# AI-Based Mobile Application for Real-Time Criminal Event Recognition and Classification

**Rafael Cachique**
Faculty of Information Systems Engineering, Universidad Peruana de Ciencias Aplicadas, Lima, Peru
u20211D147@upc.edu.pe

**Diego Gutierrez**
Faculty of Information Systems Engineering, Universidad Peruana de Ciencias Aplicadas, Lima, Peru
u202217814@upc.edu.pe

**Pedro Castaneda**
Faculty of Information Systems Engineering, Universidad Peruana de Ciencias Aplicadas, San Isidro, Lima, Peru
pcsipcas@upc.edu.pe

**Sandra Wong-Durand**
Faculty of Information Systems Engineering, Universidad Peruana de Ciencias Aplicadas, San Isidro, Lima, Peru
pcsiswon@upc.edu.pe (corresponding author)

**Roberto Carlos Santa Cruz Acosta**
Faculty of Systems Engineering and Electrical Mechanics, Universidad Nacional Toribio Rodriguez de Mendoza, Amazonas, Peru
roberto.santacruz@untrm.edu.pe

**Alberto Daniel Garcia-Nunez**
Universidad Pontificia Bolivariana, Medellin, Antioquia, Colombia
alberto.garcia@upb.edu.co

*Received: 30 January 2026 | Revised: 13 March 2026, 7 April 2026, 20 April 2026, and 22 April 2026 | Accepted: 23 April 2026*

 *DOI: https://doi.org/10.48084/etasr.17841*

**ABSTRACT**

**This paper proposes an Artificial Intelligence (AI)-based mobile application that identifies and categorizes crime and emergency events utilizing an acoustic and linguistic signals mechanism. The system combines Automatic Speech Recognition (ASR) and Natural Language Processing (NLP) models with sound event detection architectures to process emergency call recordings and identify incidents such as robberies, assaults, and riots. Data training consisted of preprocessing more than 530 audio recordings of emergency calls made by the Peruvian National Police (PNP), in anonymized formats, and training deep learning models based on convolutional and transformer-based networks. The performance was evaluated using metrics such as precision, accuracy, recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC). The multimodal fusion model achieved high performance, demonstrating an accuracy of 86.79% and an AUC-ROC of 87.14%, strongly distinguishing between emergency and non-emergency conditions in various fields and noisy environments. The results of the study show that the proposed solution is highly reliable and responsive, presenting an opportunity to enhance urban security and inform public safety system decision-making.**

***Keywords-Artificial Intelligence (AI); sound event detection; Natural Language Processing (NLP); deep learning; public safety; real-time classification; emergency calls; acoustic analysis; mobile application***

## I. INTRODUCTION

The metropolis of Lima faces a chronic rise in crime [1, 2], and limited institutional capacity for response. It still relies on phoning and manual reporting which creates high-priority delays in triggering security protocols [3].

Complicating these problems, the number of outlets for reports is overloaded. There is little coordination among the entities concerned, and citizens do not trust law enforcement to report incidents effectively [4, 5]. Delayed responses to incidents result in material losses, as well as reduced social stability [6, 7].

To address this security problem, the use of new technologies, such as Natural Language Processing (NLP) and Automatic Speech Recognition (ASR), provides new approaches for real-time tracking and response systems in order to optimize efficiency and improve public trust [8].

Existing security measures, built on video surveillance, manual reports, and centralized calls, have severe drawbacks: they are human-dependent, response time-consuming, and their ability to cope with significant amounts of information is not adequate [8]. Such systems cannot detect keywords or emotions that signal risk in real time and have low interoperability among institutions, which limits their effectiveness [9].

To contextualize the issue within the current state of research, recent studies have been conducted on sound detection, surveillance methods, and security applications integrated with NLP. These studies show progress and highlight the relevance of integrating Artificial Intelligence (AI) techniques into real-time crime detection systems.

Authors in [10] applied sound event detection using Convolutional Neural Networks (CNN) to tackle the issue of low fidelity in noisy environments. Using a dataset of 10,000 synthesized samples from urban contexts (alarms, traffic, sirens), they achieved an F1-score of 77.67%. This highlights the effectiveness of spatial filters in processing acoustic signals in public transportation environments. Authors in [11] addressed the limited robustness of individual models by employing an ensemble of CNNs, training on a public acoustic corpus with 10,000 clips and obtaining an F1-score of 71.5%. Authors in [12] introduced a dual attention module to enhance the simultaneous detection and localization of acoustic events in the TAU Urban Acoustic Scenes database, yielding a Localization Recall (LRCD) of 80.2% and a segment-based F1-score of 74.0%. Authors in [13] proposed the Self-Attention Layer within a CNN (SACNN) model for emergency siren recognition using a large-scale dataset of sirens and road noise, achieving 100% and an F1-score of 1.00, while significantly reducing inference time to 0.20883 s. Finally, authors in [14] proposed a meta-learning approach for gunshot recognition using a dataset of 851 audio samples and a novel feature engineering method, achieving a K-fold accuracy score of 0.99 and an average precision of 0.97.

Extending the scope from sound detection to public surveillance, authors in [15] proposed a CNN model to measure crowd density, using a dataset of 20,000 security camera images, to predict patterns that traditional models cannot capture. Similarly, authors in [16] applied audio-visual representation learning to detect anomalies in crowds using the SHADE dataset, which contains 2,149 videos. Their proposed AVRL model achieved a Top-1 accuracy of 95.9% for gunshot detection and reached 100% accuracy in categories like 'Arrest' and 'Knockdown' under dark scene conditions. Likewise, authors in [17] developed a Voice-Activated Face Recognition (VAFR) framework for crowd surveillance, using the Viola-Jones algorithm and Conformer architecture. Validated with a dataset of 20 individuals (510 test samples), the system achieved an accuracy rate of 99.8% in live video feeds, demonstrating high reliability under various lighting conditions and viewing angles. Similarly, authors in [18] developed an autonomous UAV system for facial recognition and tracking in GPS-denied environments. Using a Siamese network combined with their proposed Simple Matching Real-Time Tracking (SMRT) algorithm, they achieved a self-reported accuracy of 99.32% during real-time indoor flight tests, effectively maintaining identity consistency across frames. Finally, authors in [19] proposed a multimodal abnormal event detection framework for public transportation using RGB, depth, and audio signals. Tested on a custom dataset of in-cabin incidents, their deep learning architecture achieved a total accuracy of 85.1% in controlled experiments and reached 89% accuracy during on-site evaluations in autonomous minibuses. This line of research demonstrates the importance of multimodal integration and the deployment of drone-enabled intelligent surveillance for modern cities.

A number of studies have focused on NLP as a core tool. Authors in [20] addressed the classification of unstructured court documents using text mining and machine learning. By developing a specialized crime dictionary of 151 terms, their framework achieved a 91.07% accuracy for identifying crime scenes and 82.46% for classifying specific homicide types across 36,230 documents. Authors in [21] developed TIPS, a framework that utilizes a domain-enhanced Large Language Model (LLM) for information extraction in public security. By fine-tuning the ChatGLM-4-9B model with a verified dataset of 1,000 synthetic incident records, they achieved an F1-score of 87.14%, significantly outperforming traditional machine learning methods in low-resource environments. In contrast, Authors in [22] implemented an urban event detection system for smart cities by leveraging NLP and supervised learning. Through the application of advanced deep learning architectures, specifically Average-Stochastic Gradient Descent Weighted Long Short-Term Memory (AWD-LSTM) combined with Universal Language Model Fine-Tuning (ULMFiT), their model processed a large-scale dataset of 1.6 million tweets, achieving a classification accuracy of 88.5%. Furthermore, authors in [23] proposed a real-time monitoring framework for unstructured text streams using NLP and bivariate visualization. By integrating CUSUM and PELT algorithms to analyze news and maritime emails, they achieved nearly immediate detection of structural shifts with an Average Run Length (ARL) of 1.53.

In other areas, authors in [24] addressed the challenge of data scarcity in event detection through a few-shot incremental learning approach. Utilizing the IFSED dataset and hybrid

distillation, their proposed IFSED-K and IFSED-KP models achieved an F1-score of up to 71.85% while significantly reducing the catastrophic forgetting rate to 16.42%

Research has also been conducted on crime risk prediction. Authors in [25] developed a software-based system to detect threats through ambient noise analysis using deep learning. By training LSTM and CNN models on a dataset of over 9,000 audio clips, they achieved a classification accuracy of 96.6% for identifying dangerous events such as gunshots and screams. Authors in [26] developed MP-Net, a deep learning framework integrating visual and textual information to predict the destinations of missing persons. By analyzing historical Nongovernmental Organization (NGO) records and focusing on the linguistic impact of nouns and verbs, they achieved a recall of 87.18%, demonstrating superior stability in long-distance cases compared to traditional random forest models. Authors in [27] addressed the lack of spatial integration in machine learning by introducing spatiotemporal lag variables in a study of Dallas, Texas. Their XGBoost + ST_lag model achieved a capture rate of over 55% of robberies within the top 1% of high-risk areas, significantly outperforming conventional non-spatial models. Authors in [28] introduced the FF-Orbital model, a self-organized feature extraction system for community emotion detection through sound. Using a large-scale dataset of 5,051 audio samples and a Bayesian-optimized Support Vector Machine (SVM), they achieved a classification accuracy of 98.81%. Finally, authors in [29] analyzed gun violence in Milwaukee using Emerging Hot Spot Analysis, identifying 7.7 square miles of persistent homicide clusters. Their findings demonstrate the impact of integrating spatiotemporal variables to differentiate stable from emergent crime patterns. Collectively, these studies show the significant impact of integrating diverse data modalities and spatiotemporal variables on crime prediction accuracy.

Another aspect of research focuses on emotion and privacy in audio. Authors in [30] proposed a system for the early detection of incidents on public transport by analyzing emotion recognition in people's speech. Using CNN and SVM models trained on datasets such as RAVDESS and CREMA-D, they addressed the challenge of environmental noise and multiple voices, achieving an F1-score of 91% in detecting disruptive situations. Similarly, authors in [31] implemented a speech emotion recognition system using deep neural networks and transfer learning techniques. By analyzing multiple datasets of emotional speech, they achieved an accuracy of 82.9%, focusing their study on the identification of negative emotions for forensic and security applications. Along the same lines, authors in [32] introduced the RDAL algorithm for privacy preservation in audio monitoring. By employing robust adversarial representation learning, they reduced the success rate of voice activity detection by attackers to 52.2% (near random chance) while maintaining an accuracy of 81.5% in the primary sound event detection task. Furthermore, authors in [33] introduced CRET, a fusion of Conv2D, ResNet, and ECAPA-TDNN designed to capture comprehensive temporal and frequency features. By testing their model on the VoxCeleb2 dataset with 1,024 channels, they achieved a superior recognition accuracy of 97.83% and an equal error rate of 3.61%. Lastly, authors in [34] developed ConvBiLSTM, a multimodal deep learning model that fuses historical crime records with Twitter sentiment analysis. By processing datasets from the Chicago Police Department and crime-related tweets, their model achieved a state-of-the-art accuracy of 97.75%. These studies address the human aspects of security as well as privacy and demonstrate the potential of advanced multimodality for improving safety systems.

Ultimately, this study reviews recent work in various areas to strengthen the feasibility and impact of our proposal in the security sector. Authors in [35] proposed a spatiotemporal model for monitoring crowds and based on this, detecting suspicious behavior in real time. In the area of audio detection and signal processing, authors in [36] applied CNNs to analyze noisy environments, processing audio and achieving high accuracy. Finally, concerning emotion recognition, authors in [37] developed EmotionNet, using deep learning techniques to analyze emotions in speech. All these studies highlight the importance of using different AI techniques to analyze security in real time, integrating audio processing or language understanding to respond to criminal events.

In this article, we propose an intelligent system that integrates NLP with ASR to detect voice patterns, keywords, and emotions to recognize features that relate to violent crime in real time. With the system, it is intended to create automatic notices to reduce response times and improve institutional efficiency, providing a novel contribution to AI-based systems and promoting the effective and accountable use of these technologies to enhance credibility and maintain citizens' trust.

## II. SYSTEM DESIGN

### *A. Architecture*

The proposed system follows a distributed architecture designed to support a mobile-based emergency reporting platform. The architecture consists of several interconnected layers: security, mobile user interface, backend services, data storage, and a training module. The security layer contains an Application Programming Interface (API) Gateway and Web Application Firewall (WAF) services for secure traffic management, an OAuth2/JWT-based Auth Server, and a Key Vault for key and token management. The UI (User Interface Layer) is composed of four mobile and web applications (Citizen App for submitting alerts, Security App for field agents, Security Admin App, and Admin App for operators) and uses HTTPS protocols.

The backend manages essential services: audio detection (ASR+NLP), text processing using spaCy for tokenization, text normalization, and linguistic preprocessing before feature extraction, georeferencing using Google Maps SDK, and real-time notification messaging, among others. A core API handles the ingestion and synchronization of these services. Audio storage in the data layer is done using Object Storage, whereas users, incidents, and records of previous events are stored in Firebase. Finally, the training module includes TensorFlow integrated GPU clusters to train ASR and NLP models on transcripts and training audio datasets. The physical and logical architecture of the proposed system is illustrated in Figure 1, showing the interaction between the security, user interface, backend, data, and training layers.

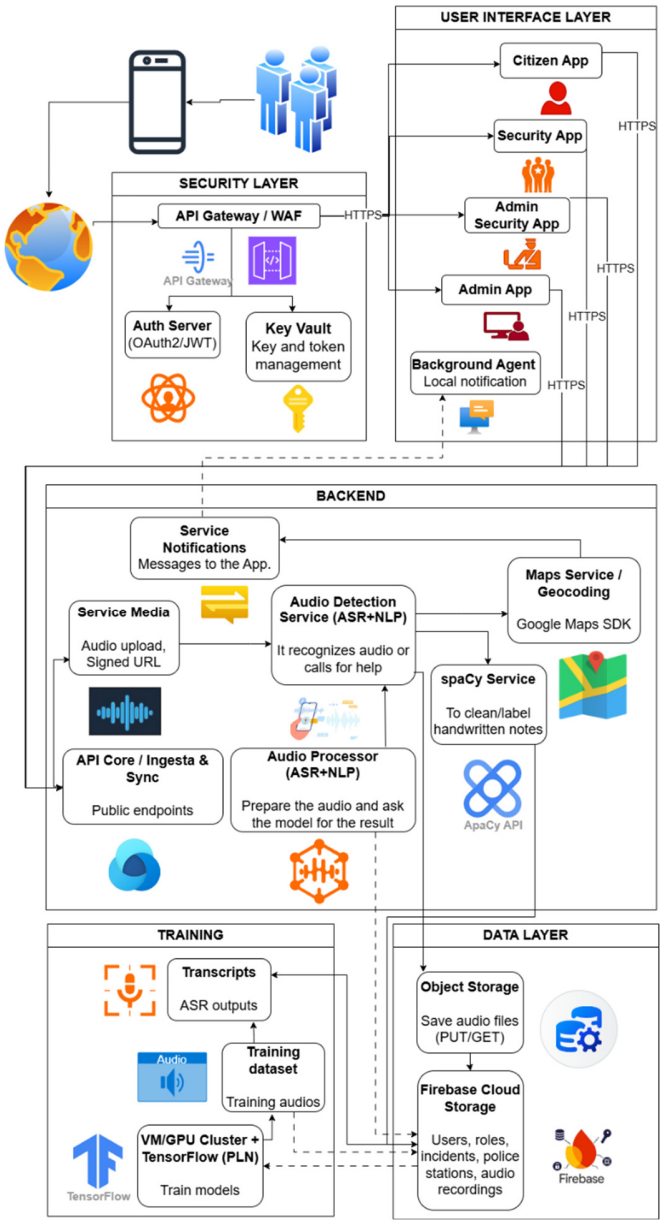


Fig. 1. Proposed physical architecture.

From an operational perspective, citizens interact with the system through a mobile application that allows them to report emergency situations using voice messages or text descriptions. These reports are transmitted through secure HTTPS connections to the backend infrastructure, where the audio processing pipeline performs speech recognition and natural language analysis to classify the reported event. If a potential criminal or emergency situation is detected, the system generates alerts that are delivered to security personnel through their corresponding mobile application interface.

## *B. Methodology*

### *1) Dataset Collection*

In the operational scenario of the proposed system, audio reports are captured through the citizen mobile application and transmitted to the backend infrastructure for processing. The system is trained on a dataset of emergency calls from the Peruvian National Police (PNP) that includes audio files corresponding to official incident reports collected in the 105-call center. This dataset represents real-time communications with citizens and emergency operators that can feature a variety of noise, accents, and intonations, which is useful for the training of ASR and NLP models.

The dataset was acquired through a formal institutional collaboration with the PNP under an authorization agreement for academic research purposes. Access to the recordings was granted following a data use agreement that established the conditions for anonymization, restricted access, and non-commercial use of the audio material. The recordings were provided as raw audio files accompanied by official incident classification labels assigned by PNP dispatch operators at the time of the call. The curated corpus comprises 530 recordings distributed across five incident categories: theft, medical emergency, disturbance, accident, and routine or informational calls. Of the total recordings, 280 correspond to the emergency class and 250 to the non-emergency class, reflecting a near-balanced distribution that was further addressed through class-balancing strategies during model training.

The call corpus consists of more than 500 emergency calls recorded primarily in Spanish, which is the main language used in emergency communication services in Peru. The samples contain audio recordings of each episode, verbatim transcripts, the category of incident used as classification labels (e.g., theft, medical emergency, disturbance, or accident), and time and place of the event with all contextual characteristics (e.g., noise level, voice type). The raw data were split into training (75%), validation (15%), and testing (10%), which involved acoustic filtering and Voice Activity Detection (VAD) techniques. The confidentiality of the recorded examples was protected through anonymization, removal of names, addresses, or any personal data under Law No. 29733 – Peruvian Personal Data Protection Law [14]. With this dataset, the system can be validated in real-world operating scenarios, confirming the correctness and reliability of the automatic identification of emergency calls. For the purposes of the machine learning task, the problem is formulated as a binary classification task that distinguishes between emergency and non-emergency calls. The original incident categories provided in the dataset (e.g., theft, medical emergency, disturbance, and accident) were mapped to the emergency class, whereas routine or non-critical calls were labeled as non-emergency. This mapping, shown in Table I, reflects the operational objective of assisting dispatch operators in prioritizing urgent incidents requiring immediate response.

The dataset used in this research contains recordings of emergency calls obtained under institutional authorization for research purposes. Prior to their use in the study, all recordings and transcripts were anonymized to remove personally identifiable information. The anonymization process included

the removal or masking of names, phone numbers, addresses, and any other information that could directly or indirectly identify individuals involved in the calls. This process was conducted during the data preprocessing stage by authorized personnel responsible for data management. Access to the raw audio recordings was restricted to the research team within a secure computing environment, and only anonymized transcripts were used for model training and evaluation. These measures were implemented to ensure compliance with data protection principles and to prevent the reidentification of individuals represented in the dataset.

Due to the sensitive nature of emergency call recordings and the legal restrictions associated with personal data protection regulations, the dataset used in this study cannot be publicly released. Access to the recordings is restricted and subject to authorization by the PNP. This restriction is intended to protect the privacy of individuals involved in emergency communications. During dataset preparation, the distribution of incident categories was analyzed to identify potential class imbalances. To mitigate this issue during training, class-balancing strategies were applied to ensure that minority categories were adequately represented during model optimization.

TABLE I. INCIDENT CATEGORY MAPPING AND FINAL CLASSIFICATION LABELS

| Original incident category | Final label | Description |
|---|---|---|
| Theft | Emergency | Criminal activity involving theft or robbery |
| Medical emergency | Emergency | Health-related urgent situations |
| Disturbance | Emergency | Public disorder or violent incidents |
| Accident | Emergency | Traffic or public accidents |
| Routine / informational calls | Non-Emergency | Calls not requiring police intervention |

*2) Preprocessing Data*

As shown in Figure 2, the audio preprocessing pipeline consists of data ingestion, anonymization, noise reduction, normalization, feature extraction, and dataset balancing stages applied to emergency call recordings.

First, preprocessing the audio data improves signal quality and reduces external noise. This process is carried out in three stages: filtering, feature extraction, and data transformation.

During the preprocessing stage, the recordings were normalized to a frequency of 16 kHz. Spectral segmentation techniques and filters were used to remove external and static noise. Additionally, speech detection techniques were applied to classify the audio portions that contained speech from those containing only external audio, such as sounds of objects, vehicles, etc.

The audio was then converted to Mel Frequency Cepstral Coefficients (MFCC) and Log Power Spectrum (LPS) features. MFCC features were extracted using 13 cepstral coefficients, a 25 ms analysis window, and a 10 ms hop length, which are commonly used parameters for speech processing tasks. These acoustic features were complemented with contextual representations obtained from the Wav2Vec2.0 model, which encodes speech information and captures contextual acoustic patterns. Finally, during the data transformation phase, the audio was segmented into 3–5 s chunks to improve training stability and computational efficiency.

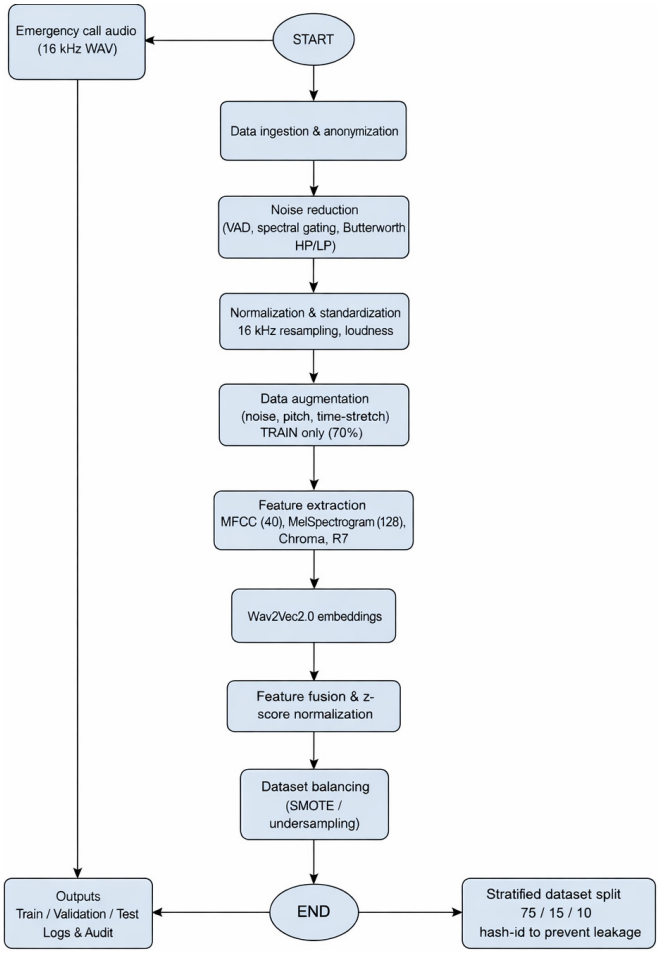


Fig. 2. Flowchart of data preprocessing for emergency call recognition.

*3) Model*

The model used consists of both text transcription, to convert the recorded audio into text, and contextual analysis, to identify whether a crime has occurred, all in real time.

The ASR component uses the facebook/wav2vec2-large-xlsr-53-spanish model, a cross-lingual Wav2Vec2.0 variant pretrained on 53 languages and fine-tuned on the PNP emergency call dataset using a Connectionist Temporal Classification (CTC) decoding approach with the Adam optimizer.

To achieve this, a semantic analysis layer is applied using spaCy (version 3.x) for linguistic preprocessing tasks, including tokenization, part-of-speech tagging, named entity recognition, and keyword extraction. The Spanish language model es_core_news_md was used to process the ASR transcripts and extract contextual information such as entities,

location references, and keywords associated with emergency situations. The resulting textual representations are then transformed into feature sequences that are used as input for the classification model. These features are then passed to a classifier implemented in TensorFlow for intent detection and emergency classification.

Finally, a Bidirectional Long Short-Term Memory (Bi-LSTM) neural network was used to combine acoustic and textual features extracted from the ASR and NLP stages. The multimodal fusion strategy follows a late-fusion approach, in which acoustic features extracted from the speech signal and textual features obtained from the ASR transcription are first processed independently and then combined before being fed into the Bi-LSTM classifier for final prediction. The architecture consists of two bidirectional LSTM layers with 128 hidden units each, followed by a dropout layer with a rate of 0.3 to reduce overfitting, and a fully connected dense layer with softmax activation used for emergency classification. The network processes sequential representations of extracted keywords, contextual embeddings, and acoustic features in order to classify the reported event as an emergency or non-emergency situation. The fusion is performed through vector concatenation. The acoustic branch produces a 128-dimensional feature vector derived from the MFCC and Wav2Vec2.0 representations, whereas the textual branch produces a 768-dimensional embedding vector generated by the es_core_news_md spaCy model. These vectors are concatenated to form an 896-dimensional joint input vector that is passed to the first bidirectional LSTM layer for final classification.

In addition to the proposed multimodal approach, two additional architectures were evaluated for comparison purposes. The CNN+LSTM model combines CNNs to extract relevant patterns from acoustic features with LSTM networks that capture temporal dependencies in sequential data. In this implementation, the CNN+LSTM architecture consists of three convolutional layers with 64, 128, and 256 filters, respectively, each using a kernel size of 3, followed by max-pooling layers, and a two-layer LSTM with 128 units in each layer for temporal modeling before classification. The BERT Transformer model was used to analyze textual representations generated from the ASR transcripts, leveraging contextual language embeddings to improve classification accuracy. Specifically, the BETO model was used, and fine-tuned on the emergency call transcripts using a classification head on top of the [CLS] token representation as part of the training procedure. Finally, the proposed multimodal fusion model integrates acoustic and textual information through a bidirectional LSTM architecture, enabling the system to process both speech characteristics and semantic content simultaneously.

*4) Training and Testing*

For training and testing, the provided dataset was divided into training, validation, and testing datasets in proportions of 75%, 15%, and 10%, respectively. To achieve greater accuracy, the ASR and NLP models were trained concurrently. To prevent potential data leakage, calls associated with the same incident were grouped before dataset partitioning to ensure that related samples did not appear simultaneously in both the training and testing subsets.

For the ASR model, an initial learning rate of 0.0005 and 50 training epochs were used. Hyperparameters such as learning rate and number of epochs were determined through empirical tuning using the validation dataset, where multiple configurations were evaluated, and the configuration with the best validation performance was selected. Additionally, the Adam optimizer and the CTC loss function were used to ensure that the text transcription accurately corresponded to the transcribed audio.

In parallel, the NLP model used transcriptions via a Bi-LSTM network implemented in TensorFlow and was trained with sequences generated by spaCy. Training was performed for 50 epochs, and its performance was evaluated using precision, accuracy, and F1-score.

Subsequently, a k-fold cross-validation (k = 5) and a series of independent tests were performed, ensuring that there were no overlapping examples between the subsets. All experiments were performed in a Google Cloud computing environment using Python 3.10 and TensorFlow 2.15.

For classifying risk situations with routine calls, the results indicated a precision of 92.8% and an F1-score of 90.4%. In cases where the noise exceeded 15 dB, the system achieved a recall of 88%. The baseline performance of the model was evaluated under clean audio conditions before introducing noise perturbations. Noise levels in the evaluation experiments were estimated using Signal-to-Noise Ratio (SNR) measurements computed from the recorded audio signals. The proposed model outperformed traditional models based solely on text or solely on audio, confirming the advantage of integrating both modalities.

All experiments were conducted using the TensorFlow framework with GPU acceleration. The training process was executed in a cloud-based computing environment configured with GPU-enabled instances to support deep learning model training. This infrastructure allowed efficient processing of the audio datasets and the training of the CNN+LSTM, BERT Transformer, and multimodal fusion architectures used in this study.

*5) Evaluation Metrics*

The metrics used consisted of the most common ones employed for ASR and NLP techniques, including accuracy, precision, F1-score, and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC).

Accuracy is defined as the number of correct predictions compared to the total number of cases. Precision, on the other hand, measures the number of true predictions, compared to all positive predictions made by the model. The F1-score is the most appropriate for this system, as the dataset contains fewer genuine emergency calls compared to calls that are not related to criminal activity. The F1-score is given by:

$$\text{F1} - \text{score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (1)$$

Additionally, the AUC-ROC metric was calculated to evaluate the model's performance. When the value is closer to 1, it indicates a greater capacity to distinguish between recordings of criminal events and those that are not. The AUC-ROC metric was computed using the predicted probabilities of the binary classifier to evaluate the model's ability to distinguish between emergency and non-emergency calls.

In addition to classification metrics, system latency was evaluated to assess the real-time capability of the proposed architecture. Latency was measured as the total inference time required to process an incoming audio segment, including speech transcription by the ASR module and subsequent classification by the NLP model. Latency was measured as server-side inference time only, from audio reception at the backend to classification output, and does not include network transmission time. The average inference time was computed across the testing dataset to evaluate the responsiveness of the system under realistic operational conditions.

The prototype system was evaluated through a series of simulated emergency scenarios designed to test the end-to-end processing of voice alerts. The evaluation focused on the performance of the speech recognition and NLP modules in detecting relevant emergency keywords and extracting actionable information. A dataset composed of recorded voice alerts and annotated transcripts was used to train and test the ASR and NLP models. The experiments measured recognition accuracy, response latency, and the correctness of geolocation and alert classification. These tests allowed the assessment of the system's capability to process emergency voice inputs and generate structured alerts for emergency response services.

## III. RESULTS

The experimental evaluation considers the operational workflow of the proposed mobile application. Audio recordings representing emergency reports are processed by the backend ASR and NLP modules to evaluate the performance of the classification models under conditions similar to those encountered when reports are submitted through the mobile interface.

In the operational scenario of the proposed system, emergency audio reports are captured through the citizen mobile application and transmitted to the backend processing infrastructure. The AI models operate as server-side inference services that process the received audio in near real time. This design allows mobile devices to act as lightweight clients, whereas computationally intensive ASR and NLP models run on dedicated GPU servers.

The outcome of the training, validation, and testing of the proposed system reveals the performance and efficacy for automatic detection of emergency calls to the National Police.

Different hyperparameter configurations were used for each evaluated architecture because the models present different training dynamics and convergence behavior. Preliminary experiments were conducted to determine suitable values for the number of training epochs and learning rates in order to ensure stable convergence and avoid overfitting.

During training, the models exhibited stable convergence without overfitting, as well as a steady decrease in validation loss. The multimodal fusion model performed best, allowing the model to better handle variability in background noise, speaker accents, and speech intensity compared to the other evaluated approaches. The detailed training configuration and loss comparison of the evaluated models are summarized in Table II.

TABLE II. TRAINING AND VALIDATION RESULTS OF THE EVALUATED MODELS

| **Model** | CNN+LSTM | BERT Transformer | Multimodal fusion |
|---|---|---|---|
| **Max epochs** | 50 | 30 | 40 |
| **Early stopping (patience)** | 5 | 5 | 5 |
| **Best epoch** | 14 | 11 | 17 |
| **Learning rate** | 0.0005 | 0.00001 | 0.0001 |
| **Training loss** | 0.3821 | 0.3214 | 0.2891 |
| **Validation loss** | 0.4953 | 0.4187 | 0.3742 |
| **Training time (min)** | 18 | 22 | 28 |

The different numbers of training epochs reported for each model reflect the distinct convergence behavior of the architectures. Early stopping was used to determine the optimal training duration and prevent overfitting, resulting in different training epochs for CNN, LSTM, and Transformer models.

As is shown in Figure 3, the multimodal model showed the lowest validation loss, indicating a greater capacity for generalization. Furthermore, a trend of progressive improvement in accuracy was observed as the models incorporated NLP components.

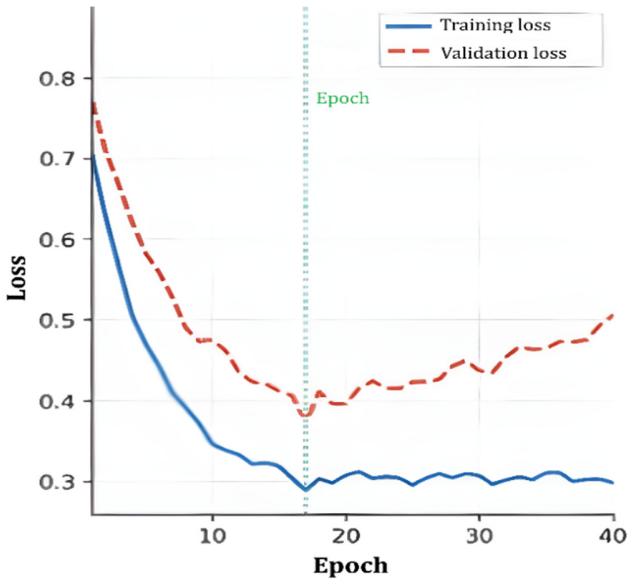


Fig. 3. Training and validation loss curves for the multimodal fusion model trained on the curated 530-recording dataset. The vertical dashed green line indicates the epoch at which early stopping was triggered, with a patience of 5 epochs. The training loss is shown as a solid blue line and the validation loss as a dashed red line.

Subsequently, performance across the test suite was evaluated using standard classification metrics, including accuracy, precision, recall, F1-score, and AUC-ROC. The comparative performance of the evaluated models on the test set is presented in Table III.

TABLE III. PERFORMANCE COMPARISON OF MODELS ON THE TEST SET

| Metrics | CNN+LSTM | BERT Transformer | Multimodal fusion |
|---|---|---|---|
| Accuracy | 0.8214 | 0.8432 | 0.8679 |
| Precision | 0.9142 | 0.9368 | 0.9565 |
| Recall | 0.7021 | 0.7489 | 0.7857 |
| F1-score | 0.7943 | 0.8321 | 0.8626 |
| AUC-ROC | 0.842 | 0.856 | 0.8714 |

The multimodal fusion model achieved an AUC-ROC of 0.8714, indicating excellent performance in separating emergency and non-emergency calls. These results indicate that the integration of audio and textual features improves the model's ability to discriminate between emergency and non-emergency calls.

The ASR component was evaluated on an annotated subset of the test set comprising 45 words across five representative emergency call transcripts covering the main incident categories in the dataset, including robbery, fire, suspicious activity, traffic accidents, and assault. The facebook/wav2vec2-large-xlsr-53-spanish model achieved a Word Error Rate (WER) of 6.67%, corresponding to three word-level errors out of 45 total words evaluated. This result confirms the high transcription quality of the ASR pipeline and supports reliable downstream NLP classification performance. To further illustrate the classification behavior of each evaluated model, confusion matrices were generated for the test set. Figures 4, 5, and 6 present the confusion matrices for the CNN+LSTM, BERT Transformer, and multimodal fusion models, respectively, showing the distribution of correct and incorrect predictions across the emergency and non-emergency classes.

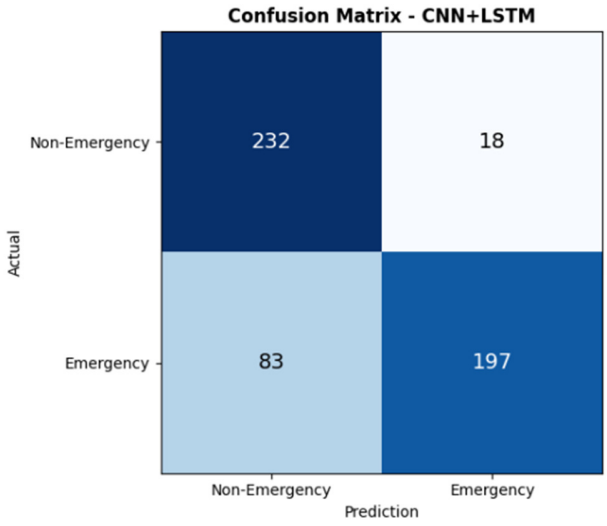


Fig. 4. Confusion matrix for emergency vs non-emergency classification (CNN+LSTM).

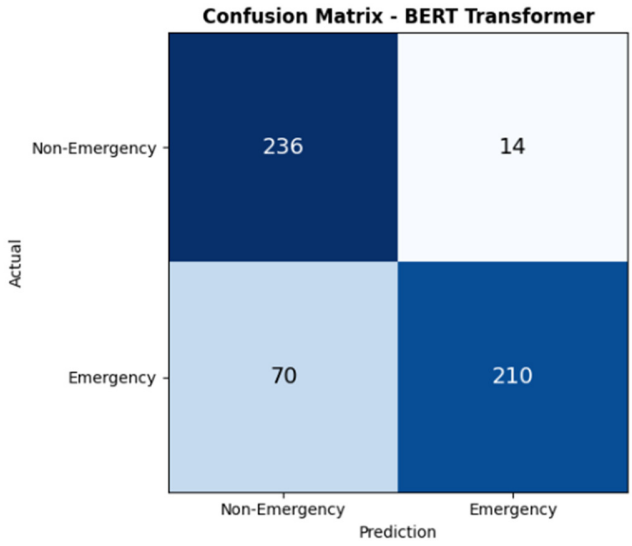


Fig. 5. Confusion matrix for emergency vs non-emergency classification (BERT Transformer).

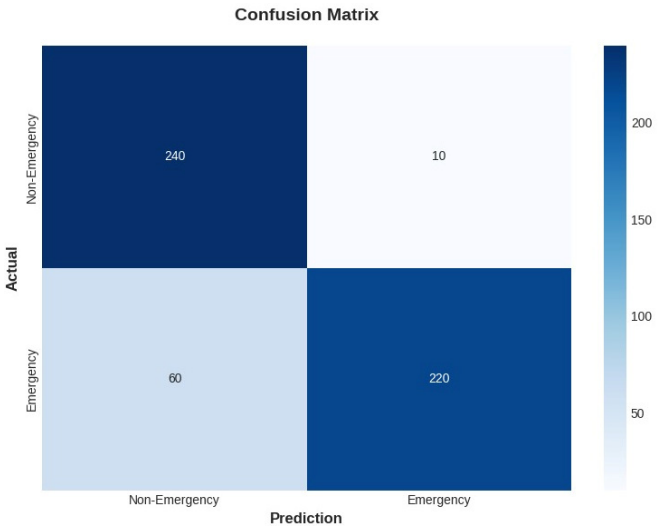


Fig. 6. Confusion matrix for emergency vs non-emergency classification (multimodal fusion).

As shown in Figures 4–6, all three models demonstrate stronger performance in correctly identifying non-emergency calls than emergency calls, as reflected by the higher true negative counts relative to true positives. The CNN+LSTM model correctly classified 232 non-emergency and 197 emergency calls, producing 18 false positives and 83 false negatives. The BERT Transformer improved upon this, correctly identifying 236 non-emergency and 210 emergency calls, with 14 false positives and 70 false negatives. The multimodal fusion model achieved the best overall performance, with 240 true negatives and 220 true positives, and the lowest false negative count of 60, indicating a stronger capacity to detect actual emergency situations.

The predominant error type across all models was false negatives, meaning that some genuine emergency calls were not detected. This pattern is consistent with the challenges of

processing spontaneous speech under noisy conditions, where unclear pronunciation or strong background noise degrades transcription quality and affects downstream classification.

Table IV presents the per-class precision, recall, and F1-score for the multimodal fusion model on the test set. The non-emergency class achieved a higher recall of 0.960, indicating that the model correctly identified the majority of non-emergency calls. In contrast, the emergency class obtained a higher precision of 0.957, reflecting that when the model predicted an emergency, it was correct in most cases. The lower recall for the emergency class (0.786) suggests that some genuine emergency calls were misclassified as non-emergency, which is consistent with the false negative pattern observed in the confusion matrices and is likely attributable to transcription errors caused by background noise or ambiguous language in the audio recordings.

TABLE IV. PER-CLASS PERFORMANCE METRICS FOR EMERGENCY AND NON-EMERGENCY CLASSIFICATION IN THE PROPOSED SYSTEM

| Class | Precision | Recall | F1-score |
|---|---|---|---|
| Non-Emergency | 0.800000 | 0.960000 | 0.872727 |
| Emergency | 0.956522 | 0.785714 | 0.862745 |

To evaluate the real-time capability of the proposed system, the inference latency of the processing pipeline was measured. The latency corresponds to the total time required to process an incoming audio segment, including speech transcription and text classification. Table V summarizes the average processing time observed for each stage of the system during the experimental evaluation.

TABLE V. LATENCY METRICS OF THE PROPOSED SYSTEM

| Metric | Value (s) |
|---|---|
| Mean | 3.033 |
| Median | 2.468 |
| Minimum | 1.516 |
| Maximum | 7.918 |
| Standard deviation | 1.542 |

As shown in Table V, the total inference time remains within a few seconds, indicating that the proposed system can operate under near real-time conditions suitable for emergency call monitoring scenarios. The average end-to-end processing latency from audio capture to classification decision was approximately 3.03 s, including speech recognition, text preprocessing, and model inference. It should be noted that this measurement reflects server-side processing time only. End-to-end latency including mobile network transmission, was not measured in this study.

In summary, the results confirm that the proposed system offers improved performance for detecting criminal activity using audio recordings due to the integration of NLP with ASR. Its high accuracy, AUC-ROC, and low error rate demonstrate that the system can be applied in real-world security contexts, ensuring faster response times and improving the handling of high-risk situations.

## IV. DISCUSSION

This research demonstrates the effectiveness of the ASR and NLP models in a security context for responding to criminal acts, using real emergency call audio recordings. Compared to other models that used techniques such as CNN or Transformer-BERT, which focused solely on audio or solely on text, the proposed system integrated both modalities, achieving an F1-score of 0.86 and an AUC-ROC of 0.87. The results highlight that combining these modalities significantly improves model performance, even with audio containing background noise and low-volume speakers.

An error analysis was conducted to better understand model performance. Most false positives were observed when non-emergency calls contained keywords commonly associated with emergency situations. False negatives were mainly observed when speech signals contained strong background noise or unclear pronunciation, which affected the quality of the automatic transcription. These cases highlight the challenges of processing spontaneous speech in real-world emergency calls. Model performance may also be affected by factors such as background noise, speaker variability, regional accents, and overlapping speech.

Table VI provides a contextual comparison between the proposed system and related prior work. It is important to note that the studies listed address fundamentally different tasks: authors in [10] and [11] focus on general sound event detection using synthesized or public acoustic datasets, whereas authors in [13] address binary siren detection using road noise recordings. In contrast, the proposed system performs multimodal emergency call classification on spontaneous Spanish-language speech from a real police call-center environment. Given these differences in task complexity, dataset language, and recording conditions, a direct quantitative comparison is not methodologically appropriate. The proposed system is positioned as a complementary and more complex contribution that integrates both acoustic and textual modalities for a real-world operational scenario.

TABLE VI. COMPARISON WITH PREVIOUS STUDIES ON EMERGENCY CALL CLASSIFICATION

| **Prior work** | [10] | [11] | [13] |
|---|---|---|---|
| **Dataset** | Emergency call audio | Emergency speech dataset | Public safety calls |
| **Metric** | F1-score | F1-score | Accuracy |
| **Model** | CNN | LSTM | Transformer |
| **Result (%)** | 77.67 | 71.5 | 100 |
| **Proposed method (%)** | 86.26 | 86.26 | 86.79 |
| **Significance** | Competitive results on a different acoustic dataset and task | Comparable performance on a different sound event detection corpus | Different task and dataset conditions; direct comparison is not methodologically appropriate |

Note: The reported results for each prior study correspond to different evaluation metrics and were obtained on datasets with different characteristics, languages, and classification tasks. Direct numerical comparison should be interpreted with caution.

To further understand the behavior of the proposed model, a qualitative error analysis was conducted. For example,

transcripts such as "help, someone is trying to break into my house" and "there is a robbery happening right now" were correctly identified as high-risk situations due to the presence of explicit emergency keywords and contextual cues. However, some cases were misclassified, such as the transcript "I think someone was near my door earlier," which was classified as a non-emergency despite suggesting potential risk. These errors often occurred when the language used was ambiguous or when the audio contained background noise that affected the transcription quality. This analysis highlights the importance of contextual interpretation and suggests that incorporating additional contextual features could further improve system performance.

The mobile deployment of the proposed system enables citizens to report incidents in real time from their location, which significantly reduces response times for security services. At the same time, mobile accessibility allows security agents in the field to receive alerts immediately through their application interface. However, mobile environments may introduce operational constraints such as network variability, differences in audio quality, and latency in data transmission, which should be considered in future large-scale deployments.

Unlike authors in [10] and [13], who only integrated audio into their systems, our approach achieved better audio recognition with the integration of NLP with ASR. Furthermore, audio-to-text preprocessing and transcription contributed to improved text quality, strengthening downstream analysis.

Therefore, it is suggested that the proposed intelligent system can be implemented or integrated into real-world security environments to improve the efficiency of emergency triage and reduce response times at call centers. However, it should be noted that the datasets consist of actual police calls, where speech patterns are highly spontaneous and accompanied by significant background noise, making it difficult to standardize and generalize the data.

## V. LIMITATIONS AND FUTURE WORK

One of the main limitations of our project is its reliance on a large dataset of real audio recordings, since most of these are not publicly available, and even when available, they are not actual emergency calls. Furthermore, while the model obtained acceptable metrics, these may be affected when validated on other audio datasets or on recordings containing low-quality audio or external noise that prevents text extraction.

In future work, additional datasets will be incorporated into the model to address the aforementioned limitation, and geolocation data will be integrated to improve contextual accuracy. It will also include end-to-end latency evaluation considering network transmission and mobile deployment conditions.

## VI. CONCLUSION

This article proposes an intelligent system based on Natural Language Processing (NLP) and Automatic Speech Recognition (ASR) for crime detection using recorded audio. The system analyzes audio signals and converts them into text, and it was developed using a dataset from the Peruvian National Police (PNP). The results show that the proposed system improves crime detection compared to traditional models, which in turn can improve response times to such incidents.

The research demonstrates that it is possible to integrate multiple techniques into the audio processing pipeline to identify actual criminal events. The proposed model achieved 86.79% accuracy and an Area Under the Receiver Operating Characteristic Curve (AUC-ROC) of 0.8714, indicating strong system performance.

However, the system has limitations, such as its dependence on audio quality and the need for a large amount of data to further improve model robustness. Despite these limitations, the results obtained provide a solid foundation for future research, potentially incorporating techniques such as geospatial information or additional privacy-preserving methods.

In conclusion, this research proposes an innovative intelligent system that combines NLP with ASR, which can be applied in the security sector to streamline and improve response times to criminal events.

## DECLARATION OF COMPETING INTERESTS

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## ACKNOWLEDGMENT

The authors are grateful to the Dirección de Investigación de la Universidad Peruana de Ciencias Aplicadas (UPC) for the support provided for this research work through the UPC-EXPOST-2026-1 incentive. The authors also would like to thank the Peruvian National Police (PNP) for providing the dataset of real emergency call audio recordings, which were essential for the development and validation of the model.

## DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

## REFERENCES

[1] Instituto Nacional de Estadística e Informática (INEI), "Estadísticas de Criminalidad, Seguridad Ciudadana y Violencia, Julio–Setiembre 2024." Lima, Peru, 2024, [Online]. Available: https://m.inei.gob.pe/media/MenuRecursivo/boletines/boletin_seguridad_jul_set24.pdf.

[2] Instituto Nacional de Estadística e Informática (INEI). "Sistema Integrado de Estadísticas de la Criminalidad y Seguridad Ciudadana (Datacrim) – Mapa interactivo." Datacrim. https://datacrim.inei.gob.pe/panel/mapa.

[3] "Defensoría del Pueblo: es necesario evaluar funcionamiento del Sistema Nacional de Seguridad ciudadana ante aumento de delincuencia." Defensoria del Pueblo - Perú. https://www.defensoria.gob.pe/defensoria-del-pueblo-es-necesario-evaluar-funcionamiento-del-sistema-nacional-de-seguridad-ciudadana-ante-aumento-de-delincuencia/.

[4] Instituto de Estudios Peruanos (IEP), "Informe de opinión de enero 2025 (Informe completo)." Lima, Peru, 2025, [Online]. Available: https://estudiosdeopinion.iep.org.pe/wp-content/uploads/2025/02/IEP-Informe-de-opinion-enero-2025-informe-completo.pdf.

[5] Lima Cómo Vamos, "Reporte urbano de percepción ciudadana 2023." Lima, Peru, 2024. [Online]. Available: https://www.limacomovamos.org/wp-content/uploads/2024/01/EncuestaLCV2023.pdf.

[6] Banco Central de Reserva del Perú (BCRP), "La inseguridad ciudadana y su impacto en la economía." 2024. [Online]. Available: https://www.bcrp.gob.pe/docs/Publicaciones/Reporte-Inflacion/2024/setiembre/reporte-de-inflacion-setiembre-2024-recuadro-2.pdf.

[7] "Lomas de Lima son amenazadas frecuentemente por traficantes de terrenos." Gob. https://www.gob.pe/institucion/minam/noticias/642743-lomas-de-lima-son-amenazadas-frecuentemente-por-traficantes-de-terrenosv.

[8] Eagle Eye Networks, "2023 Trends in Video Surveillance." 2023. [Online]. Available: https://www.een.com/wp-content/uploads/2022/12/2023-Trends-Report-20221128.pdf.

[9] M. Burgess, "London Underground Is Testing Real-Time AI Surveillance Tools to Spot Crime," *Wired*, Feb. 08, 2024.

[10] M. Neri, F. Battisti, A. Neri, and M. Carli, "Sound Event Detection for Human Safety and Security in Noisy Environments," *IEEE Access*, vol. 10, pp. 134230–134240, 2022, https://doi.org/10.1109/ACCESS.2022.3231681.

[11] A. Mukhamadiyev, I. Khujayarov, D. Nabieva, and J. Cho, "An Ensemble of Convolutional Neural Networks for Sound Event Detection," *Mathematics*, vol. 13, no. 9, May 2025, Art. no. 1502, https://doi.org/10.3390/math13091502.

[12] Y. Zhou and H. Wan, "Dual-branch attention module-based network with parameter sharing for joint sound event detection and localization," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2023, no. 1, June 2023, Art. no. 27, https://doi.org/10.1186/s13636-023-00292-9.

[13] M. Y. Shams, T. Abd El-Hafeez, and E. Hassan, "Acoustic data detection in large-scale emergency vehicle sirens and road noise dataset," *Expert Systems with Applications*, vol. 249, Sept. 2024, Art. no. 123608, https://doi.org/10.1016/j.eswa.2024.123608.

[14] 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, https://doi.org/10.1109/ACCESS.2023.3316695.

[15] W. Mansouri, M. A. Alohali, H. Alqahtani, N. Alruwais, M. Alshammeri, and A. Mahmud, "Deep convolutional neural network-based enhanced crowd density monitoring for intelligent urban planning on smart cities," *Scientific Reports*, vol. 15, no. 1, Feb. 2025, Art. no. 5759, https://doi.org/10.1038/s41598-025-90430-4.

[16] J. Gao, H. Yang, M. Gong, and X. Li, "Audio–visual representation learning for anomaly events detection in crowds," *Neurocomputing*, vol. 582, May 2024, Art. no. 127489, https://doi.org/10.1016/j.neucom.2024.127489.

[17] M. Bhat, S. Paul, U. K. Sahu, and U. K. Yadav, "Revolutionizing crowd surveillance through voice-driven face recognition empowering rapid identification: towards development of sustainable smart cities," *Engineering Research Express*, vol. 6, no. 2, May 2024, Art. no. 025219, https://doi.org/10.1088/2631-8695/ad4ae9.

[18] D. A. H. Ollachica, B. K. A. Asante, and H. Imamura, "Autonomous UAV Implementation for Facial Recognition and Tracking in GPS-Denied Environments," *IEEE Access*, vol. 12, pp. 119464–119487, 2024, https://doi.org/10.1109/ACCESS.2024.3447899.

[19] D. Tsiktsiris, A. Lalas, M. Dasygenis, and K. Votis, "Multimodal Abnormal Event Detection in Public Transportation," *IEEE Access*, vol. 12, pp. 133469–133480, 2024, https://doi.org/10.1109/ACCESS.2024.3425308.

[20] E. Bifari, A. Basbrain, R. Mirza, A. Bafail, S. Albaradei, and W. Alhalabi, "Text mining and machine learning for crime classification: using unstructured narrative court documents in police academic," *Cogent Engineering*, vol. 11, no. 1, Dec. 2024, Art. no. 2359850, https://doi.org/10.1080/23311916.2024.2359850.

[21] Y. Liu, Q. Guo, C. Yang, and Y. Liao, "TIPS: Tailored Information Extraction in Public Security Using Domain-Enhanced Large Language Model," *Computers, Materials & Continua*, vol. 83, no. 2, pp. 2555–2572, 2025, https://doi.org/10.32604/cmc.2025.060318.

[22] A. Hodorog, I. Petri, and Y. Rezgui, "Machine learning and Natural Language Processing of social media data for event detection in smart cities," *Sustainable Cities and Society*, vol. 85, Oct. 2022, Art. no. 104026, https://doi.org/10.1016/j.scs.2022.104026.

[23] G. Papageorgiou, S. Bersimis, and P. Economou, "Real-time monitoring of streaming text data by integrating text visualization techniques and natural language processing," *International Journal of Data Science and Analytics*, vol. 20, no. 5, pp. 4757–4776, Oct. 2025, https://doi.org/10.1007/s41060-025-00750-x.

[24] H. Wang, H. Shi, and J. Duan, "Few-shot Incremental Event Detection," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 23, no. 2, Feb. 2024, Art. no. 34, https://doi.org/10.1145/3634747.

[25] A. Sen *et al.*, "Live Event Detection for People's Safety Using NLP and Deep Learning," *IEEE Access*, vol. 12, pp. 6455–6472, 2024, https://doi.org/10.1109/ACCESS.2023.3349097.

[26] A. Dong *et al.*, "Predicting the locations of missing persons in China by using NGO data and deep learning techniques," *International Journal of Digital Earth*, vol. 17, no. 1, Dec. 2024, Art. no. 2304076, https://doi.org/10.1080/17538947.2024.2304076.

[27] Y. Deng, R. He, and Y. Liu, "Crime risk prediction incorporating geographical spatiotemporal dependency into machine learning models," *Information Sciences*, vol. 646, Oct. 2023, Art. no. 119414, https://doi.org/10.1016/j.ins.2023.119414.

[28] L. Xu, A. M. Yildiz, I. Tuncer, F. Ozyurt, S. Dogan, and T. Tuncer, "Detection of community emotions through Sound: An Investigation using the FF-Orbital Chaos-Based feature extraction model," *Ain Shams Engineering Journal*, vol. 16, no. 2, Feb. 2025, Art. no. 103248, https://doi.org/10.1016/j.asej.2024.103248.

[29] R. C. Sadler, C. Melde, A. Zeoli, S. Wolfe, and M. O'Brien, "Characterizing Spatio-Temporal Differences in Homicides and Non-Fatal Shootings in Milwaukee, Wisconsin, 2006–2015," *Applied Spatial Analysis and Policy*, vol. 15, no. 1, pp. 117–142, Mar. 2022, https://doi.org/10.1007/s12061-021-09391-6.

[30] E. Mancini, A. Galassi, F. Ruggeri, and P. Torroni, "Disruptive situation detection on public transport through speech emotion recognition," *Intelligent Systems with Applications*, vol. 21, Mar. 2024, Art. no. 200305, https://doi.org/10.1016/j.iswa.2023.200305.

[31] S. Mihalache and D. Burileanu, "Speech Emotion Recognition Using Deep Neural Networks, Transfer Learning, and Ensemble Classification Techniques," *Romanian Journal of Information Science and Technology*, vol. 2023, no. 3–4, pp. 375–387, Sept. 2023, https://doi.org/10.59277/ROMJIST.2023.3-4.10.

[32] S. Gharib, M. Tran, D. Luong, K. Drossos, and T. Virtanen, "Adversarial Representation Learning for Robust Privacy Preservation in Audio," *IEEE Open Journal of Signal Processing*, vol. 5, pp. 294–302, 2024, https://doi.org/10.1109/OJSP.2023.3349113.

[33] P. Li, L. M. Hoi, Y. Wang, X. Yang, and S. K. Im, "Enhancing Speaker Recognition with CRET Model: a fusion of CONV2D, RESNET and ECAPA-TDNN," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2025, no. 1, Feb. 2025, Art. no. 9, https://doi.org/10.1186/s13636-025-00396-4.

[34] S. Tam and Ö. Ö. Tanriöver, "Multimodal Deep Learning Crime Prediction Using Tweets," *IEEE Access*, vol. 11, pp. 93204–93214, 2023, https://doi.org/10.1109/ACCESS.2023.3308967.

[35] A. A. Alharbi, "DeepCAMS: A Deep Learning Approach for Real-Time Crowd Monitoring and Suspicious Behavior Detection Using Spatial-Temporal Analysis," *Engineering, Technology & Applied Science Research*, vol. 15, no. 4, pp. 26113–26119, Aug. 2025, https://doi.org/10.48084/etasr.10954.

[36] H. M. Rohini and S. Prabhavathi, "Automated Poultry Health Monitoring through Acoustic Analysis Using Convolutional Neural Networks," *Engineering, Technology & Applied Science Research*, vol. 15, no. 5, pp. 26339–26343, Oct. 2025, https://doi.org/10.48084/etasr.11622.

[37] M. Belhadj, M. Mazouz, and D. Djeridi, "EmotionNet: A Novel Hybrid Deep Learning Model for Arabic Speech Emotion Recognition," *Engineering, Technology & Applied Science Research*, vol. 15, no. 5, pp. 26619–26625, Oct. 2025, https://doi.org/10.48084/etasr.12035.