Online-Journals.org

iJIM International Journal of Interactive Mobile Technologies

iJIM | eISSN: 1865-7923 | Vol. 20 No. 2 (2026) | OPEN ACCESS

https://doi.org/10.3991/ijim.v20i02.57541

PAPER

An Artificial Intelligence-Driven Mobile Application for Real-Time Assistance in Taxi-Related Criminal Incidents Using Natural Language Processing in Metropolitan Lima

Keni Abel Sanchez Villogas[1], Paolo Manoel Pinzás Riveros[1], Pedro Castañeda[1] (✉), Alejandra Oñate-Andino[2]

[1]Universidad Peruana de Ciencias Aplicadas (UPC), Lima, Peru

[2]Escuela Superior Politécnica de Chimborazo (ESPOCH), Riobamba, Ecuador

pcsipcas@upc.edu.pe

ABSTRACT

Herein, SmartSecurity, an artificial intelligence (AI)-based mobile application designed to enhance passenger safety in taxi services across Metropolitan Lima, is proposed herein. This system detected predefined emergency keywords via real-time voice processing and automatically activated geolocation and multichannel alerts via WhatsApp and SMS. Methodologically, an applied experimental–developmental design was adopted. The application was implemented using Flutter for the front-end interface and FastAPI on Google Cloud Platform for backend services. The Whisper AI model specialized in multilingual speech recognition was fine-tuned using contextualized Spanish audio datasets. The model performance was evaluated across parameters such as accuracy, sensitivity, F1-score, false-positive rate, and latency under simulated taxi conditions. The application achieved 90% accuracy and sensitivity, an F1-score of 0.90, and a near-zero false-positive rate, with an average latency of 1 s and a total response time of 2 s. The area under the ROC curve (AUC) approached 1.00, indicating high discriminative capacity of the model even in noisy environments. The proposed model thus provides an efficient, low-cost, and robust solution to the rising issue of urban transport insecurity, contributing to safer and smarter mobility systems in Metropolitan Lima.

KEYWORDS

artificial intelligence (AI), natural language processing (NLP), speech recognition, emergency response, taxi safety, Metropolitan Lima

1 INTRODUCTION

The persistent rise in public transportation insecurity, particularly within taxi services across Metropolitan Lima, is a major social and technological challenge

Sanchez, K. A., Pinzás, P. M., Castañeda, P., Oñate-Andino, A. (2026). An Artificial Intelligence-Driven Mobile Application for Real-Time Assistance in Taxi-Related Criminal Incidents Using Natural Language Processing in Metropolitan Lima. *International Journal of Interactive Mobile Technologies (iJIM)*, 20(2), pp. 41–62. https://doi.org/10.3991/ijim.v20i02.57541

Article submitted 2025-07-05. Revision uploaded 2025-11-18. Final acceptance 2025-12-12.

© 2026 by the authors of this article. Published under CC-BY.

that affects thousands of citizens daily. Passengers face multiple risks, including robbery, kidnapping, and physical assault, due to the absence of standardized control mechanisms, insufficient driver verification systems, and expanding informal taxi operations that function without regulation or supervision. These issues jeopardize passenger safety and create a widespread perception of insecurity that erodes public confidence in transportation services. Consequently, service usage declines, urban mobility deteriorates, and the overall economic and social well-being of the city is negatively affected.

This issue must be addressed because it directly influences the quality of life of citizens and the economic sustainability of Metropolitan Lima. Public transportation perceived as unsafe causes negative effects such as reduced use of formal taxi services, greater dependence on private vehicles, heavier traffic congestion, declining urban tourism, and higher costs associated with public safety and enforcement. International evidence has shown that cities implementing innovative safety measures within their transport systems have achieved measurable improvements in economic performance, social cohesion, and citizens' sense of security [1], [2].

Thus, several recent studies have explored the use of advanced technologies to mitigate risks in transportation environments. For instance, [3] developed a real-time event detection framework based on speech recognition and contextual semantic analysis to automatically identify anomalous or hazardous situations in public spaces. This reactive capability supported by natural language understanding can be vital when applied within the context of taxis because it enables automatic alert generation without requiring manual user intervention. Similarly, [4] introduced a multimodal learning system that combines audio and video inputs to achieve robust keyword recognition within natural sentences even under noisy conditions. This system enables emergency activation via spoken commands without the need to unlock the device, which is vital in situations wherein a victim cannot act freely.

[5] comprehensively analyzed deep keyword spotting architectures, emphasizing that modern models can operate effectively on low-cost mobile devices and maintain high detection accuracy, minimal latency, and low power consumption; these characteristics are indispensable for large-scale safety applications. From a privacy perspective, [6] proposed a system based on mobile edge computing and MinHash techniques that allows users to share approximate rather than exact locations when requesting transportation services, thereby reducing the risk of exposing sensitive information.

Driver authenticity and identity assurance are other critical aspects to ensure safe transportation. [7] developed a blockchain and machine learning framework to verify drivers' identities in shared-mobility services, improving transparency and reducing impersonation; incorrect driver identity is one of the main causes of safety incidents in informal taxi networks. [8] simultaneously demonstrated that zero-shot learning methods for classifying previously unseen acoustic events allow systems to adapt dynamically to new threats without extensive retraining, improving resilience in real monitoring environments. In addition, [9] showed that emotion detection models that can recognize fear, stress, or panic from short audio segments using deep acoustic features and random forest classifiers can provide an additional layer of contextual awareness for strengthening automated alert mechanisms. In [10–12], address emotion detection and user identification using artificial intelligence (AI) and user-centered design, but applied to different media. [13] proposed ultralight neural architectures such as BC-ResNet that performed real-time voice command detection with fewer than 10,000 parameters. These efficient systems can be deployed on mobile hardware with limited computational capacity without

compromising accuracy. These studies have confirmed that AI-driven strategies can substantially enhance safety in taxi transportation. However, they have also revealed persistent challenges such as the protection of user privacy [6], high telecommunication infrastructure costs required for real-time monitoring [7], the need for robust model training to handle acoustic and linguistic variability [5], and public skepticism regarding constant tracking systems [6].

Therefore, technological solutions that are effective, accessible, scalable, and suitable for the socio-technical conditions of Metropolitan Lima are urgently needed. Such solutions must balance technical efficiency with ethical, privacy, and economic considerations while ensuring ease of adoption among the population.

The remainder of this paper is organized as follows. Section 2 describes the AI techniques used, specifically the Whisper AI model for real-time voice recognition and the integrated geolocation components that enable automatic communication with emergency services and predefined contacts via WhatsApp. Section 3 discusses the performance evaluation of the system, focusing on the effectiveness of the proposed mobile application in improving passenger safety in taxi services across Metropolitan Lima. The proposed solution aims to be both innovative and inclusive, contributing to the development of safer and more reliable urban mobility systems.

2 RELATED WORKS

The comparative analysis was based on a structured literature review. A systematic search was conducted across IEEE Xplore, Mendeley, Scopus, and SciELO databases using keywords such as “speech recognition,” “keyword spotting,” “transport safety,” and “mobile application” across reviews published in 2020–2025. The review followed inclusion criteria that required peer-reviewed studies proposing AI- or natural language processing (NLP)-based systems for safety monitoring, emergency detection, or intelligent mobility. After screening abstracts and methodologies, 28 representative studies were selected for detailed comparison across four primary dimensions: accuracy, robustness, usability, and computational resource requirements. As a result, the main technological gaps that will be addressed by the proposed SmartSecurity system were identified.

The aspects of safety, emotion analysis, and efficiency in transportation and mobility systems have been recently addressed using advanced data mining, machine learning, and blockchain techniques. For instance, [1] proposed a data mining approach to derive implications for safety policies directed at taxi drivers, emphasizing the statistical analysis of incident data to devise preventive strategies. This approach differed from that proposed by [2], i.e., Wav2KWS, a model based on transfer learning for keyword recognition via voice representations. This model enabled command activation on mobile devices. Similarly, [3] integrated blockchain and machine learning to enhance authentication and security in shared transportation services, offering a robust framework for identity verification and operational transparency. The structure of this study was conceptually similar to that of [4], where blockchain was combined with vehicle demand analysis to optimize service allocation.

In another research domain, [5] conducted a comprehensive review of sentiment analysis and its methodological evolution. It was conceptually similar to [6], who developed an audiovisual transformer to detect keywords in videos using emotional cues as triggers for alert activation. Complementary to these studies, [7] applied dynamic spatial analysis to predict high-crime areas via geospatial data mining and developed a predictive framework for safety management. In addition, [8] developed

a system for monitoring voice pathologies using parallel deep models, primarily focusing on vocal health, and employed voice signal processing techniques similar to those used by [2] and [6]. Finally, [9] proposed a privacy-preserving system for taxi services that utilized approximate location techniques and mobile edge computing to ensure data confidentiality. Although these approaches differed in their focus on incident prevention, real-time monitoring, risk prediction, and privacy protection, they shared a common goal of leveraging intelligent technologies to improve safety and operational efficiency in transportation environments.

Due to advances in natural language processing, speech synthesis, and emotion recognition, techniques for improving human–machine interaction and the interpretation of acoustic signals have been developed. For instance, [13] explored mixed-emotion speech synthesis and developed a model that could generate voices expressing combined emotional states. In contrast, [14] developed a speech conversion system that enabled intensity-controlled emotional modification. [15] examined language recognition [15] using convolutional neural networks, and [16] analyzed audio and textual data from radio broadcasts for sentiment analysis. [17] applied a similar multimodal approach [17] and developed dialog systems that were sensitive to audiovisual contexts. [18] detected self-generated voice signals in noisy environments, similar to [19], who used graph-based detection methods to address this issue. [20] reviewed deep keyword detection systems, establishing conceptual links with the voice synthesis and emotional conversion studies [13] and [14]. Although the focus of these investigations varied across dialog systems, self-voice detection, and emotional analysis, they relied on deep audio representations as their technological foundation.

The development of voice recognition and analysis technologies has also evolved rapidly. [21] proposed a method for content retrieval in multilingual applications by employing NLP for semantic classification. In addition, [22] developed a hybrid CTC plus attention model for multilingual speech recognition in low-resource contexts. Meanwhile, [23] performed NLP-based big data analysis to extract structured information from unstructured sources. Similarly, [24] proposed a post-processing technique to enhance the readability of texts generated by speech recognition systems, and [25] proposed a recommendation model for digital learning via collaborative filtering and few-shot learning. [26] developed a model for identifying Ethio-Semitic languages using audio signals [26] and recurrent networks, and [27] developed continuous speech recognition models for identifying the Uzbek language. Similarly, [28] employed broadcasted residual learning to optimize keyword detection efficiency on mobile devices. [29] used sub-word tokenization to improve recognition performance in morphologically complex languages such as Malayalam.

Recent studies have also extended NLP and multimodal data analysis toward improving personal safety via threat detection and behavioral pattern recognition. For instance, [30] introduced a self-supervised multimodal attention network for speaker tracking that integrates both audio and video streams. [31] employed BERT-based language models for detecting phishing URLs, demonstrating the applicability of NLP techniques in cybersecurity. Although these studies reported applications across domains such as public safety, traffic monitoring, forensic analysis, and supervised learning, they shared a common foundation in deep learning, NLP, and multimodal data integration to extract knowledge from complex and unstructured sources. Table 1 summarizes and compares the most relevant proposed models similar to Smart Security across core parameters such as accuracy, robustness, usability, and computational resources. This comparison highlights the strengths and limitations of previous studies in areas such as audio-based threat detection, pattern recognition, and NLP and serves as the foundation for justifying SmartSecurity.

Table 1. Comparison of existing models with SmartSecurity

Proposal	Accuracy	Robustness	Usability	Source
Threat detection software based on sound detection	Accuracy: 96.6% False Negatives: 3.4%	High	Ease of use: High	[1]
A zero-shot learning-based audio classifier using semantic embeddings	Accuracy: 52.7% False Negatives: 47.3%	High	Ease of use: Low	[2]
A cab service system based on a hybrid model that integrates machine learning for trip acceptance prediction and blockchain to ensure transaction and user data integrity	Accuracy: Not specified False Negatives: Not specified	High	Ease of use: Low	[3]
Systematic review of the field of sentiment analysis. Keyword and community detection are used to identify emerging trends and approaches	Accuracy: Not specified False Negatives: Not specified	High	Ease of use: High	[4]
An audio-based automatic Ethio-Semitic language identification system using recurrent neural networks (RNNs)	Accuracy: 98.1% False Negatives: 1.9%	High	Ease of use: Low	[5]
Application of data mining techniques to analyze survey data and accident records	Accuracy: Not specified False Negatives: Not specified	Medium	Ease of use: Low	[6]
A speech transfer keyword detection system	Accuracy: 99% False Negatives: 1%	High	Ease of use: High	[7]
A transformer-based AVKT model to detect keywords in complete sentences using audio and video inputs	Accuracy: 99% False Negatives: 1%	High	Ease of use: Medium	[8]
An LPPM privacy protocol that replaces GPS location with sets of points of interest and uses the MinHash algorithm to calculate similarities between locations without revealing the exact location	Accuracy: Not specified False Negatives: Not specified	High	Ease of use: High	[9]
A hybrid active noise cancellation system with adaptive beamforming in cabs	Accuracy: Not specified False Negatives: Not specified	High	Ease of use: Medium	[13]
Application of machine learning models to assess driver safety perception and classify risk levels in cab services	Accuracy: 91.2% False Negatives: 8.8%	High	Ease of use: Low	[14]
An emotional speech synthesis system that generates real-time speech emotions	Accuracy: 90%–96% False Negatives: Not specified	High	Ease of use: High	[15]
Sequence-to-sequence model for emotional voice conversion with explicit control of the degree of emotion	Accuracy: 80% False Negatives: 20%	High	Ease of use: Medium	[16]
A language recognition system for detecting the Persian language based on convolutional neural networks	Accuracy: 80.54% False Negatives: 19.46%	Medium	Ease of use: Medium	[17]
A combined sentiment analysis model that fuses audio and text from radio broadcasts to detect negative emotions and opinions	Accuracy: 87.12% False Negatives: 12.88%	High	Ease of use: Low	[18]
A system for detecting and classifying real-time audio signals from the victim's environment	Accuracy: 90%–96% False Negatives: Not specified	High	Ease of use: High	[19]
A hybrid CNN–LSTM model for speech emotion recognition in South Asian languages	Accuracy: 82.3% False Negatives: 17.7%	High	Ease of use: Medium	[20]
An automatic speech recognition (ASR) system for the Assamese language using CMU Sphinx	Accuracy: 91.2% False Negatives: 8.8%	Medium	Ease of use: High	[21]
An ASR architecture based on DNN–HMM using a hybrid tokenization called syllable-byte pair encoding	Accuracy: 89.4% False Negatives: 10.6%	High	Ease of use: Low	[22]
A language model for Uzbek based on RNNs and statistical models to improve continuous Uzbek language recognition	Accuracy: 91.5% False Negatives: 8.5%	High	Ease of use: Medium	[23]
A prototype software based on audio analysis to detect threats	Accuracy: 80%–90% False Negatives: Not specified	High	Ease of use: High	[24]

(Continued)

Table 1. Comparison of existing models with SmartSecurity *(Continued)*

Proposal	Accuracy	Robustness	Usability	Source
A threat identification system based on sound detection	Accuracy: 91.81%–92.63% False Negatives: Not specified	High	Ease of use: High	[25]
An end-to-end hybrid model for multilingual speech recognition	Accuracy: 81.77%–89.78% False Negatives: Not specified	High	Ease of use: High	[26]
A speech emotion recognition model based on RNNs with an attention mechanism	Accuracy: 71% False Negatives: 29%	Medium	Ease of use: Medium	[27]
A speech emotion recognition system using spectrograms as the input and a hybrid 2D CNN-LST architecture	Accuracy: 93.4% False Negatives: 6.6%	High	Ease of use: Low	[28]
A speech recognition system in Arabic that improves previous models using deep neural networks (DNNs) with supervised training	Accuracy: 84.7% False Negatives: 15.3%	Medium	Ease of use: Medium	[29]
A new network architecture, known as BC-ResNet, based on broadcasted residual learning for efficient keyword detection in resource-constrained devices	Accuracy: 98% False Negatives: 2%	High	Ease of use: Medium	[30]
A comprehensive framework for the design of sentiment analysis models	Accuracy: Not specified False Negatives: Not specified	High	Ease of use: High	[31]

Based on the comparative analysis presented in Table 1, the reviewed studies were grouped into five principal categories to clarify the technological evolution and identify the areas with limited research: voice recognition and speech processing models, multimodal detection and emotion recognition systems, blockchain-based driver verification frameworks, privacy-preserving and edge computing architectures, and geospatial predictive crime analysis. Thus, existing contributions were comprehensively examined, and how the proposed SmartSecurity system addressed the limitations of these studies was discussed.

The first category, i.e., voice recognition and speech processing models, focused on optimizing keyword detection, acoustic robustness, and computational efficiency in real-time environments. For instance, [7] and [24] introduced transfer learning strategies and deep neural architectures for improving small-footprint keyword spotting, and [19] developed voice detection models for wearable devices based on multisensor data fusion. Recent studies have explored hybrid deep learning architectures and sub-word tokenization to improve recognition accuracy in morphologically complex languages. However, many of these approaches depend on controlled acoustic conditions or specialized hardware, limiting their scalability. SmartSecurity overcomes these constraints via the contextual fine-tuning of the Whisper model; it thus enables accurate and low-latency voice command detection in realistic taxi environments that are characterized by variable accents and background noise. This adaptation ensures consistent performance using only standard mobile hardware.

The second category, i.e., multimodal detection and emotion recognition systems, integrates acoustic, visual, and affective features to improve the interpretation of human behavior. For instance, [13] and [15] developed deep learning models that could identify emotions such as fear and stress via combined audio–visual representations. Advanced approaches such as the audio–visual keyword transformer (AVKT) achieve high accuracy in controlled conditions but require cameras and specialized sensors that are impractical in mobile contexts. In contrast, SmartSecurity adopts a purely acoustic approach and infers emotional states such as panic and distress directly from voice data using NLP embeddings and spectral analysis.

This configuration enables reliable emotion inference without compromising privacy or system efficiency.

The third category, i.e., blockchain and driver verification systems, emphasizes transparency and data integrity in ride-hailing services. For instance, [3] proposed blockchain-based architectures that verified driver identities using distributed ledgers and machine learning components. Although these frameworks strengthened transaction-level trust, they operated primarily as post-incident mechanisms rather than proactive safety solutions. SmartSecurity advances this concept by integrating secure data management and real-time alert functions. It combines authentication with immediate emergency communication between passengers and authorities to prevent potential threats.

Privacy preservation and edge computing in transportation is another research category, aimed at safeguarding sensitive user information while maintaining system responsiveness. For instance, [16] combined mobile edge computing with anonymization methods such as MinHash to conceal precise geolocation data. Sun et al. proposed federated learning techniques for decentralized training under privacy constraints. These studies reinforced the necessity of designing architectures that protect user confidentiality without compromising on their performance. SmartSecurity applies this principle by complying with ISO/IEC 27001 and 27701 standards, using encryption, anonymized event storage, and controlled cloud access to ensure that data security coexists with real-time responsiveness.

Finally, studies on geospatial and predictive crime analysis have explored the application of data mining and spatial modeling to detect risk zones and behavioral anomalies. For instance, [6] and [8] demonstrated how machine learning models identified urban areas with higher crime probability by correlating taxi routes and demographic variables. Although these methods provide valuable insights into crime prevention, they are often restricted to offline analysis and lack direct intervention mechanisms. SmartSecurity enhances these approaches by incorporating real-time GPS monitoring and geofencing, which automatically activate alerts when users enter high-risk areas. This transformation converts predictive analysis into an active preventive measure that directly impacts the safety of passengers.

The aforementioned categories reveal considerable progress in isolated areas of intelligent transportation research, including speech recognition, multimodal emotion analysis, blockchain authentication, and geospatial modeling. However, the absence of integrative systems that can operate under real-world constraints remains evident. SmartSecurity consolidates these independent advances into a unified AI framework that combines NLP, cloud computing, and geospatial intelligence to provide a scalable, secure, and context-aware emergency response platform. This platform is specifically adapted to the conditions of taxi services in Metropolitan Lima.

The comparative analysis summarized in Table 1 emphasizes the strengths and limitations of existing approaches for enabling transportation safety. Although many studies achieved notable accuracy, few reported latency metrics, end-to-end response times, and integration with real-time alert mechanisms. Most models were also tested under controlled laboratory conditions, which reduced their reliability in dynamic and noisy environments such as taxi interiors. In contrast, SmartSecurity combines real-time keyword recognition with geolocation and multichannel communication via WhatsApp, SMS, and emergency calls; this makes it a practical and deployable system that offers high accuracy, responsiveness, and usability. These advantages demonstrate the relevance of SmartSecurity as a concrete and innovative contribution to intelligent transportation safety research.

3 SYSTEM DESIGN

3.1 Architecture

Figure 1 shows the integrated architecture of SmartSecurity. It combines physical, network, and logical components to provide a robust and scalable solution designed for emergency assistance, particularly for enhancing the safety of taxi users in Metropolitan Lima.

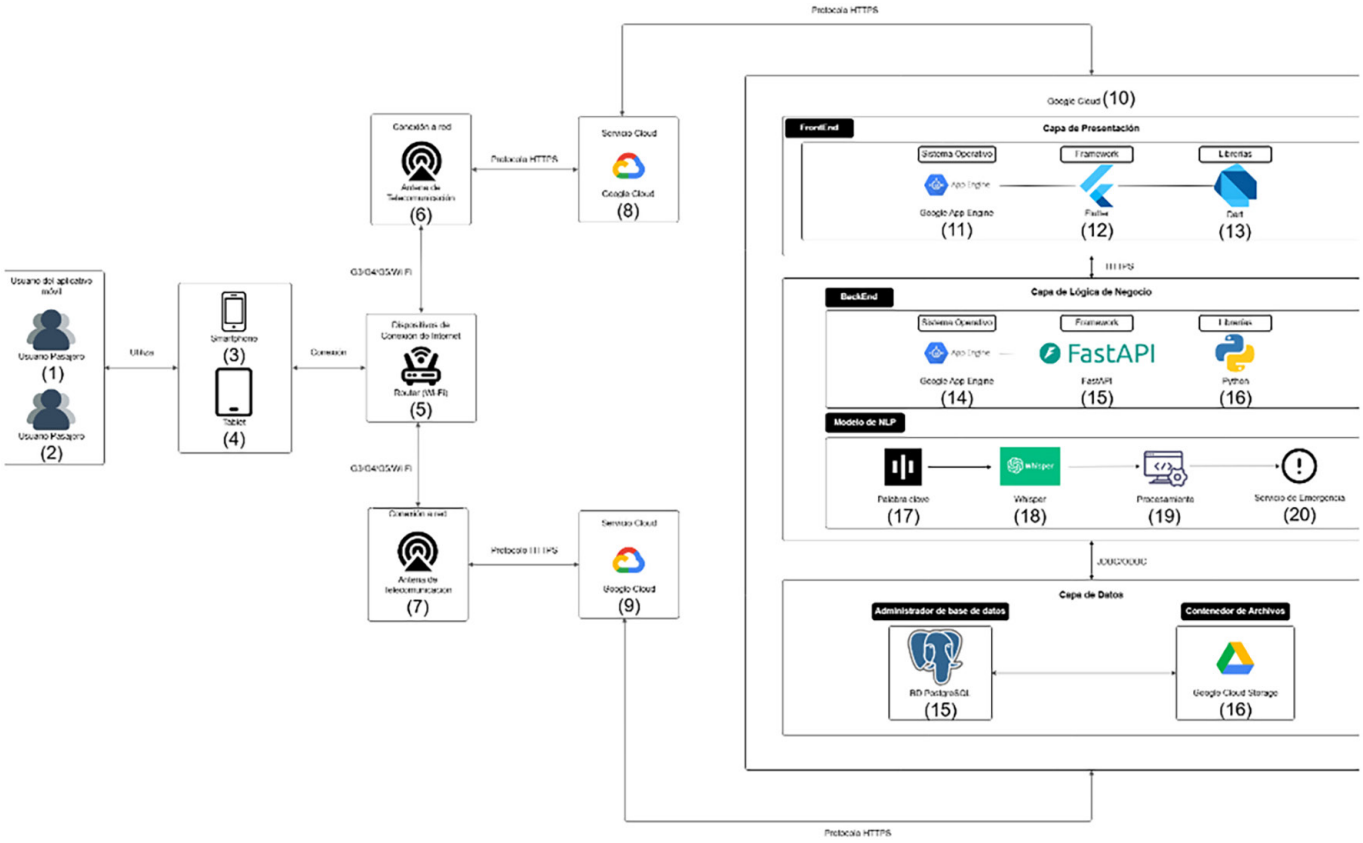

Fig. 1. Integrated architecture of SmartSecurity

At the physical layer, the interaction begins with the passengers, identified as components (1) and (2), who operate the mobile application using smartphones (3) or tablets (4). These devices are connected to the Internet either via Wi-Fi routers (5) or directly via mobile networks that support 3G, 4G, or 5G technologies, facilitated by telecommunication antennas (6) and (7). This infrastructure ensures continuous and reliable connectivity, thereby enabling uninterrupted system operation even when the user is in motion.

After the connection is established, the cloud infrastructure is accessed via Google Cloud components (8), (9), and (10); these components serve as the central platform for service deployment, data processing, and secure storage. At the presentation layer, the Google App Engine (11) functions as the execution environment for scalable cloud applications. The user interface is developed using Flutter (12), a cross-platform framework that ensures a seamless user experience. Interactive elements were added within the application using the programing language Dart (13).

Within the business logic layer, the backend is also deployed on the Google App Engine (14), and the development framework is significantly upgraded

by adopting FastAPI (15). This modern, high-performance tool built in Python (16) enables the rapid and secure exposure of RESTful services. The logic layer manages all critical operational workflows, including user authentication, secure route configuration, emergency contact administration, and the automatic coordination of alert mechanisms.

The NLP module, a fundamental component of this layer, continuously listens for and recognizes the user-defined safety keyword (17). The system employs the Whisper engine (18), which specializes in robust speech recognition under noisy environmental conditions. When a predefined keyword is detected, the system triggers the internal processing routines (19) necessary to perform emergency actions such as sending automated messages, initiating calls, and transmitting the user's real-time location to predefined emergency contacts or services (20).

At the data layer, the system ensures persistence, consistency, and traceability of all recorded events and configurations. In addition, structured information such as user profiles, alert histories, and personalized settings is stored in a PostgreSQL database (15). Concurrently, unstructured data such as audio files, activity logs, and multimedia evidence are stored within Google Cloud Storage (16). Communication between layers is established via secure JDBC, ODBC, and HTTPS protocols, thereby guaranteeing data integrity, confidentiality, and high availability throughout the system lifecycle.

This integrated architecture constitutes a modern, modular, and resilient technological ecosystem that effectively combines AI, cloud computing, and telecommunications. It thus meets contemporary public safety requirements and maintains scalability for future integration with municipal platforms, police networks, and emergency response systems. With this architecture, SmartSecurity provides a reliable and adaptive framework that can efficiently respond to risk scenarios in urban transportation environments.

3.2 Methodology

An applied research design with an experimental–developmental approach (Design Science Research Framework) was employed herein. In the design phase, the SmartSecurity prototype was developed, and controlled laboratory testing was conducted to measure the accuracy and response time of the model. Quantitative metrics were analyzed to evaluate the model performance under realistic acoustic conditions representing taxi environments.

3.3 Dataset

Data obtained from internal sources generated directly via user interaction with the mobile application formed the basis for emergency detection and real-time alert management. These data mainly comprised GPS coordinates configured by users to define safe zones or report incidents; emergency contact details, required for automated alert transmission; and personalized keywords used for activating the emergency protocol via voice commands detected by the NLP module based on the Whisper model.

Data related to user trips were collected, including timestamps, routes, mobility patterns, and emergency events reported during service usage. These data reflected the actual user behavior in high-risk contexts and enriched the system with realistic and context-aware information. Contrary to existing AI models that are trained on

public datasets, SmartSecurity relied on proprietary and contextualized data. As a result, models specifically adapted to the social, linguistic, and acoustic conditions of Metropolitan Lima and the dynamics of its taxi services were developed.

Data management was conducted securely and efficiently using the Google Cloud Platform (GCP) infrastructure. Structured data, such as user records, incident reports, and travel routes, were stored in a PostgreSQL database, and unstructured data, including emergency audio activations, system logs, and related documents, were preserved in Google Cloud Storage. This architecture ensured high availability, scalability, and resilience of the system, supported by secure JDBC and ODBC connections that mediated communication between the data and business logic layers.

The collected data were processed and divided into training, validation, and testing subsets, in accordance with best practices in AI model development. This was done to develop the speech recognition system, which enabled accurate keyword detection in noisy environments. This segmentation prevented model overfitting and promoted model generalization to real-world scenarios. The training dataset was used to optimize the parameters of the Whisper-based model, the validation dataset was used to fine-tune the model performance and select the most effective architecture, and the testing dataset was used to objectively evaluate the model performance before its deployment in production.

Throughout the data lifecycle, mechanisms for anonymization, encryption, and access control were implemented from the collection stage to storage and processing. These practices complied with international standards for information security and privacy management, specifically ISO/IEC 27001 and ISO/IEC 27701. This approach safeguarded sensitive user information and strengthened user trust in the platform by demonstrating adherence to globally recognized data protection principles.

3.4 Model

As shown in Figure 2, the SmartSecurity model integrated AI, NLP, and real-time emergency management to detect risk situations and activate automatic alerts using voice commands. The logical architecture demonstrates interactions among the user, mobile application, and cloud-based services, forming a cohesive and effective framework for providing taxi assistance within Metropolitan Lima.

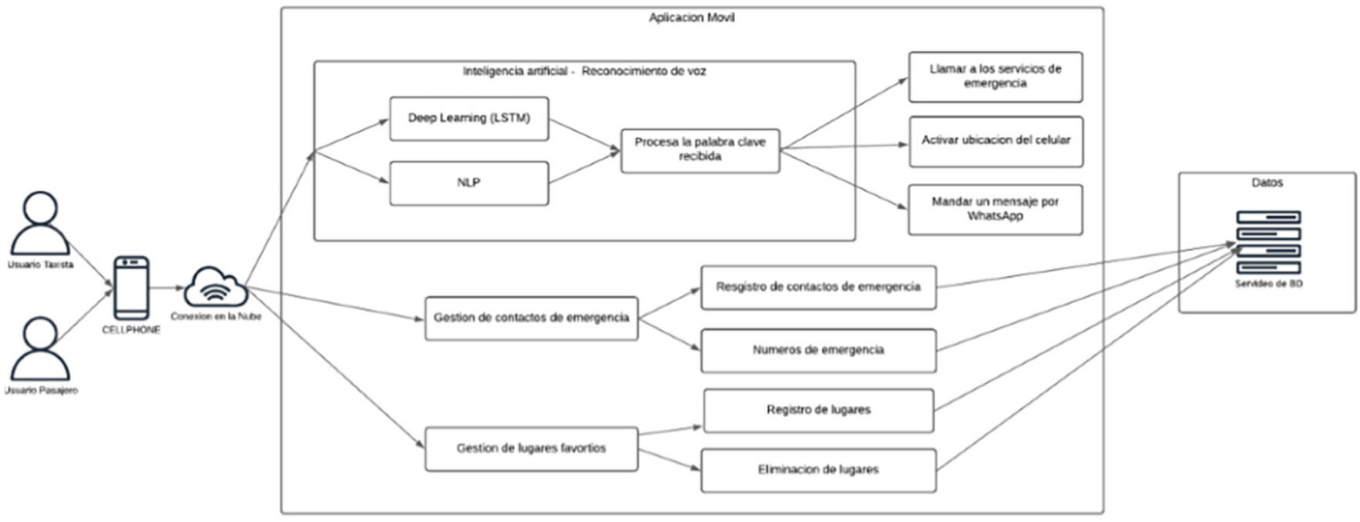

Fig. 2. Logical architecture of SmartSecurity

The interaction process begins with the passenger, who accesses the mobile application installed on their smartphone. The application operates in an active listening mode and continuously monitors the audio input via a keyword recognition module. This module employs NLP techniques and utilizes the Whisper model that is specialized in detecting activation commands in real time and functioning effectively in noisy environments.

Upon identifying the user's personalized keyword, the system automatically performs several critical operations. These include initiating a direct call to emergency services, activating the device's geolocation, and sending alert messages via WhatsApp to the user's predefined emergency contacts. These processes are executed autonomously, ensuring an immediate and reliable response without requiring manual user intervention.

This application also supports the management of secure locations, thereby enabling users to register, modify, or delete areas of interest, as well as maintain a list of emergency contacts. This functionality simplifies the administration of individuals who should be notified in the event of an incident. All the related data are centralized within the system's database, where emergency events and activity records are securely logged to ensure traceability and accountability.

Speech signal preprocessing is a crucial stage in the model's operation. This process involves normalizing audio inputs to standardize signal levels, removing background noise, and converting the speech into spectral representations (spectrograms) that allow more precise interpretation of acoustic features. The resulting data are transformed into Mel-frequency cepstral coefficients (MFCCs) and embeddings generated by the Whisper model. These data capture the phonetic and semantic characteristics of the voice signal.

To guarantee model robustness and generalization, the collected dataset is divided into training, validation, and testing subsets. The training dataset is used to adjust model parameters, the validation dataset is used to optimize the model performance and select the most effective architecture, and the testing dataset is used to objectively assess the model accuracy before its deployment to production. This process ensures a balanced learning cycle and minimizes the risk of model overfitting.

Data are managed and stored via the GCP infrastructure. Structured data, including user profiles, incident reports, and location records, are managed with PostgreSQL, while unstructured data, such as audio files and system activity logs, are stored in Google Cloud Storage. The system design ensures scalability, availability, and security, allowing for efficient and protected data access using the secure JDBC and ODBC communication channels.

Figure 3 shows the graphical interfaces of SmartSecurity, showing the user's direct interaction with the system's core functionalities. These interfaces were designed with an emphasis on usability, simplicity, and rapid responsiveness to emergency scenarios, aligning with the project's overarching objectives of safety enhancement and real-time assistance.

Fig. 3. Application screens. (a) login view, (b) registration view, and (c) start view as a registered passenger or a cab driver

The second interface is the registration screen that enables users to create new accounts. Unlike the login interface, this screen requires more detailed information, including the user's email address, password, phone number, and password confirmation to prevent authentication errors. It also includes a checkbox for accepting the terms and conditions, ensuring compliance with privacy and personal data protection policies. This process allows for the registration of essential information required for the activation of alerts and the accurate identification of users within the system.

The third interface is the main security screen that consolidates the vital operational functions of SmartSecurity. Upon access, the user is greeted with a personalized message and a direct question, "Are you safe?"; this message reinforces the preventive and reactive nature of the application. The interface also features a large and highly visible "HELP" button that enables immediate requests for assistance with a single touch. It also provides an option to activate voice recognition, allowing continuous monitoring for previously configured keywords and facilitating the automatic triggering of alerts in situations wherein manual interaction may not be possible.

This screen also enables the users to input and manage their current route, allowing for the anticipation of potential risk zones along the journey. It provides shortcuts for users to save frequently used locations such as "Home" and "Work," optimizing the management of safe zones. An interactive map displays the user's real-time location at the top of the interface, offering spatial awareness and reinforcing the perception of safety and control. The lower section contains a navigation bar that grants quick access to the "Service" and "Support" modules, enabling users to efficiently navigate and access complementary functionalities.

These interfaces collectively facilitate an intuitive, efficient, and safety-oriented user interaction experience, supporting both preventive and reactive responses to emergency situations. Each screen design directly reflects the requirements defined in the SmartSecurity model, ensuring a seamless and reliable user experience in high-risk environments.

3.5 Training

The speech recognition model developed for SmartSecurity is based on the pretrained Whisper model architecture, a robust NLP framework specialized in speech transcription under adverse acoustic conditions. Contrary to conventional approaches that require training models from the ground up, SmartSecurity employs a transfer learning strategy via contextualized fine-tuning. This adaptation process allows the Whisper model to align with the linguistic, environmental, and operational characteristics of Metropolitan Lima and to recognize specific emergency commands issued within taxi environments.

This approach primarily employed a proprietary dataset composed of user recordings collected while configuring personalized keywords within the application. These recordings were obtained under realistic conditions, incorporating variations in accents, intonations, speech rates, and the presence of ambient noise such as traffic, music, and background conversations. To enhance phonetic diversity and improve model robustness, the dataset was supplemented with a filtered subset obtained from Mozilla's Common Voice Corpus 19.0 [32]; it contained Spanish samples with phonetic features similar to those prevalent in the Peruvian context.

Audio data preprocessing was a critical step for ensuring training quality and consistency. Each audio sample was resampled to a frequency of 16 kHz for ensuring compatibility with the Whisper model while balancing acoustic resolution and computational efficiency. Subsequent processing included amplitude normalization, adaptive spectral noise reduction, and conversion of signals into log-Mel spectrograms for capturing the spectro-temporal representations of speech. Data augmentation techniques, including time masking, frequency masking, and slight pitch shifts, were employed to simulate various acoustic conditions and enhance the generalization capability of the model.

Feature extraction combined MFCCs, which captured the spectral envelope of speech, with embeddings generated by the intermediate layers of the Whisper model. These embeddings encapsulated both phonetic and semantic features relevant to keyword recognition. By combining these representations, data dimensionality was reduced while preserving critical acoustic patterns required for accurate detection.

During fine-tuning, the lower layers of the Whisper encoder were frozen to retain the phonetic knowledge acquired during pretraining. The upper layers of the decoder and classification head were adjusted to optimize task-specific performance. A sigmoid activation module was incorporated to enable the binary detection of the presence or absence of the activation keyword. This eliminated the need for complete text decoding during each inference, thereby improving response latency.

The model was trained on the GCP using Google Compute Engine instances equipped with NVIDIA T4 GPUs. The Adam optimizer, known for its high efficiency in handling noisy gradients, was employed with an initial learning rate of 0.0005; the learning rate was progressively reduced via a cosine annealing schedule after the validation improvements were stabilized. Model training was performed with a batch size of 32, balancing convergence speed and gradient stability. The process was limited to 15 epochs, and an early stopping criterion was triggered if the F1-score did not improve over three consecutive iterations.

To ensure the validity of evaluation, the dataset was partitioned into 70% for training, 15% for validation, and 15% for testing; these subsets were maintained independently to preserve the integrity of model evaluation. In the validation phase, hyperparameters were adjusted to mitigate overfitting risks. The model was finally

evaluated on the test subset to objectively determine model generalization under real-world operational conditions.

Table 2 summarizes the technical parameters established during the training of the speech recognition model, such as the base architecture, dataset composition, acoustic features, preprocessing procedures, training configuration, and evaluation metrics. These specifications form the technical foundation of the speech recognition system integrated within the SmartSecurity environment, thereby ensuring high accuracy, resilience, and real-time responsiveness in emergency scenarios.

Table 2. Technical parameters of voice model training

Parameter	Value/Configuration
Base model	Whisper model (base version pre-trained in multiple languages)
Dataset used	User-generated proprietary data (custom keywords) + subset of Common Voice Corpus 19.0
Sampling frequency	16,000 Hz
Preprocessing	Audio normalization, noise reduction, spectrogram generation, and MFCCs.
Feature extraction	MFCCs + Whisper embeddings
Transcription tokenization	Byte pair encoding with domain-specific vocabulary for short commands
Optimization	Adam optimizer
Learning rate (LR)	0.0005
Number of epochs	10–15 (with early stopping in validation)
Batch size	32
Dataset division	70% training, 15% validation, and 15% testing
Training environment	GCP (App Engine + Cloud Storage)
Metrics evaluated	Accuracy, recall, F1score, word error rate, and inference latency
Average inference time	500 ms per sample on a mobile device (in controlled tests)

3.6 Evaluation and statistical analysis

The effectiveness of the speech recognition model implemented in SmartSecurity was assessed by employing several key statistical metrics to quantify its performance in detecting activation keywords under emergency conditions. These metrics provide a comprehensive evaluation of the accuracy, detection capability, and operational reliability of the model. In critical scenarios such as those targeted by this application, where timely activation may be decisive for the user's safety, establishing robust indicators that validate the system's effectiveness is crucial.

The model achieved an outstanding accuracy of 90%. This metric represents the proportion of correctly detected keywords among those detected by the model. In practical terms, accuracy measures how precisely the system identifies legitimate activation commands while avoiding erroneous alerts. Such high accuracy ensures that nearly all alerts generated by the model are genuine keyword utterances, thereby minimizing false activations. This result is particularly relevant for maintaining user confidence because it reduces the likelihood of false alarms that can desensitize users or induce unnecessary stress.

Sensitivity, or the true positive rate, reflects the capacity of the model to correctly identify the keywords spoken by users. The system achieved a sensitivity of 90%, indicating that it successfully detected most keywords under real-world conditions, even in acoustically challenging environments such as the interior of a moving taxi. This performance is critical in emergency contexts, as a missed detection can compromise user safety. Such a high sensitivity demonstrates the robustness of the model in recognizing commands spoken under stress, at low volume, or with variable noise levels.

In addition, the false-positive rate was found to be nearly zero. This measure quantifies the frequency with which the model incorrectly classifies non-keywords as activation commands. By maintaining this rate at minimal levels, unnecessary interruptions or the inappropriate triggering of emergency protocols can be prevented. A high false activation rate can diminish user trust and lead to neglect of genuine alerts. Thus, the extremely low rate confirms the reliability of the model in accurately distinguishing activation keywords from unrelated speech.

The F1-score was calculated to obtain a balanced indicator that integrates both precision and sensitivity into a single performance measure. The model achieved an F1 score of 0.90, representing an optimal equilibrium between its ability to correctly identify keywords and its capacity to minimize detection errors. The F1-score reinforces the overall robustness and consistency of the system across multiple evaluation dimensions. It is particularly meaningful in safety-critical applications, wherein both false positives and false negatives can considerably affect users.

Table 3 summarizes the primary metrics used to evaluate the performance of the speech recognition model. These metrics confirmed that the system operated with high accuracy, strong sensitivity, and negligible false activation rates. This ensured dependable real-time keyword recognition and contributed to the overall reliability of SmartSecurity in emergency scenarios.

Table 3. Metrics table

#	Metrics	Equation
1	Accuracy	$Prec. = \frac{TP}{TP + FP}$
2	Sensitivity (true positive rate)	$Sens. = \frac{TP}{TP + FP}$
3	False positive rate	$FPR = \frac{FP}{FP + TN}$
4	F1-score	$F1 = 2 \times \frac{Prec. \times Sens.}{Prec. + Sens.}$

Statistical analysis revealed that the voice recognition model integrated into SmartSecurity was highly efficient, accurate, and reliable, as confirmed by the accuracy and sensitivity of 90%, a false-positive rate close to zero, and an F1 score of 0.90. These values also indicated that the system can function effectively in real emergency scenarios by providing timely, precise, and error-free responses. These outcomes validate the inclusion of the model as a core element of the security architecture of SmartSecurity, ensuring alignment with rigorous standards of quality, reliability, and technological excellence.

4 RESULTS

Table 4 summarizes the most relevant criteria of SmartSecurity, emphasizing its technological capabilities and performance under critical operating conditions.

Table 4. Proposal criteria

Criteria	NPL to Request Assistance in the Event of Criminal Acts
Accuracy and error rate	Accuracy: 90% False Negative: 10%
Response time	Latency: 1 s Total response time: 2 s
Algorithms employed	Algorithm accuracy: 90% Robustness: High
Usability and experience	Ease of Use: Very High
Alert the media	Reliability: 100% Multichannel: Yes
Implementation cost	Initial Cost: Low Maintenance Cost: Low

The integration of the speech recognition model in SmartSecurity, developed to assist users in cases of criminal activity during taxi services in Metropolitan Lima, yielded highly promising results in terms of accuracy, response speed, robustness, and ease of use. The system was built using a combination of NLP techniques and advanced deep learning models, supported by the cloud infrastructure of GCP. The system was trained via a fine-tuning approach applied to the pretrained Whisper model, which was optimized for keyword detection in noisy environments with diverse accents.

Real and contextualized recordings were utilized, preprocessed, and standardized during training via resampling at a frequency of 16 kHz. These data were enriched with MFCCs, enabling precise and efficient acoustic representation. The data were tokenized into short, context-aware commands such as “help” or “distress.” The Adam optimizer was employed with a batch size of 32, which yielded a stable and efficient learning curve even under challenging auditory conditions such as those inside a moving taxi.

The performance outcomes (Table 4) confirmed the robustness of SmartSecurity. It achieved an accuracy of 90%, with a false negative rate of 10%, ensuring a high detection rate for actual emergency commands. In terms of temporal performance, the model demonstrated a latency of ~1 s from the moment of keyword detection and a total response time of 2 s, including processing and activation of alert mechanisms. This performance positions SmartSecurity within the ideal parameters for critical real-time applications.

In terms of algorithmic performance, the model consistently achieved 90% accuracy and demonstrated strong resilience to acoustic interference. The user experience was reported as highly intuitive and responsive, which is essential in stressful or high-risk contexts. The alerting system also achieved full compatibility with multiple communication channels, thereby enabling simultaneous activation of calls, messages, and notifications, including via WhatsApp.

Finally, the implementation cost was low both during the initial deployment and in subsequent maintenance phases. This efficiency was achieved using scalable cloud services, open-source frameworks, and optimized computational resources. These results indicate that SmartSecurity satisfies the performance and speed standards required for a reliable emergency assistance platform.

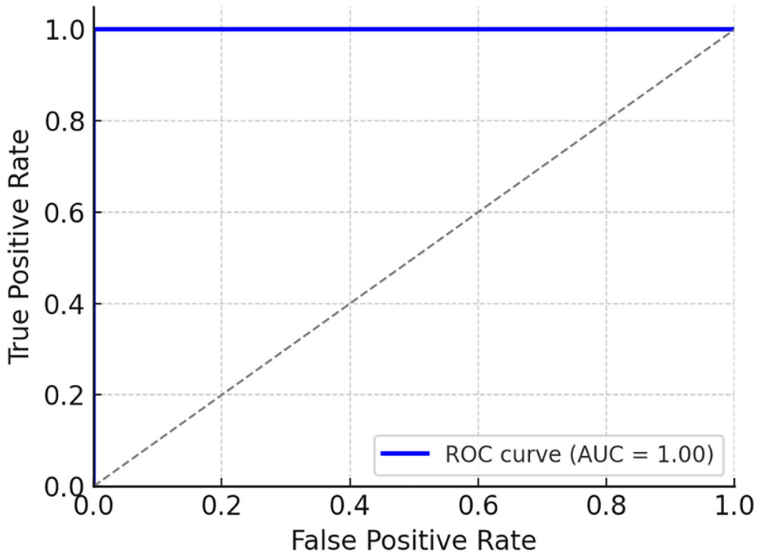

Fig. 4. Receiver operating characteristic (ROC)

The model performance was further evaluated using the receiver operating characteristic (ROC) curve (see Figure 4), which depicts the relationship between the true positive rate (sensitivity) and the false-positive rate across various decision thresholds. In this evaluation, the area under the curve (AUC) reached a value close to 1.00; this indicated that the model could efficiently differentiate between positive cases, represented by spoken activation keywords, and negative cases, represented by irrelevant sounds or phrases. This result was consistent with the sensitivity values obtained during practical testing and with the recorded false negative rate of 10%, demonstrating a strong detection capability. However, the model remained subject to potential enhancement in extremely noisy environments or when processing atypical pronunciations.

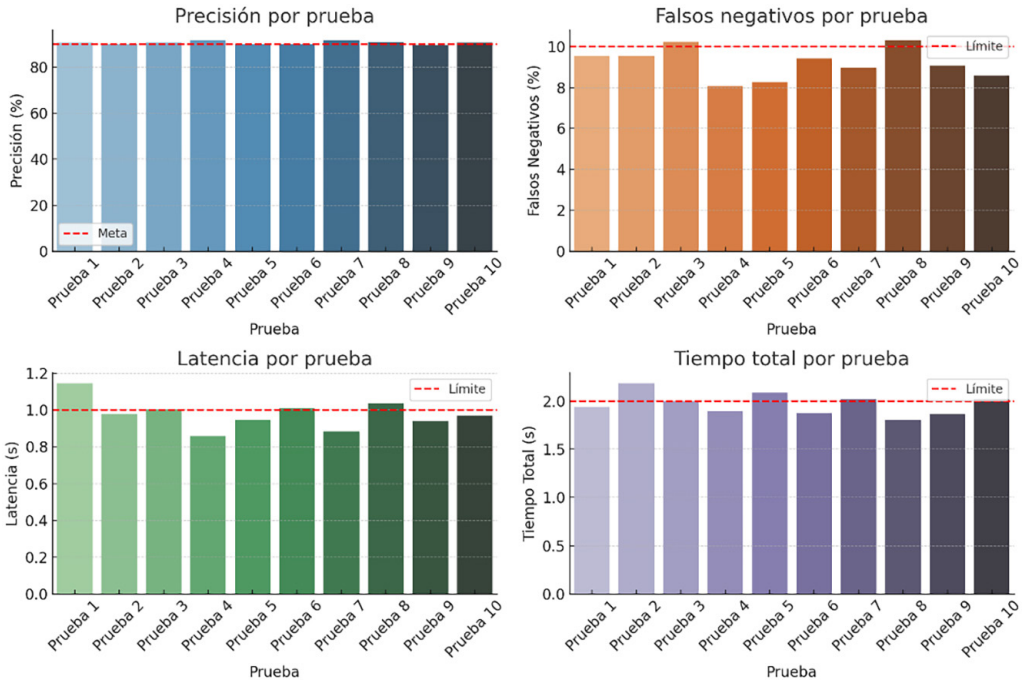

Fig. 5. Total time per test

Figure 5 shows a set of graphs that simulate the performance of SmartSecurity across 10 consecutive test iterations, verifying compliance with the criteria established in the proposed design. Results showed that model accuracy consistently remained ~90% and the false negative rate remained ~10%, demonstrating its strong ability to detect activation keywords. In addition, the latency values averaged to ~1 s, and the total system response time did not exceed 2 s, ensuring rapid and efficient activation of emergency alerts. These findings indicated that SmartSecurity operated with reliability, speed, and precision, fulfilling the technical standards required for deploying the proposed system in safety-critical and high-risk environments.

5 DISCUSSION

The results represent a significant advancement in keyword detection for emergency situations within taxi environments characterized by background noise, constant movement, and limited privacy. The implementation of a model based on Whisper AI, an advanced speech recognition system developed by OpenAI and trained with real voice data, demonstrated technical improvement over previously proposed approaches and a highly applicable solution for urban contexts where passenger safety remains a critical concern. The model achieved an accuracy of 90%, a latency of 1 s, and a false negative rate of 10%. Thus, the model is the most robust and reliable alternative for real-time deployment, particularly on mobile devices operating under noisy and resource-constrained conditions.

From an analytical perspective, this study strengthened the field of transportation safety and introduced a new approach of using AI models trained with open datasets, such as Common Voice Corpus 19.0, rather than closed or less representative collections. By including diverse accents and speech patterns in this dataset, a more inclusive and adaptive model was developed, and detection bias was reduced; this ensured broader applicability of the model across different user profiles. Contrary to previous studies, which often relied on controlled recordings or small populations, this study highlights the importance of training systems with heterogeneous and context-rich data to enhance real-world performance.

Compared with previously proposed models, SmartSecurity presents clear advantages across several critical areas. Studies such as [1], [13], and [14] provided solid technical frameworks; however, they did not report detailed latency or total response time metrics, limiting their evaluation as viable solutions for time-sensitive emergencies. In contrast, SmartSecurity delivered precise and transparent results. The latency consistently remained below 1 s, and the total response time was 2 s, enabling the near-instantaneous activation of emergency assistance. Even more sophisticated approaches, such as that reported in [16], which relied on pretrained models, do not provide comprehensive information regarding processing time. This shortcoming reinforces the technical and practical superiority of SmartSecurity in this dimension.

Another essential area of differentiation is the usability and customization of the proposed system. Unlike existing systems [18], which lack flexibility in configuring alert channels or command parameters, SmartSecurity enables users to define their own activation keywords. As a result, it achieved high recognition accuracy and user familiarity. Its integration with WhatsApp and GPS offers an additional advantage by enabling immediate communication with emergency contacts; in contrast, existing models rely solely on internal or single-channel alerts. This multichannel capacity enhances the overall reliability of SmartSecurity and aligns with a fundamental principle in emergency management: ensuring that alerts are delivered rapidly and effectively.

Beyond its technical contributions, this study offers tangible benefits for clinical, public safety, and social applications. The proposed system can be adapted to support vulnerable populations, including older adults, individuals with disabilities, and women at risk. Its adaptable design also allows implementation in other contexts such as public transport, residential monitoring, or healthcare environments, where quick voice activation can prevent or mitigate critical events. Owing to its lightweight architecture and deployment on the GCP, this system remains scalable and accessible for low-cost mobile devices without requiring specialized hardware. This feature enhances both affordability and feasibility of SmartSecurity, surpassing other models [15] that require costlier components.

From a research perspective, this study establishes a solid foundation for the future exploration of speech models operating in complex acoustic conditions by proposing a reproducible experimental framework. It also opens the possibility of integrating contextual attention mechanisms or leveraging synthetic data augmentation to expand model generalization.

In conclusion, this study demonstrates a tangible improvement in keyword detection for high-risk situations and considerably contributes to existing literature by presenting a practical, adaptable, and efficient solution. The findings validate the relevance and applicability of SmartSecurity in real-world and clinical settings. The study also offers opportunities for further advancements aimed at consolidating the effectiveness of intelligent safety systems in diverse social and urban contexts.

6 CONCLUSIONS

In this study, the technical and practical feasibility of an AI-based mobile application designed to assist users during criminal activities in taxi services within Metropolitan Lima was demonstrated. By employing a voice detection approach, the proposed system achieved an accuracy of 90% and a response latency of ~1 s, enabling the real-time activation of alerts. These results confirmed that SmartSecurity could respond rapidly and reliably to emergency situations, meeting the essential safety requirements demanded in urban environments.

Its main advantage was the integration of Whisper AI, a speech recognition model developed by OpenAI, which enabled precise and robust detection of vocal commands even under noisy conditions. Due to this architecture comprising multilingual encoders and decoders trained on diverse acoustic datasets, Whisper AI facilitated the creation of a system adaptable to various environmental contexts, such as those found inside a moving taxi. SmartSecurity was developed with a strong commitment to data privacy, adhering to international standards such as ISO/IEC 27001. This ensured that functionality and ethical compliance were maintained throughout its operation.

SmartSecurity enhanced passenger safety during transportation, setting an important precedent for the responsible application of AI in crime prevention and real-time citizen assistance. Its scalable architecture, supported by technologies such as GCP, Flutter, and Python, enabled it to adapt seamlessly to new scenarios without requiring extensive structural modification. The ability to customize activation keywords and operate across multiple communication channels, including WhatsApp, considerably the usability and user acceptance of SmartSecurity.

Despite these promising results, certain limitations were identified with the proposed system. Although a high level of accuracy was achieved, the false negative rate of 10% indicated potential for further improvement, particularly under extreme noise conditions or when processing atypical pronunciations. In the

future, these challenges can be addressed by employing advanced techniques such as attention-based mechanisms, synthetic data generation, and deeper neural architectures.

In addition, expanding the linguistic coverage of SmartSecurity can facilitate its implementation across different regions and cultural contexts. Similarly, the framework can be adapted to other risk scenarios, such as domestic violence, street assaults, and urgent medical emergencies, by making only minimal adjustments to its existing architecture. Integrating the solution with official public safety platforms can further enhance its social impact by establishing a direct communication channel between citizens and emergency authorities.

This study demonstrates that AI can be applied responsibly, efficiently, and effectively to address real-world public safety challenges. SmartSecurity is technologically sound and cost-effective as well as ethically sustainable and socially relevant, marking a significant step toward the proactive and conscious use of intelligent systems for the collective well-being of the society.

7 ACKNOWLEDGMENT

The authors are grateful to the Dirección de Investigación of the Universidad Peruana de Ciencias Aplicadas for the support provided for this study work through the A-202-2025 incentive.

8 REFERENCES

[1] 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

[2] H. Xie and T. Virtanen, "Zero-shot audio classification via semantic embeddings," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 29, pp. 1233–1242, 2021. https://doi.org/10.1109/TASLP.2021.3065234

[3] Z. Shahbazi and Y.-C. Byun, "Blockchain and machine learning for intelligent multiple factor-based ride-hailing services," *Computers, Materials & Continua*, vol. 70, no. 3, pp. 4429–4446, 2022. https://doi.org/10.32604/cmc.2022.019755

[4] J. Cui, Z. Wang, S.-B. Ho, and E. Cambria, "Survey on sentiment analysis: Evolution of research methods and topics," *Artificial Intelligence Review*, vol. 56, pp. 8469–8510, 2023. https://doi.org/10.1007/s10462-022-10386-z

[5] A. A. Alemu, M. D. Melese, and A. O. Salau, "Towards audio-based identification of Ethio-Semitic languages using recurrent neural network," *Scientific Reports*, vol. 13, no. 1, p. 19346, 2023. https://doi.org/10.1038/s41598-023-46646-3

[6] J. Park, S. Lee, C. Oh, and B. Choe, "A data mining approach to deriving safety policy implications for taxi drivers," *Journal of Safety Research*, vol. 76, pp. 238–247, 2021. https://doi.org/10.1016/j.jsr.2020.12.017

[7] S. Seo, H. Oh, and Y. Jung, "Wav2KWS: Transfer learning from speech representations for keyword spotting," *IEEE Access*, vol. 9, pp. 80682–80691, 2021. https://doi.org/10.1109/ACCESS.2021.3078715

[8] G. Hajela, M. Chawla, and A. Rasool, "Crime hotspot prediction based on dynamic spatial analysis," *ETRI Journal*, vol. 43, no. 6, pp. 1058–1080, 2021. https://doi.org/10.4218/etrij.2020-0220

[9] M. Alhussein and G. Muhammad, "Automatic voice pathology monitoring using parallel deep models for smart healthcare," *IEEE Access*, vol. 7, pp. 46474–46479, 2019. https://doi.org/10.1109/ACCESS.2019.2905597

[10] I. D. Wahyono *et al.*, "Emotion detection based on column comments in material of online learning using artificial intelligence," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 16, no. 3, pp. 82–91, 2022. https://doi.org/10.3991/ijim.v16i03.28963

[11] A. Helen Victoria, P. Supraja, V. M. Gayathri, M. Sukeri Khalid, and N. H. B. Sulaiman, "Visualization for emotion detection in mobile-based land monitoring using user-centred design," *International Journal of Interactive Mobile Technologies (iJIM),* vol. 17, no. 4, pp. 21–36, 2023. https://doi.org/10.3991/ijim.v17i04.37805

[12] T. M. Taha, Z. Ben Messaoud, and M. Frikha, "Convolutional neural network architectures for gender, emotional detection from speech and speaker diarization," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 18, no. 3, pp. 88–103, 2024. https://doi.org/10.3991/ijim.v18i03.43013

[13] Y. Li, J. Ren, Y. Wang, G. Wang, X. Li, and H. Liu, "Audio–visual keyword transformer for unconstrained sentence-level keyword spotting," *CAAI Transactions on Intelligence Technology*, vol. 9, no. 1, pp. 142–152, 2024. https://doi.org/10.1049/cit2.12212

[14] X. Shen, L. Wang, Q. Pei, Y. Liu, and M. Li, "Location privacy-preserving in online taxi-hailing services," *Peer-to-Peer Networking and Applications*, vol. 14, pp. 69–81, 2021. https://doi.org/10.1007/s12083-020-00982-7

[15] K. Zhou, B. Sisman, R. Rana, B. W. Schuller, and H. Li, "Speech synthesis with mixed emotions," *IEEE Transactions on Affective Computing*, vol. 14, no. 4, pp. 3120–3134, 2023. https://doi.org/10.1109/TAFFC.2022.3233324

[16] H. Choi and M. Hahn, "Sequence-to-sequence emotional voice conversion with strength control," *IEEE Access*, vol. 9, pp. 42674–42687, 2021. https://doi.org/10.1109/ACCESS.2021.3065460

[17] L. Khosravani Pour and A. Farrokhi, "Language recognition by convolutional neural networks," *Scientia Iranica*, vol. 30, no. 1, pp. 116–123, 2023. https://doi.org/10.24200/sci.2022.59110.6064

[18] N. Dhariwal, S. C. Akunuri, Shivama, and K. S. Banu, "Audio and text sentiment analysis of radio broadcasts," *IEEE Access*, vol. 11, pp. 126900–126916, 2023. https://doi.org/10.1109/ACCESS.2023.3331226

[19] P. Pertilä, E. Fagerlund, A. Huttunen, and V. Myllylä, "Online own voice detection for a multi-channel multi-sensor in-ear device," *IEEE Sensors Journal*, vol. 21, no. 24, pp. 27686–27697, 2021. https://doi.org/10.1109/JSEN.2021.3122936

[20] I. López-Espejo, Z-H. Tan, J. H. L. Hansen, and J. Jensen, "Deep spoken keyword spotting: An overview," *IEEE Access*, vol. 10, pp. 4169–4199, 2022. https://doi.org/10.1109/ACCESS.2021.3139508

[21] M. M. Rezapour Mashhadi and K. Osei-Bonsu, "Speech emotion recognition using machine learning techniques: Feature extraction and comparison of convolutional neural network and random forest," *PLoS ONE*, vol. 18, no. 11, p. e0291500, 2023. https://doi.org/10.1371/journal.pone.0291500

[22] K. Manohar, A. R. Jayan, and R. Rajan, "Improving speech recognition systems for the morphologically complex Malayalam language using subword tokens for language modelling," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2023, 2023. https://doi.org/10.1186/s13636-023-00313-7

[23] A. Mukhamadiyev, M. Mukhiddinov, I. Khujayarov, M. Ochilov, and J. Cho, "Development of language models for continuous Uzbek speech recognition system," *Sensors*, vol. 23, no. 3, p. 1145, 2023. https://doi.org/10.3390/s23031145

[24] R. Jahangir, Y. W. Teh, H. F. Nweke, G. Mujtaba, M. Ali Al-Garadi, and I. Ali, "Speaker identification through artificial intelligence techniques: A comprehensive review and research challenges," *Expert Systems with Applications*, vol. 171, p. 114591, 2021. https://doi.org/10.1016/j.eswa.2021.114591

[25] S. Baskar, S. Dhote, T. Dhote, G. Jayanandini, D. Akila, and S. Doss, "A predictive typological content retrieval method for real-time applications using multilingual natural language processing," *Expert Systems*, vol. 41, no. 6, 2024. https://doi.org/10.1111/exsy.13172

[26] S. Liang and W. Q. Yan, "A hybrid CTC+Attention model based on end-to-end framework for multilingual speech recognition," *Multimedia Tools and Applications*, vol. 81, pp. 41295–41308, 2022. https://doi.org/10.1007/s11042-022-12136-3

[27] M. Pirnau *et al.*, "Content analysis using specific natural language processing methods for big data," *Electronics*, vol. 13, no. 3, p. 584, 2024. https://doi.org/10.3390/electronics13030584

[28] J. Liao, Y. Shi, and Y. Xu, "Automatic speech recognition post-processing for readability: Task, dataset and a two-stage pre-trained approach," *IEEE Access*, vol. 10, pp. 117053–117066, 2022. https://doi.org/10.1109/ACCESS.2022.3219838

[29] J. Li, "A recommendation model for college English digital teaching resources using collaborative filtering and few-shot learning technology," *Computational Intelligence and Neuroscience*, vol. 2022, 2022. https://doi.org/10.1155/2022/1233057

[30] B. Kim, S. Chang, J. Lee, and D. Sung, "Broadcasted residual learning for efficient keyword spotting," *arXiv preprint arXiv:2106.04140*, 2021. https://doi.org/10.21437/Interspeech.2021-383

[31] M. Bordoloi and S. K. Biswas, "Sentiment analysis: A survey on design framework, applications and future scopes," *Artificial Intelligence Review*, vol. 56, pp. 12505–12560, 2023. https://doi.org/10.1007/s10462-023-10442-2

[32] J. Rose, "Common Voice 19.0 dataset release," *Mozilla Discourse*, 2024. [Online]. Available: https://discourse.mozilla.org/t/common-voice-19-0-dataset-release/135857 [Accessed: June 18, 2025].

9 AUTHORS

Keni Abel Sanchez Villogas is a Systems Engineering at the Universidad Peruana de Ciencias Aplicadas in Lima, Peru (E-mail: U202018789@upc.edu.pe).

Paolo Manoel Pinzás Riveros is a Systems Engineering at the Universidad Peruana de Ciencias Aplicadas in Lima, Peru (E-mail: U201910787@upc.edu.pe).

Pedro Castañeda is a RENACYT Researcher and holds a PhD in Systems Engineering, a master's degree in management and information technology management from UNMSM, and a master's degree in business administration (MBA) – ESAN. He has completed doctoral studies in Public Policy and State Management at the Centro de Altos Estudios Nacionales (CAEN). He leads e-brokerage projects, software development, and process improvement, using agile and traditional methodologies. He has the following certifications: Project Management Professional (PMP), Scrum Certified Developer (CSD), IBM Certified Professional in Rational Unified Process, and ORACLE Certifications. Areas of Interest: Artificial Intelligence, Software Productivity, Business Intelligence, Data Analytics, Machine Learning, Software Engineering (E-mail: pcsipcas@upc.edu.pe).

Alejandra Oñate-Andino holds a degree in Computer Systems Engineering from Escuela Superior Politécnica de Chimborazo (Ecuador), a Master in Network Interconnectivity from Escuela Superior Politécnica de Chimborazo (Ecuador), and a PhD in Systems Engineering and Computer Science from Universidad Mayor de San Marcos (Peru) (E-mail: monate@espoch.edu.ec).