

Real-Time Edge-Based Gunshot Detection Using CNN14 Audio Embeddings and Attention-Guided Temporal Modeling

P. Praveen Kumar

Abstract—Acoustic gunshot detection is a critical component of modern public safety and surveillance systems, requiring high accuracy, low latency, and reliable operation under noisy real-world conditions. Conventional approaches based on handcrafted audio features exhibit limited robustness and are unsuitable for on-device deployment. This paper presents a lightweight deep learning framework for real-time gunshot detection on edge platforms using log-Mel spectrograms and pretrained CNN14 embeddings. Temporal dynamics are modeled using a bidirectional LSTM augmented with an attention mechanism to selectively emphasize impulsive gunshot frames while suppressing background noise. A balanced dataset of 17,746 audio segments was curated from multiple open-source repositories. The proposed system achieves a validation accuracy of 97.57% and real-time inference latency of 100–200 ms per one-second audio segment on a Raspberry Pi 4. The results demonstrate that embedding-based audio representation combined with attention-guided temporal modeling enables accurate, efficient, and privacy-preserving gunshot detection on resource-constrained edge devices.

Index Terms—Gunshot detection, edge computing, audio embeddings, CNN14, attention mechanism, real-time inference.

I. INTRODUCTION

Rapid detection of firearm-related acoustic events is essential for timely emergency response and situational awareness in public safety applications. Gunshot sounds are characterized by short-duration, high-energy impulses whose acoustic signatures vary significantly due to firearm type, propagation distance, reverberation, and environmental noise. Traditional detection systems relying on handcrafted features such as MFCCs and shallow classifiers demonstrate limited generalization and high false alarm rates in real-world deployments.

Recent advances in deep learning have enabled data-driven feature learning from time–frequency representations; however, most existing systems depend on cloud-based inference, introducing latency, privacy concerns, and network dependency. This work addresses these limitations by proposing an edge-optimized gunshot detection framework that combines pretrained deep audio embeddings with lightweight temporal modeling, enabling real-time on-device inference.

II. RELATED WORK

Early gunshot detection systems utilized signal energy, zero-crossing rates, and MFCC features combined with classical classifiers. Subsequent studies incorporated temporal models

such as HMMs and LSTMs to better capture transient acoustic patterns. More recently, convolutional neural networks trained on spectrogram representations have demonstrated improved performance.

Transfer learning using large-scale pretrained audio models such as YAMNet and PANNs has further enhanced detection accuracy. Nevertheless, many reported systems remain computationally intensive or cloud-dependent, limiting their suitability for edge deployment. In contrast, this work focuses on embedding-only inference using CNN14, significantly reducing computational overhead while maintaining high detection performance.

III. SYSTEM OVERVIEW

The proposed system follows a modular pipeline consisting of audio acquisition, preprocessing, feature extraction, embedding generation, temporal classification, and decision logic. An overview of the architecture is shown in Fig. 1.

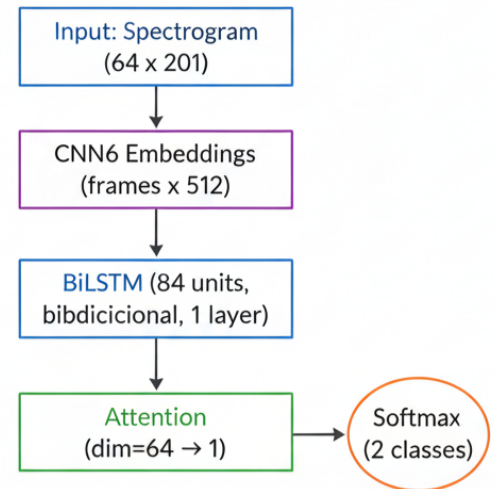


Fig. 1: Overall architecture of the proposed CNN14–BiLSTM–Attention gunshot detection system.

IV. DATASET AND PREPROCESSING

Gunshot samples were collected from multiple open-source datasets including Kaggle, Mendeley, and Zenodo. Non-

gunshot samples were sourced from UrbanSound8K and ESC-50 to incorporate acoustically similar impulsive events. The final dataset consists of 17,746 audio segments equally distributed between gunshot and non-gunshot classes.

All audio recordings were resampled to 32 kHz, converted to mono, and segmented into one-second windows. Low-energy segments were removed using RMS-based filtering to eliminate silence and irrelevant background noise. The preprocessing pipeline is illustrated in Fig. 2.

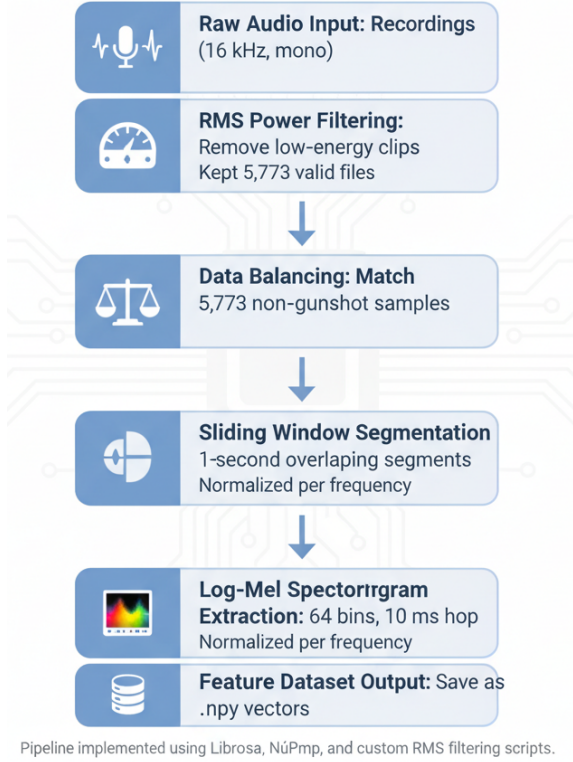


Fig. 2: Audio preprocessing pipeline including RMS filtering, segmentation, and normalization.

V. FEATURE EXTRACTION AND EMBEDDING GENERATION

Each audio segment is transformed into a log-Mel spectrogram using 64 Mel filters. The spectrogram is processed by the pretrained CNN14 encoder from the PANNs framework. Only the convolutional backbone up to the global pooling layer is retained, producing a fixed-length 2048-dimensional embedding vector. This embedding captures high-level spectral-temporal characteristics while enabling efficient inference.

VI. ATTENTION-BASED TEMPORAL MODELING

Temporal dependencies in the embedding sequence are modeled using a bidirectional LSTM. An additive attention mechanism is applied to assign higher weights to frames corresponding to impulsive gunshot events. The attention-weighted context vector is passed through a compact fully connected layer with sigmoid activation for binary classification.

VII. EXPERIMENTAL RESULTS

The dataset was split into training, validation, and testing sets using a 60/20/20 ratio. The proposed model achieved a peak validation accuracy of 97.57%. Fig. 3 presents the confusion matrix obtained on the test set, demonstrating low false positive and false negative rates.

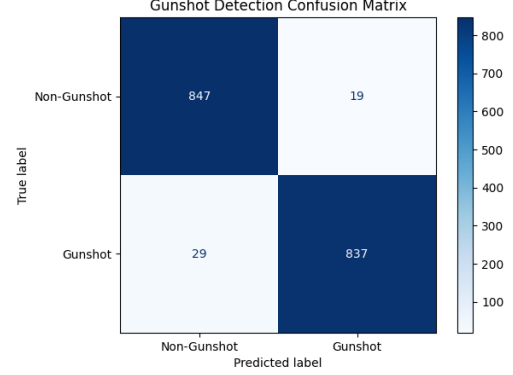


Fig. 3: Confusion matrix for gunshot vs. non-gunshot classification.

VIII. EDGE DEPLOYMENT AND LATENCY ANALYSIS

The trained model was deployed on a Raspberry Pi 4 (4 GB RAM) using CPU-only inference. End-to-end inference latency ranged between 100 and 200 ms per one-second audio segment. The real-time deployment workflow is illustrated in Fig. 4.



Fig. 4: Real-time edge deployment workflow on Raspberry Pi.

IX. CONCLUSION AND FUTURE WORK

This paper presented an efficient and noise-robust gunshot detection system optimized for edge deployment. By combining CNN14 audio embeddings with attention-guided temporal modeling, the proposed approach achieves high accuracy with low inference latency. Future work will focus on multi-class firearm classification, direction-of-arrival estimation, and FPGA-based acceleration for ultra-low-power deployments.

REFERENCES

- [1] Q. Kong *et al.*, “PANNs: Large-scale pretrained audio neural networks,” *IEEE/ACM TASLP*, 2020.
- [2] Google Research, “YAMNet,” GitHub repository.
- [3] T. Giannakopoulos, “Audio feature extraction using MFCCs,” Elsevier, 2015.
- [4] X. Xiong, “Real-time gunshot detection system,” Pennsylvania State University, 2023.