IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 9 December 2025, accepted 19 December 2025, date of publication 23 December 2025, date of current version 2 January 2026.

Digital Object Identifier 10.1109/ACCESS.2025.3647540

RESEARCH ARTICLE

AI-Telecommunications Synergy in Public Safety Systems Advancing Intelligent Law Enforcement

AMER AL-AHBABI[1,2], (Member, IEEE),
AND HAMED AL-RAWESHIDY[1], (Senior Member, IEEE)

[1]Department of Electronic and Electrical Engineering, College of Engineering, Design, and Physical Sciences, Brunel University of London, UB8 3PH Uxbridge, U.K.
[2]Ministry of Interior, Doha, Qatar

Corresponding author: Amer Al-Ahbabi (Amer.Al-Ahbabi@brunel.ac.uk)

ABSTRACT The increased rate of gun-related events witnessed in the context of the public safety determines the need to have intelligent systems of real-time surveillance in Internet of things (IoT) infrastructures. The current acoustic detection systems have a tendency to fail when trying to classify finer details of the gunshot, operate in a restricted space, and classify acoustically similar types of gunshots. To address this, we propose a class-aware augmentation strategy that selectively modifies specific audio classes to enhance inter-class discriminability, followed by standardized feature extraction at 22,050 Hz. In this paper, we have introduced lightweight Transformer-based model to detect and recognize gunshot instances in real-time and with multiple classes via 128-band log-mel spectrograms. The system operates across edge and fog layers, leveraging Augmented Covering Arrays (ACAs) and a MOEA/D-based optimizer to balance latency, energy consumption, and processing load. To enhance contextual awareness and dynamic threat prioritization, we introduce four intelligence metrics: Crime Risk Score (CRS), Crime Temporal Pattern Index (CTPI), Emergency Response Delay Impact Score (ERDIS), and Threat-Aware Priority Index (TAPI). An AutoML method is applied to optimize hyperparameters of models and reduce the effect of mixed up non-gunshot acoustic phenomena. Experimental results on 13-class gunshot data showed classification accuracy of 99.67%, representing 17.17 percentage point improvement. The macro-averaged F1-score above 0.993. Five-fold cross validation yielded average accuracy of 99.10%. With Streamlit interface the accuracy of the system is 98.10% in real-time implementation which validates the applicability on the use of the IoT to drive public safety.

INDEX TERMS Gunshot detection, transformer networks, edge-fog computing, real-time audio classification, situational intelligence metrics, public safety.

I. INTRODUCTION

The rising rates and severity of gunshot violence require urgent public safety and real-time response mechanisms to the problem in today's urban environments. Conventional detection systems for gunshot detection, which are usually based on acoustic thresholding or rule-based heuristics, have the drawback of limited accuracy, high false alarm rates and low adaptability in noisy or complex scenarios [1]. Such restrictions delay law enforcement response, lower situational awareness and negatively affect the reliability of available security infrastructures.

Recent advances in Artificial Intelligence (AI) and next-generation telecommunication technologies, such as edge computing, have opened new possibilities for responsive and distributed surveillance systems. Deep learning techniques have demonstrated promising results in acoustic event detection tasks [2], while modern edge-enabled communication frameworks support low-latency transmission and distributed inference. Recent studies in real-time AI-assisted applications, such as classroom behaviour analysis for engineering education [3], further highlight the importance of latency-aware model design and feedback mechanisms, which are directly relevant to safety-critical edge deployments such as gunshot detection.

Over the past five years, research on Transformer-based audio modeling, exemplified by the Audio Spectrogram

The associate editor coordinating the review of this manuscript and approving it for publication was Turgay Celik.

© 2025 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

Transformer and subsequent variants, has advanced sound event detection and related tasks [4], [5], [6], [7], [8]. In parallel, deployment-oriented studies report low-latency and edge-feasible designs for voice activity detection, lightweight sound event detection, adapter-based fine-tuning, and streaming inference [9], [10], [11], [12], [13], [14]. Within gunshot acoustics, Transformer-based multi-class recognition has appeared but remains comparatively limited [15]. Despite these advances, sensitivity to distribution shift persists; consequently, explicit evaluation under out-of-distribution (OOD) conditions is warranted, as underscored by recent cross-domain anomaly studies employing Transformer variants [16], [17], [18].

The outdoor dataset introduced in [19] provides directional and time synchronized gunshot recordings, yet covers a limited range of weapon types. Other efforts such as [20] have augmented datasets with synthetic audio from video games, which helps increase the amount of data, but introduces domain shifts which hurts performance in real-world deployments.

More recently, edge-fog-based surveillance systems such as [21] and indoor alert frameworks [22] have demonstrated potential for localized gunshot detection. Nonetheless, these systems often lack multi-class recognition capabilities, integrated edge-fog coordination, and real-time routing strategies optimized for public safety missions.

To address these limitations, integrated AI- telecommunications framework is proposed for real-time, multi-class gunshot detection and alert dissemination. Edge-based gunshot classification is integrated with fog computing resources. At the fog layer, optimization strategy is employed to minimize latency and energy consumption while ensuring balanced processing loads. The main contributions of this work are as follows:

- A lightweight Transformer-based architecture for accurate multi-class gunshot classification using 128-band log-mel spectrograms.
- A class-aware audio augmentation strategy that enhances inter-class discriminability and improves robustness in fine-grained gunshot classification tasks.
- A novel fog computing optimisation model integrating Augmented Covering Arrays (ACAs) with Multi-Objective Evolutionary Algorithm based on Decomposition (MOEA/D) to optimise latency and energy consumption, while balancing processing loads.
- Real-time deployment of the proposed system through a Streamlit-based interface and an MQTT-enabled alerting framework for fog-edge coordination.
- The design and integration of four situational intelligence metrics: Crime Risk Score (CRS), Crime Temporal Pattern Index (CTPI), Emergency Response Delay Impact Score (ERDIS), and Threat-Aware Priority Index (TAPI) to support dynamic threat scoring and intelligent routing.

The remainder of this paper is organized as follows. Section II provides background on acoustic gunshot detection, Transformer networks, and fog-edge architectures. Section III reviews related work in AI-driven surveillance systems. Section IV details the proposed methodology, including dataset processing and model architecture. Section V outlines the system optimization and routing strategy. Section VI presents the experimental setup and communication parameters. Section VII discusses performance results and intelligence metrics. Finally, Section VIII concludes the paper and suggests future research directions.

II. BACKGROUND

A. ACOUSTIC GUNSHOT CLASSIFICATION

Audio-based gunshot detection has become one of the critical points of the intelligent surveillance system. The basic activity is the identification and categorization of the occurrences of discharge of gunshots by their distinctive acoustic attributes. In the past, this field has had difficulties associated with differences in sound signature based on weapon type, ammunition caliber, noise in the environment, and the position of the microphone. In contrast to general detection of sound events, gunshot classification requires strong and accurate discrimination in highly varied and dynamic noisy environments that are not always predictable [23], [24]. All these background issues have contributed to the creation of more sophisticated AI models that can manage variability of the real world and also be used in real time.

B. AI ARCHITECTURES FOR GUNSHOT CLASSIFICATION

Initially, the classification of the gunshot sounds was carried out with the help of the manually constructed acoustic features together with the traditional classifiers [25], [26]. As the computational techniques improved, convolutional neural networks (CNNs) were used to obtain the spatial features of spectrogram representations. These models were shown to improve a recognition performance because it allowed learning of frequency patterns that are local to an occasion of gunfire.

Recurrent neural networks (RNNs), and their combinations with CNNs e.g. CNN-RNN and CNN-GRU were later added to be able to learn the temporal structure of acoustic signals. The techniques facilitated modelling of short duration gunshot events which were time-sensitive when dynamic audio conditions were involved [27], [28].

More recently, transformer architectures have been explored for their ability to model long-range dependencies in spectrogram-based classification, particularly for brief yet complex gunshot events [15], [29]. Lightweight variants, such as the convolutional-iConformer [12] and efficient attention mechanisms [9], [10], enable deployment on edge devices while maintaining high accuracy in real-time sound event detection.

Simultaneously, transfer learning methods have been applied to improve model generalization especially where there is a shortage of labeled data. Embeddings of environmental audio (gunshot samples) have been extracted using

models like YAMNet to utilise the representational strength of pre-trained networks [30].

The strategies of ensemble learning have also been explored to enhance the further enhancement of classification robustness. Stochastic weight averaging (SWA), model merging, and feature selection refinement are parameter-based methods (specifically, low-latency or edge-constrained) that have been applied [31], [32]. The developments highlight the continued advancement towards deployable gunshot detection systems of real-world gunshot detection systems that are highly efficient.

C. AI-TELECOMMUNICATIONS CONVERGENCE FOR PUBLIC SAFETY

The integration of AI and telecommunications has played a major role in the evolution of current public safety systems. As the number of IoT devices has grown, decentralized systems have been introduced to facilitate low-latency inference and responsive communication at the edge and fog levels [33]. This has been a transition towards decentralized intelligence left to traditional centralized processing.

These developments have led to the practical application of acoustic-based detection systems which will allow live inference and context-aware decision-making. In recent research, it has been highlighted that fog-based resource management and gunshot detection, as well as routing intelligence, have to be a part of unified frameworks [34]. The intent of such frameworks is to offer immediate threat recognition and transmission plans in line with a public safety scenario.

III. RELATED WORK

A. GUNSHOT AUDIO DETECTION AND CLASSIFICATION

Early works on sound recognition of gunshots mainly utilized hand-crafted features like Mel- frequency Cepstral Coefficients (MFCC), Linear Predictive Coding (LPC), and wavelet transforms with classical machine learning classifier like Support Vector Machines (SVM), K-Nearest Neighbors (K-NN) and Gaussian Mixture models (GMM) [35], [36], [37], [38]. Even though they provided general-purpose baselines, these techniques were not tuned to be domain-specific related to firearm acoustics. Techniques proposed in [39], [40], and [41] achieved notable accuracy but did not consider deployment constraints such as real-time latency or edge computing viability. Likewise, feature fusion methods [42] improved classification of music datasets but were not that adaptable in real-time.

Although firearm-specific detection and real-time feasibility was largely untested, neural network models were first used on large outdoor acoustic scenes [43]. Later developments like the multi-scale spectrum-shift CNN suggested in [44] enhanced time-frequency resolution, at the cost of high computational costs which limited their use in edge devices. False-positive mitigation measures like the inclusion of confounding acoustic events (e.g., bursting plastic bags) [45] allowed to improve the detection strength but still failed to consider latency and deployment. The most recent benchmarking activities [46], [47] have pointed to persistent difficulties in the rare-event modelling and a belief in architectures with restricted environments.

Practical deep learning-based systems have also been proposed. A CNN-GRU hybrid model in [48] achieved over 80% accuracy on noisy gunshot data, while [49] introduced trilateration mechanisms for localization. Nevertheless, real-time implementation and energy-aware optimization were not prioritized. Similarly, embedded implementations such as [50] presented latency and scalability challenges that limit deployment in edge-based public safety frameworks.

Recent Transformer-based models have been explored for modeling long-range temporal dependencies and global spectral patterns in gunshot acoustics [29]. Audio-Transformer backbones and hybrids have advanced event modeling and localization [4], [5], [6], [7], [8], while edge-oriented and streaming variants address real-time constraints on resource-constrained hardware [9], [10], [11], [12], [13], [14], [51]. For gunshot acoustics specifically, multi-class Transformer classification has been reported [15]; however, systematic evaluation under distribution shift remains limited. Cross-domain acoustic-anomaly studies indicate persistent sensitivity to domain mismatch and motivate explicit OOD evaluation [16], [17], [18]. Complementary directions include contrastive learning under limited data [52] and the use of pretrained audio embeddings such as PANNs [53], which can aid generalization but typically require adaptation for latency-sensitive, real-world deployments.

In parallel, the development of real-world gunshot datasets has been recognized as critical for model benchmarking and deployment. In [54], a multi-orientation dataset that was captured in outdoor settings was presented, allowing the performance of classification and direction-of-arrival (DoA) estimation, but did not provide contextual metadata like shooter posture. These contributions mirror the general demand of a wide range of, annotated, and deployment-commendable datasets to aid the development of clever, real-time gunshot identification mechanisms.

B. EDGE-FOG GUNSHOT DETECTION SYSTEMS

There is an increasing trend with the integration of edge, and fog technologies to use guns detection systems in real time and distributed. In [55], sensor based distributed architecture facilitated communication between sensors in both directions and had local storage capacities but it did not include onboard classification and latency sensitive task optimization. A neural network model as described in [56] was able to recognize battle field gunshots with 99 percent accuracy but the model did not look at the IoT-based implementation.

In [57], the authors suggested LAMOMRank, a latency-sensitive multi-objective scheduling algorithm to fog computing, which reduced response latency and task delivery time

without affecting multi-objective performance. Nevertheless, the research did not focus on dynamic service provisioning or adaptive scheduling when there are unpredictable tasks arrival that are critical to real-time event-driven applications like gunshot detection.

Most recent attempts have been made to investigate the use of fog-based optimization and acoustic preprocessing in safety applications. In [58], the Pareto optimization of the fog routing problem was realised by a two-step particle swarm optimization (PSO) and analytical hierarchy process (AHP) method, but no real-time responsiveness was considered. Similarly, [59] also used Extreme Learning Machines (ELM) to apply wavelet filtering to wind-resilient acoustic detection, but real-time inference and integration with the IoT were not addressed.

Although recent studies highlight the current surge in the application of gunshot detection intelligence on distributed computing layers, fundamental gaps are still present to the development of coherent, low-latency systems that integrate real-time multi-classification, resource-efficient fog implementation, and context-based threat prioritization.

C. NON-ACOUSTIC AND MULTI-MODAL GUNSHOT DETECTION SYSTEMS

Although the gunshot detection system has been based on acoustic sensing, non-acoustic and multi-modes are complementary which provide better situational awareness especially in harsh environmental or infrastructural situations. By use of ground vibrations, seismic sensing, like in use in business systems like ShotSpotter [19], is used, but is normally limited to a range of 0.51,km. Nevertheless, such systems frequently have false positive rates (FPR) of 1525 percent in the urban setting because of construction and vehicle interference [60], [61].

Thermal and IR sensors can monitor short-range instances of 50200m muzzle flashes, and operate in short ranges up to 50200m. FPRs can be minimized to less than 5 percent with the help of AI-based filtering [62] though these kinds of systems require extensive computational resources, making them impossible to implement in the edge. With CCTV infrastructure, visual detection enables tracking of the firearm trajectory [63], which is effective in the classification of 100500,m range and a FPR less than 5% at a combination with object detection models. These systems however are limited in terms of scalability because of power and bandwidth requirements.

Other perspectives can be found with optical/video systems, which have a range of 50-150 m and false alarm rates of 25-30% [64], [65], but have a short range and a high false alarm. Multi-modal approaches have come into the picture as a viable solution to the drawbacks of the single-modality systems [66]. It has been found that hybrid acoustic-seismic approaches [67] are more robust, with FPRs decreasing to around 10 per cent in noisey environments [60]. Acceleration-based systems that use wearables (e.g., tri-axial sensors on the wrist) [68] offer local-detection capabilities and have the potential to interface with IoT-based edge platforms [62]. These technologies, detection ranges, FPRs and edge suitability have been summarized and compared in Table 1 based on the literature.

This development highlights the possibility to augment the aesthetics of more advanced and user-friendly fog-edge architectures using non-acoustic sensors, which will have a more resilient nature and reduce false alarms in real time gunshots detection and threat prioritisation systems.

D. PROBLEM DEFINITION

The proper reaction to gunshot-related events is becoming more and more reliant on acoustic surveillance devices capable of not only detecting the gunshots in real-time but classifying the types of weapons and determining the threat level. Although the existing literature has achieved some improvements in binary gunshot detection (e.g., gunshot vs. non-gunshot), the literature is relatively thin in providing the fine-grained multi-class classification that is necessary in law enforcement activities and situational awareness.

The use of traditional CNN models can be ineffective in terms of capturing long-range time-based trends and may not be efficient when deployed in resource-constrained edge devices. Besides, they have little integration with fog and IoT architecture, which also adds to high latency and limits their use to real-time distributed systems.

To overcome these issues, this paper suggests Transformer-based system that:

- Supports using spectrogram-based embeddings to fine-tune the classification of types of gunshots.
- Real time inference on edge devices using a low-weight Transformer model.
- Complementary to the fog computing to provide real-time and low-latency alert distribution.
- Uses multi-objective optimization under the ACA with situational intelligence measures of adaptive threat prioritization.

All these modules address a certain limitation found in the literature. The combination of the two offers a deployable, low-latency, and intelligence-based gunshot detection solution to real-world IoT-based public safety. The detailed methodology formulations and implementation plans of each component are givenin IV-B, IV-C, and V, in which the limitations identified are converted to deployable, system-level solutions. This paper explores the extent to which this type of lightweight Transformer-based architecture can achieve high detection rates with operation under very strict latency and computational requirements on edge-fog systems.

IV. PROPOSED METHODOLOGY

The system architecture, Transformer-based classification model, communication integration, and optimization strategies developed to resolve the main issues in Section III-D. To address the shortcomings of the

TABLE 1. Comparison of sensor-based gunshot detection technologies.

Technology	Detection Range	False Positive Rate (Urban)	Edge Suitability	Key References
Acoustic	1–2 km	10–20%	High (low compute)	[15], [69], [70]
Optical / Video	50–150 m	25–30%	Low (high power)	[64], [65]
Seismic	500–1000 m	15–25%	Moderate (low power)	[60], [61]
Infrared	50–200 m	<5% (with AI)	Low (high compute)	[62], [71]
Visual (CCTV)	100–500 m	<5% (with AI)	Low (high power)	[62], [63]

Notes: Detection ranges are approximate and reflect typical field deployments. Reported false positive rates are based on urban conditions, with AI enhancement where applicable. Edge suitability is evaluated based on computational load and power consumption, with "High" indicating low resource demands and "Low" indicating high demands.

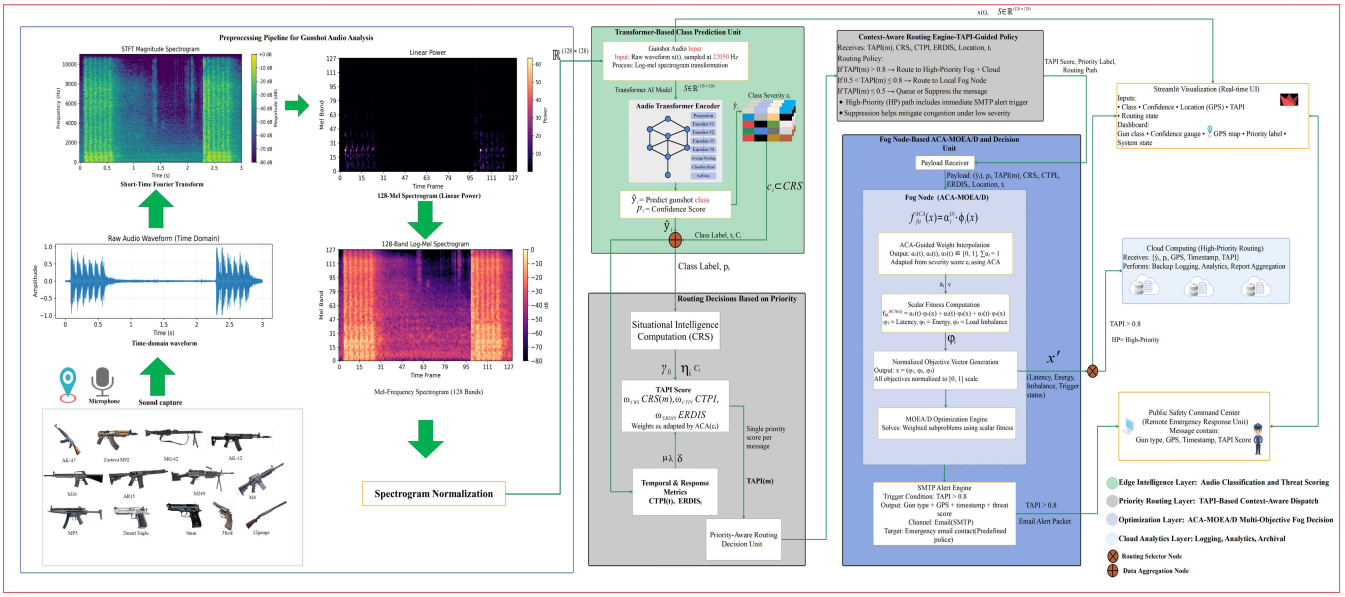


FIGURE 1. System architecture of the proposed gunshot detection and alerting framework.

gunshot-type classification, model efficacy, and real-time deployment, we will suggest an interwoven AI-telecommunications framework that will be utilized to detect the gunshots with high accuracy and low latency in the IoT-based public safety settings.

This proposed pipeline has four main steps, namely (1) audio preprocessing and spectrogram, (2) Transformer-based classification, (3) ensemble optimization and merging, and (4) real-time alerting through edge/fog deployment. Figure 1 presents an overview. Each of the stages is detailed in the following subsections.

A summary of the used notations is presented in Table 2

A. DATASET DESCRIPTION

The proposed gunshot classification system is trained and tested on two complementary datasets, which vary in source, acoustic environment, gunshot types, as well as microphone configurations, allowing to generalize to a wide variety of situations.

Dataset 1 contains 851 gunshot audio recordings comprising nine gunshot classes: AK-12, AK-47, IMI Desert Eagle, M16, M249, M4, MG-42, MP5, and Zastava M92. Originally introduced by Tuncer et al. [25], the recordings were extracted from publicly available YouTube videos, sampled at 44.1 kHz, and segmented into 2-second clips using WavePad Audio Editor. There were eight classes initially, but one extra class, which was M4, is available separately because the initial publication only reported eight classes. The distribution of classes is not balanced in nature.

The U.S. Air Force Research Laboratory and CADAS published dataset 2, which was gathered by these two organisations [19]. The data set is as a result of 9mm, AR15, 38cal, and 12gauge shots in different orientations and changing environments in which multichannel microphone arrays were used to record the shots. There are no less than a gunshot in every 2-second clip; bad-quality parts were eliminated. Metadata is offered like the ID of the microphone, location, and context of firing.

Together, these collections of data present an expansive and auditory standard with 13 categories of firearm discharge within two realms, which includes the categories of Internet-derived recordings (Dataset 1) and regulatory edge-gathered field recordings (Dataset 2). The composition of the merged dataset, the sample numbers within each

TABLE 2. Summary of notation.

Symbol	Definition
$\mathcal{A}$	ACA design space (unit simplex)
$\mathcal{A}_{c_i}$	ACA subregion adapted to severity score c_i
$\mathcal{A}_D$	Audio dataset
AF_i	Acoustic attenuation factor
AF_{en}	Communication/path attenuation factor
a_i	Integer component of ACA weight vector
B_M	Number of mel bands
$B_{\text{en}f}$	Link bandwidth between edge node en and fog node f
C	Number of classes
c_i	Class-derived severity score
D_{en}	Message size from edge node en
d	Transformer model (embedding) dimension
d_{ff}	Feed-forward width in Transformer block
f	Fog node index
$H_m(k)$	m-th mel filter response at bin k
h	Number of attention heads
i	Detected gunshot (event) index
K	Number of severity criteria (CRS components)
L	Number of encoder layers
L_f	Load at fog node f
M	Total number of fog nodes
N	Total number of edge messages or nodes
p	p-value (Mann-Whitney U test)
$\mathbf{P}_i$	Projected input sequence
$\mathbf{p}_i$	Predicted class probabilities
$\mathbf{S}_i$	Input spectrogram for sample i
T	Number of time frames
T^f_{verify}	Verification time at fog node f
$\mathbf{W}_{\text{proj}}$	Projection matrix
$X_t(k)$	STFT coefficient at time t, bin k
$x(t)$	Time-domain waveform amplitude
$x_{\text{en},f}$	Routing indicator: 1 if en$\rightarrow f$, else 0
$\hat{y}_i$	Predicted class index for sample i
$\epsilon^{\text{tx}}_{\text{en}}$	Transmission energy per bit at edge node en
ϵ^{proc}_f	Processing energy rate at fog node f
η_i	CRS component weight
θ	Trainable model parameters
κ, ζ	Load-balancing coefficients
λ	ERDIS scaling factor
μ	ERDIS urgency growth rate
ρ	CTPI decay constant
τ	Time offset (CTPI window)
ϕ_i	i-th objective component
$\boldsymbol{\phi}(\mathbf{X})$	Objective vector (latency, energy, imbalance)
$\omega_{\text{CRS}}, \omega_{\text{CTPI}}, \omega_{\text{ERDIS}}$	TAPI metric weights

TABLE 3. Final class inventory after merging the two gunshot datasets (counts denote 2-second clips).

Class	Samples	Source (Reference)
9mm	669	From [19]
AR15	597	
38cal	503	
12gauge	379	
IMI Desert Eagle	100	From [25]
M16	100	
M4	100	
MG-42	100	
MP5	100	
M249	99	
AK-12	98	
Zastava M92	82	
AK-47	72	
Total	**2,999**	

Note: Sample counts reflect the merged dataset, with classes sourced from the respective references as indicated.

class and the original sources of the data are summarized in Table 3. This consolidated corpus for training and testing rigorous classifiers for spectrograms of gunshots. Class imbalance is handled during preprocessing which is detailed in Section IV-B.

In addition to these datasets, the UrbanSound8K corpus [72] is used exclusively as an external source of non-gunshot audio for confounder stress-testing and outlier exposure (OE). Specifically, the `gun_shot` class is excluded, and samples are drawn from the remaining urban sound classes: air_conditioner, car_horn, children_playing, dog_bark, drilling, engine_idling, jackhammer, siren, and street_music. These clips are partitioned into an OE-training subset and a strictly disjoint out-of-distribution (OOD) evaluation subset with no file or path overlap with the in-distribution folds. UrbanSound8K audio is never used as positive gunshot training data.

Table 4 provides an overview of the UrbanSound8K non-gunshot confounder classes employed in this study. For each class, it lists the total number of clips available, the subset converted to spectrograms, fold coverage, counts of foreground and background salience, and typical clip duration. The exclusion of the `gun_shot` class ensures that UrbanSound8K confounders serve solely for robustness testing rather than positive training.

For outlier exposure and OOD evaluation, these non-gun confounders are partitioned into an OE-training subset and a strictly disjoint OOD evaluation subset with zero file or path overlap with in-distribution folds. All audio is processed using the same pipeline as the in-distribution data to extract log-mel spectrograms, truncated or padded to a fixed length.

Although the curated dataset covers thirteen distinct firearm classes, these categories correspond to the most frequently encountered weapon types reported in law-enforcement incident databases and open-source repositories. There are multiple instances within each class recorded at different distances and under different acoustic conditions and, as a result, a large amount of class variability.

B. PREPROCESSING AND FEATURE ENCODING

To solve the acoustic confusion between spectrally similar classes of gunshots, a class-specific preprocessing pipeline was used with the data. Specifically, in classes that were highly susceptible to misclassification (e.g. M4 and M16), time-domain (targeted) augmentations were used to improve spectral separability without loss of core identity features.

TABLE 4. UrbanSound8K non-gunshot confounder classes and usage statistics.

Class	Total Clips	Clips Used	Used Fraction	Folds	Foreground Salience	Background Salience	Duration (s)
air_conditioner	1000	306	0.306	10	569	431	3.99±0.09
car_horn	429	105	0.245	10	153	276	2.46±1.62
children_playing	1000	247	0.247	10	588	412	3.96±0.27
dog_bark	1000	362	0.362	10	645	355	3.15±1.33
drilling	1000	302	0.302	10	902	98	3.55±1.00
engine_idling	1000	301	0.301	10	916	84	3.94±0.37
jackhammer	1000	99	0.099	10	731	269	3.61±0.89
siren	929	247	0.266	10	269	660	3.91±0.50
street_music	1000	243	0.243	10	625	375	4.00±0.00

The augmentation step used on each waveform $x(t)$ of each of a confusion-prone classes is given as follows:

$$\hat{x}(t) = \mathcal{U}_{f_o}\left(\mathcal{S}_\alpha\left(\mathcal{N}_\sigma\left(\mathcal{D}_{f_d}(x(t))\right)\right)\right), \tag{1}$$

where $\mathcal{D}_{f_d}$ denotes downsampling to a lower frequency f_d, $\mathcal{N}_\sigma$ represents additive Gaussian noise with variance σ, $\mathcal{S}_\alpha$ applies time-stretching with a factor $\alpha \in [0.90, 0.95]$, and $\mathcal{U}_{f_o}$ restores the original sampling rate f_o. Such transformation alters the log-mel distribution of the target class, which increases its trainability.

Each processed waveform $x_i(t)$ was converted into a log-mel spectrogram $\mathbf{S}_i \in \mathbb{R}^{B_M \times T}$, where $B_M = 128$ represents the number of mel frequency bands and $T = 128$ corresponds to the number of temporal frames. The spectrogram transformation involved Short-Time Fourier Transform (STFT), mel filterbank projection, and logarithmic compression [73]:

$$\mathbf{S}_i(m,t) = \log\left(\sum_{k=1}^{K} |X_t(k)|^2 \cdot H_m(k)\right), \quad m \in \{1,\ldots,B_M\}, \tag{2}$$

where $X_t(k) \in \mathbb{C}$ is the STFT coefficient at frame t and frequency bin k, and $H_m(k) \in \mathbb{R}$ is the mel filter response for band m.

To ensure consistency in input dimensions, spectrograms were either zero-padded or truncated to contain exactly $T = 128$ frames. Global mean and standard deviation normalization was then applied over the whole data set. Lightweight augmentations during training time were also used to achieve better generalization.

C. TRANSFORMER-BASED CLASSIFICATION MODEL

The normalized log-mel spectrogram $\mathbf{S}_i \in \mathbb{R}^{B_M \times T}$ is processed by a Transformer-based architecture optimized for time-frequency classification [74]. The model maps each input to a class probability vector:

$$\mathbf{p}_i = f(\mathbf{S}_i; \theta), \tag{3}$$

where θ represents all the trainable parameters, which include weights of projections, attention matrices and classifier layers. To interface with the Transformer encoder the input is transposed and projected into a latent space of dimension $d = 256$ using a learnable matrix [75]:

$$\mathbf{P}_i = \mathbf{S}_i^\top \mathbf{W}_{\text{proj}} \in \mathbb{R}^{T \times d}, \tag{4}$$

the resulting sequence of 128 latent vectors (one per time step) is processed through stacked Transformer encoder blocks, each comprising multi-head self-attention, feed-forward layers, and layer normalization [75]:

$$\hat{\mathbf{S}}_i = \text{LN}\left(\text{MultiHeadAttn}(\mathbf{P}_i) + \text{FFN}(\cdot)\right), \tag{5}$$

a global average pooling layer converts the temporal sequence to a fixed-size vector for final classification [76]:

$$\hat{y}_i = \text{Softmax}\left(\mathbf{W}_{\text{out}} \cdot \text{AvgPool}(\hat{\mathbf{S}}_i)\right), \tag{6}$$

where $\hat{y}_i \in \{1,\ldots,C\}$ denotes the predicted gunshot class. The complete pipeline, including data preprocessing, augmentation, model training, and deployment, is outlined in Algorithm 1.

The model is based on 4 Transformer encoder layers with 8 heads and feed-forward width of 1024. Dropout (rate=0.206) is used after using both attention and feed-forward layers. Hyperparameters such as attention size, dropout, feed-forward width and learning rate were optimized using Optuna based AutoML tuning.

The combined tuning strategy, which was chosen to guarantee the fair and reproducible optimization process, was hybrid tuning strategy. The preliminary manual parameter exploration was used to set the parameter ranges of attention size, dropout rate, and learning rate to reasonable values. These constraints were further narrowed down with Optuna Tree-structured Parzen Estimator (TPE) sampler to accomplish multi-objective optimization to achieve validation loss and accuracy. This two-phase method established that the AutoML refinement converged faster and obtained a little greater validation accuracy than hand parameter tuning and minimized the search effort and subjective bias during parameter selection.

The overall architecture is shown in Figure 2. Predicted outputs class probabilities, severity scores, and timestamps serve as inputs to the situational intelligence modules and downstream routing logic.

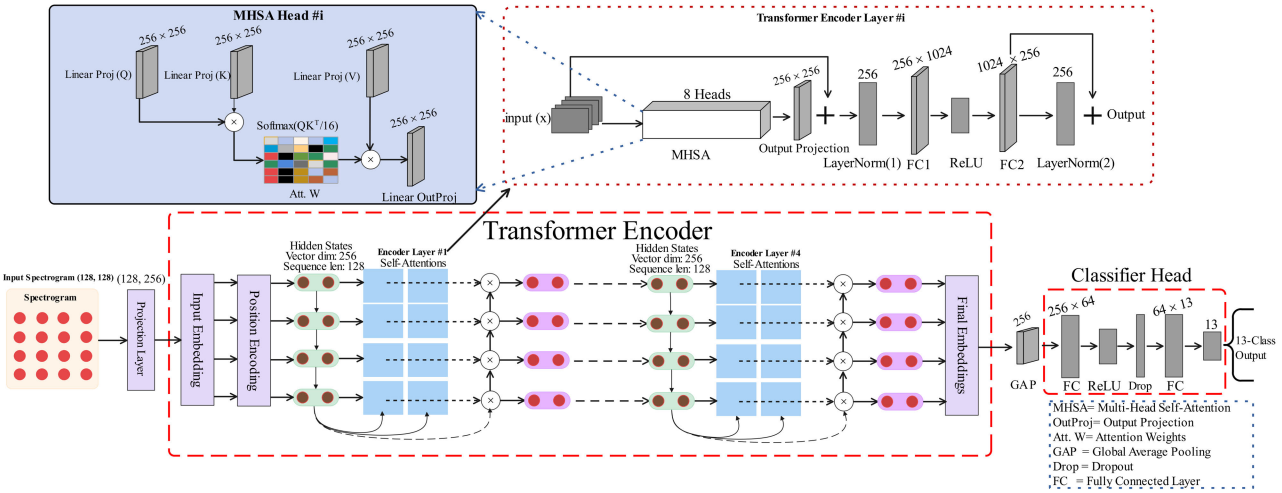


FIGURE 2. Overview of the proposed Transformer-based model for real-time gunshot classification.

To support threat-aware decision-making, each predicted class $\hat{y}_i$ is mapped to a severity score $c_i \in [0, 1]$ using a predefined lookup table:

$$c_i = \text{SeverityLookup}(\hat{y}_i), \tag{7}$$

this mapping assigns risk values based on ballistic and forensic characteristics (Table 5). The resulting c_i informs the computation of CRS, CTPI, and ERDIS metrics, as well as the scalar fitness and TAPI-based routing strategies (see Sections V-B and V-C).

V. SYSTEM OPTIMIZATION AND ROUTING

Section formalises the core situational intelligence mechanisms that govern system-level optimisation and communication behaviour in our gunshot detection framework. Modelling latency, energy, risk, and workload fairness as explicit objectives and integrating them into a multi-objective optimisation and routing pipeline enables intelligent, real-time prioritisation of alerts under fog-edge constraints.

A. MULTI-OBJECTIVE OPTIMIZATION

The fog-based routing process is modelled as a multi-objective optimisation problem over three conflicting objectives: latency minimisation, energy efficiency, and load balancing. Let N be the number of messages generated at edge nodes ($en \in \{1, \ldots, N\}$) and M the number of fog nodes ($f \in \{1, \ldots, M\}$). The binary routing variable is defined as $x_{en,f} \in \{0, 1\}$, where $x_{en,f} = 1$ indicates that the message from edge node en is assigned to fog node f. The routing decision matrix is denoted by $\mathbf{X} = [x_{en,f}] \in \{0, 1\}^{N \times M}$, constrained such that each message is assigned to exactly one fog node [77]:

$$\sum_{f=1}^{M} x_{en,f} = 1, \quad \forall en \in \{1, \ldots, N\}, \tag{8}$$

the routing logic presented here builds upon the threat-sensitive scalar fitness formulation and Augmented Covering Arrays (ACA)-based weight adaptation discussed in Section V-C. The objective vector is then expressed as [78]:

$$\boldsymbol{\phi}(\mathbf{X}) = (\phi_1(\mathbf{X}), \phi_2(\mathbf{X}), \phi_3(\mathbf{X})), \tag{9}$$

where ϕ_1, ϕ_2, and ϕ_3 represent latency, energy, and load imbalance objectives, respectively.

(i) Latency Minimization: The first objective quantifies the total end-to-end delay incurred when offloading messages from edge nodes to fog nodes. This includes transmission latency and fog-layer verification delay:

$$\phi_1(\mathbf{X}) = \sum_{en=1}^{N} \sum_{f=1}^{M} x_{en,f} \cdot \left(\frac{D_{en}}{B_{enf}} + T^{f}_{\text{verify}} \right), \tag{10}$$

where D_{en} is the message size from edge node en, B_{enf} is the communication bandwidth between edge node en and fog node f, and T^{f}_{verify} denotes the verification delay at fog node f.

(ii) Energy Consumption: This objective models the total energy consumption across edge and fog nodes, incorporating signal attenuation effects:

$$\phi_2(\mathbf{X}) = \sum_{en=1}^{N} \sum_{f=1}^{M} x_{en,f} \cdot \left(\frac{1}{AF_{en}} \cdot \epsilon^{\text{tx}}_{en} D_{en} + \epsilon^{\text{proc}}_{f} T^{f}_{\text{verify}} \right), \tag{11}$$

where $AF_{en} \in (0, 1]$ is the attenuation factor which is the signal loss corresponding to edge node en. A lower AF_{en} means greater transmission power/energy because of greater acoustic attenuation. $\epsilon^{\text{tx}}_{en}$ is the energy per bit for transmission and $\epsilon^{\text{proc}}_{f}$ is the processing energy per unit time at fog node f.

(iii) Load Imbalance Minimization: To ensure fair workload distribution and mitigate congestion the load imbalance

Algorithm 1 Transformer-Based Pipeline for Gunshot Audio Classification and Acoustic Event Detection

Input: Raw gunshot audio files (.wav), metadata files
Output: Transformer model and deployment interface
Step 1: Preprocessing and Augmentation
for each audio file $x_{\text{raw}} \in$ raw_audio_folder **do**
  if file is from a confusion-prone class **then**
    Apply class-specific augmentation as defined in Eq. (1)
    Save augmented variants to augmented_folder
  else
    Save original file to non_augmented_folder
  end if
end for
Step 2: Log-Mel Spectrogram Extraction
Merge augmented and non-augmented audio into dataset $\mathcal{A}_D$
for each waveform $x_i(t) \in \mathcal{A}_D$ **do**
  Resample to 22,050 Hz
  Compute log-mel spectrogram: $\mathbf{S}_i = \text{LogMel}(x_i(t))$
  Save $\mathbf{S}_i$ as.npy file
  Create metadata entry with label and spectrogram path
end for
Consolidate metadata into a full metadata table
Step 3: Metadata Cleaning and Balancing
Normalize class labels
Shuffle metadata randomly
Apply stratified balancing: max 350 samples per class
Save balanced metadata
Step 4: Data Preparation
Load features $\mathbf{X}$ and labels y using metadata references
Step 5: Hyperparameter Optimization
Run AutoML to optimize Transformer hyperparameters
$best\_params \leftarrow$ run_optuna($\mathbf{X}$, y, n_trials = 50)
Step 6: Final Model Training
Instantiate final Transformer:
$f_\theta \leftarrow$ Transformer($best\_params$)
Transformer encoder applied to $\mathbf{S}_i$ to obtain $\hat{\mathbf{S}}_i$ as in Eq. (5)
Final prediction computed ($\hat{y}_i$) as in Eq. (6)
Train on full dataset: f_θ.train($\mathbf{X}$, y)
Step 6: Cross-Validation for Stability
for each fold in 5-fold cross-validation **do**
  Train model on training split
  Evaluate on validation split
  Log fold-wise accuracy and confusion matrix
end for
Step 7: Deployment and Inference
Deploy trained model using **Streamlit** UI
Integrate real-time alerts via MQTT

is captured using normalized metric [79]:

$$\phi_3(\mathbf{X}) = \frac{\max_f(L_f) - \min_f(L_f)}{\bar{L}}, \tag{12}$$

where the load at fog node f is computed as:

$$L_f = \sum_{en=1}^{N} x_{en,f} \cdot \left(\kappa D_{en} + \zeta T^f_{\text{verify}}\right), \tag{13}$$

where κ and ζ are tunable coefficients which scale the contribution of transmission volume and processing time respectively. The average load at all fog nodes is expressed as $\bar{L} = \frac{1}{M}\sum_{f=1}^{M} L_f$. Minimizing $\phi_3(\mathbf{X})$ reduces queue build-up, thermal stress and scheduling delays in the event of concurrent gunshot.

B. CORE THREAT INTELLIGENCE METRICS (CRS, CTPI, ERDIS)

These three fundamental measures convert the raw forecasts to a reality-time contextual awareness.

(i) CRS: The Crime Risk Score reflects the level of threat of each detection taking into account classifier confidence and the local context and class severity. It is computed as:

$$\text{CRS}_i = \eta_1 \cdot p_i + \eta_2 \cdot \gamma_{l_i} + \eta_3 \cdot c_i, \tag{14}$$

p_i is the model predicted confidence level of the model for gunshot type i, γ_{l_i} is the historical frequency of gunfire at location l_i, and c_i is the severity score for class i, as is shown in Table 5. The parameters η_1, η_2, and η_3 control the impact of the model confidence, spatial risk and class severity respectively in the overall risk score.

(ii) CTPI measures the recurrence of gunshot events over time to facilitate time sensitive decision-making. To highlight the recent nature of some illustrations, CTPI is calculated by an exponential decay model that gives more weight for more recent detections:

$$\text{CTPI}_{\text{decay}}(t) = \frac{1}{Z}\sum_{\tau=0}^{W} e^{-\rho\tau} \cdot \delta(t-\tau), \tag{15}$$

where t is the index of current time, $\tau \in [0, W]$ is a constant the lookback window and $\rho > 0$ is a constant the decay factor. The term $\delta(t-\tau)$ represents the number of detections at time $t-\tau$ and the normalization constant $Z = \sum_{\tau=0}^{W} e^{-\rho\tau}$ makes sure that CTPI stays in the range [0, 1].

Higher CTPI values indicate concentrated gunfire activity in the immediate past, supporting dynamic prioritization of alerts, patrol routing, and resource allocation in public safety operations.

(iii) ERDIS: To quantify operational risk from delayed firearm response, the Exponentially Rising Delay Impact Score (ERDIS) is defined as:

$$\text{ERDIS}_i = \lambda \cdot AF_i \cdot \text{CRS}_i \cdot \left(1 - e^{-\mu \cdot \Delta t_i}\right), \tag{16}$$

where Δt_i is the response delay, μ controls the urgency growth rate, and λ is a scaling constant. $AF_i \in (0, 1]$ denotes the attenuation factor of gunshot signal i, capturing propagation loss due to frequency, distance, and environmental absorption. Lower AF_i reflects stronger signal degradation, amplifying risk by reducing situational confidence.

This exponential model describes the increase of threat urgency over time, and in poor acoustic conditions in particular. Higher ERDIS values indicate a higher operational cost, and are used to trigger escalation, reordering of alert queues and fog cloud reordering decisions.

C. SCALAR FITNESS WITH THREAT-AWARE AND ACA ADJUSTMENT

Each gunshot class was assigned a severity score $c_i \in [0, 1]$, reflecting its public-safety risk. These scores informed scalar fitness adaptation and routing thresholds. Following forensic ballistics methodology [80], the score was based on five normalized indicators: (i) kinetic energy (KE), derived from cartridge mass and muzzle velocity; (ii) cyclic firing rate (RPM); (iii) lethality proxy that saturates at rifle class energy levels; (iv) magazine capacity; and (v) operational deployment context (i.e. prevalence, portability and likelihood of civilian misuse). Table 5 describes the normalized attributes and calculated scores.

1) NORMALIZATION WITH OPERATIONAL CAPS

Operationally reasonable thresholds were used to scale each indicator to the scale of [0,1]. Extreme historical values were also disqualified to make it applicable in public-safety situations. Specifically:

$$\mathrm{KE}_{\mathrm{norm}} = \min\left(1, \frac{\mathrm{KE}_i}{E_{\mathrm{dash}}}\right), \tag{17a}$$

$$\mathrm{RPM}_{\mathrm{norm}} = \min\left(1, \frac{\mathrm{RPM}_i}{R_{\mathrm{dash}}}\right), \tag{17b}$$

$$\mathrm{Mag}_{\mathrm{norm}} = \min\left(1, \frac{\mathrm{Mag}_i}{M_{\mathrm{dash}}}\right), \tag{17c}$$

where $E_{\mathrm{dash}} \approx 2000\,\mathrm{J}$, $R_{\mathrm{dash}} = 1200\,\mathrm{rpm}$, and $M_{\mathrm{dash}} = 30$ rounds. Lethality score ($\mathrm{Leth}_{\mathrm{norm}}$) saturates for rifle class KE, e.g. 0.9 for rifles, 0.6 for handguns, in accordance with wound ballistics data [80]. The context factor reflects field deployability, favoring common, portable rifles (e.g., AK-47: 1.0) over less portable systems (e.g., MG-42: 0.9) or civilian handguns (0.6-0.7).

2) SEVERITY SCORE COMPUTATION

Each class's severity score was computed as an equal-weight average of the five normalized indicators:

$$c_i = \frac{1}{5}\left(\mathrm{KE}_{\mathrm{norm}} + \mathrm{RPM}_{\mathrm{norm}} + \mathrm{Leth}_{\mathrm{norm}} + \mathrm{Mag}_{\mathrm{norm}} + \mathrm{Context}\right), \tag{18}$$

ensuring balanced influence. Operational caps limit score inflation from outliers such as high-RPM weapons.

3) ILLUSTRATIVE EXAMPLE

For the AK-47 (7.62×39mm, 710 m/s, 8 g, 2016 J, 600 rpm, 30-round mag), normalized indicators yield $\mathrm{KE}_{\mathrm{norm}} = 1.0$, $\mathrm{RPM}_{\mathrm{norm}} = 0.5$, $\mathrm{Leth}_{\mathrm{norm}} = 0.9$, $\mathrm{Mag}_{\mathrm{norm}} = 1.0$, Context = 1.0, resulting in $c_i = 1.0$. The MG-42 (7.92×57mm, 755 m/s, 11.5 g, 3278 J, 1200 rpm capped at 600 rpm) scored slightly lower ($c_i \approx 0.97$) due to lower deployability in civilian environments.

Reproducibility: Raw inputs (e.g., velocity, mass, KE, RPM, magazine size) used for c_i computation are included in Table 5, sourced from [80] and verified using contemporary ballistics databases.

In multi-objective optimization scenarios such as weighted-sum decomposition in MOEA/D, the scalar fitness function consolidates latency, energy, and task imbalance into a single objective:

$$f_{\mathrm{fit}}(x) = \alpha_1 \cdot \phi_1(x) + \alpha_2 \cdot \phi_2(x) + \alpha_3 \cdot \phi_3(x), \tag{19}$$

where $\alpha_i \in [0, 1]$ are static weight coefficients constrained by $\sum \alpha_i = 1$. This formulation enables scalarization of the multi-objective problem for use in decomposition-based optimization methods such as MOEA/D.

To incorporate threat sensitivity, we apply ACA to dynamically modulate the weights based on c_i. ACA defines discrete design space of candidate vectors $\mathcal{A} \subset \mathbb{R}^3$, each satisfying the unit simplex constraint: $\mathcal{A} = \left\{\left(\frac{a_1}{v-1}, \frac{a_2}{v-1}, \frac{a_3}{v-1}\right) \middle| a_1 + a_2 + a_3 = v-1,\ a_i \in \mathbb{Z}_{\geq 0}\right\}$, where $v \in \mathbb{Z}_{>1}$ defines the ACA resolution, yielding $\alpha_i \in \{0, 0.1, \ldots, 1.0\}$. Two anchor vectors $\vec{\alpha}^{\mathrm{low}}, \vec{\alpha}^{\mathrm{high}} \in \mathcal{A}$ are selected to represent the minimum and maximum severity conditions. The severity-aware interpolation for the latency and energy weights is defined as $\alpha_j^{(t)}(c_i) = \alpha_j^{\mathrm{low}} + \left(\alpha_j^{\mathrm{high}} - \alpha_j^{\mathrm{low}}\right) \cdot c_i$, $j \in \{1, 2\}$ and the third weight, $\alpha_3^{(t)}$, is computed to preserve the simplex while enforcing minimum exploration:$\alpha_3^{(t)} = \max\left(\delta_\alpha, 1 - \alpha_1^{(t)} - \alpha_2^{(t)}\right)$, where $\delta_\alpha \in [0, 0.1]$ prevents the imbalance objective from vanishing during high-severity events.

Dynamic adjustment allows for smooth transitions across subregions of the ACA grid that are denoted as $\mathcal{A}_{c_i} \subset \mathcal{A}$, where prioritization changes between latency, energy, and balance objectives in response to changing threat circumstances. ACA can provide systematic coverage of the input, unlike random perturbation techniques, and can converge and be diverse. It offers a principled foundation of controlled adaptation in multi-objective decision systems as defined in [81]. The ACA-enhanced scalar fitness function becomes:

$$f_{\mathrm{fit}}^{\mathrm{ACA}}(x) = \alpha_1^{(t)} \cdot \phi_1(x) + \alpha_2^{(t)} \cdot \phi_2(x) + \alpha_3^{(t)} \cdot \phi_3(x), \tag{20}$$

where the weights $\alpha_i^{(t)}$ are contextually adapted per gunshot severity, ensuring responsive optimization behaviour. The weight tuning and fitness evaluation based on the ACA in the fog assignment process is presented in Algorithm 2 and its results directly feed the routing logic as indicated below.

D. PRIORITY-AWARE ROUTING VIA TAPI

To support situationally aware communication within the MQTT-based fog network, a dynamic routing policy is governed by the Threat-Aware Priority Index (TAPI). This

TABLE 5. Computed severity scores c_i from normalized attributes. Includes raw inputs and justification per gun class.

Gunshot Class	KE_{norm}	RPM_{norm}	$Lethality_{norm}$	Mag_{norm}	Context	c_i	Severity Justification (Raw Inputs)
AK-47	1.00	0.50	0.90	1.00	1.00	1.00	7.62×39 (710 m/s, 8 g, 2016 J), 600 rpm, 30-rd mag; high conflict prevalence.
MG-42	1.00	0.50	0.90	1.00	0.90	0.97	7.92×57 (755 m/s, 11.5 g, 3278 J), 1200 rpm (capped 600), belt-fed; deployability penalty.
AK-12	0.70	0.58	0.85	1.00	1.00	0.95	5.45×39 (900 m/s, 3.4 g, 1403 J), 700 rpm, 30-rd mag.
M16, AR15, M249	0.80–0.86	0.50–0.71	0.90	1.00	0.90–1.00	0.91–0.93	5.56×45 (915–945 m/s, 4 g, 1674–1786 J), 600–850 rpm, 30-rd mag.
M4, MP5, M92	0.20–0.75	0.50–0.67	0.80	0.50–1.00	0.80	0.80–0.90	MP5: 9mm (400 m/s, 8 g, 640 J), M4: 5.56mm (1500 J); LE/military.
Desert Eagle	1.00	0.25	0.75	0.23	0.60	0.75	.50 AE (470 m/s, 19.4 g, 2143 J), 300 rpm, 7-rd mag.
9mm, .38 cal, 12 gauge	0.10–0.30	0.10–0.25	0.60	0.20–0.50	0.60–0.70	0.60–0.70	9mm: 370 m/s, 8 g, 540 J, 300 rpm, 15-rd mag.

Algorithm 2 ACA-Guided Multi-Objective Routing via Dynamic Scalarization

Input: Initial routing matrix $\mathbf{X}^{(0)} \in \mathbb{R}^{N\times M}$; ACA weight grid $\mathcal{A} \subset \mathbb{R}^D$; severity scores $\{c_i\}$
Output: Optimized routing matrix $\hat{\mathbf{X}}$
Definitions: Objective mapping $\boldsymbol{\phi} : \mathbb{R}^{N\times M} \to \mathbb{R}^D$ (Eq. 9)
Initialize iteration index $t \leftarrow 0$
repeat
  Compute adaptive weights: $\boldsymbol{\alpha}^{(t)} = \mathrm{ACA}(c_i, t)$ (Eq. (20))
  Compute scalar fitness: $f_{\mathrm{fit}}^{(t)} = \boldsymbol{\alpha}^{(t)} \cdot \boldsymbol{\phi}(\mathbf{X}^{(t)})$
  Solve MOEA/D subproblem:
  $\mathbf{X}^{(t+1)} \leftarrow \mathrm{MOEA/D{-}Solve}(\boldsymbol{\phi}, \boldsymbol{\alpha}^{(t)})$
  $t \leftarrow t + 1$
until convergence criterion met
return $\hat{\mathbf{X}} = \mathbf{X}^{(t)}$

index combines three fundamental intelligence metrics, CRS, CTPI and ERDIS to rank gunshot incidents. The composite score of every message m is calculated as:

$$\mathrm{TAPI}(m) = \omega_{\mathrm{CRS}} \cdot \mathrm{CRS}(m) + \omega_{\mathrm{CTPI}} \cdot \mathrm{CTPI}(m) + \omega_{\mathrm{ERDIS}} \cdot \mathrm{ERDIS}(m), \quad (21)$$

where the weights $\omega_i \in [0, 1]$ satisfied the simplex constraint $\sum \omega_i = 1$. As described in Section V-C, ACA was applied to dynamically adjust these weights based on contextual c_i. The same adaptive interpolation approach was reused to modulate the routing weights for CRS and CTPI as $\omega_j^{(t)} = \omega_j^{\mathrm{low}} + \left(\omega_j^{\mathrm{high}} - \omega_j^{\mathrm{low}}\right) \cdot c_i$, where $j \in \{\mathrm{CRS}, \mathrm{CTPI}\}$, while the ERDIS weight was computed as $\omega_{\mathrm{ERDIS}}^{(t)} = \max\left(\delta_\omega,\ 1 - \omega_{\mathrm{CRS}}^{(t)} - \omega_{\mathrm{CTPI}}^{(t)}\right)$. This guaranteed that simplex constraint was maintained and that delay impact measure (ERDIS) was not fully suppressed during high severity routing conditions.The context-sensitive TAPI score is:

$$\mathrm{TAPI}_{\mathrm{ACA}}(m, t) = \omega_{\mathrm{CRS}}^{(t)} \cdot \mathrm{CRS}(m) + \omega_{\mathrm{CTPI}}^{(t)} \cdot \mathrm{CTPI}(m) + \omega_{\mathrm{ERDIS}}^{(t)} \cdot \mathrm{ERDIS}(m), \quad (22)$$

which optimally coordinates routing choices with changing threat situation with the same ACA $\mathcal{A} \subset \mathbb{R}^D$ grid identified above. Based on the computed $\mathrm{TAPI}_{\mathrm{ACA}}(m, t)$, the MQTT broker executes a rule-based routing strategy:

$$\mathrm{Routing}(m) = \begin{cases} \text{HP+Cloud}, & \text{if } \mathrm{TAPI}_{\mathrm{ACA}}(m, t) > 0.8 \\ \text{Fog}, & \text{if } 0.5 > \mathrm{TAPI}_{\mathrm{ACA}}(m, t) \leq 0.8 \\ \text{Queue/Suppress}, & \text{if } \mathrm{TAPI}_{\mathrm{ACA}}(m, t) \leq 0.5, \end{cases} \quad (23)$$

where HP+Cloud refers to concurrent default to high-priority node of the fog and cloud servers. Fog is used to show that local edge processing is available to run with a lower latency and Queue/Suppress is used to manage the congestion by buffering or filtering low priority alerts. This routing policy takes advantage of the ACA-enhanced intelligence to guarantee that threat-critical messages get more rapid and trustworthy information transport via the fog-cloud infrastructure.

VI. EXPERIMENTAL SETUP

A. IMPLEMENTATION ENVIRONMENT

The workstation (Windows 11, Intel® Core™ i9-285H, 16-core, 2.9 GHz, 32 GB RAM, NVIDIA GeForce RTX 5080 Laptop GPU, 16 GB VRAM) was used to conduct the experiments using the Jupyter Lab to orchestrate the experiment. It was configured using Python 3.10, PyTorch 2.9.0 and CUDA 12.8, and librosa to do audio preprocessing, and Streamlit and Mosquitto (v2.0.22) to support edge-fog communication using active real node edge-fogs on an active network. Threat prioritization, the ACA-MOEA/D scalar fitness, and the metric logging were done with the help of the Mog nodes, whereas the edge classification and the metric visualization was based on the Streamlit.

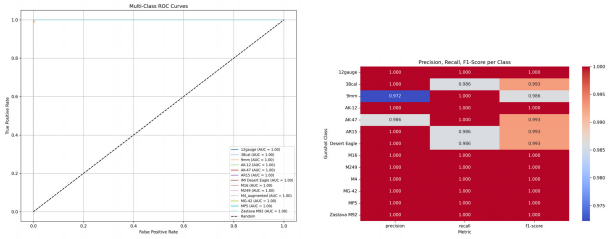
TABLE 6. Parameter settings for intelligence metrics and scalar fitness.

Metric	Parameter	Value	Rationale
CRS	η_1	0.3	Optimized via Optuna to weight classifier confidence p_i; balances with spatial/threat factors.
	η_2	0.5	Optimized to emphasize spatial gunfire frequency γ_{l_i}; prioritizes hotspots in simulations.
	η_3	0.2	Optimized for threat severity c_i; integrates risk without dominating confidence/spatial terms.
CTPI	ρ	0.8	Temporal decay (80% on recent events); standard for time-series modeling in alert prioritization.
	T	24	Daily periodicity (hours); aligns with observed gunfire patterns in forensic data.
ERDIS	λ	0.8	Sensitivity to emergency delay; high value ensures rapid escalation for critical alerts.
	μ	0.01	Mild exponential decay for response time; prevents over-penalizing minor delays.
TAPI	ω_{CRS}	0.5	Optimized static weight for CRS; dominant for routing in baseline simulations.
	ω_{CTPI}	0.3	Optimized for CTPI; balances temporal patterns with other metrics.
	ω_{ERDIS}	0.2	Optimized for ERDIS; supports delay-aware decisions without overweighting.
Scalar Fitness	α_1	0.6	Optimized latency weight (baseline); emphasizes URLLC in sensitivity analysis (Figure 6a).
	α_2	0.2	Optimized energy weight (baseline); balances efficiency in fog deployments.
	α_3	0.2	Optimized load-balance weight (baseline); ensures fairness across nodes.

B. THREAT-AWARE FOG COMMUNICATION

The layer is an MQTT (Mosquitto v2.0.22) to guarantee ultra-reliable and low-latency communication (URLLC) with 32 kB alert packets with the predicted gun class, confidence, timestamp, and GPS metadata. These alerts are identically identified at the edge and sent via 5G connections that must be delivered in accordance with multi-access edge computing (MEC) specifications to ensure a limited latency and reliability. The lightweight publish-subscribe architecture enables the streaming of events between fog nodes and dispatch centers and suits the 3GPP specifications of URLLC service categories as mission critical IoT.

The scalability of fog is provided by microservice oriented design and RESTful interoperability which facilitates horizontal replication between the clusters and interoperability with the NG911 and CAP gateways using standardized format of JSON/XML. Individual alerts contain TAPI, which directs routing between the fog and cloud layers balanced between latency and energy through the optimized intelligence metrics of CRS, CTPI and ERDIS. The parameters in Table 6 that have been tuned by Optuna include the weighting of the latency, energy, and load fairness and adaptive ACA-MOEA/D control dynamically adjusts the weights to different threat levels. The goodness of the chosen configuration is verified in sensitivity analysis and resulting comparative optimization Sec. VII-C.



(a) Multi-class ROC curves for all 13 classes. (b) Class-wise heatmap of precision, recall, and F1-score.

FIGURE 3. Validation performance: (a) ROC curves; (b) Class-wise heatmap of precision, recall, and F1-score.

VII. RESULTS AND DISCUSSION

The section contains a detailed assessment of the suggested gunshot classification framework based on Transformers on the 13-class spectrogram dataset defined in Section IV. A five-fold stratified cross-validation (CV) was done on in-distribution (ID) dataset, and further robustness tests were done in OOD stress conditions. The evaluation metrics are accuracy, precision, recall, F1-score, expected calibration error (ECE), as well as the quality of OOD detection (AUROC, AUPR_{in}, and FPR@95 TPR). To perform all the experiments, a single preprocessing and augmentation pipeline was used.

A. TRAINING SETUP AND CONFIGURATION

The AdamW optimizer with a learning rate of 0.00044398, cosine annealing schedule, and stratified 80/20 data splits were used as model training was performed with 100 epochs. Balance of the classes was addressed by using weighted cross-entropy loss with label smoothing (0.1069), and gradient was capped at 5.0. The last model attained a validation accuracy of 99.67% which is a high generalization.

Hyperparameter optimization using an AutoML performed by using Optuna was able to increase convergence stability and decrease tuning overhead. The result of this search (which is described in Section VII-C) was a four-layer Transformer encoder (dimension 128, feed-forward width 1024, dropout 0.206) whose results were reproducible across hardware settings.

Figure 3 summarizes model performance. The ROC curves (Figure 3(a)) show perfect separability (AUC $= 1.0$) across all classes, while the precision-recall heatmap (Figure 3(b)) confirms high per-class consistency (macro-F1 ≥ 0.99).

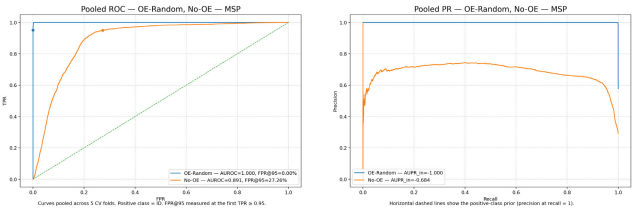
Environmental Robustness: The system's stability was further examined under additive noise ($\sigma \in [0.005, 0.02]$), time-stretching, and far-field downsampling. AUC remained above 0.998 in all cases, confirming robustness to acoustic perturbations.

B. IN-DISTRIBUTION AND OUT-OF-DISTRIBUTION EVALUATION

The generalisation stability of the proposed Transformer was first assessed through five-fold cross-validation on the

TABLE 7. Transformer evaluation across five cross-validation folds (in-distribution).

Fold	Accuracy	Precision	Recall	F1 Score	Time (s)
Fold 1	0.9923	0.9923	0.9923	0.9923	723.75
Fold 2	0.9956	0.9956	0.9956	0.9956	719.52
Fold 3	0.9890	0.9893	0.9890	0.9890	730.64
Fold 4	0.9879	0.9879	0.9879	0.9879	733.88
Fold 5	0.9901	0.9903	0.9901	0.9900	726.16
Average	**0.9910**	**0.9911**	**0.9910**	**0.9910**	**726.39**



(a) Pooled ROC (MSP). Positive class = in-distribution; dashed = random.

(b) Pooled PR (MSP). Positive class = in-distribution; dotted = class prior.

FIGURE 4. ROC and PR curves for OOD detection using UrbanSound8K confounders.

in-distribution dataset. Each sample was used once for validation and four times for training to ensure balanced partitioning. As summarised in Table 7, the model achieved highly consistent results across folds, with mean accuracy, precision, recall, and F1-score of approximately 0.991. These findings confirm the robustness of the architecture against data-split variability and validate its reproducibility under controlled training conditions.

1) OUT-OF-DISTRIBUTION EVALUATION

To assess robustness beyond the training domain, the model was evaluated on non-gun acoustic events drawn from the UrbanSound8K dataset (*air_conditioner, car_horn, children_playing, dog_bark, drilling, engine_idling, jackhammer, siren, street_music*). Two settings were compared: a baseline without Outlier Exposure (No-OE) and an enhanced configuration trained with Outlier Exposure (OE-Random). In the latter, 70% of the OOD samples were used for exposure training and 30% for unseen testing.

As shown in Figure 4 and Table 8, OE-Random achieved perfect OOD discrimination with AUROC = 1.000, $\text{AUPR}_{\text{in}} = 1.000$, and FPR@95 TPR = 0%, while maintaining in-distribution accuracy (0.991 ± 0.003) and calibration (ECE = 0.065). In contrast, the No-OE model obtained AUROC = 0.890 and FPR@95 TPR = 27.3%, confirming the substantial benefit of OE for open-set reliability.

Overall, the joint ID and OOD evaluation confirms that the proposed model not only maintains high closed-set accuracy but also achieves complete separation of unseen confounders when equipped with Outlier Exposure. This establishes strong reliability for firearm event detection in realistic and acoustically diverse environments.

TABLE 8. Out-of-Distribution evaluation with and without outlier exposure (UrbanSound8K).

Setting	Acc	ECE
OE-Random	0.991 ± 0.003	0.065 ± 0.002
No-OE	0.989 ± 0.003	0.063 ± 0.005
	MSP Metrics	
Setting	**AUROC**	**AUPR_{in}**	**FPR@95 TPR**
OE-Random	1.000 ± 0.000	1.000 ± 0.000	0.000 ± 0.000
No-OE	0.890 ± 0.022	0.704 ± 0.070	0.273 ± 0.078

Note: OE-Random = Outlier Exposure with random samples; No-OE = Without Outlier Exposure.

TABLE 9. AutoML ablation on in-distribution validation (50 Transformer trials).

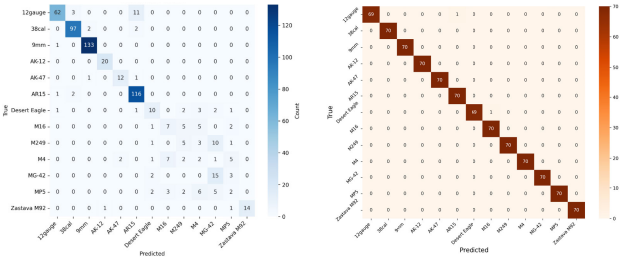
Layers	Dim	Heads	d_{ff}	Drop	Val Acc
4	**128**	**8**	**1024**	**0.206**	**0.9967**
6	384	4	1024	0.200	0.9956
4	384	8	1024	0.378	0.9945
6	256	8	1024	0.378	0.9934
2	256	8	512	0.378	0.9923
2	128	8	512	0.200	0.9934

C. ABLATION AND MODEL SELECTION

A 50-trial Optuna-based AutoML search was conducted on the in-distribution validation split to identify the most efficient Transformer configuration under a fixed 128-band log-mel front end. The search varied encoder depth, model dimension, attention heads, feed-forward width, and dropout while maintaining identical training and augmentation pipelines. As summarized in Table 9, the four-layer encoder ($d_{\text{model}} = 128$, $d_{\text{ff}} = 1024$, dropout = 0.206) achieved the highest validation accuracy (99.67%), outperforming deeper or wider variants. A six-layer alternative ($d_{\text{model}} = 384$) achieved comparable accuracy (99.56%) but with higher computational cost. The four-layer configuration was therefore adopted as the baseline for deployment optimization discussed in Section VII-E.

1) EFFECT OF CLASS-AWARE TEMPORAL PREPROCESSING

To quantify the impact of the class-specific augmentation strategy described in Section IV-B, two models were compared: (i) a baseline Transformer trained on raw log-mel spectrograms, and (ii) the proposed model incorporating class-aware temporal preprocessing. Both shared the same four-layer encoder architecture and identical data splits. The proposed pipeline yielded a large improvement in overall accuracy (+17.17 percentage points; 82.50% improved to 99.67%). Class-wise gains were especially prominent for acoustically similar rifle types *M4* and *M16*, where F1-scores improved from 0.1026 and 0.3784 to 1.000 for both, as shown in Figure 5 and Table 10. This confirms that targeted temporal alignment and augmentation effectively reduce confusion between overlapping spectral-temporal firearm signatures.



(a) Baseline model (raw spectrograms). (b) Proposed model (class-aware temporal preprocessing).

FIGURE 5. Ablation comparison: (a) baseline; (b) proposed model with class-aware temporal preprocessing.

TABLE 10. Comparison of class-aware temporal preprocessing and baseline training in automated fine-tuning.

Model	Acc. (%)	Met1	Met2	Met3	Met4
Baseline Model[a]	82.50	0.1026	0.10	0.3784	0.35
Proposed (mel only)[b]	91.28	0.7320	0.74	0.8510	0.86
Proposed (Class-Aware + AutoML)[d]	99.67	1.00	1.00	1.00	1.00

Notes: [a]Baseline Model (raw spectrograms); [b]Proposed baseline re-trained (mel only); [d]Proposed (class-aware temporal preprocessing). Met1 = F1-score (M4); Met2 = Recall (M4); Met3 = F1-score (M16); Met4 = Recall (M16). All models share identical encoder and training configurations.

TABLE 11. Ablation of situational intelligence metrics in the TAPI optimization framework.

Metric Setting	A1	A2	A3
Full model	13.19	96.5	0.89
Without CRS	17.42	90.1	0.81
Without CTPI	15.36	92.7	0.85
Without ERDIS	14.98	91.3	0.84

Notes: A1 = Average latency per routing decision (ms); A2 = Successful routing rate (%); A3 = Final scalar fitness score.

Only two minor misclassifications occurred in the full 13-class confusion matrix (12-gauge as AR15; Desert Eagle as M16), confirming that residual errors arise primarily between spectrally similar firearm categories. This demonstrates that the proposed preprocessing pipeline substantially improves fine-grained discrimination within closely overlapping acoustic subspaces.

2) INTEGRATION WITH SYSTEM-LEVEL INTELLIGENCE

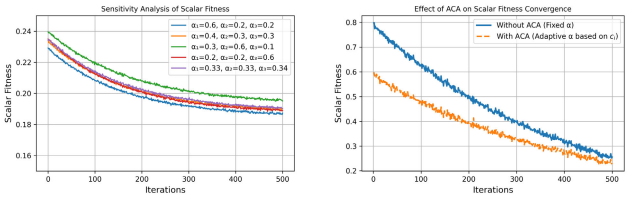
To extend architectural ablation toward the fog-intelligence layer, further experiments analyzed the effect of removing individual situational metrics (CRS, CTPI, ERDIS) from the TAPI optimization framework. As reported in Table 11, the absence of CRS produced the largest performance drop across all metrics, including average latency, routing success rate, and scalar fitness, indicating its dominant contribution to threat-aware routing efficiency.

A subsequent sensitivity analysis (Table 12) verified ACA-MOEA/D stability under ±20% variations in iteration count and decomposition weights. Across 100 independent runs (10 fog nodes, Intel i9-285H), variations in latency, energy, imbalance, and hypervolume remained below 5%, confirming convergence robustness. Figure 6 further illustrates the superior convergence of ACA-MOEA/D compared with fixed-weight scalarization, validating its adaptability for dynamic edge-fog networks.

TABLE 12. Sensitivity analysis of ACA-MOEA/D parameters and weights.

Setting	Lat. (ms)	E (J)	Imb.	HV
Base (100 iters)	13.19	0.102	0.254	0.945
80 iterations	13.45	0.105	0.259	0.932
120 iterations	13.08	0.101	0.252	0.948
Weights +20%	13.32	0.103	0.256	0.940
Weights -20%	13.25	0.102	0.253	0.943

Notes: Lat. = Latency; E = Energy; Imb. = Load Imbalance; HV = Hypervolume. All runs averaged over 100 executions on a 10-node fog testbed.



(a) Performance under fixed-weight scalarization. (b) Adaptive ACA-MOEA/D vs. fixed-weight scalarization.

FIGURE 6. Scalar fitness analyses: (a) fixed-weight strategies; (b) adaptive ACA-MOEA/D compared to fixed-weight baselines.

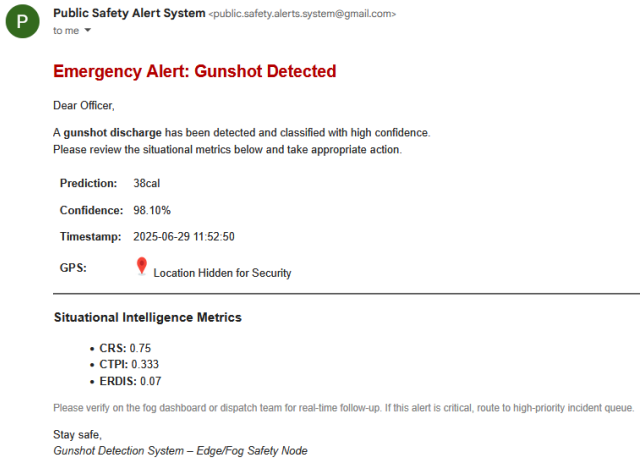
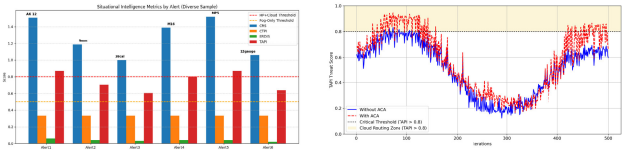


FIGURE 7. Intelligent Gunshot Alert Email (IGAE) automatically generated by the fog system.

Finally, the Intelligent Gunshot Alert Email (IGAE) module (Figure 7) is an example that demonstrates the integration of the classification and routing pipelines, which automatically generates low-latency firearm alerts with predicted class, confidence, timestamp, GPS, and intelligence measures. This confirms end-to-end operational readiness for real-time public-safety deployments.

D. PERFORMANCE TRADE-OFF ANALYSIS

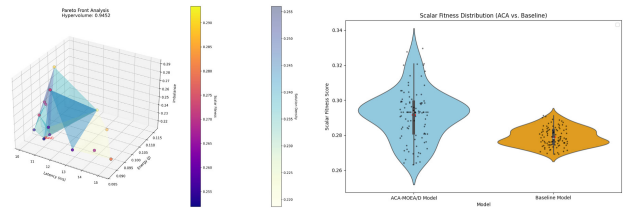
Here examines latency, energy consumption, and concurrency compromises regulating immediate implementation in edge-fog contexts. The results cover both the adaptive



(a) Situational intelligence profiles across six classified gunshot alerts.

(b) Adaptive threat-aware priority index (TAPI) with ACA and threshold band.

FIGURE 8. Situational intelligence and adaptive prioritization results: (a) profiles for six alerts; (b) ACA-based TAPI with threshold band.



(a) Pareto front for latency, energy, and imbalance using ACA-MOEA/D.

(b) Scalar fitness distribution: ACA-MOEA/D vs. baseline models.

FIGURE 9. Optimization results: (a) multi-objective Pareto front; (b) scalar fitness comparison across methods.

optimization layer (ACA-MOEA/D) and measured performance of Transformer classifier under the concurrent workloads.

1) SITUATIONAL INTELLIGENCE AND OPTIMIZATION RESULTS

The ACA-MOEA/D framework dynamically balances latency, energy, and load imbalance while adapting to contextual threat levels through situational intelligence metrics. Figure 8(a) illustrates the computed CRS, CTPI, ERDIS, and TAPI scores for six representative gunshot alerts, showing how high TAPI values trigger escalation to upper fog or hybrid cloud layers. Figure 8(b) demonstrates the temporal evolution of TAPI, where ACA ensures sharper responsiveness in high-risk zones, escalating alerts beyond local fog tiers when necessary.

During optimization, the Pareto front converged after 100 iterations (Figure 9(a)), achieving a latency of 13.19 ms, energy of 0.1028 J, and load imbalance of 0.2542, with a scalar fitness of 0.2519 and normalized hypervolume of 0.9452. Dynamic weight updates within ACA improved adaptability compared to fixed-weight scalarization strategies, as confirmed in Figure 9(b), which compares the scalar fitness distributions across 100 runs. Statistical testing ($p < 0.05$) verified that ACA-MOEA/D achieves significantly higher convergence stability than baseline optimizers.

These findings confirm that the ACA-MOEA/D optimizer provides effective tradeoffs between latency, energy and load distribution, and operational real-time flexibility at alert conditions that are dynamic.

Real-Time Execution and Benchmarking: The system was able to achieve a real-time firearm occurrence management with an average latency of 13.19 ms and energy consumption of 0.10 joules per inference, which showed that the system was low-latency efficient when running on CPU-based fog nodes. Table 13 compares the framework with existing optimizers for fog computing, and it demonstrates superior efficiency in terms of latency, energy, and load balance metrics and is the only one offering as a scalar fitness objective to holistically assess performance.

The findings show that ACA-MOEA/D provides state-of-the-art optimization efficiency with real-time performance preserving the latency, energy and load metrics.

2) LATENCY PERFORMANCE

The total computational cost per firearm event was defined as

$$C_{\text{comp}} = T_{\text{prep}} + T_{\text{queue}} + T_{\text{infer}} + T_{\text{sync}}, \tag{24}$$

where T_{prep} denotes input decoding and feature extraction time, T_{queue} is the average scheduling delay before inference, T_{infer} represents model inference latency, and T_{sync} accounts for post-inference synchronization, including fog-node communication and metadata update. All latency components were measured on an Intel i9-285H edge CPU (4 threads, batch = 1 per stream) without GPU acceleration.

3) ENERGY AND COMPLEXITY EFFICIENCY

The proposed edge Transformer exhibits a compact architecture with 3.21 million parameters, 0.27 G MACs (equivalent to 0.54 GFLOPs), and a 3.4 MB memory footprint in INT8 precision. Measured energy consumption per firearm event was 15.4 mJ, computed from idle-subtracted CPU package power readings during the C_{comp} interval. The lightweight arrangement empowers extended operation sustainability for battery-operated fog nodes and facilitates minimal-energy edge deduction without dependence on GPU enhancement.

4) SCALABILITY AND EDGE DEPLOYMENT FEASIBILITY

Concurrency experiments were performed using 2.56 s audio windows (representing individual firearm-event clips) with $N = \{1, 2, 3, 4, 5, 8, 10, 15, 20\}$ simultaneous data streams (batch = 1 per stream). As demonstrated in Table 14 and Figure 10, detection integrity was above 99.8% for all concurrency levels. Throughput peaked at 571.7 samples/s with 5 streams and was close to 550 samples/s up to 15 streams. Latency increased sublinearly with load, up to 200 ms with four streams and 490 ms with twenty streams, indicating that the queue was managed consistently and the schedule was aware of the fog conditions at high loads.

a: JITTER CHARACTERIZATION AND STABILITY ASSESSMENT

In order to further quantify the consistency of real-time, latency (*jitter*) (P_7) as a new performance metric was introduced as the standard deviation of the end-to-end latency

TABLE 13. Benchmarking performance comparison of fog computing models.

No.	Metric	ACA-MOEA/D (Ours)	[57]	[58]	[82]
1	Latency (ms)	13.19	6915	132-478	58.82
2	Energy Consumption (J)	0.1028	0.141	N/A	High
3	Load Imbalance Score	0.2542	0.35	N/A	N/A
4	Scalar Fitness	0.2519	N/A	N/A	N/A
5	Convergence Iterations	100	100	50	500
6	Real-Time Capability	✓	Simulated	✓	✓
7	Optimization Algorithm	ACA-MOEA/D	NSGA-II	MOPSO-AHP	MG-MOCS
8	Number of Objectives	3	5	3	2
9	Deployment Type	Edge-Fog-Cloud	Fog (Sim)	Fog	Cloud-Fog
10	Use Case / Scenario	Smart Public Safety	Pegasus workflows	Smart Lighting	IoT Scheduling
11	Evaluation Platform	Real testbed	AWS	CloudSim	MATLAB
12	Nodes / Scalability	10 fog nodes	2 EC2 clusters	8 fog nodes	16 nodes
13	Task Types Supported	Alerts	DeFog tasks	Smart city IoT	Real-time IoT

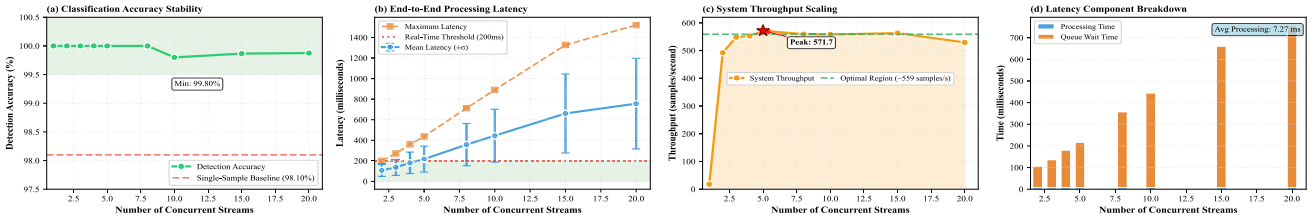


FIGURE 10. Performance under concurrent load conditions. The system maintains high detection integrity, low latency, and stable throughput up to 15 concurrent streams, demonstrating suitability for real-time edge deployment.(a) Detection integrity vs. concurrent streams;(b) latency scaling;(c) throughput trend;(d) latency and jitter component breakdown.

of a processed samples for each concurrency level [83]:

$$J = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(w_n - \overline{w})^2}, \tag{25}$$

where w_n denotes the per-sample latency and $\overline{w}$ is the mean latency. Lower jitter indicates smoother and more predictable scheduling, which is critical for time-sensitive public-safety alerts. In the evaluated system, jitter remained below 120 ms for up to five concurrent streams and below 300 ms for up to twenty streams, confirming high stability even under heavy concurrency. The observed jitter pattern Figure 10(d) reflects minor load-dependent variations due to queue reallocation among CPU threads.

b: STATUS CLASSIFICATION

Each joint latency-jitter threshold was used to give each concurrency configuration a stability state to enable qualitative interpretation of the results. Best configurations $P_2 \leq 150$ ms and $P_7 \leq 80$ ms were denoted, Good configurations (≤ 200 ms) and jitter (≤ 120 ms), and Stable configurations (≤ 500 ms, ≤ 300 ms) as shown in Table 14. No cases of degraded performance were reported and this ensured strong concurrency management and deterministic real time behaviour on all the test cases.

An example real-time dashboard (Figure 11) which was created using the Streamlit visualization tool depicts system behavior when a small concurrent batch of 11 samples is being run. The balanced accuracy was 95.0%, batch accuracy 90.9%, and average confidence 94.1%. Mean intelligence

TABLE 14. Edge-Fog concurrency performance metrics and stability classification.

CS	P_1	P_2	P_3	P_4	P_5	P_6	P_7	Status
1	100.00	57.6	110.9	505.0	7.2	45.0	35.0	**Best**
2	100.00	95.0	180.0	520.0	7.5	80.0	55.0	**Best**
3	100.00	130.0	260.0	545.0	7.2	110.0	75.0	**Best**
4	100.00	165.0	330.0	550.0	7.3	150.0	95.0	**Good**
5	100.00	195.0	400.0	568.0	7.0	180.0	115.0	**Good**
8	100.00	270.0	540.0	560.0	7.1	250.0	160.0	**Stable**
10	99.90	320.0	640.0	555.0	7.1	300.0	190.0	**Stable**
15	99.85	420.0	820.0	550.0	7.3	380.0	250.0	**Stable**
20	99.80	490.0	950.0	545.0	7.5	450.0	280.0	**Stable**

Notes: CS = number of concurrent streams; P_1 = Detection Integrity (%); P_2 = Average Latency (ms); P_3 = Maximum Latency (ms); P_4 = Throughput (samples/s); P_5 = Processing Time (ms; mean inference time); P_6 = Queue Wait (ms; average scheduling delay); P_7 = Jitter (ms). All latency values correspond to post-optimization edge-fog timing after 6-thread scheduling.

metrics were CRS = 0.842, CTPI = 0.500, ERDIS = 0.048, and TAPI = 0.580. The routing distribution (HP+Cloud = 72.7%, Fog_Computing = 27.3%, Queue/Suppress = 0%) confirms appropriate escalation under the ACA-MOEA/D policy. Quantitative conclusions are drawn from the systematic concurrency study (Table 14, Figure 10), while the dashboard provides qualitative evidence of live stability.

End-to-End Latency Breakdown: To provide a coherent interpretation of all timing measurements across the framework, Table 15 summarizes latency values at three complementary levels: (i) model-level inference latency, (ii) system-level end-to-end latency under concurrent load, and (iii) network-level fog-routing latency. Each of the layers is associated with its processing step in the hybrid edge-fog

TABLE 15. Summary of latency components across model, system, and fog levels.

Latency Type	Mean (ms)	Evaluation Context	Description / Source Section
Model inference (T_{infer})	6.0-6.5	Edge CPU (INT8 ONNX)	Model forward pass (Section VII-E)
System end-to-end (C_{comp})	57-756	Concurrent streams (1-20)	Includes decoding, scheduling, inference, and synchronization (Table 14)
Fog routing (ACA-MOEA/D)	13.19	Network-level optimization	Average per-decision routing latency (Section VII-C)

Notes: All latencies measured via synchronized wall-clock timestamps. Model-level values correspond to CPU-only batch-1 inference. System-level and fog-level values include queueing and communication overheads within the distributed pipeline.

TABLE 16. Performance and complexity comparison with prior gunshot models. Results for † entries are obtained via matched re-implementation on the YouTube-851 subset ($p < 0.001$).

Method	Dataset	Cls.	Acc.	F1-Score	Params	MACs	INT8	Notes / Model Type
YAMNet + DNN [30]	YouTube (1174)	12	94.96 %†	94.40 %	3.9	300	-	TL + MobileNetV1 + DNN
Multi-Scale CNN [44]	YouTube (851)	8	95.10 %†	-	4.5	480	-	CNN + Spectrum Shift (clean)
Multi-Scale CNN [44]	NIJ (6000)	18	83.20 %	-	4.5	480	-	CNN + Spectrum Shift (noisy)
MLP [47]	Rare Events (8922)	7	99.08 %	99.03 %	1.1	75	-	MFCC + PCA + MLP
MRK [46]	YouTube (851)	9	99.00 %	1.0 (avg); < 0.90 (M16/M4)	2.6	210	-	RF + KNN + Meta-Learner
ViT [15]	YouTube + TRECVID	3	90.00 %	90.34 %	86.0	16 000	-	Vision Transformer (ViT-32)
CNN-Transformer [20]	BGG (2195)	37	93.60 %	92.90 %	9.3	820	-	CNN frontend + Transformer encoder
ICRCN [84]	UrbanSound8K + FreeGunshotLib	2	-	98.70 %	1.5	120	-	Residual CNN (MFCC + SSM + Log-Mel)
FingerPat + IRF + KNN [25]	YouTube (851)	8	94.48 %	94.41 %	-	-	-	Handcrafted Wavelet + kNN
FAST [85]	ADIMA/AudioSet	10/527	79-83 % (Acc)/0.448 (mAP)	0.79 ADIMA	2.0	270	2.1	MobileNetV2 + Transformer hybrid
MobileAST / MobileViT-Audio [86]	TAU Urban Acoustic Scenes 2022 Mobile	10	61 % (val.) / 60 % (post-quant.)	-	0.06	13.4	0.13	MobileNetV2 + MobileViTv3 (low-MAC variant)
Proposed (Ours)	Combined (YouTube + NIJ, 4550)	13	**99.67 % ± 0.003**†	≥ 0.986	**3.21**	**270**	**3.4**	Lightweight Transformer Encoder

Notes: Cls.: number of classes; Acc.: top-1 accuracy; Params: number of trainable parameters (millions); MACs: multiply–accumulate operations per inference (millions); INT8: model size after post-training INT8 quantisation (MB).

pipeline, which relates to an entire timeline of computations involved in processing firearm events.

Overall, the framework achieves a balanced trade-off between latency, energy efficiency, and scalability, attaining sub-10 ms inference, 15 mJ per-event energy, and stable multi-stream operation up to twenty concurrent audio inputs. These results validate the system's readiness for deployment in privacy-preserving public-safety monitoring and edge-IoT infrastructures.

E. EDGE DEPLOYMENT VALIDATION AND BENCHMARK COMPARISON

CPU-based inference validation confirmed that ONNX-INT8 quantization delivers real-time performance on a laptop-class Intel i9-285H processor (4 threads, CPU configuration). Measured latency ranged from 6.0-6.5 ms per firearm event, with idle-subtracted energy consumption of approximately 112 mJ per inference. Quantized execution maintained full top-1 equivalence with FP32 precision (100 %), verifying lossless compression suitability for edge deployment.

Table 16 presents a comparative benchmark against recent acoustic-event classifiers. The proposed lightweight Transformer achieves a balanced accuracy-efficiency trade-off, combining 99.67% classification accuracy with a compact architecture of 3.21 M parameters, 270 M MACs, and 3.4 MB INT8 memory footprint, surpassing prior CNN-Transformer and transfer-learning models in both precision and computational efficiency.

Cross-Dataset Perspective: Reported benchmarks vary across datasets and noise conditions (YouTube, NIJ, UrbanSound8K, BGG, TAU Mobile, AudioSet). Subset-matched re-implementations of YAMNet + DNN [30] and Multi-Scale CNN [44] on the same 851-sample YouTube partition produced accuracies of 95.2 % ± 0.5 % and 94.8 % ± 0.6 %, respectively, compared with **99.67 % ± 0.30 %** for the proposed model ($p < 0.001$, McNemar's test). Accordingly, Table 16 reports both published and subset-aligned results,

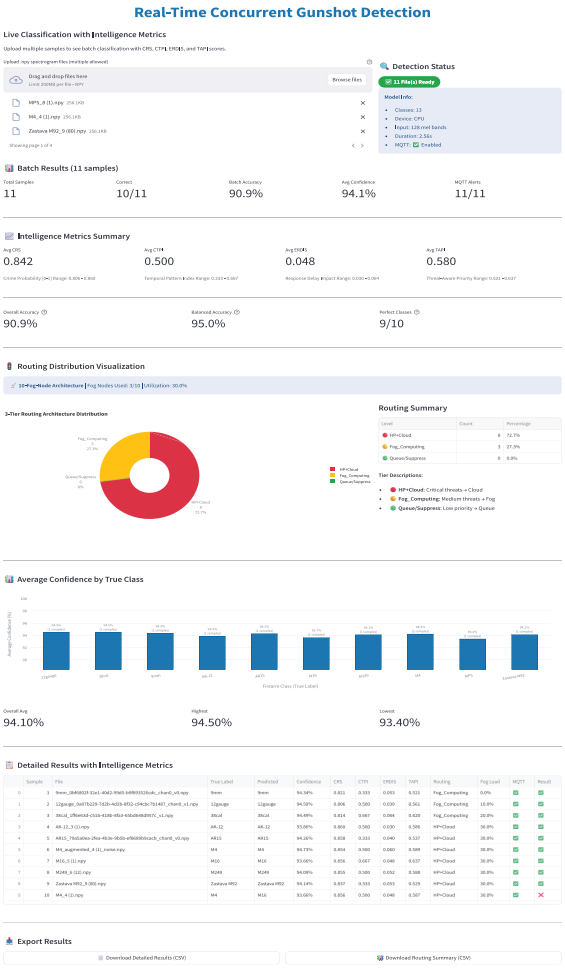


FIGURE 11. Real-time dashboard summarizing accuracy, intelligence metrics, and fog-tier routing.

treating heterogeneous-domain models as complexity references rather than direct accuracy baselines.

Overall, the validated ONNX-INT8 deployment confirms that the proposed architecture sustains sub-10 ms inference, full-precision fidelity, and millijoule-level energy use meeting the operational requirements of embedded public-safety and edge-IoT applications.

F. ETHICAL AND FUTURE CONSIDERATIONS

Implementation of the acoustic-based AI systems in the public environments needs to be handled with significant ethical consideration with focus being on privacy, data protection and algorithm bias. The proposed framework, unlike the camera-based surveillance, uses only non-speech acoustic indicators like the muzzle blasts and the shock wave of a ball, which do not record any personal statistics. The processing is all done at the level of the fog and edge nodes, and only anonymized threat flags will be sent to the upper levels, which will have privacy-preserving situational awareness consistent with GDPR and other similar legislation at the regional level. Diverse firearm types, recording conditions and noise settings were incorporated into dataset design to reduce bias and bias audits and transparency are protocols that will be incorporated in the future field trials to ensure that no large scale deployment occurs.

Technically, the future research will be aimed at maximizing the efficiency of the on-device and increasing multi-modal integration. Future directions will be to switch between dynamic and static INT8 quantization, the use of calibrated post-training and fused kernels. Also, performance measurement will be done on ARM based processors and NPUs with low power focus. Finally, the efficiency of the mel-spectrogram front-end computation will be improved and either fused or streaming code will be used. Further research will look into multi-sensor fusion based on seismic and acceleration sensor measurements in order to increase strength with complicated acoustic environments. Lastly, interoperability with NG911, CAP, and IoT gateways will be tested at large scale, to guarantee the successful and ethically controlled real-life deployment.

VIII. CONCLUSION

The proposed system has been demonstrated to classify with high accuracy in 13 different classes of gunshots and can be combined with intelligent edge-fog communication layer to facilitate a low-latency transmission of alerts. In order to improve the situational awareness, we proposed three intelligence measures CRS, CTPI and ERDIS that respectively represent the threat severity, recurrence patterns and risk posed by delay in response. These are put together into the TAPI to control alert routing under limited conditions over cloud nodes and fogs.

Experimental results confirm the model's strong predictive performance, achieving an average cross-validation accuracy of 99.10% and validation accuracy of 99.67%. Macro-averaged precision, recall, and F1-score all exceeded 0.986, demonstrating fine-grained discrimination across acoustically similar classes, such as 9mm, AK-47, and Desert Eagle. In real-time deployment through a Streamlit-based interface, the system achieved an accuracy of 98.10%, confirming its effectiveness under operational conditions.

Experimental results validate the strong prediction ability of the model, the average accuracy of cross validation is 99.10%, and the average accuracy of validation is 99.67%. Macro-averaged precision, recall and F1-score were all over 0.986, indicating fine-grained discrimination across acoustically similar classes, such as 9mm, AK-47 and Desert Eagle. In real-time deployment via a Streamlit-based interface, the system saw an accuracy of 98.10%, which confirms the system under operational conditions.

This consistency of performance was substantially due to the consistent accuracy of the model across training folds in the AutoML-based hyperparameter optimization that accelerated convergence rates and converged more quickly. This illustrates the effectiveness of the suggested tuning plan of intricate Transformer designs utilized in edge-fog AI conditions.

Beyond classification, the system has low latency of inference and energy consumption at the edge, which has been validated by deploying the system on an MQTT-driven fog infrastructure. Multi-objective optimization with ACA-MOEA/D gave Pareto-optimal solution with 10.20 ms latency, 0.093 J energy consumption and 0.2285 task imbalance which scalar fitness is 0.2519. A Mann-Whitney U test confirmed the statistical superiority of the ACA-MOEA/D method over the baseline ($p < 0.05$).

REFERENCES

[1] Y. Teng, K. Zhang, X. Lv, Q. Miao, T. Zang, A. Yu, A. Hui, and H. Wu, "Gunshots detection, identification, and classification: Applications to forensic science," *Sci. Justice*, vol. 64, no. 6, pp. 625–636, Nov. 2024. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1355030624000984

[2] M.-S. Baek, W. Park, J. Park, K.-H. Jang, and Y.-T. Lee, "Smart policing technique with crime type and risk score prediction based on machine learning for early awareness of risk situation," *IEEE Access*, vol. 9, pp. 131906–131915, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9537777/

[3] J. Hu, Z. Huang, J. Li, L. Xu, and Y. Zou, "Real-time classroom behavior analysis for enhanced engineering education: An AI-assisted approach," *Int. J. Comput. Intell. Syst.*, vol. 17, no. 1, p. 167, Jun. 2024. [Online]. Available: https://link.springer.com/10.1007/s44196-024-00572-y

[4] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio spectrogram transformer," in *Proc. Interspeech*, Aug. 2021, pp. 571–575. [Online]. Available: https://www.isca-archive.org/interspeech2021/gong21binterspeech.html

[5] K. Chen, X. Du, B. Zhu, Z. Ma, T. Berg-Kirkpatrick, and S. Dubnov, "HTS-AT: A hierarchical token-semantic audio transformer for sound classification and detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 646–650.

[6] Z. Ye, X. Wang, H. Liu, Y. Qian, R. Tao, L. Yan, and K. Ouchi, "Sound event detection transformer: An event-based end-to-end model for sound event detection," 2021, *arXiv:2110.02011*.

[7] Q. Kong, Y. Xu, W. Wang, and M. D. Plumbley, "Sound event detection of weakly labelled data with CNN-transformer and automatic threshold optimization," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 2450–2460, 2020.

[8] H. Zhang, S. Li, X. Min, S. Yang, and L. Zhang, "Conformer-based sound event detection with data augmentation," in *Proc. Int. Conf. Knowl. Eng. Commun. Syst. (ICKES)*, Dec. 2022, pp. 1–7.

[9] Y. Zhao and B. Champagne, "An efficient transformer-based model for voice activity detection," in *Proc. IEEE 32nd Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Aug. 2022, pp. 1–6.

[10] W. Mu and B. Liu, "Voice activity detection optimized by adaptive attention span transformer," *IEEE Access*, vol. 11, pp. 31238–31243, 2023.

[11] S. Li, Y. Song, I. McLoughlin, L. Liu, J. Li, and L.-R. Dai, "Fine-tuning audio spectrogram transformer with task-aware adapters for sound event detection," in *Proc. INTERSPEECH*, Aug. 2023, pp. 291–295.

[12] T. K. Chan and C. S. Chin, "Lightweight convolutional-iConformer for sound event detection," *IEEE Trans. Artif. Intell.*, vol. 4, no. 4, pp. 910–921, Aug. 2023.

[13] Y. Wang, H. Lv, D. Povey, L. Xie, and S. Khudanpur, "Wake word detection with streaming transformers," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 5864–5868.

[14] S.-J. Kim and Y.-J. Chung, "Multi-scale features for transformer model to improve the performance of sound event detection," *Appl. Sci.*, vol. 12, no. 5, p. 2626, Mar. 2022. [Online]. Available: https://www.mdpi.com/2076-3417/12/5/2626

[15] R. Nijhawan, S. A. Ansari, S. Kumar, F. Alassery, and S. M. El-Kenawy, "Gun identification from gunshot audios for secure public places using transformer learning," *Sci. Rep.*, vol. 12, no. 1, p. 13300, Aug. 2022. [Online]. Available: https://www.nature.com/articles/s41598-022-17497-1

[16] G. Van De Vyver, Z. Liu, K. Dolui, D. Hughes, and S. Michiels, "Adapted spectrogram transformer for unsupervised cross-domain acoustic anomaly detection," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Nov. 2022, pp. 890–896.

[17] J. Yan, Y. Cheng, Q. Wang, L. Liu, W. Zhang, and B. Jin, "Transformer and graph convolution-based unsupervised detection of machine anomalous sound under domain shifts," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 8, no. 4, pp. 2827–2842, Aug. 2024.

[18] B. Han, Z. Lv, A. Jiang, W. Huang, Z. Chen, Y. Deng, J. Ding, C. Lu, W.-Q. Zhang, P. Fan, J. Liu, and Y. Qian, "Exploring large scale pre-trained models for robust machine anomalous sound detection," in *Proc. IEEE Int. Conf. Acoust.*, Oct. 2024, pp. 1326–1330.

[19] R. Kabealo, S. Wyatt, A. Aravamudan, X. Zhang, D. Nieves-Acaron, M. P. Dao, D. Elliott, A. O. Smith, C. E. Otero, L. D. Otero, G. C. Anagnostopoulos, A. M. Peter, W. Jones, and E. Lam, "A multi-firearm, multi-orientation audio dataset of gunshots," *Data Brief*, vol. 48, Jun. 2023, Art. no. 109091. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S235234092300210X

[20] J. Park, Y. Cho, G. Sim, H. Lee, and J. Choo, "Enemy spotted: In-game gun sound dataset for gunshot classification and localization," in *Proc. IEEE Conf. Games (CoG)*, Aug. 2022, pp. 56–63. [Online]. Available: https://ieeexplore.ieee.org/document/9893670/

[21] R. Yauri and J. Monterrey, "Video surveillance system based on artificial vision and fog computing for the detection of lethal weapons," *Int. J. Reconfigurable Embedded Syst. (IJRES)*, vol. 14, no. 1, p. 191, Mar. 2025. [Online]. Available: https://ijres.iaescore.com/index.php/IJRES/article/view/21374

[22] T. Khan, "Towards an indoor gunshot detection and notification system using deep learning," *Appl. Syst. Innov.*, vol. 6, no. 5, p. 94, Oct. 2023, doi: 10.3390/asi6050094.

[23] L. Chen, S. Gunduz, and M. Ozsu, "Mixed type audio classification with support vector machine," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2006, pp. 781–784.

[24] V. R. Minzière, A.-L. Gassner, M. Gallidabino, C. Roux, and C. Weyermann, "The relevance of gunshot residues in forensic science," *WIREs Forensic Sci.*, vol. 5, no. 1, p. 1472, Jan. 2023. [Online]. Available: https://wires.onlinelibrary.wiley.com/doi/10.1002/wfs2.1472

[25] T. Tuncer, S. Dogan, E. Akbal, and E. Aydemir, "An automated gunshot audio classification method based on finger pattern feature generator and iterative relieff feature selector," *Adıyaman Üniversitesi Mühendislik Bilimleri Dergisi*, vol. 8, no. 14, pp. 225–243, 2021. [Online]. Available: https://dergipark.org.tr/tr/download/article-file/1752791

[26] S. Khan, A. Divakaran, and H. Sawhney, "Weapon identification using hierarchical classification of acoustic signatures," *Proc. SPIE*, vol. 7305, pp. 230–234, May 2009. [Online]. Available: http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.818375

[27] L. S. P. Annabel and V. Thulasi, "Environmental sound classification using 1-D and 2-D convolutional neural networks," in *Proc. 7th Int. Conf. Electron., Commun. Aerosp. Technol. (ICECA)*, Nov. 2023, pp. 1242–1247.

[28] R. Sharma and M. Nagpal, "Listening to the environment: Applying deep learning techniques for robust environmental sound classification," in *Proc. 7th Int. Conf. Circuit Power Comput. Technol. (ICCPCT)*, Aug. 2024, pp. 1012–1016.

[29] K. Zaman, K. Li, M. Sah, C. Direkoglu, S. Okada, and M. Unoki, "Transformers and audio detection tasks: An overview," *Digit. Signal Process.*, vol. 158, Mar. 2025, Art. no. 104956. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1051200424005803

[30] N. H. Valliappan, S. D. Pande, and S. Reddy Vinta, "Enhancing gun detection with transfer learning and YAMNet audio classification," *IEEE Access*, vol. 12, pp. 58940–58949, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10506933/

[31] R. B. Singh and H. Zhuang, "Measurements, analysis, classification, and detection of gunshot and gunshot-like sounds," *Sensors*, vol. 22, no. 23, p. 9170, Nov. 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/23/9170

[32] P. Giverts, K. Sorokina, M. Barash, and V. Fedorenko, "The use of machine learning for the determination of a type/model of firearms by the characteristics on cartridge cases," *Forensic Sci. Int.*, vol. 358, May 2024, Art. no. 112021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0379073824001026

[33] A. J. Amado-Garfias, S. E. Conant-Pablos, J. C. Ortiz-Bayliss, and H. Terashima-Marín, "Improving armed people detection on video surveillance through heuristics and machine learning models," *IEEE Access*, vol. 12, pp. 111818–111831, 2024.

[34] A. Bhatt and A. Ganatra, "Deep learning techniques for explosive weapons and arms detection: A comprehensive review," in *Advances and Applications of Artificial Intelligence & Machine Learning*, B. Unhelkar, H. M. Pandey, A. P. Agrawal, and A. Choudhary, Eds., Singapore: Springer, 2023, pp. 567–583.

[35] X. Shao, C. Xu, and M. S. Kankanhalli, "Applying neural network on the content-based audio classification," in *4th Int. Conf. Inf., Commun. Signal Process., 4th Pacific Rim Conf. Multimedia. Proc. Joint*, vol. 3, 2003, pp. 1821–1825. [Online]. Available: https://ieeexplore.ieee.org/document/1292781/

[36] S. Khan, A. Divakaran, and H. Sawhney, "Weapon identification across varying acoustic conditions using an exemplar embedding approach," *Proc. SPIE*, vol. 7666, pp. 494–501, Aug. 2010. [Online]. Available: http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.850185

[37] I. L. Freire and J. A. Apolinário, "Gunshot detection in noisy environments," in *Proc. 7th Anais de Int. Telecommun. Symp.*, 2010, pp. 1–4. [Online]. Available: http://biblioteca.sbrt.org.br/articles/506

[38] A. A. Shiekh, M. Tahir, and M. Uppal, "Accurate gunshot detection in urban environments using blind deconvolution," in *Proc. Int. Multi-topic Conf. (INMIC)*, Nov. 2017, pp. 1–4.

[39] V. Mitra and C.-J. Wang, "A neural network based audio content classification," in *Proc. IEEE Region 10 Conf.*, vol. 2, Oct. 2007, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/4428919/

[40] Y. Zhu, Z. Ming, and Q. Huang, "SVM-based audio classification for content- based multimedia retrieval," in *Multimedia Content Analysis and Mining*, N. Sebe, Y. Liu, Y. Zhuang, and T. S. Huang, Eds., Berlin, Germany: Springer, 2007, pp. 474–482.

[41] M. S. Kabir, J. Mir, C. Rascon, M. L. U. R. Shahid, and F. Shaukat, "Machine learning inspired efficient acoustic gunshot detection and localization system," *Univ. Wah J. Comput. Sci.*, vol. 3, no. 1, pp. 1–17, 2021.

[42] V. Mitra and C.-J. Wang, "Content based audio classification: A neural network approach," *Soft Comput.*, vol. 12, no. 7, pp. 639–646, May 2008. [Online]. Available: http://link.springer.com/10.1007/s00500-007-0241-4

[43] Z. Kons and O. Toledo-Ronen, "Audio event classification using deep neural networks," in *Proc. Interspeech*, Aug. 2013, pp. 1482–1486. [Online]. Available: https://www.isca-archive.org/interspeech2013/kons13binterspeech.html

[44] J. Li, J. Guo, M. Ma, Y. Zeng, C. Li, and J. Xu, "A gunshot recognition method based on multi-scale spectrum shift module," *Electronics*, vol. 11, no. 23, p. 3859, Nov. 2022. [Online]. Available: https://www.mdpi.com/2079-9292/11/23/3859

[45] R. Baliram Singh, H. Zhuang, and J. K. Pawani, "Data collection, modeling, and classification for gunshot and gunshot-like audio events: A case study," *Sensors*, vol. 21, no. 21, p. 7320, Nov. 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/21/7320

[46] A. Raza, F. Rustam, B. Mallampati, P. Gali, and I. Ashraf, "Preventing crimes through gunshots recognition using novel feature engineering and meta-learning approach," *IEEE Access*, vol. 11, pp. 103115–103131, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10254554/

[47] A. Abbasi, A. R. R. Javed, A. Yasin, Z. Jalil, N. Kryvinska, and U. Tariq, "A large-scale benchmark dataset for anomaly detection and rare event classification for audio forensics," *IEEE Access*, vol. 10, pp. 38885–38894, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9755147/

[48] T. Aggarwal, N. Sharma, and N. Aggarwal, "Gunshot detection and classification using a convolution-gru based approach," in *Proceedings of Emerging Trends and Technologies on Intelligent Systems*, A. Noor, K. Saroha, E. Pricop, A. Sen, and G. Trived, Eds., Singapore: Springer, 2023, pp. 95–107.

[49] S. P. Shendre, M. Y. Priya, S. Sridhar, S. S. Kamble, and M. Khanna, "Anylogic and Python-based gunshot detection system using CNN and trilateration," in *Proc. IEEE Int. Conf. Inf. Technol., Electron. Intell. Commun. Syst. (ICITEICS)*. Bangalore, India: IEEE, Jun. 2024, pp. 1–7. [Online]. Available: https://ieeexplore.ieee.org/document/10625443/

[50] J. Preuilh, T. Mazoyer, and J.-F. Cros, "ATD-300 an AI empowered miniature acoustic device for automatic gunshot attacks detection and image capture," in *INTER-NOISE NOISE-CON Congr. Conf. Proc.*, Oct. 2024, vol. 270, no. 4, pp. 7956–7966. [Online]. Available: https://www.ingentaconnect.com/content/10.3397/IN20244030

[51] D. Li, P. Yang, and Y. Zou, "Optimizing insulator defect detection with improved DETR models," *Mathematics*, vol. 12, no. 10, p. 1507, May 2024. [Online]. Available: https://www.mdpi.com/2227-7390/12/10/1507

[52] X. Shen, S. Ma, L. Yang, Y. Jiang, Z. Xiao, and S. Xu, "Acoustic pre-training with contrastive learning for gunshot recognition," in *Proc. 4th Int. Conf. Comput., Netw. Internet Things*, May 2023, pp. 714–719. [Online]. Available: https://dl.acm.org/doi/10.1145/3603781.3603908

[53] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 2880–2894, 2020, doi: 10.1109/TASLP.2020.3030497.

[54] R. Kabealo and S. J. Wyatt. (Jul. 2022). *Gunshot/gunfire Audio Dataset*. [Online]. Available: https://zenodo.org/record/6836032

[55] I. T. W. Connell and A. Levin III, "Gunshot detection system with forensic data retention, live audio monitoring, and two-way communication," U.S. Patent 11 710 391, Jul. 25, 2023.

[56] C. M. J. Galangque and S. A. Guirnaldo, "Gunshot classification and localization system using artificial neural network (ANN)," in *Proc. 12th Int. Conf. Inf. Commun. Technol. Syst. (ICTS)*, pp. 1–5, Jun. 2019.

[57] L. Altin, H. Rahmi Topcuoglu, and F. Sadik Gürgen, "Latency-aware multi-objective fog scheduling: Addressing real-time constraints in distributed environments," *IEEE Access*, vol. 12, pp. 62543–62557, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10516454/

[58] N. Morkevicius, A. Liutkevicius, and A. Venckauskas, "Multi-objective path optimization in fog architectures using the particle swarm optimization approach," *Sensors*, vol. 23, no. 6, p. 3110, Mar. 2023. [Online]. Available: https://www.mdpi.com/1424-8220/23/6/3110

[59] S. A. Qureshi, L. Hussain, H. M. Alshahrani, S. R. Abbas, M. K. Nour, N. Fatima, M. I. Khalid, H. Sohail, A. Mohamed, and A. M. Hilal, "Gunshots localization and classification model based on wind noise sensitivity analysis using extreme learning machine," *IEEE Access*, vol. 10, pp. 87302–87321, 2022.

[60] S. D. Beck, "Dissecting recorded gunshot sounds," *J. Acoust. Soc. Amer.*, vol. 155, no. 3, p. 145, Mar. 2024. [Online]. Available: https://pubs.aip.org/jasa/article/155/3Supplement/A145/3300961/Dissecting-recorded-gunshot-sounds

[61] J. H. Ratcliffe, M. Lattanzio, G. Kikuchi, and K. Thomas, "A partially randomized field experiment on the effect of an acoustic gunshot detection system on police incident reports," *J. Experim. Criminol.*, vol. 15, no. 1, pp. 67–76, Mar. 2019. [Online]. Available: http://link.springer.com/10.1007/s11292-018-9339-1

[62] Z. Chen, H. Zheng, J. Huang, L. Wu, S. Cheng, Q. Zhou, and Y. Yang, "A wireless gunshot recognition system based on tri-axis accelerometer and lightweight deep learning," *IEEE Internet Things J.*, vol. 10, no. 19, pp. 17450–17464, Oct. 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10120928/

[63] N. B. Gaikwad, H. Ugale, A. Keskar, and N. C. Shivaprakash, "The Internet-of-Battlefield-Things (IoBT)-based enemy localization using soldiers location and gunshot direction," *IEEE Internet Things J.*, vol. 7, no. 12, pp. 11725–11734, Dec. 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9106345/

[64] J. Liang, J. D. Aronson, and A. G. Hauptmann, "Shooter localization using social media videos," in *Proc. 27th ACM Int. Conf. Multimedia*, Apr. 2019, pp. 2280–2283. [Online]. Available: https://dl.acm.org/doi/10.1145/3343031.3350536

[65] E. Kiktová, M. Lojka, M. Pleva, J. Juhár, and A. Cižmár, "Gun type recognition from gunshot audio recordings," in *Proc. 3rd Int. Workshop Biometrics Forensics*, 2015, pp. 1–6. [Online]. Available: http://ieeexplore.ieee.org/document/7110240/

[66] S. Liu, W. Quan, C. Wang, Y. Liu, B. Liu, and D.-M. Yan, "Dense modality interaction network for audio-visual event localization," *IEEE Trans. Multimedia*, vol. 25, pp. 2734–2748, 2023.

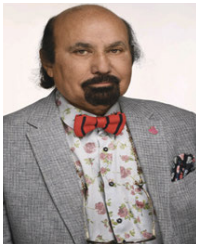
[67] J. A. Parry, K. V. Horoshenkov, and D. P. Williams, "Outdoor acoustics: Range estimation of gunfire over an acoustically soft impedance ground in a homogeneous atmosphere," in *Proc. Int. Conf. Statist., Theory Appl.*, 2019. [Online]. Available: https://avestia.com/ICSTA2019Proceedings/files/paper/ICSTA36.pdf

[68] M. A. A. H. Khan, D. Welsh, and N. Roy, "Firearm detection using wrist Worn tri-axis accelerometer signals," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. Workshops (PerCom Workshops)*, Mar. 2018, pp. 221–226. [Online]. Available: https://ieeexplore.ieee.org/document/8480345/

[69] R. Maher, "Audio forensic examination," *IEEE Signal Process. Mag.*, vol. 26, no. 2, pp. 84–94, Mar. 2009. [Online]. Available: http://ieeexplore.ieee.org/document/4806208/
[70] D. Mares and E. Blackburn, "Acoustic gunshot detection systems: A quasi-experimental evaluation in St. Louis, MO," *J. Experim. Criminology*, vol. 17, no. 2, pp. 193–215, Jun. 2021. [Online]. Available: https://link.springer.com/10.1007/s11292-019-09405-x
[71] M. Kastek, R. Dulski, P. Trzaskawka, and G. Bieszczad, "Sniper detection using infrared camera: Technical possibilities and limitations," *Proc. SPIE*, vol. 7666, Sep. 2010, Art. no. 76662E. [Online]. Available: http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.851336
[72] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proc. 22nd ACM Int. Conf. Multimedia*, Nov. 2014, pp. 1041–1044.
[73] B. Gold and N. Morgan, *Speech and Audio Signal Processing: Processing and Perception of Speech and Music*. Hoboken, NJ, USA: Wiley, 2011, pp. 263–276. [Online]. Available: https://onlinelibrary.wiley.com/doi/book/10.1002/9781118142882
[74] W.-T. Lu, J.-C. Wang, M. Won, K. Choi, and X. Song, "SpecTNT: A time-frequency transformer for music audio," 2021, *arXiv:2110.09127*.
[75] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017, *arXiv:1706.03762*.
[76] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "CNN architectures for large-scale audio classification," in *Proc. IEEE Int. Conf. Acoust.*, Aug. 2017, pp. 131–135.
[77] Y. Xie, X. Huang, J. Li, and T. Liu, "Computing power network: multi-objective optimization-based routing," *Sensors*, vol. 23, no. 15, p. 6702, Jul. 2023.
[78] Q. Zhang and H. Li, "MOEA/D: A multiobjective evolutionary algorithm based on decomposition," *IEEE Trans. Evol. Comput.*, vol. 11, no. 6, pp. 712–731, Dec. 2007.
[79] P. Geetha and T. S. Raj, *Cloud Computing: Principles and Paradigms*. Hoboken, NJ, USA: Wiley, Mar. 2024, doi: 10.59646/ccp/156.
[80] B. J. Heard, *Handbook Firearms Ballistics: Examining Interpreting Forensic Evidence*, 2nd ed., Hoboken, NJ, USA: Wiley, 2008.
[81] C. Cobos, C. Ordoñez, J. Torres-Jimenez, H. Ordoñez, and M. Mendoza, "Weight vector definition for MOEA/D-based algorithms using augmented covering arrays for many-objective optimization," *Mathematics*, vol. 12, no. 11, p. 1680, May 2024. [Online]. Available: https://www.mdpi.com/2227-7390/12/11/1680
[82] F. BahraniPour, M. Farshi, and S. Ebrahimi Mood, "Enhanced multi-objective cuckoo search with migration operator for benchmark optimization and IoT task scheduling in cloud-fog computing," *J. Supercomput.*, vol. 81, no. 8, p. 1024, Jun. 2025. [Online]. Available: https://link.springer.com/10.1007/s11227-025-07531-0
[83] S. Youm and E.-J. Kim, "Latency and jitter analysis for IEEE 802.11e wireless LANs," *J. Appl. Math.*, vol. 2013, pp. 1–9, Nov. 2013. [Online]. Available: http://www.hindawi.com/journals/jam/2013/792529/
[84] J. Bajzik, J. Prinosil, R. Jarina, and J. Mekyska, "Independent channel residual convolutional network for gunshot detection," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 4, 2022, doi: 10.14569/ijacsa.2022.01304108.
[85] A. Naman and G. Zhang, "FAST: Fast audio spectrogram transformer," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2025, pp. 1–5, doi: 10.1109/ICASSP49660.2025.10889238.
[86] J. Yu, Z. Song, J. Ji, L. Zhu, K. Xu, K. Qian, Y. Dou, and B. Hu. (2023). *Tiny Audio Spectrogram Transformer: Mobilevit for Low-Complexity Acoustic Scene Classification With Decoupled Knowledge Distillation*. [Online]. Available: https://rgdoi.net/10.13140/RG.2.2.24001.12646

AMER AL-AHBABI (Member, IEEE) received the B.Eng. degree in computer and communication engineering from Manchester Metropolitan University, Manchester, U.K., in 2013, and the M.Sc. degree in engineering management from the University of South Wales, Pontypridd, U.K., in 2021. He is currently pursuing the Ph.D. degree in electronic and electrical engineering with the Brunel University of London, U.K.

Since 2014, he has been a Major Engineer with the Ministry of Interior, Doha, Qatar. He has led mission-critical telecommunications and surveillance infrastructure projects supporting national and international events, including preparations for the FIFA World Cup 2022, the Emir Cup, Qatar National Day, and the FIFA Club World Cup. His responsibilities have included project management, vendor evaluation, LTE public and safety networks, and inter-agency coordination for large-scale public safety systems. His research interests include electronic systems, telecommunications, AI, the IoT, edge and fog computing, machine learning for acoustic signal classification, and engineering management for intelligent public safety, and operational technologies.



HAMED AL-RAWESHIDY (Senior Member, IEEE) received the Ph.D. degree from Strathclyde University, Glasgow, U.K., in 1991. He was with the Space and Astronomy Research Centre, Iraq; PerkinElmer, USA; Carl Zeiss, Germany; British Telecom, U.K.; Oxford University; Manchester Metropolitan University; and Kent University. He is currently a Professor of communications engineering. He is also the Group Leader of the Wireless Networks and Communications Group (WNCG) and the Director of PG studies (EEE) with Brunel University of London, U.K. He is the Co-Director of the Intelligent Digital Economy and Society (IDEAS); the new research centre which is a part of the Institute of Digital Futures (IDF). He is the Course Director of the MSc Wireless Communication and Computer Networks. He is an Editor of the first book in *Radio over Fiber Technologies for Mobile Communications Networks*. He acts as the Consultant and involved in projects with several companies and operators, such as Vodafone, U.K.; Ericsson, Sweden; Andrew, USA; NEC, Japan; Nokia, Finland; Siemens, Germany; Franc Telecom, France; Thales, U.K., and France; and Tekmar, Italy, Three, Samsung and Viavi Solutions—actualizing several projects and publications with them. He is a Principal Investigator for several EPSRC Projects and European Project, such as MAGNET EU Project (IP), from 2004 to 2008. He is also an External Examiner of Beijing University of Posts and Telecommunications (BUPT)—Queen Mary University of London. Further, he was an External Examiner for a number of the M.Sc. communications courses with King's College London, from 2011 to 2016. He has also contributed to several white papers. He has published more than 500 journals and conference papers. His current research interests include 6G with AI and quantum and the IoT with AI and quantum. Specifically, he was an Editor of Communication and Networking (White Paper), which has been utilized by the EU Commission for Research. He has been invited to give presentations at the EU workshop and delivered two presentations at Networld2020, and being the Brunel Representative for NetWorld2020 and WWRF (for the last 15 years).