

## Audio Event Detection for People's Safety Using Deep Learning

**Dr. Ayesha Ameen**

Professor and Head of Department  
Department of CSE  
Deccan College of Engineering and  
Technology  
Affiliated to Osmania University  
Hyderabad, Telangana.  
[hod\\_cse@deccancollege.ac.in](mailto:hod_cse@deccancollege.ac.in)

**Zulekha Zafreen**

PG Scholar  
Department of CSE  
Deccan College of Engineering and  
Technology  
Affiliated to Osmania University  
Hyderabad, Telangana  
[zulekhagaffar65@gmail.com](mailto:zulekhagaffar65@gmail.com)

**Abstract:** Criminal activities such as robbery, assault, and homicide pose significant risks, particularly to individuals working alone in remote areas at night, with women being especially vulnerable. Conventional image-based AI detection systems often suffer from limited accuracy. This study proposes a solution using sound feature extraction to detect potential criminal activities in real-time. We applied machine learning techniques to audio data from the URBAN8K dataset, which includes ten sound classes, and incorporated additional classes, such as 'crackling\_fire' and 'glass\_breaking' from the ESC-50 dataset. Audio files were converted to mono and downsampled to 16,000 Hz for preprocessing, and Mel-spectrogram features were extracted to identify abnormal audio patterns distinguishing normal from alert sounds. Experimental results show that the Bi-Directional GRU model outperformed other models, achieving 96.77% accuracy, 96.60% precision, 93.50% recall, and a 94.88% F1-score. This approach highlights the potential of sound-based analysis in improving the accuracy and reliability of real-time event detection systems, offering significant improvements in public safety.

*“Index Terms - Natural language processing (NLP), deep learning, audio, recording, CNN, LSTM, classification, prediction”.*

### 1. INTRODUCTION

In the physical world, every event is accompanied by a distinct sound. Whether it's the sound of a stone falling, a bird chirping, or the noise of a road being constructed, these sounds are unique to each event. Positive events, like crowds cheering or fireworks, have sounds associated with them, just as negative events, such as road accidents, gunshots, or natural disasters, generate their own specific sounds. The ability to detect these sounds and classify them as either positive or negative could revolutionize security systems. The potential for a system that can

automatically distinguish between ambient sounds, identifying whether they are linked to a positive or negative event, offers significant promise for enhancing public safety. Although such systems are not yet widely implemented in everyday life, the underlying technology already exists in various forms.

One of the most notable technologies in this domain is Natural Language Processing (NLP), which has already made significant strides in speech and text-based analysis. NLP is widely used in modern smartphones, which are equipped with AI-powered

voice recognition systems. For instance, Google's Voice Assistant, available on every Android smartphone, can process voice commands to search the web and retrieve necessary information without the user having to type a single search query [1]. Similarly, Apple's Siri, which functions as a personal assistant, has been integrated into the iPhone series and evolved with various versions, improving voice recognition and usability over time [2]. Amazon's Alexa has also become a prominent example of AI-powered NLP, providing smart home control through voice commands, while Samsung's Bixby offers its own voice assistant service [3]. These systems have already demonstrated the significant potential of NLP in simplifying daily tasks.

NLP has also made an impact in the realm of textual data analysis and classification. Google, for example, uses NLP in its search engine to offer search query suggestions, as well as to determine the relevance of search results [4]. Social media platforms like Twitter and Instagram use NLP for sentiment analysis, helping to detect the emotions expressed in posts and tailoring users' timelines based on this analysis. This approach has transformed how we interact with content online, offering users a more personalized experience by analyzing both the tone and context of the text [5]. These advancements in NLP technology, particularly in voice recognition and text-based analysis, lay the groundwork for more complex applications such as automatic event detection through sound, which could be applied to improve security and public safety.

## 2. RELATED WORK

The integration of technology in improving safety, particularly for women and vulnerable individuals,

has gained significant attention in recent years. Various approaches have been explored to enhance security using sensors, mobile applications, and advanced algorithms. One such system, the IoT-based women safety device proposed by Suma and Rekha [6], focuses on detection and real-time video capturing in the event of distress. Their study emphasizes the importance of smart safety solutions in responding to dangerous situations, where the device automatically sends alerts and records video footage, ensuring a swift response to potential threats. This system highlights how IoT and real-time monitoring can contribute to personal safety, and it aligns with growing efforts to harness technology for protection in dangerous environments.

In parallel, the role of sound detection in safety systems has garnered significant interest. The work by Ciaburro and Iannace [7] investigates the potential of deep neural network algorithms for detecting sound events to improve safety in smart cities. Their study demonstrates how advanced sound event detection systems can be employed to identify abnormal sounds, such as screams or crashes, which may indicate dangerous situations. By utilizing neural networks for sound classification, their approach helps to automatically recognize potentially hazardous events in public spaces. This research shows how leveraging AI and deep learning can enhance security systems by improving the accuracy and efficiency of sound event detection, which is crucial for real-time responses.

Monisha et al. [8] proposed a women safety device, FEMME, that integrates mobile applications with wearable technology. This system focuses on detecting critical sounds, such as screams, and provides automatic alerts to pre-configured contacts.

The device also records audio and video during distress situations, offering critical evidence that can be used in investigations. The FEMME system illustrates the convergence of mobile and wearable technologies for proactive safety solutions, reinforcing the importance of quick alert mechanisms in high-risk scenarios.

In another study, Kendzhaeva et al. [9] explore a system that detects anomalous sounds in urban environments to improve safety for both citizens and tourists. The study presents a sound detection system embedded in a smart city infrastructure to identify sounds that deviate from normal patterns. The system is designed to flag abnormal occurrences, such as accidents, fights, or other emergencies, enabling authorities to respond quickly. By combining sound recognition with other urban monitoring systems, this approach creates a safer environment, proving that sound detection can be a valuable tool in public safety systems.

Neri et al. [10] examine sound event detection (SED) techniques in the context of human safety and security in noisy environments. The research focuses on the challenges of detecting sounds in environments with high ambient noise, such as urban streets or crowded places, where distinguishing between normal and abnormal sounds can be difficult. The authors propose a combination of machine learning techniques to enhance sound event recognition, improving the system's ability to detect specific security-related events like accidents or disturbances. This work highlights the importance of refining SED systems to function effectively in diverse and noisy environments, a challenge often faced in urban safety applications.

Deep learning has also been applied to detect criminal activities and promote safety, as discussed

by Mathur et al. [11]. They developed a system that employs deep learning models to analyze audio and video feeds for detecting criminal activities. The system uses sound and visual data to identify suspicious behavior, such as arguments or fights, and can alert law enforcement authorities. By combining multiple forms of data and advanced learning models, the system improves detection accuracy and provides a foundation for future applications in public safety, especially in urban settings where criminal activities are more likely to occur.

Villegas-Ch and Govea [12] further expanded on the use of deep learning for early detection of emergency situations and security monitoring in public spaces. They propose an approach that combines sound detection with visual monitoring to identify potential threats before they escalate. By employing convolutional neural networks (CNNs) and other advanced deep learning techniques, their system can process sound and image data simultaneously to detect events such as accidents or criminal activities. This multi-modal approach allows for more accurate and reliable detection, offering an advantage in real-world security applications where sound alone may not always provide sufficient context.

Overall, the literature on sound event detection and its applications in safety and security systems highlights the potential for deep learning and machine learning techniques to enhance the effectiveness of these systems. Sound recognition can play a pivotal role in monitoring environments, alerting authorities to unusual or dangerous events in real time. As demonstrated by various studies, the integration of advanced algorithms, AI, and IoT devices can significantly improve safety in urban spaces and for individuals, providing a more

proactive approach to security. With continued advancements in technology, the accuracy and reliability of these systems are expected to improve, leading to smarter, safer cities and better protection for individuals in high-risk situations.

### 3. MATERIALS AND METHODS

The proposed system aims to enhance the detection of criminal activities in real-time by leveraging audio-based analysis rather than relying solely on traditional image-based methods. The system extracts sound features using Mel-spectrograms from audio files to distinguish between normal and alert sounds. Two datasets, URBAN8K [13] and ESC-50 [14], are utilized, comprising classes such as 'gun\_shot,' 'glass\_breaking,' and 'crackling\_fire,' which are preprocessed by converting audio to mono and downsampling to 16,000 Hz. Extracted features are trained using various deep learning algorithms, including CNN1D, CNN2D, and LSTM [16], to evaluate their ability to classify abnormal sounds accurately. Additionally, the system incorporates a Bidirectional-GRU layer to further enhance performance. Upon detecting a threat, the system triggers an alert by sending a notification to a predefined email address, making it an efficient and proactive solution for improving public safety.

The image (Fig.1) depicts a typical machine learning pipeline for threat detection using audio features. It starts with a dataset that undergoes pre-processing steps like visualization, data processing, and shuffling. This pre-processed data is then fed into various machine learning models such as CNN1D, CNN2D, LSTM [16], and Bi-directional GRU. These models are trained on the data, and their performance is evaluated using metrics like accuracy, precision, recall, and F1-score. The trained models are then used to detect threats using audio NLP features, ultimately aiming to provide a robust and effective solution for audio-based threat identification.

#### i) Dataset Collection:

The proposed system uses two audio datasets, UrbanSound8K and ESC-50, to train models for detecting criminal activities through sound analysis.

#### UrbanSound8K Dataset

UrbanSound8K [13] contains 8,732 audio samples categorized into ten classes, including 'air\_conditioner,' 'car\_horn,' 'children\_playing,' 'dog\_bark,' 'drilling,' 'engine\_idling,' 'gun\_shot,' 'jackhammer,' 'siren,' and 'street\_music.' Each audio file is annotated with metadata such as slice\_file\_name, fsID, start, end, salience, fold, classID, and class. The dataset is organized into 10-second audio slices and grouped into ten folds, supporting cross-validation for machine learning applications.



Fig.1 Proposed Architecture

id	src_file_name	fold	start	end	balance	fold	classID	class
0	100003-3-0-0.wav	100003	0.000000	0.317561	1	5	0	dog_bark
1	100263-2-0-117.wav	100263	98.900000	100.900000	1	5	2	children_playing
2	100263-2-0-121.wav	100263	98.900000	104.900000	1	5	2	children_playing
3	100263-2-0-126.wav	100263	98.900000	107.000000	1	5	2	children_playing
4	100263-2-0-137.wav	100263	98.900000	12.900000	1	5	2	children_playing
...	...	...	...	...	...	...	...	...
8727	06612-1-2-0.wav	99912	159.522205	160.522205	2	7	1	car_horn
8728	06612-1-3-0.wav	99912	181.542431	180.284976	2	7	1	car_horn
8729	06612-1-4-0.wav	99912	242.891902	246.197886	2	7	1	car_horn
8730	06612-1-5-0.wav	99912	253.239850	255.741946	2	7	1	car_horn
8731	06612-1-6-0.wav	99912	332.289203	334.831302	2	7	1	car_horn

8732 rows x 9 columns

Fig.2 Dataset Collection Table – UrbanSound8K

### ESC-50 Dataset

ESC-50 [14] is a curated dataset of 80 audio samples spanning classes like 'crackling\_fire' and 'glass\_breaking.' Each sample is annotated with filename, fold, target, category, esc10, src\_file, and take. It provides a rich diversity of environmental sounds, which are especially useful for distinguishing abnormal events. The dataset is well-suited for audio classification tasks and contributes additional critical sound classes to the system.

	filename	fold	target	category	esc10	src_file	take
0	1-17150-A-12.wav	1	12	crackling_fire	True	17150	A
1	1-17565-A-12.wav	1	12	crackling_fire	True	17565	A
2	1-17742-A-12.wav	1	12	crackling_fire	True	17742	A
3	1-17808-A-12.wav	1	12	crackling_fire	True	17808	A
4	1-17808-B-12.wav	1	12	crackling_fire	True	17808	B
...	...	...	...	...	...	...	...
75	5-233607-A-39.wav	5	39	glass_breaking	False	233607	A
76	5-257642-A-39.wav	5	39	glass_breaking	False	257642	A
77	5-260432-A-39.wav	5	39	glass_breaking	False	260432	A
78	5-260433-A-39.wav	5	39	glass_breaking	False	260433	A
79	5-260434-A-39.wav	5	39	glass_breaking	False	260434	A

80 rows x 7 columns

Fig.3 Dataset Collection Table – Crackle fire and glass breaking

### ii) Pre-Processing:

The pre-processing phase involves loading, cleaning, and extracting essential features from audio files to prepare them for analysis.

**a) Visualization:** Visualization provides insights into the dataset's structure and extracted features. Class label distributions are represented on a graph where the x-axis displays class labels, and the y-axis indicates the number of audio files per class. Mel-spectrograms are visualized to show voice activity, with red areas indicating sound and other areas showing silence. Additionally, audio signal plots depict dense lines for sound portions and thin lines for silent segments, offering an intuitive understanding of the audio data.

**b) Data Processing:** During data processing, the system loops through all audio files, extracts Mel-spectrogram features, and organizes them into X and Y arrays for feature and label storage. Each file is read and transformed into meaningful representations, enabling deep learning algorithms to distinguish between normal and abnormal sounds. The number of processed audio files is tracked and displayed, ensuring all data is accurately handled for training. This stage is critical for creating a robust foundation for classification tasks.

**c) Shuffling:** Shuffling ensures the dataset is randomized before training, avoiding bias from sequential patterns. The processed audio data and corresponding labels are shuffled to mix examples from different classes uniformly. This step improves the model's ability to generalize by exposing it to diverse combinations of data during training, thereby minimizing overfitting risks. Shuffling is a fundamental pre-training operation that enhances the overall effectiveness of the learning process.

### iii) Training & Testing:

After pre-processing, the dataset is split into training and testing sets, with 80% of the data used for training the model and 20% reserved for testing.

This split ensures that the model learns effectively from a substantial portion of the data while leaving sufficient unseen examples for evaluation. The training set is utilized to fit the deep learning algorithms, while the testing set evaluates performance metrics like accuracy, precision, recall, and F1-score. This 80:20 ratio balances learning and validation for optimal performance.

#### iv) Algorithms:

**CNN1D** (One-Dimensional Convolutional Neural Network[15]) is utilized for analyzing sequential data, making it suitable for audio signal processing. CNN1D extracts patterns from Mel-spectrogram features of audio files, effectively distinguishing between normal and alert sounds. This algorithm's convolutional layers allow it to learn spatial hierarchies, enhancing its ability to detect specific audio characteristics, such as gunshots or glass breaking, contributing to real-time threat detection and alert generation.

**CNN2D** (Two-Dimensional Convolutional Neural Network) [15] processes 2D data, such as images or spectrograms, to identify spatial patterns. CNN2D analyzes Mel-spectrogram representations of audio files, enabling it to learn complex features and relationships between different audio classes. This algorithm enhances detection accuracy by leveraging its ability to capture both temporal and frequency information, allowing the system to effectively classify various sounds associated with potential threats, such as sirens or engine idling.

**LSTM** (Long Short-Term Memory) networks are designed for sequence prediction tasks, particularly useful for time-series data. [16] LSTM processes the sequential data extracted from audio files, capturing temporal dependencies crucial for identifying

patterns in sound over time. By analyzing audio features, LSTM improves the system's ability to detect abnormal sounds indicative of criminal activity. Its effectiveness in managing long-range dependencies enhances the accuracy of threat detection in real-time scenarios.

The **Extension Bi-directional GRU (Gated Recurrent Unit)** enhances LSTM's capabilities by processing data in both forward and backward directions. It leverages audio feature sequences to identify patterns indicative of threats. By considering past and future context, the bi-directional GRU improves the model's understanding of temporal relationships in audio data. This results in more accurate detection of critical sounds, such as gunshots or breaking glass, leading to timely alerts for potential threats.

## 4. RESULTS & DISCUSSION

**Accuracy:** The accuracy of a test is its ability to differentiate the patient and healthy cases correctly. To estimate the accuracy of a test, we should calculate the proportion of true positive and true negative in all evaluated cases. Mathematically, this can be stated as:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

**Precision:** Precision evaluates the fraction of correctly classified instances or samples among the ones classified as positives. Thus, the formula to calculate the precision is given by:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (2)$$

**Recall:** Recall is a metric in machine learning that measures the ability of a model to identify all

relevant instances of a particular class. It is the ratio of correctly predicted positive observations to the total actual positives, providing insights into a model's completeness in capturing instances of a given class.

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

**F1-Score:** F1 score is a machine learning evaluation metric that measures a model's accuracy. It combines the precision and recall scores of a model.

The accuracy metric computes how many times a model made a correct prediction across the entire dataset.

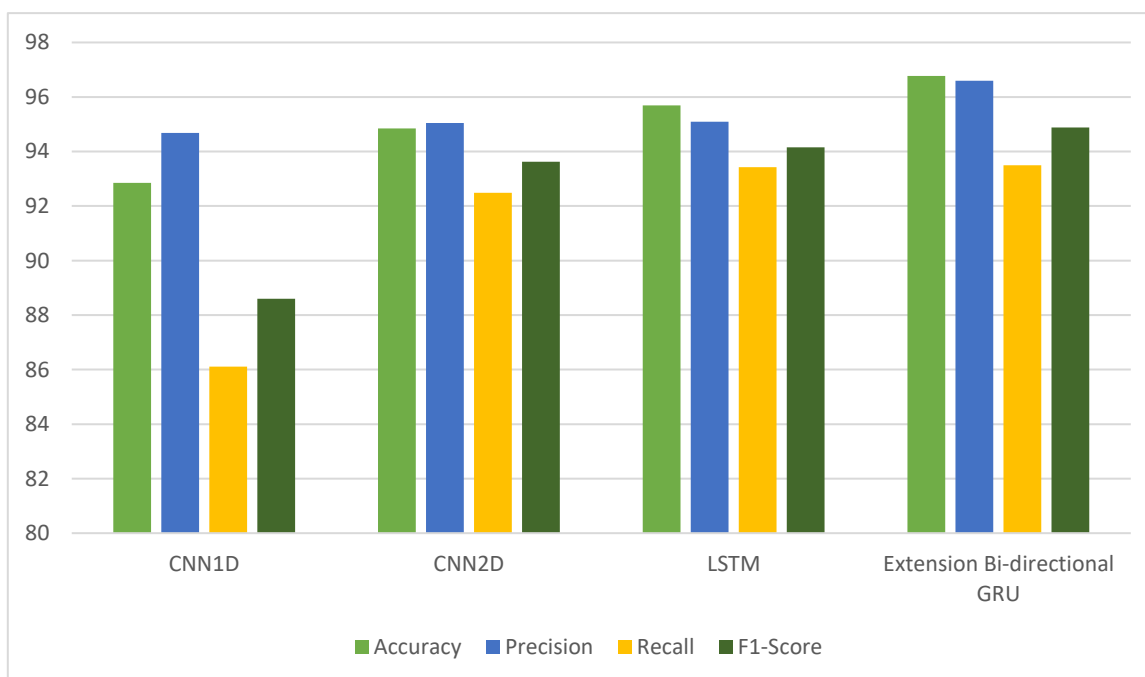
$$F1\ Score = 2 * \frac{Recall * Precision}{Recall + Precision} * 100 \quad (1)$$

We evaluate the performance metrics—accuracy, precision, recall, and F1-score—for each algorithm in Table 1. The Bi-directional GRU achieves the highest scores. The table below also presents the metrics of other algorithms for comparison.

Table.1 Performance Evaluation Metrics

ML Model	Accuracy	Precision	Recall	F1-Score
CNN1D	92.85	94.68	86.11	88.60
CNN2D	94.84	95.04	92.48	93.62
LSTM	95.69	95.09	93.43	94.15
<b>Extension Bi-directional GRU</b>	<b>96.77</b>	<b>96.60</b>	<b>93.50</b>	<b>94.88</b>

Graph.1 Comparison Graphs



Graph 1 displays accuracy in light green, precision in blue, recall in light yellow, and the F1 score in green. The bi-directional GRU outperforms the other algorithms in all metrics, with the highest values compared to the remaining models. The above graph visually represents these details.



Fig.4 Home Page

In above fig.4 user interface dashboard with navigation and a welcome message.

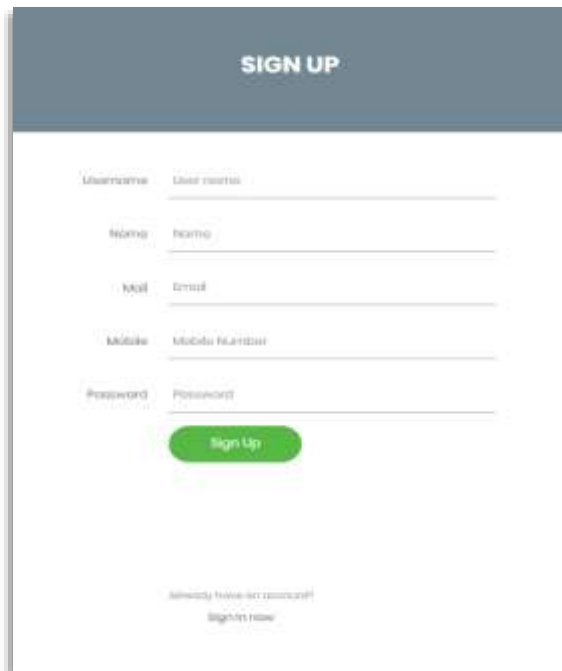


Fig.5 Registration Page

In above fig.5 sign-up form with fields for username, name, email, mobile number, and password buttons.

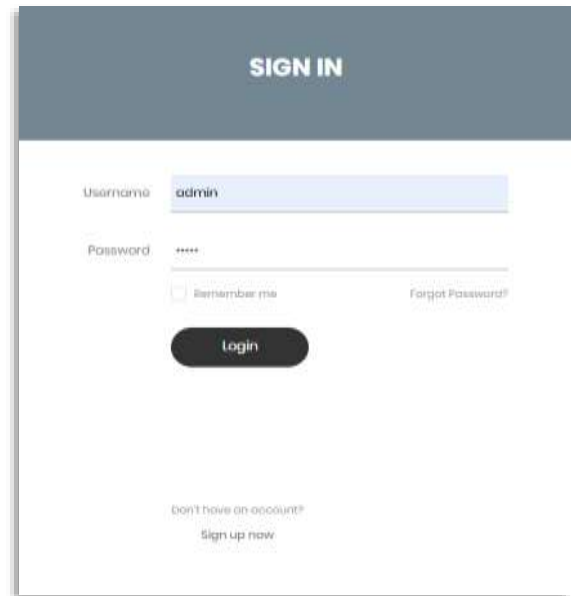


Fig.6 Login Page

In above fig.6 Sign-in form with username and password fields, "Remember Me," "Forgot Password,".

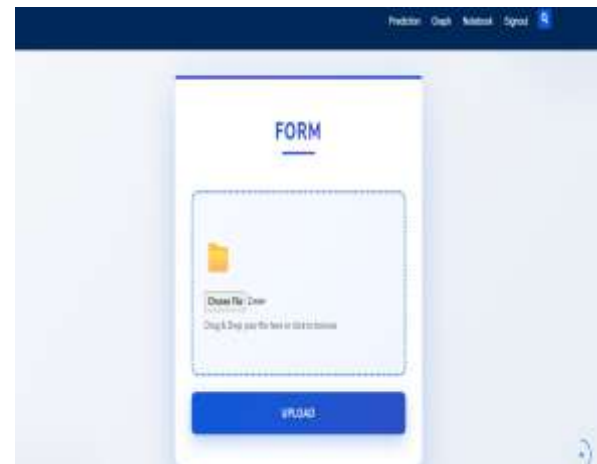


Fig.7 Upload Input Page

In above Fig.7 form with coordinate input field and upload button.

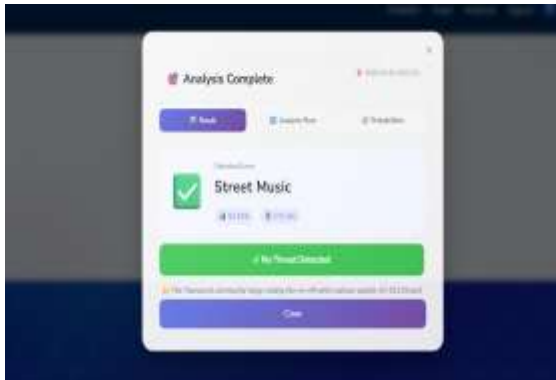


Fig.8 Predict Result for given input

In above Fig.8 Predicted result based on the input test data.

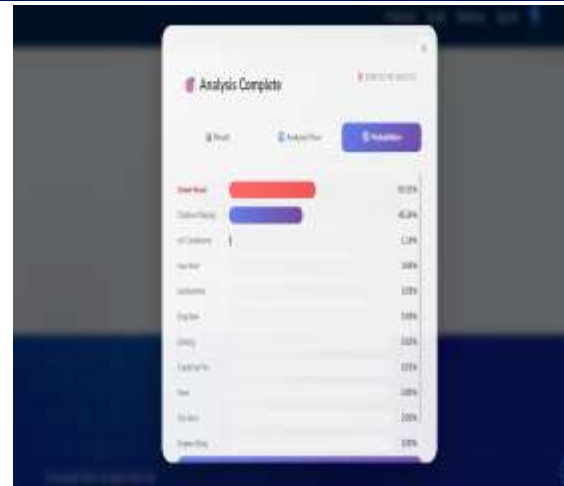


Fig 10 : predicted classes

## 5. CONCLUSION

In conclusion, a real-time threat detection system is developed using deep learning to analyze surrounding audio signals and send automated alerts via email, SMS, and WhatsApp. By utilizing audio features such as Mel-spectrograms, the system accurately identifies potential threats such as gunshots, glass breaking, and sirens. Extensive experimentation was conducted using audio data from the Urban8K and ESC-50 datasets, representing a wide range of sound classes. The system demonstrates significant improvement in detecting abnormal sounds compared to existing image-based detection methods. The high-performing algorithm, enhanced by the inclusion of the Bidirectional GRU layer, achieves an impressive accuracy of 96.77%, ensuring a reliable and robust solution for detecting threats in real-time. This software-based approach eliminates the need for additional hardware components, making it a cost-effective and easily deployable solution for improving personal safety, particularly for individuals working alone in vulnerable or remote locations. The system offers a promising real-time



Fig 9: procedure Analysis

solution to mitigate potential dangers and enhance security.

In the future, the proposed system can be enhanced by incorporating advanced techniques like attention mechanisms, transformers, and hybrid deep learning models to improve threat detection accuracy. Additionally, integrating real-time audio data augmentation methods and adaptive noise filtering could further refine the system's performance in varied environments. Expanding the system to support multiple languages and dialects, as well as enhancing mobile integration for faster response times, can broaden its effectiveness and scope.

## REFERENCES

- [1] Google Search by Voice: A Case Study. Accessed: Dec. 1, 2023. [Online]. Available: <https://research.google.com/pubs/archive/36340.pdf>
- [2] M. Assefi, G. Liu, M. P. Wittie, and C. Izurieta, "An experimental evaluation of apple Siri and Google speech recognition," in Proc. ISCA SEDE, 2015, p. 118.
- [3] A. L. Nobles, E. C. Leas, T. L. Caputi, S.-H. Zhu, S. A. Strathdee, and J. W. Ayers, "Responses to addiction help-seeking from alexa, siri, Google assistant, cortana, and bixby intelligent virtual assistants," *npj Digit. Med.*, vol. 3, no. 1, p. 11, Jan. 2020.
- [4] S. Zad, M. Heidari, J. H. J. Jones, and O. Uzuner, "Emotion detection of textual data: An interdisciplinary survey," in Proc. IEEE World AI IoT Congr. (AIoT), May 2021, pp. 0255–0261.
- [5] W. Graterol, J. Diaz-Amado, Y. Cardinale, I. Dongo, E. Lopes-Silva, and C. Santos-Libarino, "Emotion detection for social robots based on NLP transformers and an emotion ontology," *Sensors*, vol. 21, no. 4, p. 1322, Feb. 2021.
- [6] T. P. Suma and G. Rekha, "Study on IoT based women safety devices with screaming detection and video capturing," *Int. J. Eng. Appl. Sci. Technol.*, vol. 6, no. 7, pp. 257–262, 2021.
- [7] G. Ciaburro and G. Iannace, "Improving smart cities safety using sound events detection based on deep neural network algorithms," *Informatics*, vol. 7, no. 3, p. 23, Jul. 2020.
- [8] D. G. Monisha, M. Monisha, G. Pavithra, and R. Subhashini, "Women safety device and application-FEMME," *Indian J. Sci. Technol.*, vol. 9, no. 10, pp. 1–6, Mar. 2016.
- [9] Kendzhaeva, B., Omarov, B., Abdiyeva, G., Anarbayev, A., Dauletbek, Y., & Omarov, B. (2021). Providing safety for citizens and tourists in cities: a system for detecting anomalous sounds. In *Advanced Informatics for Computing Research: 4th International Conference, ICAICR 2020, Gurugram, India, December 26–27, 2020, Revised Selected Papers, Part I 4* (pp. 264-273). Springer Singapore.
- [10] Neri, M., Battisti, F., Neri, A., & Carli, M. (2022). Sound event detection for human safety and security in noisy environments. *IEEE Access*, 10, 134230-134240.
- [11] Mathur, R., Chintala, T., & Rajeswari, D. (2022, January). Detecting criminal activities and promoting safety using deep learning. In *2022 international conference on advances in computing, communication and applied informatics (ACCAI)* (pp. 1-8). IEEE.

[12] Villegas-Ch, W., & Govea, J. (2023). Application of deep learning in the early detection of emergency situations and security monitoring in public spaces. *Applied System Innovation*, 6(5), 90.

[13] Urban Sound Datasets. Accessed: Dec. 1, 2023. [Online]. Available: <https://urbansounddataset.weebly.com/download-urbansound8k.html>

[14] Environmental Sound Classification 50. Accessed: Dec. 1, 2023. [Online]. Available: <https://www.kaggle.com/datasets/mmoreaux/environmental-sound-classification-50?select=audio>

[15] J. Cao, M. Cao, J. Wang, C. Yin, D. Wang, and P.-P. Vidal, "Urban noise recognition with convolutional neural network," *Multimedia Tools Appl.*, vol. 78, no. 20, pp. 29021–29041, Oct. 2019.

[16] M. A. Sit, C. Koylu, and I. Demir, "Identifying disaster-related tweets and their semantic, spatial and temporal context using deep learning, natural language processing and spatial analysis: A case study of hurricane irma," *Int. J. Digit. Earth*, vol. 12, no. 11, pp. 1205–1229, Nov. 2019.