on Electrical Information and Communication Technology (EICT)*. IEEE, 1–6.

[3] Abdullah Atalar. 1979. A physical model for acoustic signatures. *Journal of Applied Physics* 50, 12 (1979), 8237–8239.

[4] Andrea Bernardini, Federica Mangiatordi, Emiliano Pallotti, and Licia Capodiferro. 2017. Drone detection by acoustic signature identification. *electronic imaging* 29 (2017), 60–64.

[5] Sung-Woo Byun and Soek-Pil Lee. 2016. A design of dangerous sound detection engine of wearable device for hearing impaired persons. *The Transactions of the Korean Institute of Electrical Engineers* 65, 7 (2016), 1263–1269.

[6] Pasquale Foggia, Nicolai Petkov, Alessia Saggese, Nicola Strisciuglio, and Mario Vento. 2015. Audio surveillance of roads: A system for detecting anomalous sounds. *IEEE transactions on intelligent transportation systems* 17, 1 (2015), 279–288.

[7] Ameer Hamza, Abdul Rehman Rehman Javed, Farkhund Iqbal, Natalia Kryvinska, Ahmad S Almadhor, Zunera Jalil, and Rouba Borghol. 2022. Deepfake audio detection via MFCC features using machine learning. *IEEE Access* 10 (2022), 134018–134028.

[8] Zeinel Momynkulov, Zhandos Dosbayev, Azizah Suliman, Bayan Abduraimova, Nurzhigit Smailov, Maigul Zhekambayeva, and Dusmat Zhamangarin. 2023. Fast Detection and Classification of Dangerous Urban Sounds Using Deep Learning. *CMC-COMPUTERS MATERIALS & CONTINUA* 75, 1 (2023), 2191–2208.

[9] Zeinel Momynkulov, Nurzhan Omarov, and Yerkebulan Uxikbayev. 2024. Detection of Dangerous Situations by Sounds in Real-Time Using Deep Learning. In *2024 IEEE 4th International Conference on Smart Information Systems and Technologies (SIST)*. IEEE, 278–283.

[10] Muhammad Awais Saba Nazir, Sheraz Malik, and Fatima Nazir. 2018. A review on scream classification for situation understanding. *Int. J. Adv. Comput. Sci. Appl.* 9 (2018), 63–75.

[11] Yoojeong Seo, Beomhui Jang, and Sungbin Im. 2018. Drone detection using convolutional neural networks with acoustic STFT features. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 1–6.

[12] Nurzhigit Smailov, Zhandos Dosbayev, Nurzhan Omarov, Bibigul Sadykova, Maigul Zhekambayeva, Dusmat Zhamangarin, and Assem Ayapbergenova. 2023. A novel deep CNN-RNN approach for real-time impulsive sound detection to detect dangerous events. *International Journal of Advanced Computer Science and Applications* 14, 4 (2023).

[13] Hyewon Suh, Seungwook Seo, and Young H Kim. 2018. Deep Learning-Based Hazardous Sound Classification for the Hard of Hearing and Deaf. In *2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*. IEEE, 073–077.

[14] Yuji Tokozume, Yoshitaka Ushiku, and Tatsuya Harada. 2017. Learning from between-class examples for deep sound recognition. *arXiv preprint arXiv:1711.10282* (2017).

[15] Haniza Yazid, Shafriza Nisha Basah, Saufiah Abdul Rahim, Muhammad Juhairi Aziz Safar, and Khairul Salleh Basaruddin. 2022. Performance analysis of entropy thresholding for successful image segmentation. *Multimedia Tools and Applications* 81, 5 (2022), 6433–6450.