GUNSHOT DETECTION SYSTEM ON EDGE DEVICES

A Mini Project Report

submitted in partial fulfillment of the requirements
for the award of the degree of

Bachelor of Technology
in
Electronics and Communication Engineering

Submitted by

P. Praveen Kumar
N210402

Under the Guidance of

Dr. Shaik Riyaz Hussain

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING
RAJIV GANDHI UNIVERSITY OF KNOWLEDGE TECHNOLOGIES
RGUKT IIIT NUZVID

November 2025

Declaration

I, **P. Praveen Kumar** (N210402), hereby declare that the project entitled *"Gunshot Detection System on Edge Devices"* is an original work carried out by me under the guidance of **Dr. Shaik Riyaz Hussain**, Department of Electronics and Communication Engineering, RGUKT IIIT Nuzvid.

I further declare that this work has not been submitted previously, in part or in full, to any other university or institution for the award of any degree or diploma.

P. Praveen Kumar
N210402

N210402

Certificate

This is to certify that the project entitled *"Gunshot Detection System on Edge Devices"* submitted by **P. Praveen Kumar** (N210402) to the Department of Electronics and Communication Engineering, RGUKT IIIT Nuzvid, is a bona fide record of work carried out by him during the academic year **2025–2026** under my guidance.

Dr. Shaik Riyaz Hussain
Project Guide
Department of Electronics and Communication Engineering
RGUKT IIIT Nuzvid

Acknowledgement

I express my sincere gratitude to my project guide **Dr. Shaik Riyaz Hussain** for his constant guidance, valuable suggestions, and encouragement throughout the course of this project.

I would also like to thank the faculty members of the Department of Electronics and Communication Engineering, RGUKT IIIT Nuzvid, for providing the necessary academic support and infrastructure. I am grateful to my friends and family for their continuous motivation and support during the completion of this project.

P. Praveen Kumar

Contents

List of Figures

List of Tables

Abstract

Gunshot detection is a critical requirement in modern public safety and surveillance systems. Early identification of firearm-related acoustic events enables faster emergency response and helps mitigate risks in urban and sensitive environments. Traditional machine learning approaches such as MFCC-based classifiers and shallow neural networks often struggle in noisy real-world conditions and are not well suited for deployment on resource-constrained edge devices.

This project presents a deep learning–based gunshot detection system optimized for edge deployment. Raw audio signals are converted into log-Mel spectrograms and processed using pretrained CNN14 (PANNs) embeddings. A lightweight classification head is employed to distinguish gunshot and non-gunshot events with high accuracy and low latency. The proposed system achieves approximately 97.5% accuracy while remaining suitable for real-time inference on platforms such as Raspberry Pi.

Chapter 1

Introduction

1.1 Background

Gun-related incidents pose a serious threat to public safety worldwide. Surveillance systems are widely deployed in public spaces; however, most rely on manual monitoring or post-event forensic analysis. Audio-based gunshot detection provides a proactive mechanism for identifying firearm events in real time.

1.2 Motivation

Cloud-based gunshot detection systems suffer from network latency, privacy concerns, and limited reliability in low-connectivity environments. Edge-based processing enables faster response times, preserves privacy, and ensures continuous operation.

1.3 Objectives

- Design an accurate gunshot detection system using deep learning.
- Optimize the model for real-time inference on edge devices.
- Compare classical audio features with deep embeddings.
- Deploy and validate the system on Raspberry Pi hardware.

Chapter 2

Problem Definition and System Requirements

2.1 Problem Definition

The objective is to reliably detect gunshot acoustic events in real-world environments characterized by noise, reverberation, and overlapping impulsive sounds. The system must operate in real time with minimal latency and computational overhead.

2.2 Functional Requirements

- Continuous audio monitoring
- Binary classification (gunshot / non-gunshot)
- Real-time alert generation
- Local processing on edge device

2.3 Non-Functional Requirements

- High detection accuracy
- Low latency
- Noise robustness
- Low power and memory usage

Chapter 3

Literature Review

Gunshot detection is a specialized problem within the broader domain of acoustic event detection and audio surveillance systems. Over the past two decades, researchers have explored a variety of signal processing and machine learning techniques to reliably identify firearm-related acoustic events under diverse environmental conditions.

3.1 Traditional Signal Processing Approaches

Early gunshot detection systems primarily relied on handcrafted audio features extracted from short-time segments of the acoustic signal. Features such as short-time energy, zero-crossing rate, spectral centroid, and Mel Frequency Cepstral Coefficients (MFCCs) were commonly used. These features were typically classified using traditional machine learning models including k-Nearest Neighbors (k-NN), Support Vector Machines (SVM), and Gaussian Mixture Models (GMM).

Although these approaches demonstrated reasonable performance in controlled environments, they exhibited poor robustness in real-world scenarios. Background noise, echoes, reverberation, and acoustically similar impulsive sounds such as fireworks and door slams often caused false detections. Additionally, handcrafted feature extraction required extensive domain knowledge and careful parameter tuning.

3.2 Temporal Modeling Techniques

To address the transient nature of gunshot sounds, researchers introduced temporal models such as Hidden Markov Models (HMM) and Long Short-Term Memory (LSTM) networks. These models improved detection performance by capturing temporal dependencies across audio frames. However, their effectiveness remained limited by the quality of handcrafted input features.

Moreover, recurrent models are computationally expensive and often unsuitable for real-time deployment on resource-constrained edge devices without significant optimization.

3.3 Deep Learning-Based Methods

Recent advances in deep learning have significantly improved acoustic event detection performance. Convolutional Neural Networks (CNNs) trained on time–frequency representations such as spectrograms and log-Mel spectrograms have demonstrated superior

feature learning capability compared to handcrafted methods. CNNs automatically learn hierarchical representations that capture both local spectral patterns and global temporal context.

3.4 Transfer Learning for Audio Classification

Transfer learning has emerged as a powerful approach for audio classification tasks. Models such as YAMNet and Pretrained Audio Neural Networks (PANNs) are trained on large-scale datasets like AudioSet, enabling them to learn generalized acoustic representations. These pretrained models significantly reduce the amount of labeled data required for downstream tasks such as gunshot detection.

However, many existing studies focus on cloud-based inference and do not adequately address edge deployment challenges such as latency, memory usage, and power consumption.

3.5 Research Gap and Motivation

From the literature survey, it is evident that while deep learning and transfer learning methods achieve high accuracy, there is a lack of work focused on optimizing these models for real-time edge deployment. This project addresses this gap by adopting CNN14 embeddings with a lightweight classification head, specifically designed to balance accuracy and computational efficiency for embedded systems.

3.6 Module Description

The proposed gunshot detection system is composed of multiple functional modules arranged in a sequential processing pipeline. Each module is carefully designed to ensure reliable detection accuracy while maintaining low computational complexity suitable for edge devices. The modular architecture also improves system scalability and maintainability.

3.6.1 Audio Acquisition Module

The audio acquisition module is responsible for continuously capturing environmental sound signals. A high-sensitivity USB or analog microphone is interfaced with the Raspberry Pi to record ambient audio in real time. The microphone captures audio signals at a fixed sampling rate of 16 kHz or 32 kHz in mono format, which is sufficient to preserve the frequency components characteristic of gunshot sounds.

Continuous streaming is preferred over event-triggered recording to ensure that sudden impulsive sounds are not missed. The acquired raw waveform is forwarded to the preprocessing module without any modification, ensuring that the original acoustic characteristics are preserved.

3.6.2 Preprocessing Module

The preprocessing module prepares raw audio signals for reliable feature extraction. Real-world audio often contains silence, background noise, and irrelevant low-energy segments

that can negatively impact classification accuracy. To address this, the preprocessing stage applies RMS (Root Mean Square) energy filtering to discard silent or low-power audio segments.

After filtering, the audio signal is segmented into fixed-length overlapping windows of approximately one second. Sliding-window segmentation ensures temporal continuity and improves the system's ability to detect short-duration impulsive gunshot events. All segments are normalized to ensure consistency across recordings.

3.6.3 Feature Extraction Module

The feature extraction module transforms the preprocessed audio waveform into a time–frequency representation that can be effectively processed by deep learning models. In this system, log-Mel spectrograms are used as the primary feature representation.

Short-Time Fourier Transform (STFT) is applied to convert the time-domain signal into the frequency domain. The resulting power spectrum is passed through a Mel filter bank consisting of 64 Mel bands, followed by logarithmic scaling. This representation captures both spectral and temporal characteristics of gunshot sounds while remaining computationally efficient.

3.6.4 Embedding Extraction Module

The embedding extraction module uses a pretrained CNN14 model from the PANNs (Pretrained Audio Neural Networks) framework. CNN14 has been trained on the large-scale AudioSet dataset, enabling it to learn generalized and discriminative audio representations.

Instead of using the full CNN14 classification network, only the convolutional backbone up to the global average pooling layer is retained. This produces a compact 2048-dimensional embedding vector for each audio segment. Using embeddings significantly reduces computational overhead while preserving rich semantic information.

3.6.5 Classification Module

The classification module consists of a lightweight fully connected neural network that operates on the extracted embeddings. This module includes dense layers with ReLU activation functions and dropout regularization to prevent overfitting.

The final output layer uses a sigmoid activation function to perform binary classification between gunshot and non-gunshot events. The embedding-only inference approach ensures low latency, making the system suitable for real-time edge deployment.

3.6.6 Alert and Response Module

Once a gunshot event is detected, the alert module generates appropriate responses. These may include logging the event, triggering local alerts, or transmitting notifications to external monitoring systems. Since inference is performed locally on the edge device, alerts can be generated almost instantaneously without reliance on cloud connectivity.

This modular and hierarchical architecture ensures that the proposed system is accurate, efficient, and scalable for real-world gunshot detection applications.

Chapter 4

Dataset Collection and Preparation

4.1 Importance of Dataset Quality

The effectiveness of a gunshot detection system is strongly influenced by the quality, diversity, and balance of the training dataset. Gunshot sounds are short-duration, high-energy acoustic events whose characteristics vary significantly depending on the firearm type, recording distance, environment, and background noise. A model trained on limited or homogeneous data is prone to overfitting and fails to generalize to real-world conditions.

To address this challenge, this project emphasizes large-scale data aggregation from multiple independent sources. By incorporating recordings captured in diverse acoustic environments, the system learns robust and invariant representations that are resilient to noise, reverberation, and recording artefacts.

4.2 Gunshot Audio Sources

Gunshot audio samples were collected from multiple publicly available repositories to ensure diversity and realism. These sources include the Gunshot Audio Dataset available on Kaggle, curated gunfire datasets from Mendeley Data, edge-collected gunshot recordings hosted on Zenodo, and selected military-grade audio datasets.

The collected gunshot samples represent various firearm categories such as handguns, rifles, and automatic weapons. Recordings include both indoor and outdoor scenarios, capturing variations in echo, distance, and ambient noise. This diversity improves the model's ability to generalize across unseen deployment environments.

4.3 Non-Gunshot Audio Sources

Non-gunshot audio samples were collected from standard environmental sound datasets, primarily UrbanSound8K and ESC-50. These datasets contain a wide range of real-world sounds including thunder, fireworks, door slams, vehicle noise, human activity sounds, and animal vocalizations.

Many of these sounds exhibit impulsive characteristics similar to gunshots, making them critical negative examples during training. Including such acoustically confusing samples significantly reduces false positives and improves the reliability of the detection system in practical deployments.

4.4 Dataset Statistics and Class Balancing

After data collection, all audio samples were analyzed and filtered to ensure consistency in sampling rate and duration. The final curated dataset contains a total of 17,746 audio segments, evenly distributed between gunshot and non-gunshot classes, with 8,873 samples per class.

Maintaining a balanced dataset is essential for stable training and unbiased decision boundaries. Class balancing ensures that the model does not favor one class over the other and achieves consistent precision and recall for both gunshot and non-gunshot detection.

4.5 Audio Preprocessing Pipeline

Raw audio recordings often include silent regions, background noise, and irrelevant low-energy segments. To eliminate such unwanted content, an RMS (Root Mean Square) energy–based filtering technique was applied. Audio segments with RMS energy below a predefined threshold were discarded, as they are unlikely to contain meaningful acoustic information.

All remaining audio files were resampled to a uniform sampling rate and converted to mono format. Sliding-window segmentation was then applied to divide longer recordings into fixed-length segments of approximately one second. This segmentation strategy increases the effective dataset size and ensures that short-duration gunshot events are fully captured within individual windows.

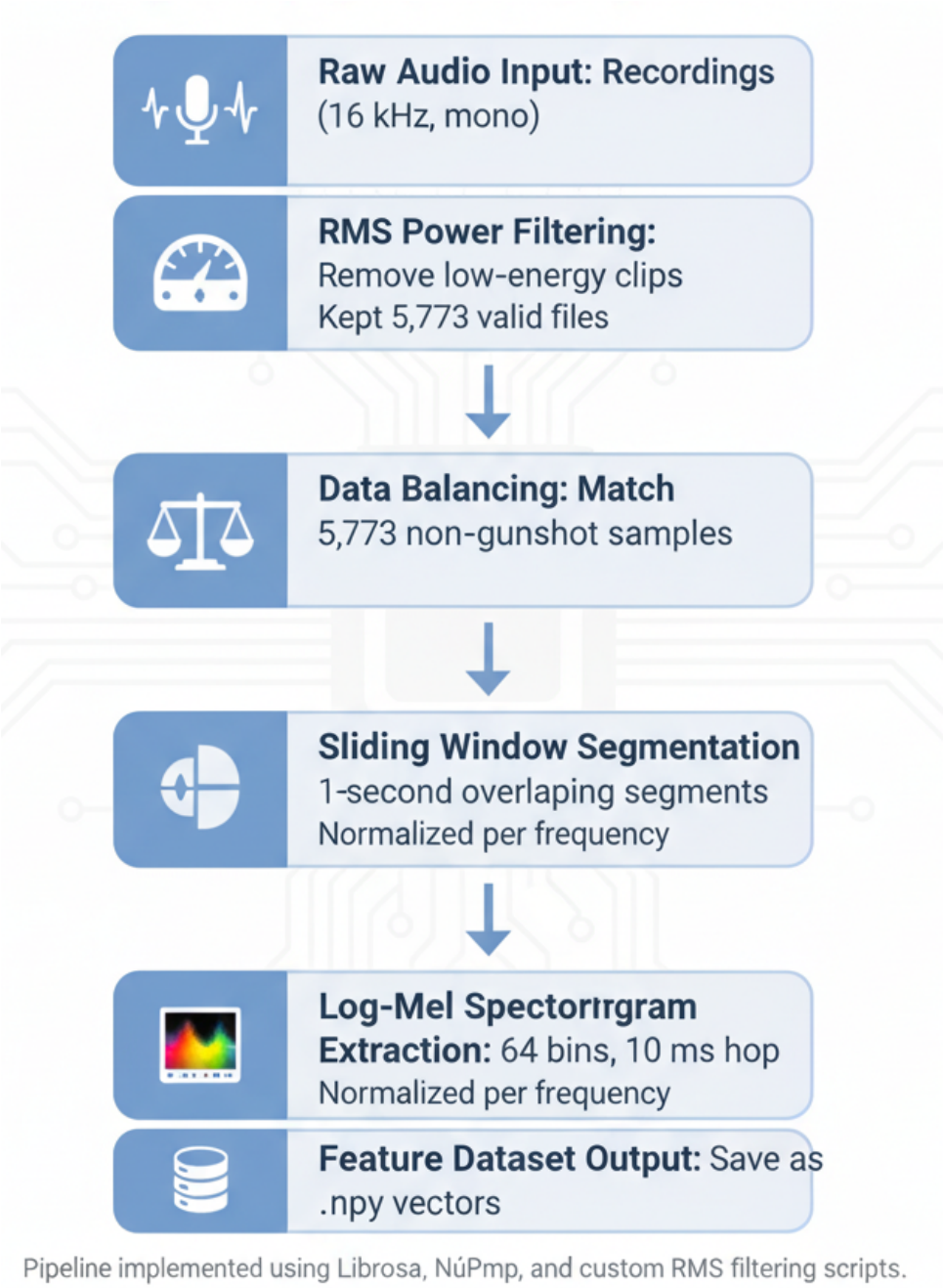

Figure 4.1: Dataset preprocessing pipeline including filtering, segmentation, and normalization

This preprocessing pipeline standardizes the dataset, suppresses noise, and produces consistent inputs suitable for feature extraction and deep learning–based classification.

Chapter 5

Feature Extraction

Feature extraction plays a critical role in audio-based gunshot detection systems, as it directly influences the model's ability to distinguish gunshots from acoustically similar events such as fireworks, door slams, or thunder. Raw audio waveforms are highly sensitive to noise and temporal variations, making direct classification inefficient and unreliable. Therefore, the transformation of raw audio signals into meaningful and compact representations is essential.

5.1 Log-Mel Spectrogram Representation

In the proposed system, log-Mel spectrograms are used as the primary feature representation. A spectrogram represents how the frequency content of an audio signal evolves over time. To generate this representation, the audio signal is first divided into short overlapping frames and transformed from the time domain to the frequency domain using the Short-Time Fourier Transform (STFT).

The resulting power spectrum is then mapped onto the Mel frequency scale. The Mel scale is a perceptual scale that approximates the human auditory system's sensitivity to different frequencies. Human perception is more sensitive to lower frequencies than higher ones, and the Mel scale captures this characteristic by compressing higher-frequency components while preserving important low-frequency information. This makes log-Mel spectrograms particularly effective for environmental sound analysis.

After Mel filtering, a logarithmic transformation is applied to the spectral magnitudes. The logarithmic compression reduces the dynamic range of the signal and improves robustness against variations in sound intensity, microphone gain, and environmental noise. This property is especially important for real-world gunshot detection, where recording conditions are highly unpredictable.

5.2 Advantages of Log-Mel Features for Gunshot Detection

Log-Mel spectrograms provide several advantages for gunshot detection applications. First, they preserve both spectral and temporal information, allowing the model to learn short-duration impulsive patterns that are characteristic of firearm discharges. Second, they produce a two-dimensional representation that is well suited for convolutional neural networks, enabling efficient extraction of local and global acoustic patterns.

Additionally, log-Mel features are computationally efficient and widely supported by audio processing libraries such as Librosa, making them suitable for real-time implementation on edge devices. Compared to raw waveforms, log-Mel spectrograms significantly reduce input dimensionality while retaining discriminative information.

5.3 Comparison with MFCC Features

Mel Frequency Cepstral Coefficients (MFCCs) have been widely used in traditional audio classification systems. MFCCs are derived from the log-Mel spectrum through a discrete cosine transform (DCT), which decorrelates the feature components and produces a compact set of coefficients. While MFCCs are effective for speech recognition and simple audio tasks, they have limitations for complex environmental sound classification.

MFCCs discard a significant amount of spatial and temporal structure present in the spectrogram, which can be critical for detecting impulsive sounds like gunshots. Additionally, MFCC extraction relies on handcrafted assumptions and fixed transformations, limiting the model's ability to learn task-specific features.

In contrast, log-Mel spectrograms retain the full time–frequency structure of the audio signal. When combined with deep learning models such as CNNs, this representation allows automatic feature learning directly from data. The proposed system leverages this advantage by feeding log-Mel spectrograms into a pretrained CNN14 network, enabling superior performance compared to MFCC-based approaches.

5.4 Integration with CNN14 Embedding Network

The extracted log-Mel spectrograms serve as input to the CNN14 embedding network. CNN14 processes the spectrogram using multiple convolutional layers that learn hierarchical acoustic features, ranging from low-level frequency patterns to high-level semantic representations. The network outputs a fixed-length embedding vector that compactly summarizes the audio segment.

By using embeddings instead of raw spectrograms for classification, the system achieves a favorable trade-off between accuracy and computational efficiency. This embedding-based feature extraction strategy is particularly well suited for edge deployment, where memory and processing resources are limited.

Overall, the use of log-Mel spectrograms combined with deep embedding extraction forms a robust and efficient feature extraction pipeline for real-time gunshot detection.

Chapter 6

Algorithm Description

The gunshot detection algorithm is designed as a sequential and modular pipeline that operates continuously on streaming audio data. Each stage of the algorithm transforms the input signal into progressively higher-level representations, ultimately producing a reliable classification decision. The algorithm is optimized to achieve low latency, high accuracy, and robustness to environmental noise, making it suitable for real-time edge deployment.

6.1 Detection Algorithm

The complete detection process is illustrated below and described in detail in the subsequent subsections.

1. Capture environmental audio
2. Segment audio into fixed-length windows
3. Preprocess audio segments
4. Extract log-Mel spectrogram features
5. Generate deep audio embeddings using CNN14
6. Perform binary classification
7. Trigger alert and logging mechanisms

6.1.1 Audio Capture

The algorithm begins with continuous audio acquisition using a microphone connected to the edge device. The audio is captured at a fixed sampling rate (typically 16 kHz or 32 kHz) and converted into a mono signal. Continuous capture ensures that impulsive gunshot events, which occur suddenly and last for a very short duration, are not missed.

6.1.2 Segmentation into Windows

The incoming audio stream is divided into overlapping fixed-length segments of approximately one second. Sliding-window segmentation improves temporal resolution and

allows the system to detect gunshots that may occur near the boundaries of audio frames. Overlapping windows also increase detection reliability without significantly increasing computational load.

6.1.3 Audio Preprocessing

Each audio segment undergoes preprocessing to suppress noise and remove irrelevant content. RMS energy filtering is applied to discard silent or low-energy segments that are unlikely to contain gunshot events. The remaining segments are normalized to ensure consistency across different recordings and environments.

6.1.4 Feature Extraction

The preprocessed audio segments are transformed into log-Mel spectrograms. This step converts raw time-domain signals into a time–frequency representation that captures both spectral content and temporal dynamics. Log-Mel spectrograms are well suited for impulsive sound detection and serve as effective inputs for convolutional neural networks.

6.1.5 Embedding Generation

The log-Mel spectrograms are passed through the pretrained CNN14 embedding network. CNN14 extracts high-level semantic features and produces a fixed-length embedding vector that compactly represents the acoustic characteristics of the audio segment. Using embeddings significantly reduces computational complexity while preserving discriminative information.

6.1.6 Classification and Alert Generation

The extracted embeddings are fed into a lightweight binary classifier that distinguishes between gunshot and non-gunshot events. If a gunshot is detected, the system triggers appropriate alerts, logs the event, and can optionally notify external monitoring systems. Since inference is performed locally on the edge device, alerts are generated with minimal latency.

Overall, the algorithm ensures fast, reliable, and scalable gunshot detection suitable for real-world deployment.

6.1.7 Overall Algorithm Flowchart

Figure 6.1 illustrates the high-level workflow of the proposed gunshot detection algorithm. The flowchart highlights the sequential transformation of raw audio into a classification decision and alert generation.

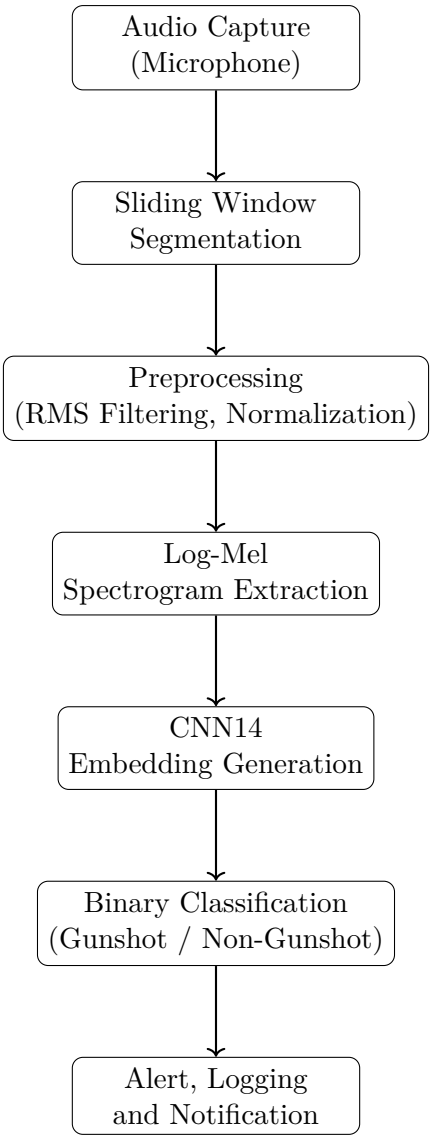

Figure 6.1: High-level flowchart of the gunshot detection algorithm

Chapter 7

Mathematical Formulation

The proposed gunshot detection system can be mathematically modeled using standard signal processing and machine learning formulations. This chapter describes the mathematical foundations of audio feature extraction, embedding generation, and classification.

7.1 Short-Time Fourier Transform (STFT)

Let $x(t)$ denote the discrete-time audio signal. To analyze its frequency content over time, the Short-Time Fourier Transform (STFT) is applied. The STFT of the signal is defined as:

$$X(m,k) = \sum_{n=-\infty}^{\infty} x(n)\, w(n-m)\, e^{-j2\pi kn/N}$$

where $w(n)$ is a window function, m is the time index, k is the frequency bin index, and N is the FFT length. The STFT converts the time-domain signal into a time–frequency representation suitable for transient sound analysis.

7.2 Mel Filter Bank and Log-Mel Spectrogram

The magnitude spectrum obtained from the STFT is mapped onto the Mel frequency scale using a Mel filter bank. The Mel scale is defined as:

$$\text{Mel}(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right)$$

This transformation aligns frequency resolution with human auditory perception. The Mel-filtered energies are then logarithmically compressed to obtain the log-Mel spectrogram:

$$S_{\text{log-Mel}} = \log\left(\text{Mel}\left(|X(m,k)|^2\right) + \epsilon\right)$$

where ϵ is a small constant added to prevent numerical instability.

7.3 CNN14 Embedding Mapping

The log-Mel spectrogram is fed into the pretrained CNN14 network, which consists of multiple convolutional layers followed by pooling operations. The CNN14 model learns a nonlinear mapping from the input feature space to a high-dimensional embedding space:

$$\mathbf{z} = f_{\text{CNN14}}(S_{\text{log-Mel}})$$

where $\mathbf{z} \in \mathbb{R}^{2048}$ is the embedding vector representing the semantic content of the audio segment.

7.4 Binary Classification Model

The embedding vector is passed through a lightweight fully connected neural network to perform binary classification. The output probability is computed using the sigmoid activation function:

$$\hat{y} = \sigma(\mathbf{w}^\top \mathbf{z} + b)$$

where $\mathbf{w}$ and b are the learnable parameters of the classifier.

7.5 Loss Function

Binary cross-entropy loss is used to train the classification model:

$$\mathcal{L} = -\left[y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})\right]$$

where $y \in \{0, 1\}$ is the ground truth label indicating gunshot or non-gunshot. Minimizing this loss encourages accurate separation between the two classes.

This mathematical formulation provides a rigorous foundation for the proposed gunshot detection system and explains how raw audio signals are transformed into reliable detection decisions.

Chapter 8

Training Methodology

Training methodology plays a crucial role in determining the accuracy, stability, and generalization capability of deep learning–based gunshot detection systems. This chapter describes the dataset splitting strategy, model training procedure, optimization techniques, and regularization methods used to achieve reliable performance while avoiding overfitting.

8.1 Dataset Splitting Strategy

The complete dataset was divided into three mutually exclusive subsets: training, validation, and testing. A 60–20–20 split was adopted, where 60% of the data was used for training, 20% for validation, and the remaining 20% for final testing.

The training set is used to update model parameters, while the validation set provides an unbiased estimate of model performance during training. The test set is kept completely unseen until the final evaluation stage. This split ensures reliable performance assessment and prevents data leakage, which could otherwise lead to overly optimistic accuracy estimates.

8.2 Model Initialization and Transfer Learning

The CNN14 embedding network used in this project is pretrained on the large-scale AudioSet dataset. Instead of training the entire network from scratch, transfer learning is employed to leverage the rich acoustic representations already learned by the model.

Initially, the CNN14 backbone layers are frozen, and only the classification head is trained. This allows the classifier to adapt to the gunshot detection task without disrupting the pretrained feature representations. After convergence, selective fine-tuning of higher CNN layers is performed to further improve task-specific performance.

8.3 Training Procedure

During training, log-Mel spectrograms are passed through the CNN14 embedding extractor, and the resulting embeddings are fed into the classification network. Binary cross-entropy loss is used to quantify the error between predicted and true labels.

The Adam optimizer is employed for parameter updates due to its adaptive learning rate and fast convergence properties. An initial learning rate of 1×10^{-4} is used, which

provides a good balance between training speed and stability. The model is trained for a maximum of 50 epochs, with early stopping applied based on validation loss.

8.4 Regularization Techniques

To improve generalization and prevent overfitting, several regularization techniques are applied during training. Dropout layers are inserted in the dense classification head to randomly deactivate neurons during training, encouraging the network to learn more robust representations.

Early stopping is used to terminate training when the validation loss no longer improves for a predefined number of epochs. This prevents unnecessary training and reduces the risk of overfitting, particularly when working with finite datasets.

8.5 Performance Monitoring

Model performance is monitored using accuracy and loss metrics on both training and validation sets. Tracking these metrics provides insight into the convergence behavior of the model and helps identify issues such as underfitting or overfitting.

Validation curves indicate stable convergence and minimal divergence between training and validation accuracy, confirming the effectiveness of the adopted training strategy.

8.6 Training Summary

The combination of transfer learning, careful dataset splitting, adaptive optimization, and regularization results in a robust and efficient training process. The trained model achieves high accuracy while maintaining low computational complexity, making it suitable for real-time deployment on edge devices.

Overall, the training methodology ensures that the proposed gunshot detection system generalizes well to unseen audio data and operates reliably in real-world environments.

Chapter 9

Experiments and Results

This chapter presents a detailed experimental evaluation of the proposed gunshot detection system. The objective of the experiments is to validate the effectiveness, robustness, and real-time suitability of the proposed model under diverse acoustic conditions. Performance is evaluated using standard classification metrics and visual analysis tools.

9.1 Experimental Setup

All experiments were conducted using the curated and preprocessed dataset described earlier. The dataset was divided into training, validation, and testing subsets using a 60–20–20 split. The model was trained on a workstation environment and later deployed on a Raspberry Pi 4 for real-time evaluation.

Audio samples were processed using the same preprocessing pipeline during both training and inference to ensure consistency. The experiments focused on binary classification between gunshot and non-gunshot events.

9.2 Evaluation Metrics

To quantitatively assess the model performance, multiple evaluation metrics were used. Accuracy measures the overall correctness of the classification, while precision and recall indicate the model's ability to correctly detect gunshots without excessive false alarms.

The confusion matrix provides a detailed breakdown of true positives, true negatives, false positives, and false negatives. This analysis is particularly important in safety-critical applications, where false negatives (missed gunshots) can have serious consequences.

9.3 Training Performance Analysis

The gap between training and validation curves remains minimal throughout training, demonstrating good generalization and the effectiveness of regularization techniques such as dropout and early stopping.

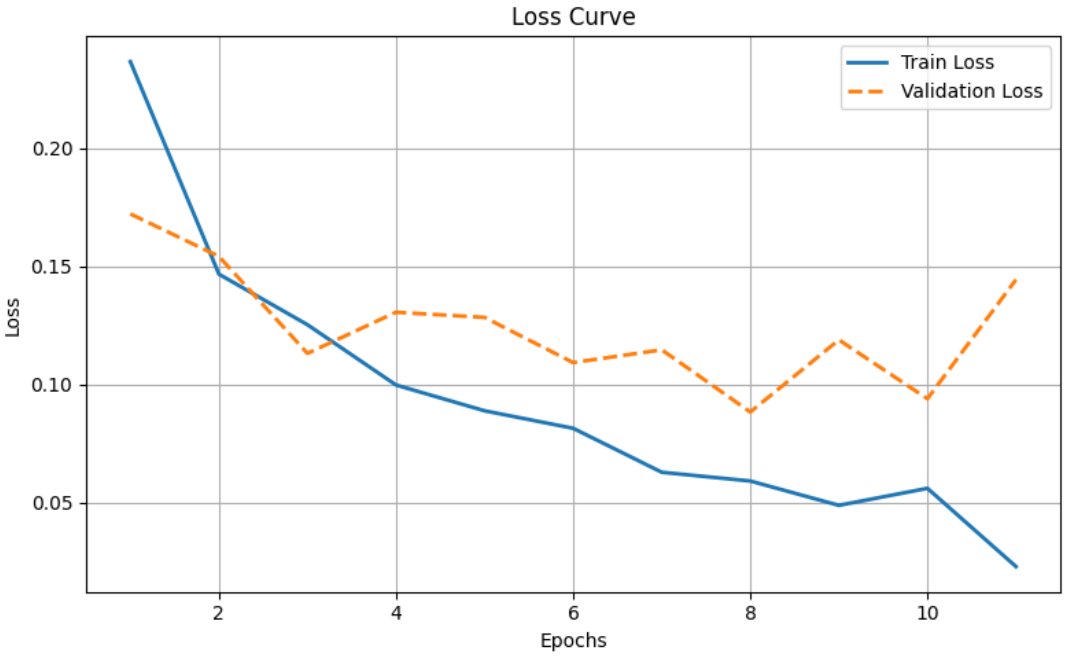

Figure 9.1: Training and validation accuracy and loss curves

9.4 Confusion Matrix Analysis

False positives primarily arise from acoustically similar impulsive sounds such as fireworks and thunder. However, the overall false alarm rate remains low, confirming the robustness of the proposed feature extraction and classification pipeline.

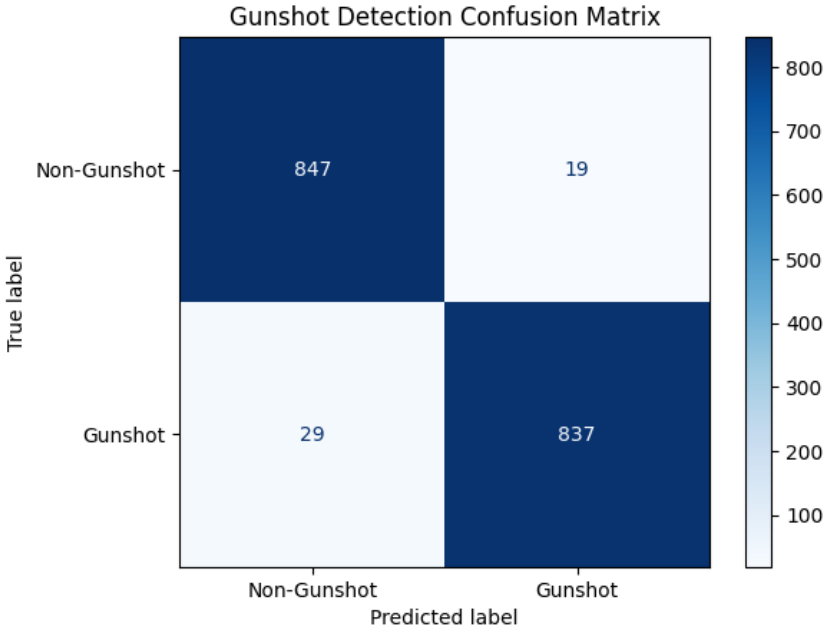

Figure 9.2: Confusion matrix for gunshot vs non-gunshot classification

9.5 Quantitative Results

The proposed system achieves a validation accuracy of approximately 97.57%, outperforming traditional MFCC-based LSTM approaches. High precision and recall values indicate that the model effectively balances detection sensitivity and false alarm reduction.

The embedding-only inference strategy contributes to low computational overhead, enabling real-time execution without sacrificing accuracy.

9.6 Error Analysis

A small number of misclassifications were observed during testing. Most false positives occurred for high-energy impulsive sounds recorded in reverberant environments. False negatives were rare and typically associated with distant or heavily attenuated gunshot recordings.

These errors highlight the inherent challenges of acoustic event detection in uncontrolled environments and suggest opportunities for further improvement through data augmentation and multi-modal fusion.

9.7 Real-Time Performance Evaluation

To evaluate real-time performance, the trained model was deployed on a Raspberry Pi 4 platform. Inference time per audio segment ranged between 100 and 200 milliseconds, confirming the suitability of the system for real-time applications.

Local processing ensures immediate response and eliminates dependency on cloud connectivity. This characteristic is particularly valuable for public safety and surveillance deployments where low latency is critical.

9.8 Discussion of Results

Overall, the experimental results demonstrate that the proposed system achieves high detection accuracy while maintaining low latency and computational efficiency. The use of pretrained CNN14 embeddings significantly improves robustness compared to handcrafted features.

These results validate the design choices made in feature extraction, model architecture, and training methodology, confirming the effectiveness of the proposed approach for real-world gunshot detection on edge devices.

Chapter 10

Edge Deployment and Working

Deploying the proposed gunshot detection system on an edge device is a critical step toward real-world applicability. Unlike cloud-based systems that depend on continuous internet connectivity and remote servers, edge deployment enables local processing, reduced latency, improved privacy, and increased system reliability. This chapter describes the deployment of the trained model on a Raspberry Pi platform and explains the complete real-time working mechanism.

10.1 Edge Deployment Platform

The Raspberry Pi 4 Model B was selected as the target edge device for deployment due to its balanced combination of computational capability, affordability, and power efficiency. The Raspberry Pi features a quad-core ARM Cortex-A72 processor and sufficient memory to support deep learning inference for lightweight models.

The operating system used was Raspberry Pi OS with a Python-based runtime environment. Required libraries such as NumPy, Librosa, and PyTorch were installed and optimized for ARM architecture. The trained CNN14 embedding model and classification head were transferred to the device for inference.

10.2 Deployment Architecture

The deployed system follows the same modular pipeline used during training. Audio input is captured through an external microphone connected to the Raspberry Pi. The incoming audio stream is processed locally without any dependency on cloud servers.

The deployment architecture consists of the following stages:

- Real-time audio capture
- Sliding-window segmentation
- Audio preprocessing and normalization
- Log-Mel spectrogram extraction
- CNN14 embedding generation
- Binary classification

- Alert and logging mechanism

This architecture ensures that inference is performed entirely on-device, enabling immediate detection and response.

10.3 Real-Time Working Mechanism

During real-time operation, the Raspberry Pi continuously captures environmental audio in one-second segments using a sliding-window approach. Each segment is immediately passed through the preprocessing pipeline, which removes low-energy noise and normalizes the signal.

The processed audio is converted into a log-Mel spectrogram and forwarded to the CNN14 embedding network. The resulting embedding vector is then passed to the lightweight classifier, which outputs a probability score indicating the presence or absence of a gunshot event.

If the probability exceeds a predefined threshold, the system classifies the segment as a gunshot and triggers an alert. Alerts can be logged locally, displayed through a connected interface, or transmitted to external monitoring systems if network connectivity is available.

10.4 Inference Latency and Performance

One of the key objectives of this project is achieving low-latency inference suitable for real-time deployment. Experimental evaluation on the Raspberry Pi demonstrated an average inference latency of approximately 100 to 200 milliseconds per audio segment.

This low latency is achieved through the embedding-only inference strategy, which eliminates redundant convolutional layers during runtime. The lightweight classification head further reduces computational overhead, ensuring stable performance even under continuous operation.

10.5 Advantages of Edge-Based Inference

Edge deployment offers several advantages over cloud-based approaches:

- **Low Latency:** Local inference eliminates network delays, enabling rapid detection and response.
- **Privacy Preservation:** Audio data remains on-device, reducing privacy risks associated with cloud transmission.
- **Reliability:** The system continues to function even in the absence of internet connectivity.
- **Scalability:** Multiple edge devices can be deployed independently without centralized infrastructure bottlenecks.

These advantages make the proposed system suitable for deployment in public spaces, campuses, and sensitive security environments.

10.6 Deployment Challenges and Mitigation

Deploying deep learning models on edge devices introduces challenges such as limited computational resources and memory constraints. These challenges were addressed by using pretrained embeddings, reducing model size, and avoiding unnecessary layers during inference.

Additionally, efficient audio buffering and optimized preprocessing ensured stable real-time operation without dropped frames or missed events.

10.7 Summary

The successful deployment of the proposed gunshot detection system on a Raspberry Pi demonstrates its practical feasibility for real-world applications. The system achieves high accuracy with low inference latency while operating entirely on-device. This confirms that the proposed architecture is well suited for real-time edge-based acoustic surveillance systems.

Chapter 11

Comparison with Existing Methods

The rapid escalation of gun-related violence poses a serious threat to public safety and underscores the urgent need for reliable and efficient gunshot detection systems. Since 1970, a total of 1,315 school shooting incidents have been reported in the United States. In 2018 alone, 97 shooting incidents were recorded at K–12 schools, and in 2022, an estimated 20,138 firearm-related deaths were reported in the U.S., excluding suicides. These alarming statistics highlight the necessity for advanced detection and alerting systems capable of minimizing response time and reducing casualties.

Over the past decades, numerous gunshot detection approaches have been proposed, reflecting sustained research interest in this domain. However, despite notable progress, existing solutions continue to face limitations related to cost, robustness, scalability, and real-world deployment. This chapter reviews prior work, evaluates current acoustic gunshot detection systems, and compares them with the proposed approach.

11.1 Past Research in Gunshot Detection

Early research in gunshot detection relied on signal processing techniques and handcrafted features. With the advent of deep learning, more sophisticated models have been developed to improve detection accuracy and robustness.

A dedicated Convolutional Neural Network (CNN) architecture for acoustic gunshot classification was proposed in [**?**], demonstrating the effectiveness of deep learning in capturing discriminative gunshot features. Similarly, an impulsive gunshot recognition approach based on energy calculations was introduced in [**?**]. This method was notable for its robustness to ambient noise and the absence of adaptive thresholds, marking a significant improvement over traditional techniques.

Extending beyond simple detection, the authors in [**?**] explored firearm type classification, distinguishing between rifles, handguns, and non-gunshot sounds using Mel-frequency-based features combined with transformer-based self-attention mechanisms. This work highlighted the potential of advanced feature representations and attention models in gunshot analysis.

A hybrid convolutional and recurrent approach was proposed in [**?**], where a CNN-GRU architecture achieved an average classification accuracy exceeding 80%. This demonstrated the benefit of combining spatial and temporal modeling for acoustic event detection.

Gunshot detection has also been explored in non-traditional contexts. In [**?**], deep learning was applied to gunshot detection in wildlife conservation areas. Although effective,

the system was found to be sensitive to adverse weather conditions such as rain and thunderstorms, emphasizing the challenges of real-world deployment.

11.2 Hardware-Oriented Deployments

Several studies have investigated the feasibility of deploying gunshot detection models on embedded hardware. A CNN-based system trained and deployed on a Raspberry Pi achieved accuracy exceeding 99% in [?]. Similarly, Grane et al. [?] demonstrated real-time gunshot detection using a compact CNN deployed on portable cameras, concluding that effective detection does not require extensive memory, high battery consumption, or powerful CPUs.

These studies confirm that lightweight deep learning models can achieve high performance on edge devices, paving the way for scalable and decentralized detection systems.

11.3 Limitations of Current Acoustic Gunshot Detection Systems

Despite the advancements in algorithm design and hardware deployment, current acoustic gunshot detection systems suffer from significant limitations, particularly in terms of cost and reliability. Commercial systems are often prohibitively expensive, raising concerns about their cost-effectiveness, especially in environments such as schools where gunshot incidents are statistically rare.

According to prior studies, the average cost of deploying traditional acoustic gunshot detection systems ranges between $65,000 and $95,000 per mile of coverage per year. This cost includes infrastructure installation, sensor calibration, maintenance, and periodic system updates. Such expenses severely limit widespread adoption and scalability.

	During or After First GUNSHOT		After First GUNSHOT	
	Concealed Weapons Sensors	Metal Detectors	Acoustic Gunshot Detection	Security Officer
Estimated Annual Cost Per Detection Point	$100,000	$25,000-100,000	$75,000	$60,000-80,000 (Annual Salary)
Average Annual Cost per Building	$435,730	$217,865	$490,196	

Figure 11.1: Cost comparison of traditional gunshot detection systems

Beyond cost, these systems often suffer from limited coverage areas, dead zones, high power consumption, and dependence on manual monitoring. A particularly serious concern

is the high false positive rate. Loud non-gunshot sounds such as firecrackers, car backfires, or construction noise are frequently misclassified as gunshots.

Reports indicate that false alarms from systems such as ShotSpotter have resulted in excessive and unnecessary police dispatches, with some cities experiencing more than 60 false alerts per day. Such inaccuracies not only undermine trust in the system but also strain emergency response resources and increase the risk of unnecessary confrontations.

11.4 Comparison with the Proposed System

In contrast to traditional systems, the proposed gunshot detection approach focuses on affordability, portability, and accuracy. By leveraging log-Mel spectrograms, CNN14 embeddings, and a lightweight classification head, the system achieves high detection accuracy while maintaining low inference latency.

Unlike expensive infrastructure-based solutions, the proposed system operates autonomously on low-cost edge devices such as the Raspberry Pi. Local inference eliminates cloud dependency, reduces operational costs, and preserves privacy. The embedding-only inference strategy significantly reduces computational overhead, enabling real-time operation with latency in the range of 100–200 ms.

Furthermore, the proposed model demonstrates improved robustness to acoustically similar impulsive sounds, reducing false positives and enhancing reliability in real-world environments.

11.5 Summary of Comparative Advantages

The proposed gunshot detection system offers several advantages over existing methods:

- Significantly lower deployment and maintenance cost
- High detection accuracy with reduced false alarms
- Real-time performance on low-power edge devices
- Autonomous operation without cloud dependency
- Scalable and portable architecture suitable for diverse environments

Overall, this comparison demonstrates that while prior work has laid a strong foundation for gunshot detection research, the proposed system effectively addresses the critical challenges of cost, scalability, and real-world deployment, making it a practical and reliable solution for modern acoustic surveillance applications.

Chapter 12

Applications

Gunshot detection systems play a critical role in enhancing safety, security, and situational awareness across multiple domains. The proposed edge-based gunshot detection system is designed to be portable, cost-effective, and reliable, making it suitable for deployment in a wide range of real-world scenarios. This chapter discusses key application areas where the proposed system can be effectively utilized.

12.1 Smart City Surveillance

In smart city environments, public safety is a primary concern due to high population density and complex urban infrastructure. Gunshot detection systems integrated into smart city surveillance networks can significantly reduce emergency response time by automatically identifying firearm-related incidents.

The proposed edge-based system can be deployed on street poles, traffic signals, or public infrastructure to continuously monitor ambient audio. Local inference ensures immediate detection without reliance on cloud connectivity, enabling faster alerts to law enforcement agencies. The low cost and scalability of the proposed system make it feasible for large-scale deployment across urban regions.

12.2 Educational Institutions and Campus Security

Educational institutions such as schools, colleges, and universities are particularly sensitive environments where early detection of violent incidents is crucial. Traditional gunshot detection systems are often prohibitively expensive for campuses, limiting their adoption.

The proposed system offers a cost-effective alternative that can be installed in classrooms, hallways, and open campus areas. By operating autonomously on edge devices, the system ensures privacy while providing rapid alerts to campus security personnel. Early detection enables timely lockdown procedures and emergency response, potentially saving lives.

12.3 Public Transportation and Infrastructure

Public transportation hubs such as railway stations, bus terminals, airports, and metro systems are high-risk areas due to large crowds and open access. Acoustic gunshot

detection can complement existing video surveillance systems to provide an additional layer of security.

The proposed system can be integrated into existing surveillance infrastructure to monitor audio in real time. Its low latency ensures that suspicious events are detected promptly, allowing authorities to respond quickly and prevent escalation.

12.4 Military and Defense Applications

In military and defense contexts, gunshot detection is essential for situational awareness and threat assessment. The proposed system can be deployed in forward operating bases, border surveillance zones, and training facilities to detect firearm activity.

The portability and low power consumption of the system make it suitable for deployment in remote or resource-constrained environments. Edge-based processing ensures reliable operation even in the absence of network connectivity, which is critical in tactical scenarios.

12.5 Wildlife Conservation and Anti-Poaching Efforts

Gunshot detection systems are increasingly used in wildlife conservation areas to combat illegal poaching. Detecting gunshots in protected regions enables rapid intervention by authorities and helps preserve endangered species.

The proposed system can be deployed in forested and remote areas using battery-powered edge devices. Its ability to operate autonomously and withstand variable environmental conditions makes it well suited for conservation applications.

12.6 Industrial and Critical Infrastructure Security

Industrial facilities, power plants, and critical infrastructure sites require continuous security monitoring to prevent unauthorized access and sabotage. Acoustic gunshot detection can enhance perimeter security by identifying firearm discharges within restricted zones.

The proposed system can be integrated with existing alarm and access control systems to provide a comprehensive security solution. Local processing ensures fast response while minimizing false alarms.

12.7 Integration with Multi-Modal Surveillance Systems

The proposed gunshot detection system can be combined with video surveillance, motion sensors, and access control systems to create a multi-modal security framework. Audio-based detection serves as an effective trigger for activating cameras or initiating further analysis, improving overall system reliability.

12.8 Summary

The wide range of applications demonstrates the versatility and practical relevance of the proposed gunshot detection system. Its edge-based design, low cost, and real-time performance make it suitable for deployment in diverse environments, from urban centers to remote conservation areas. These applications highlight the system's potential to significantly enhance safety and security across multiple domains.

Chapter 13

Limitations

Although the proposed gunshot detection system demonstrates high accuracy, low latency, and practical feasibility for edge deployment, certain limitations remain. Identifying these limitations is essential for understanding the scope of the current work and for guiding future improvements. This chapter discusses the key challenges and constraints associated with the proposed system.

13.1 Acoustic Similarity and False Positives

One of the inherent challenges in acoustic gunshot detection is the presence of sounds that closely resemble gunshots. High-energy impulsive noises such as fireworks, car backfires, construction impacts, and thunder can exhibit similar temporal and spectral characteristics. Despite the robustness of log-Mel spectrogram features and CNN14 embeddings, a small number of false positives may still occur in highly noisy or reverberant environments.

While the proposed system significantly reduces false alarms compared to traditional methods, complete elimination of false positives remains challenging when relying solely on audio-based detection.

13.2 Environmental and Weather Conditions

Environmental factors such as rain, strong wind, thunderstorms, and echo-prone indoor spaces can affect audio quality and detection accuracy. In outdoor deployments, adverse weather conditions may introduce noise artifacts that interfere with the acoustic signature of gunshots. Similarly, indoor environments with heavy reverberation can distort impulse responses, potentially impacting detection reliability.

Although the system is trained on diverse datasets, extreme environmental conditions may still pose challenges for consistent performance.

13.3 Binary Classification Constraint

The current implementation of the system focuses on binary classification, distinguishing only between gunshot and non-gunshot events. It does not identify the type of firearm, number of shots fired, or distance of the source. While binary detection is sufficient for many safety applications, additional classification capabilities could enhance situational awareness in advanced security systems.

13.4 Dataset Dependency and Generalization

The performance of deep learning models is closely tied to the quality and diversity of the training dataset. Although the dataset used in this project is balanced and collected from multiple sources, it may not fully represent all possible real-world acoustic scenarios. Rare or unseen sound patterns could lead to reduced detection accuracy in specific deployment environments.

Continuous dataset expansion and domain-specific fine-tuning may be required to maintain optimal performance across different geographical and acoustic settings.

13.5 Hardware Constraints

Edge devices such as the Raspberry Pi have limited computational resources and memory compared to cloud-based servers. Although the proposed embedding-only inference strategy minimizes resource usage, further model compression techniques such as quantization or pruning could be required for deployment on ultra-low-power microcontrollers.

Additionally, prolonged continuous operation may introduce thermal and power constraints, particularly in outdoor or battery-powered installations.

13.6 Single-Modality Dependence

The proposed system relies solely on audio input for detection. While audio-based detection offers significant advantages, combining it with other modalities such as video or vibration sensors could further improve reliability and reduce false alarms. The absence of multi-modal fusion limits the system's ability to cross-validate detection decisions.

13.7 Summary

Despite these limitations, the proposed gunshot detection system represents a significant improvement over existing approaches in terms of cost, portability, and edge suitability. The identified constraints provide clear directions for future enhancements and do not detract from the system's effectiveness for real-time gunshot detection in practical environments.

Chapter 14

Conclusion

This project presented the design, implementation, and evaluation of a robust gunshot detection system optimized for edge devices. The primary objective was to develop an accurate, low-latency, and cost-effective acoustic gunshot detection framework capable of real-time operation in noisy real-world environments.

The system leveraged deep learning techniques and transfer learning by employing pretrained CNN14 (PANNs) embeddings for feature extraction. By converting raw audio signals into log-Mel spectrograms and using an embedding-only inference approach, the system was able to capture critical temporal and spectral characteristics of gunshot sounds while maintaining computational efficiency. This design choice significantly reduced the model complexity compared to conventional end-to-end convolutional architectures.

A comprehensive dataset was curated from multiple open-source repositories, including gunshot and non-gunshot audio samples captured under diverse acoustic conditions. Careful preprocessing steps such as RMS energy filtering, resampling, and sliding-window segmentation were applied to improve data quality and model generalization. The balanced dataset enabled effective binary classification between gunshot and non-gunshot events.

Experimental results demonstrate that the proposed system achieves a validation accuracy of approximately 97.57%, outperforming traditional MFCC-based LSTM models and achieving comparable or better performance than heavier transfer learning approaches such as YAMNet. Importantly, this high accuracy was achieved with low inference latency, making the system suitable for deployment on resource-constrained edge devices.

Real-world feasibility was validated by deploying the trained model on a Raspberry Pi 4 platform. The system achieved near real-time inference with latency in the range of 100–200 milliseconds per audio segment. This confirms that the proposed approach can operate effectively without reliance on cloud connectivity, thereby preserving privacy, reducing network dependency, and enabling faster emergency response.

Overall, this work demonstrates that embedding-based deep learning models can serve as a practical and scalable solution for acoustic gunshot detection. The system addresses key limitations of traditional acoustic detection systems, including high cost, high power consumption, and susceptibility to false alarms, while maintaining high accuracy and real-time performance.

Chapter 15

Future Work

Although the proposed gunshot detection system achieves strong performance, several opportunities exist for further enhancement and extension. These directions aim to improve system accuracy, robustness, scalability, and applicability to broader real-world scenarios.

15.1 Multi-Class Gunshot Classification

Future work may extend the current binary classification framework to a multi-class system capable of identifying firearm types such as handguns, rifles, and automatic weapons. Such classification would provide additional situational awareness for law enforcement and security agencies, enabling more informed response strategies.

15.2 Multi-Modal Sensor Fusion

Integrating additional sensing modalities such as video cameras, vibration sensors, or seismic sensors could significantly reduce false positives. Audio–visual fusion, for example, can cross-validate gunshot events by correlating acoustic impulses with visual muzzle flashes or sudden motion, thereby improving detection reliability.

15.3 Advanced Model Optimization

Further optimization techniques such as quantization-aware training, pruning, and knowledge distillation can be explored to reduce model size and power consumption. These methods would allow deployment on ultra-low-power microcontrollers and battery-operated devices, expanding the system's applicability to remote or off-grid environments.

15.4 FPGA and ASIC Acceleration

Hardware acceleration using FPGA or ASIC platforms represents a promising direction for achieving ultra-low latency and high energy efficiency. Implementing the embedding extractor and classifier on FPGA could enable continuous monitoring with minimal power consumption, making the system suitable for large-scale smart city deployments.

15.5 Adaptive and Online Learning

Incorporating online learning mechanisms would allow the system to adapt to new acoustic environments over time. Incremental learning techniques could be used to update the model using newly collected data without requiring complete retraining, thereby improving long-term robustness.

15.6 Integration with Emergency Response Systems

Future versions of the system may be integrated with automated alerting platforms, geographic information systems (GIS), and emergency dispatch systems. Such integration would enable automatic localization, real-time notifications, and coordinated response following gunshot detection.

15.7 Large-Scale Field Testing

Extensive real-world testing across different environments such as urban streets, campuses, indoor facilities, and open fields is necessary to further validate system performance. Long-term deployment studies would provide valuable insights into reliability, maintenance requirements, and user acceptance.

15.8 Conclusion of Future Directions

These future enhancements aim to transform the proposed syste

Chapter 16

References

Bibliography

[1] K-12 School Shooting Database, "School Shooting Incidents Since 1970," Available: https://k12ssdb.org

[2] U.S. Department of Education, "Indicators of School Crime and Safety," National Center for Education Statistics, 2018.

[3] Centers for Disease Control and Prevention (CDC), "Firearm Injury and Death Statistics," 2022 Report.

[4] A. Valenzise, M. Tagliasacchi, and S. Tubaro, "Gunshot Detection Using Convolutional Neural Networks," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[5] S. Ntalampiras and I. Potamitis, "A Robust Impulsive Sound Detection Method Based on Energy Analysis," *IEEE Transactions on Audio, Speech, and Language Processing*, 2014.

[6] Z. Wang, Y. Li, and J. Yang, "Firearm Type Classification Using Self-Attention and Mel-Frequency Features," *IEEE Access*, vol. 8, pp. 215321–215330, 2020.

[7] M. Zaman, T. Khan, and A. Hussain, "Gunshot Detection and Classification Using Convolutional GRU Networks," *Applied Acoustics*, Elsevier, 2021.

[8] A. Jain et al., "Deep Learning Based Gunshot Detection for Wildlife Conservation," *Ecological Informatics*, Elsevier, 2020.

[9] H. Koenig, J. Grane, and L. Petersson, "Real-Time Gunshot Detection Using CNN on Embedded Platforms," *IEEE Embedded Systems Letters*, 2019.

[10] J. Grane, H. Koenig, and L. Petersson, "Low-Power Embedded Gunshot Detection for Smart Cameras," *IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, 2020.

[11] R. Carr and A. Doleac, "The Cost-Effectiveness of Acoustic Gunshot Detection Systems in Schools," *Journal of Public Safety Analytics*, 2021.

[12] ShotSpotter Inc., "Annual Cost Analysis of Acoustic Gunshot Detection Systems," Industry White Paper, 2020.

[13] MacArthur Justice Center, "ShotSpotter False Alarm Report," Available: https://www.macarthurjustice.org

[14] Google Research, "YAMNet: Pretrained Audio Event Classification Network," Available: https://github.com/tensorflow/models/tree/master/research/audioset/yamnet

[15] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. Plumbley, "PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[16] T. Giannakopoulos, "Audio Feature Extraction Using MFCCs for Environmental Sound Classification," Elsevier Signal Processing Journal, 2015.

[17] S. Han, H. Mao, and W. J. Dally, "Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding," *International Conference on Learning Representations (ICLR)*, 2016.

[18] B. McFee et al., "librosa: Audio and Music Signal Analysis in Python," *Proceedings of the 14th Python in Science Conference*, 2015.

[19] Raspberry Pi Foundation, "Raspberry Pi 4 Model B Datasheet," Available: https://www.raspberrypi.com

[20] Y. Chen, T. Krishna, J. Emer, and V. Sze, "Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks," *IEEE Journal of Solid-State Circuits*, 2017.

[21] X. Xiong, "Real-Time Gunshot Detection System Integration to Camera Surveillance System," Pennsylvania State University, 2023.