Computers, Materials & Continua
DOI: 10.32604/cmc.2023.036205
Article

Fast Detection and Classification of Dangerous Urban Sounds Using Deep Learning

Zeinel Momynkulov[1], Zhandos Dosbayev[2,3,*], Azizah Suliman[4], Bayan Abduraimova[5], Nurzhigit Smailov[2], Maigul Zhekambayeva[2] and Dusmat Zhamangarin[6]

[1]International Information Technology University, Almaty, Kazakhstan
[2]KazNRTU named after K. I. Satbayev, Almaty, Kazakhstan
[3]U. A Joldasbekov Institute of Mechanics and Engineering, Almaty, Kazakhstan
[4]Faculty of Data Science and Information Technology, INTI International University, Putra Nilai, Malaysia
[5]L. N. Gumilyov Eurasian National University, Astana, Kazakhstan
[6]Kazakh University of Technology and Business, Astana, Kazakhstan
*Corresponding Author: Zhandos Dosbayev. Email: zhandosdosbayev@gmail.com
Received: 20 September 2022; Accepted: 23 December 2022

Abstract: Video analytics is an integral part of surveillance cameras. Compared to video analytics, audio analytics offers several benefits, including less expensive equipment and upkeep expenses. Additionally, the volume of the audio datastream is substantially lower than the video camera datastream, especially concerning real-time operating systems, which makes it less demanding of the data channel's bandwidth needs. For instance, automatic live video streaming from the site of an explosion and gunshot to the police console using audio analytics technologies would be exceedingly helpful for urban surveillance. Technologies for audio analytics may also be used to analyze video recordings and identify occurrences. This research proposed a deep learning model based on the combination of convolutional neural network (CNN) and recurrent neural network (RNN) known as the CNN-RNN approach. The proposed model focused on automatically identifying pulse sounds that indicate critical situations in audio sources. The algorithm's accuracy ranged from 95% to 81% when classifying noises from incidents, including gunshots, explosions, shattered glass, sirens, cries, and dog barking. The proposed approach can be applied to provide security for citizens in open and closed locations, like stadiums, underground areas, shopping malls, and other places.

Keywords: Deep learning; urban sounds; CNN; RNN; classification; impulsive sounds

1 Introduction

Human daily activities have been significantly permeated by automation technologies (AT) [1,2]. One of the main objectives of applying AT in daily life is to maintain civic security, which

is achieved by keeping an eye out in public spaces and spotting any danger signs. The studies on autonomous surveillance systems are primarily concerned with the utilization of computer vision to identify occurrences [3]. It is surmised that the use of acoustic monitoring in conjunction with closed-circuit television (CCTV) systems would improve the efficiency of action recognition and provide an additional source of data [4,5]. When compared to video analysis software, audio analysis offers considerable advantages that may help to solve monitoring tasks more successfully. These advantages include: (a) minimal processing requirements; and (b) independence from visibility circumstances.

Law enforcement organizations may use dangerous urban sound detection systems as a helpful tool to handle weapons offenses quickly and effectively. Due to the wide implementation of information technologies in modern society [6], even unknown events may be tracked down in audio recordings by using a threshold-based energy detector, accompanied by a recognizing block, which is considered a relatively straightforward and efficient method. The issue of identifying and detecting impulsive sounds is discussed in [7]. The authors utilized two statistical classifiers together with a median filter to identify events in a noisy background. The system performed well above 0 dB of the signal-to-noise ratio, but in a real-world setting with repercussions, the efficiency rapidly degraded. Initially, the performance of the system was enhanced by calculating the energy using significant wavelet coefficients [8]. In addition, the study in [9] provided a technique for analytically determining the thresholds for emission detectors [10], based on certainly missed recognition rates. Moreover, a significant amount of early research was focused on the development of efficient artificial intelligence techniques for the detection of potentially hazardous urban sounds. These techniques have been demonstrated to perform well even in cases with a low signal-to-noise ratio; some instances of these approaches were highlighted in state-of-the-art literature [11–13].

It is quite challenging to detect dangerous urban soundscapes in confined spaces. The principal cause of such an issue, in addition to ambient noises, is the natural resonances from nearby surfaces, which alter the impulsive sound signature's structure, properties, and envelopes [14]. Simple matched filtering or other linear signal processing methods cannot reduce the nonlinear dynamics in sound propagation [15]. To create a reliable and robust gunshot detection system for urban environments, the reverberation effects of the acoustic channel need to be reduced before classification algorithms can be employed. Alternatively, machine learning techniques are set to be designed with features that are preserved even in the presence of reverberations. The majority of the research, previously published in the literature, is concerned with the detection and identification of guns in noise-free environments. There are some examples in recent studies [16–18]. Only two contributions, as far as the present authors are aware, deal with dangerous urban sound detection systems in reverberant environments. The researchers in [19] distinguished between signals that pose a danger and those that do not, using parameters such as sound energy and entropy. A book banging on a table, a paper bag bursting with air, and a wrench slamming against a metal ladder are examples of signals that are considered to be non-threat indicators. Nevertheless, they are quite restricted in scope and do not always have the same impulsive quality as a gunshot signal. Bine et al. [20] suggested a method for detecting gunshots in enclosed spaces, which includes several parameters. For example, frequency, onset time, and sound pressure level are parameters that may be utilized to tell a gunshot from a false alarm. However, it should be noted that the strength of these properties against reverberations was not examined.

There are some studies on combined models, such as convolutional neural network and recurrent neural network (CNN-RNN), and CNN and long short-term memory (CNN-LSTM), which coincide with the proposed model [21–25]. Some research used the CNN-LSTM model for medicine [21,22], the agrarian sphere [23], and the financial sector [24,25]. Khan et al. proposed CNN-LSTM for the COVID-19 hotspots prediction problem [21]. Researchers have provided a hybrid architecture

that can extract characteristics from CNN's convolutional layers that span many time scales. After that, the many time-scale characteristics are combined and fed into a two-layer LSTM model. The representation of time-series data is learned by the LSTM model, which then recognizes short, medium, and long-term relationships.

The objective of this study is to provide a deep learning-based technique for identifying noises associated with critical circumstances in the audio stream. The phrase "critical circumstance" refers to an occurrence in this work that may include acoustic artifacts as indicated by its distinctive sound indicators (a shot, a scream, a glass crash, an explosion, a siren, etc.). A technique was created that permits identifying crucial instances in the audio signal as part of the job, and the minimal collection of features for the deep learning model was chosen. Training and test samples were also constructed.

The remainder of this paper is as follows: the next section reviews literature related to the research problem describing state-of-the-art research results in the area of dangerous urban sound detection. Section 3 explains materials and methods, including data collection, applied dataset, and architecture of the proposed deep learning model. Section 4 demonstrates obtained results, training and test results of the proposed CNN-RNN model for dangerous urban sound detection problems. Section 5 discusses the given problem, proposed approach, and open challenges in real-time impulsive sound detection. Section 6 concludes the study by demonstrating its results and discussing the proposed model.

2 Related Works

In the course of this research, large-scale preparations have been made to implement the "secure city" platform, which is comprised of a network of video analytical techniques. This service's primary purpose is to rapidly recognize and respond to a variety of emergency scenarios as well as incidences of law enforcement misconduct, with the end goal of ensuring the safety of citizens.

As an increasing number of modern video cameras come with built-in microphones and audio analytics, a method of identifying unusual or critical circumstances is now undergoing rapid development. In contrast to video analytics, audio analysis can be quickly completed and does not require devices or techniques for high processing.

The American ShotSpotter system is one of the renowned breakthroughs in audio analytics [21]. Since its installation in underserved regions of Washington in 2006, this system has localized 39,000 gunshots, promptly notifying authorities of the start of an incident.

The creation of Audio Analytical Ltd from the United Kingdom (UK) is another example of adopting a sound detection system [22]. The "smart house" sensors, developed by the company, can detect noises like gunshots, angry yells, screaming children, automobile alarms, shattering glass, etc. The system notifies the user and security authorities of further action when an incident is registered.

The AudioAnalytics project [23] offers many solutions at once for distinct use cases. The suggested solution's architecture is as follows: On the user's device, the CoreLogger software enables the receiving, displaying, and saving of alarm events. It works in tandem with Sound Packs, a component of the entire system that is nothing more than a collection of several audio analytics modules [24]. These modules' primary functions include the recognition of various audio events, such as aggression (high-pitched conversation, shouting), car alarm system, breaking glass, search for keywords ("police", "help", etc.), shots fired, cry of a child, and explosion.

These technologies precisely pinpoint the location of an event by employing triangulation techniques and microphone locations [25–28]. The inexpensive cost of loudspeakers, as opposed to camcorders, the lack of "blind spots", and the lesser density of territorial coverage compared to video surveillance are all undeniable benefits of audio analytics systems. Moreover, it is easier to transfer and process the audio signals from the microphone since it is less in size than the visual stream.

Nowadays, smartphones linked to data channels of cellular networks are considered to be the most convenient tools for accomplishing sound detection. The key benefit of this approach is that smartphones are becoming more and more common in the world. The most recent statistics show that globally, more than 30% of people between the ages of 18 and 50 own smartphones. Every smartphone contains a microphone, a channel for exchanging information, a location sensor, and the ability to install other software [29–32].

The smartphone is an alternative worth considering as a technique for the timely identification of an emergency as well as a way of personally informing the user about the existence of an emergency in their immediate vicinity. To put it differently, if the cellphone is equipped with software applications and a transmitting data channel that is fully operational, it will be capable of recognizing audio that can classify a disaster and send data on the registered sounds, as well as information about its destination and the exact moment, to a framework for analysis in distance. In other words, the device will be capable of detecting sounds that indicate an emergency and sending this information to a system for distant location analysis.

Terrorist threats, breaches of public order, and numerous human-caused calamities that are accompanied by loud explosions, sirens, and other sonic artifacts are examples of crises that may be detected by this approach. There are certain knowledge gaps in the identification of impulsive sounds:

(1) The lack of an impulse sound database. To fill this knowledge vacuum, this study created a dataset of impulsive noises. Ten thousand sounds were included in the collection, which was divided into eight classes: gunshot, glass breaking, fire, explosion, scream, siren, dog barking, and fire alarm bell. Noises from the predefined categories might be used to train machine learning models to identify impulsive sounds quickly and accurately.
(2) Data input noise. The input data must be noise-free to achieve high accuracy in the detection of impulsive sounds. To enhance the accuracy of the detection procedure, this study eliminated all the noises from the dataset.
(3) Deep learning model. To achieve high detection accuracy, the proposed deep learning model is crucial. Deep learning techniques can use different hidden parameters, hyperparameters, and features, and give high classification and detection rates.

3 Materials and Methods

This section describes the materials and methods that were employed in this study, the dataset used to identify dangerous urban sounds, as well as the data collection, preprocessing stage, and development of a deep learning model for dangerous urban sound detection. Fig. 1 demonstrates the overall architecture of the proposed framework for dangerous urban sound detection in real time. The following subsections explain the materials and methods step-by-step by demonstrating datasets that were applied during the research, data collection process, model overview, impulsive sound detection problem, and demonstration of the proposed deep CNN-RNN architecture.

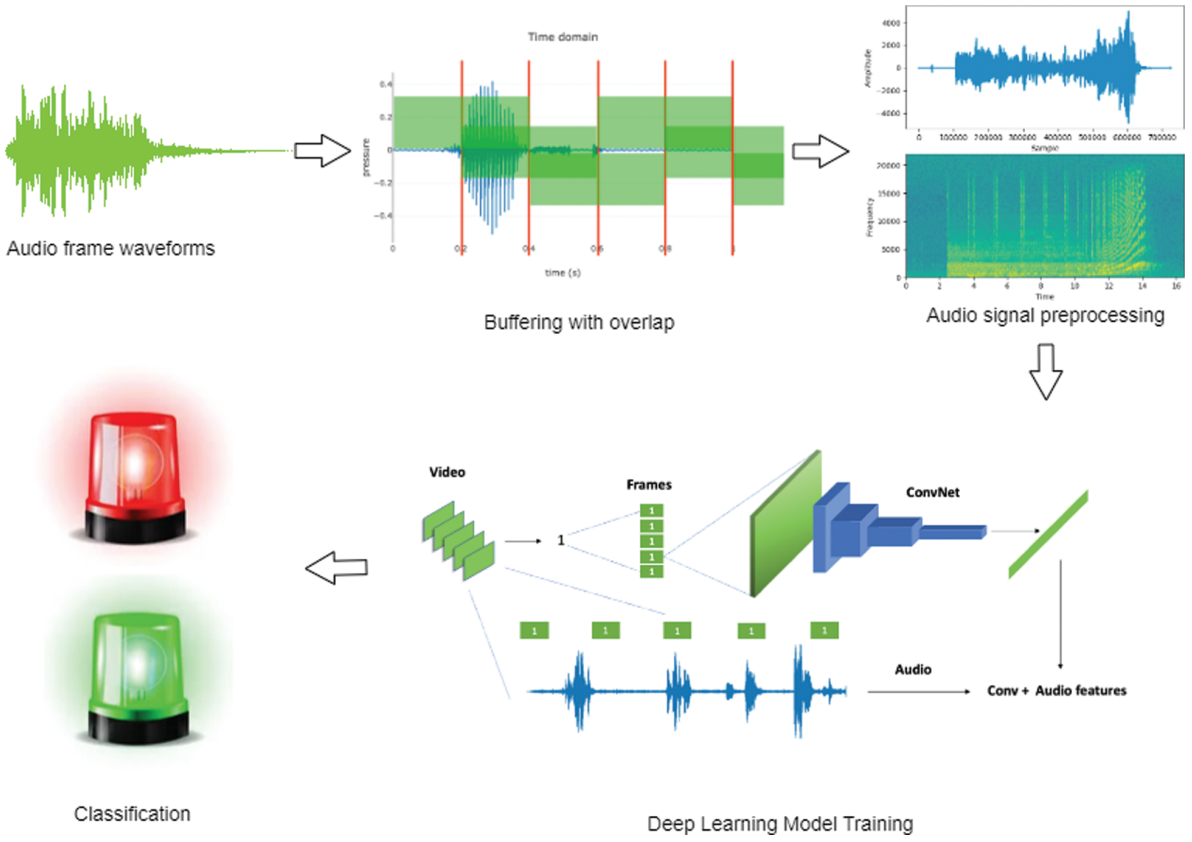


Figure 1: Architecture of the proposed framework

3.1 Dataset

The first stage of the study comprised data collection, since conducting any research requires a huge amount of information. Several large-scale datasets were used to analyze the so-called "dangerous" sounds. For testing the program, the environmental sound classification (ESC-50) dataset was chosen (and out of 2,000 sounds, the study selected about ~300 sounds) [33].

The list of sounds for this dataset included categories such as animal sounds, natural sounds, human sounds, household or everyday sounds, and dangerous sounds.

- Animal sounds (dogs, cats, cows, pigs)
- Natural sounds or sounds of nature (rain, sea, birds, lightning)
- Human sounds (baby crying, footsteps, coughing, breathing)
- Household or everyday sounds (knocking on the door, typing on the keyboard, alarm clock, breaking glass)
- Dangerous sounds (police siren, train, engine, chainsaw, airplane, fireworks, explosion, dog barking gunshots, and other dangerous impulsive sounds)

Despite such an impressive amount of information, this study utilized only dangerous sounds in the initial stage, cutting out the rest. Technical parameters of the developed dataset in comparison with the original one are demonstrated in Table 1.

Table 1: Technical parameters of the developed dataset

Characteristics	Accuracy
Overall size	661 MB
Size after preprocessing	45 MB
Number of files	2000
Number of files after preprocessing	301
Extension of files	.ogg

Shots, screaming, and smashed windows were among the occurrences that were deemed "odd" in the area being studied. Therefore, the performance of the suggested system was assessed for an automated surveillance application.

To achieve this goal, this study created a dataset using several audio samples captured in different settings at train stations. The dataset included 8,000 different dangerous urban sounds of eight classes in audio format. The proposed dataset can be applied to train and test the machine learning and deep learning models for the detection and classification of dangerous urban sounds.

The dataset mainly comprised background sounds such as pick, shot, and shattered glass signals. To take into consideration the features of various application contexts, background sounds were collected both internally and externally.

For research purposes, the signals were split into intervals of one second (the typical duration of each event of interest), and each interval was further split into frames of 200 MS, overlapping by 50%. Specifically, each interval was made up of nine frames.

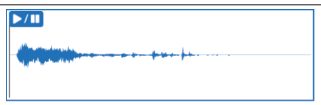
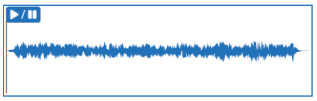
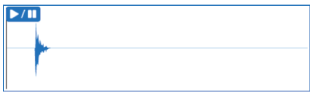
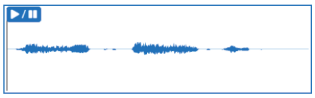
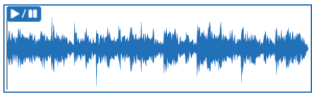
Table 2 provides a summary of the signals, frames, and intervals that made up the dataset. This table demonstrated the description of distinct dangerous urban sounds and their spectrograms. Table 2 explained spectrograms of several samples of impulsive sounds, such as automobile glass shattering, dog barking, police siren, ambulance siren, constant wail from police siren, single gunshot, explosion, baby crying, burglar alarm, fire alarm beeping, fire alarm bell, fire alarm yelp, and smoke alarm. Therefore, the table can express the importance of the proposed dataset and the proposed CNN-RNN deep learning architecture.

3.2 Model Overview

The next stage was logic programming. The main part of this project stage was to figure out ways to pick up sounds in general. The challenge of identifying various frightening sound occurrences may be broken down into two separate subtasks [34]:

- The detection and separation, within the audio data stream, of distinct pulse signals that are distinct from background noise.
- The categorization (identification) of the received signal as belonging to one of the several kinds of acoustic occurrences.

Table 2: Samples of impulsive sounds in the developed dataset

Sound	Time (sec)	Spectrograms
Automobile glass shattering	3.84	
Dog barking	22.15	
Police siren	24.19	
Ambulance siren	15.41	
Constant wail from police siren	56.87	
Single gunshot	3.84	
Explosion	7.78	
Baby crying	6.66	
Burglar alarm	11.13	
Fire alarm beeping	1.41	
Fire alarm bell	1.59	
Smoke alarm	0.99	
Fire alarm yelp	2.3	

3.3 Detection of Impulsive Sound Events

Determining the power for a collection of successive audio signal blocks that do not overlap is the foundation for several different ways [9]. The formula used to calculate the power of the kth signal block, which is composed of N samples, is as follows (1):

$$e(k) = \frac{1}{N}\sum_{n=0}^{N-1} x^2(n + kN) \qquad k = 0, 1, \ldots \tag{1}$$

For example, the sound of a shot happened in about 4.6 s and the range of power values for blocks of N = 4,000 records, which relates to the duration of each block, is around 90 milliseconds. The block that is autonomously identified in response to a quick pulse noise is done so in numerous different ways, depending on the method:

- According to the standard deviation of the data that have been standardized for power units;
- Using the middle value of the median filter for the power units;
- Establishing dynamic limits for the power units. A closer look at the method suggests that the methodology relies on the standard deviation of the normalized values of power units as its foundation. It has been determined that the normalized values of the power blocks that fall within the range [0,1] are the most important component of this strategy.

$$e_{norm}(j) = \frac{e_{win}(j) - \min_j(e_{win}(j))}{\max_j\left(e_{win}(j) - \min_j(e_{win}(j))\right)} \tag{2}$$

The next step was to compute the standard deviation, often known as the variance, of a given group of values:

$$\text{var}(k) = \frac{1}{L-1}\sum_{j=0}^{L-2}\left[e_{norm}(j,k) = \overline{e}_{norm}(k)\right]^2 \tag{3}$$

Whenever ambient noise is present, the block powers tend to be uniformly distributed between [0,1]. (seen on the left). As the new power value for the audio unit is the re-normalized values within the specified range, the unit is automatically detected with a pulse signal if a significantly higher power level occurs, in comparison to the previously set values for the power of the background units. A slow-changing signal may be detected by examining the mean value of normalized power sources, and the approach is robust against variations in noise levels.

3.4 Proposed Model

This study proposed CNN in combination with RNN, where RNN would not be a recurrence for the CNN itself, but rather as a separate layer with ReLU activation for information. The dimension of RNN would be 128. Its architecture is shown in Fig. 2 as follows.

3.5 Feature Extraction

It is worth noting that the feature extraction process with sounds took about an hour and a half, provided that it was only a 6.6 GB dataset. A 6.6 GB dataset was used so far as the study attempted to try it on a relatively simple dataset. In the second one, the study would use achieved techniques and run the data only once. Moreover, after parsing all the sounds, only 8,674 would be left in a duration

of 5,439 s or 90 min. By looking at the feature for getting part of the code, then everything can be visualized in a diagram as displayed in Fig. 3.

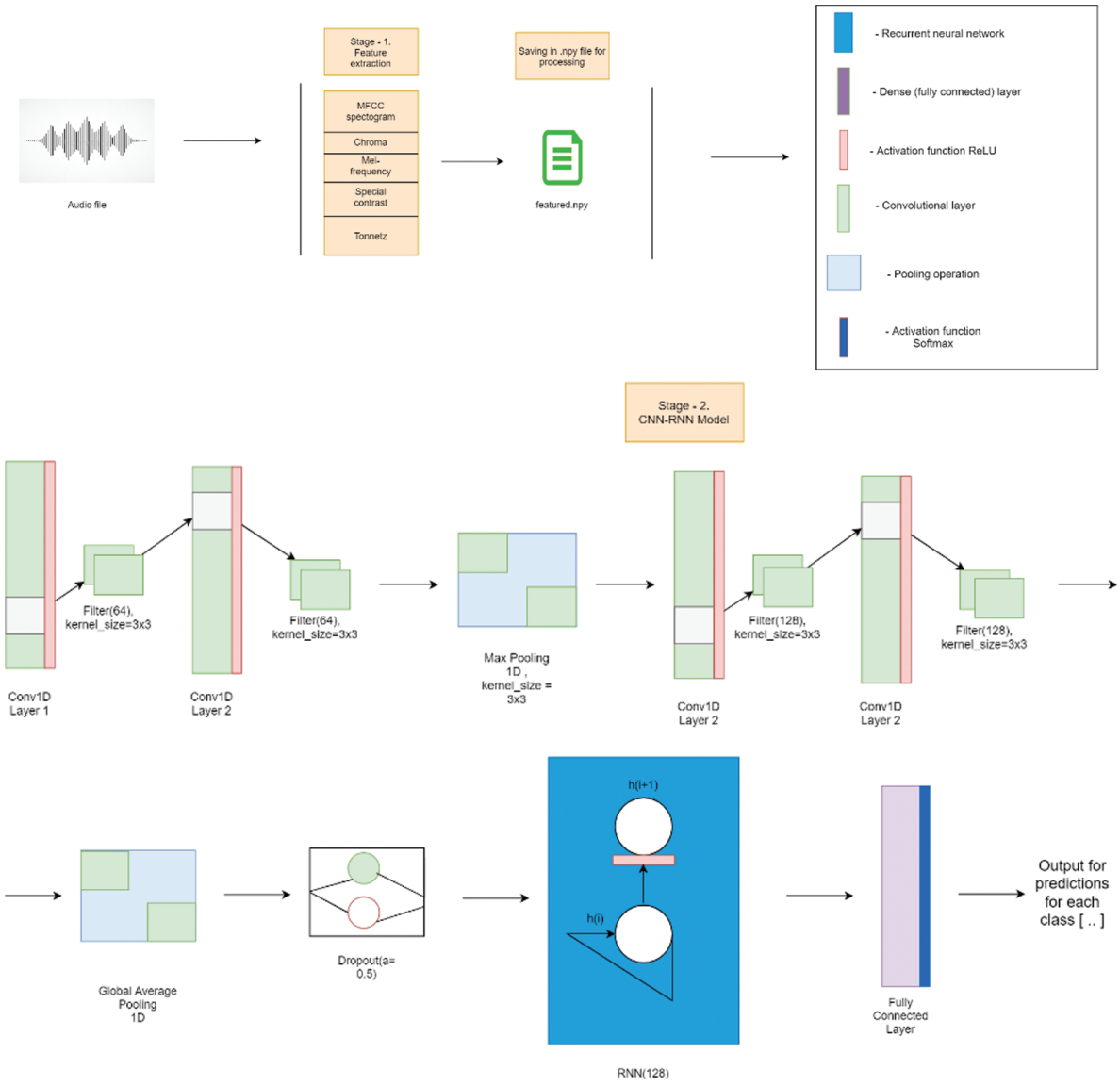


Figure 2: Architecture of the proposed model

There are four functions, three of which were responsible for obtaining features and for further saving. Specifically: "load_files_and_save()"—called "load_data()", which in turn iterated through the data and obtained the individual sound and then called "get_features_from()" to get the features of the specified sound. Afterward, everything would be saved into two files for further work (features, labels) in ".npy" format After saving, data were obtained through the load_featured_files() function. The training phase started with the RNN-CNN model as described in Fig. 2. It consisted of two convolution layers: one from a global maximum pool and another from a global averaging pool.

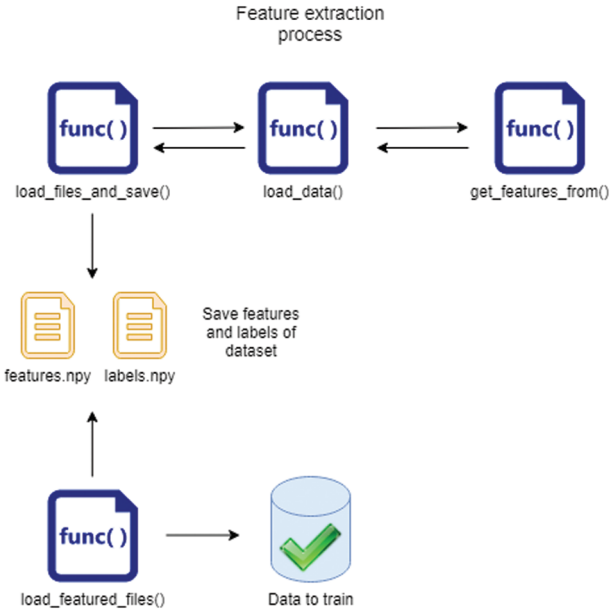


Figure 3: Architecture of the proposed model

4 Experiment Results

This section highlights the experiment results of the proposed CNN-RNN approach for dangerous urban sound detection problems. Firstly, the evaluation metrics to assess the proposed CNN-RNN deep learning model are provided. Subsequently, training and test results are introduced, including the accuracy and losses of the proposed model, and the confusion matrix for each impulsive sound class. In addition, the study demonstrates the accuracy of each class in Table 3 by giving the percentage of each category, including accuracy, precision, recall, F-score, and area under the curve receiving operating characteristics (AUC-ROC) curve.

Table 3: Samples of impulsive sounds in the developed dataset

Sound	Accuracy	Precision	Recall	F-score	AUC-ROC
Gunshot	0.8937	0.9012	0.9206	0.8754	0.9557
Broken glass	0.9131	0.9534	0.9004	0.8961	0.9334
Fire	0.9203	0.9112	0.9097	0.9139	0.9345
Siren	0.9306	0.9223	0.9669	0.9427	0.9402
Explosion	0.7901	0.8025	0.8141	0.7912	0.9117
Cry	0.8404	0.8309	0.8633	0.8533	0.9228
Dog barking	0.8219	0.8108	0.8322	0.8026	0.9264
Fire alarm	0.8413	0.8221	0.8347	0.8214	0.9253

4.1 Evaluation Metrics

In impulsive urban sound detection, the study used accuracy, precision, recall, F-measure, and AUC-ROC curve as evaluation parameters. Eqs. (1)–(4) demonstrated the formula of each evaluation parameter. Besides, the study employed train loss, test loss, and number of epochs to explain the efficiency of the proposed CNN-RNN deep model. Here, TP stands for true positive, TN stands for true negative, FP stands for false positive, FN stands for false negative, P stands for positive values, and N is negative values.

$$accuracy = \frac{TP + TN}{P + N} \tag{4}$$

$$precision = \frac{TP}{TP + FP} \tag{5}$$

$$recall = \frac{TP}{TP + FN} \tag{6}$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \tag{7}$$

4.2 Experiment Results

This section illustrates the experiment results of dangerous urban sound detection. Figs. 4 and 5 show the training and test results of the CNN-RNN model for impulsive sound detection. Fig. 4 demonstrates the training and test results for the first dataset by the research group. The CNN-RNN model achieved 95% accuracy in training for about 80 epochs. The number of parameters achieved was about 87,822. Thus, the training lasted about ~267 s, which was over four minutes.

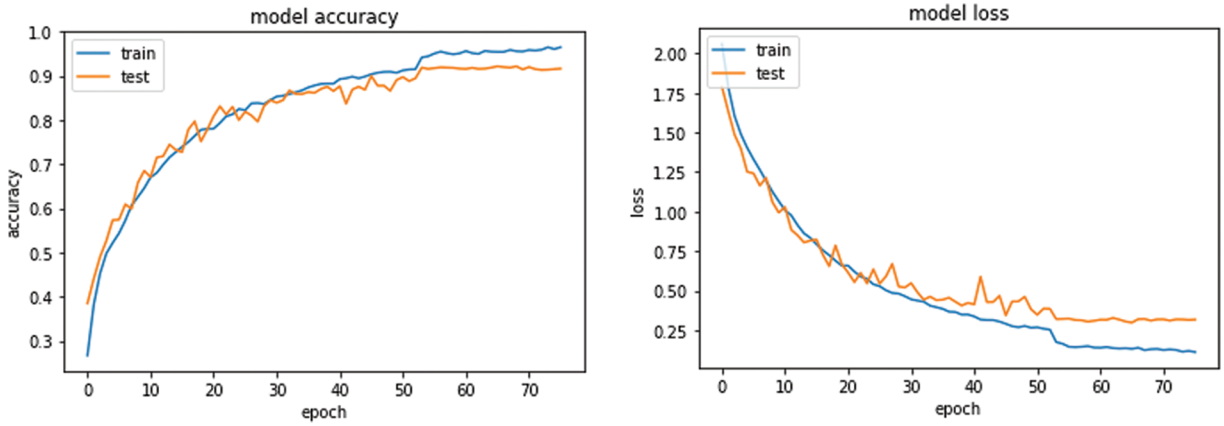


Figure 4: CNN-RNN training (left—accuracy, right—loss)

Fig. 5 exhibits the training and test results for the second dataset that was taken from an open source. The applied CNN-RNN model required about 110 epochs and gave an accuracy of about 92%. The second part of Fig. 5 demonstrates training and test losses. As illustrated in the figure, the test results remained permanent after 60 epochs.

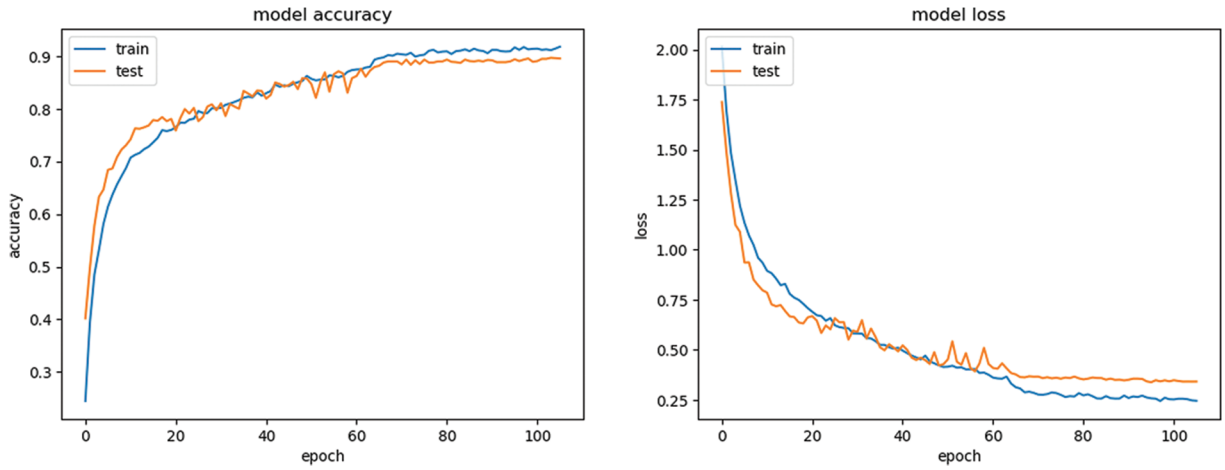


Figure 5: CNN-RNN training (left—accuracy, right—loss)

The trained model allowed for the acquisition of the confusion matrix to identify the accuracy of false positive, false negative, true positive, and true negative samples depending on various types of urban sounds and the prediction percentage. Fig. 6 illustrates the confusion matrix for the classification of impulsive sounds. The applied CNN classified eight types of dangerous urban sounds.

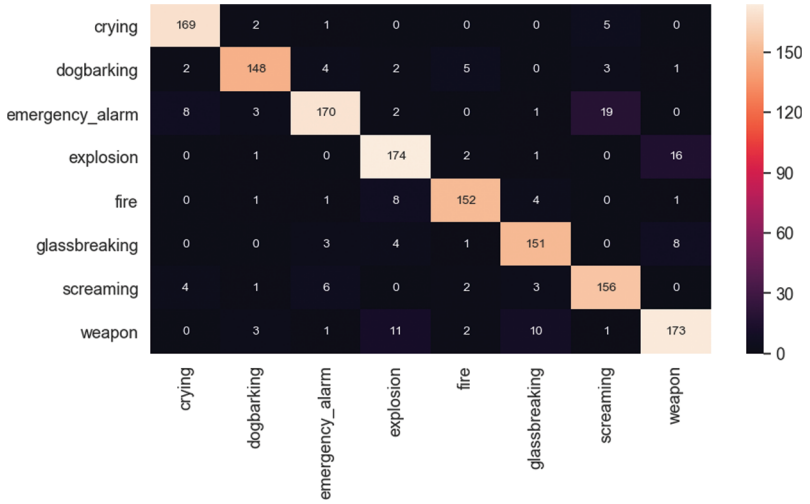


Figure 6: Confusion matrix

Fig. 7 demonstrates the AUC-ROC curve in the dangerous sound event detection problem. This roughly showed how the classifier output was affected by changes in the training data. The obtained results revealed that the proposed CNN-RNN model classified dangerous sound events with high accuracy. According to the graph, a stable result can be seen, ensuring that the algorithm was well-trained to identify dangerous sound events.

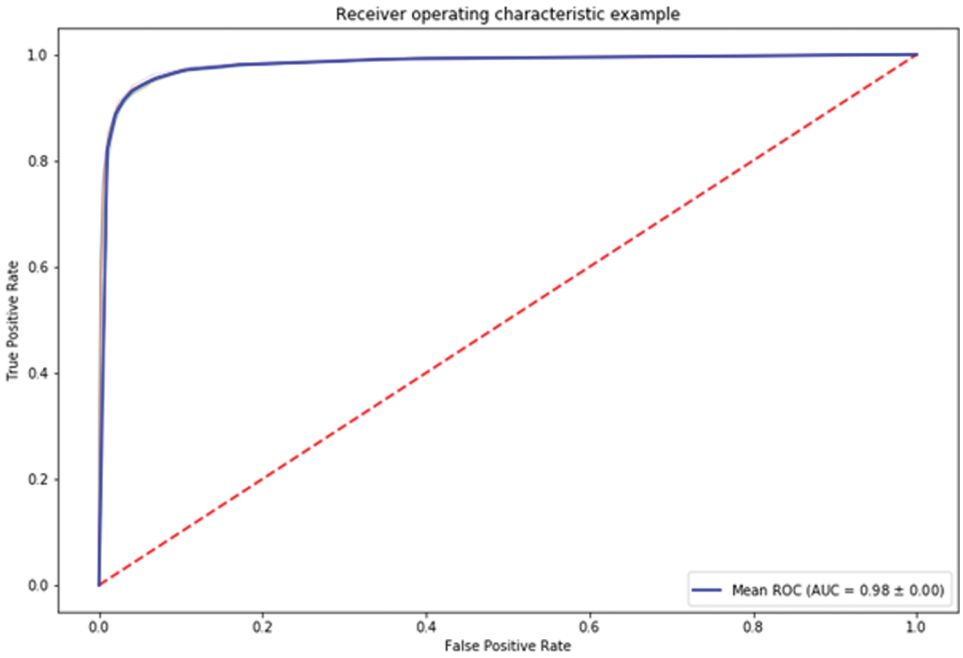


Figure 7: AUC-ROC curve

As can be seen from the graphs, the results were rather good, 83% minimum in emergency alarm and 95% in crying sound predictions. Table 3 demonstrates the accuracy of the applied CNN for the impulsive sound detection problem and allows for the evaluation of each dangerous impulsive urban sound class by different parameters, including accuracy, precision, recall, F-score, and AUC-ROC value.

As a result, the proposed deep neural network has the greatest performance for consistently identifying dangerous urban sounds across all assessment criteria. The application of the suggested deep RNN-CNN for weight and bias tuning as well as a reduction in training time may be responsible for the proposed approach's successful outcomes. The results demonstrated that the suggested deep neural network may be easily modified to support both short and lengthy texts in the present.

5 Discussion

One of the most crucial requirements for urban management is crime prevention. In this context, video monitoring is an integral part of crime detection; however, it can only provide a visual element. Environment-related noises that may improve spatial awareness should be included in a more comprehensive approach.

Both streaming streams and archival files may be processed using audio analytics. The technique may be employed in certain scenarios in lieu of video surveillance since it can detect noises in full darkness and because microphones are less expensive and need less maintenance than cameras [35–37]. It is possible to identify specific sounds in an audio stream, remove background noise from audio recordings, identify people by their voices, improve the clarity of the speaker's voice, and detect issues with the operation of mechanisms by using sound recognition technology [38].

To combat crime, several cities use video surveillance systems. Nevertheless, prior research claimed that using video surveillance alone is insufficient for deterring and averting crimes [39–41]. Today, the most crucial elements of urban security solutions are threat detection and data analysis technology, such as motion sensors, thermal imaging systems, and license plate recognition software. However, they mainly focus on visual aspects, while sound detection technology should be a part of any strategy that addresses urban security in its entirety, according to experts [42]. By notifying offenders via a loudspeaker, a security system with audio transmission will enable operators to hear whether a person is in danger, offer them information, or deter them from breaking into a building.

According to recent studies, verbal aggressiveness occurs 90% of the time before physical aggression [43–48]. According to the paper, the advantage of the aggressive sound detection system is that it enables security officers to hear the tension in voices and other noises connected to verbal aggression, fear, and other negative emotions. Security and law enforcement professionals can distinguish between noises that are relevant to them and those that are unrelated to them with the use of audio analytics. The software used to identify violence analyzes the generated sounds using sophisticated algorithms and compares them to patterns. The program quickly alerts the security crew if it determines that the sound is notable. Table 4 compares the proposed model with the state-of-the-art research results. As it is illustrated in the table, the proposed model can classify different types of dangerous sounds.

Table 4: Samples of impulsive sounds in the developed dataset

Model	Sound event	Feature	Results	Reference
Proposed model	Gunshot, cry, broken glass, dog barking, fire, siren, explosion, fire alarm	Spectrogram	0.79–0.93 accuracy; 0.80–0.95 precision; 0.83–0.96 recall; 0.79–0.94 f-score; 0.91–0.95 AUC-ROC	–
Bagged tree	Gunshot	LPC, MFCC, GTCC	0.97 true positive rate	Kabir et al. [49]
Fire	Gunshot	Antilog energy features	0.931 true positive	Rahman et al. [50]
Neural network	Gunshot	MFCC	0.95	Sigmund and Hrabina [51]
Machine learning	Fire alarm	–	0.96 accuracy	Gupta [52]

The sounds of hostility (such as verbal insults) and the sounds of weapons are the two types of noises that are most crucial to be able to assess in maintaining security in cities. Law enforcement organizations will be able to combat crime more successfully with the use of systems for detecting noises of hostility and the usage of guns.

6 Conclusion

In conclusion, it is worth noting that the current study presents prospects for future research and is aimed to attract the attention of relevant companies in deep learning and data classification. This research employed available technologies in the deep learning area and emphasized opportunities to work with the model for production purposes.

The study analyzed the technologies that can be used to obtain the required outcome. After all, the field of deep learning is immense and has not yet been thoroughly explored. It can be assumed that this project shows only a small fragment of deep learning's capabilities. Nevertheless, this is a great start for a full study of this area and further development in different applications (for instance, security organizations, public address systems, and in general, as a private device for sound analysis).

Another feasible employment of deep learning concerns alarm systems, where it could set up constant monitoring and subsequently record on a trigger, or simply display a sound graph, thus providing more convenient and accessible observation, especially at nighttime.

Funding Statement: The paper is funded by the project, "Design and implementation of real-time safety ensuring system in the indoor environment by applying machine learning techniques". IRN: AP14971555.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] J. Bajzik, J. Prinosil, R. Jarina and J. Mekyska, "Independent channel residual convolutional network for gunshot detection," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 4, pp. 950–958, 2022.

[2] K. M. Nahar, F. Al-Omari, N. Alhindawi and M. Banikhalaf, "Sounds recognition in the battlefield using convolutional neural network," *International Journal of Computing and Digital Systems*, vol. 11, no. 1, pp. 189–198, 2022.

[3] I. Estévez, F. Oliveira, P. Braga-Fernandes, M. Oliveira, L. Rebouta *et al.,* "Urban objects classification using Mueller matrix polarimetry and machine learning," *Optics Express*, vol. 30, no. 16, pp. 28385–28400, 2022.

[4] Z. Peng, S. Gao, Z. Li, B. Xiao and Y. Qian, "Vehicle safety improvement through deep learning and mobile sensing," *IEEE Network*, vol. 32, no. 4, pp. 28–33, 2018.

[5] Y. Wei, L. Jin, S. Wang, Y. Xu and T. Ding, "Hypoxia detection for confined-space workers: Photoplethysmography and machine-learning techniques," *SN Computer Science*, vol. 3, no. 4, pp. 1–11, 2022.

[6] Y. Arslan and H. Canbolat, "Sound based alarming based video surveillance system design," *Multimedia Tools and Applications*, vol. 81, no. 6, pp. 7969–7991, 2022.

[7] K. Pawar and V. Attar, "Deep learning approaches for video-based anomalous activity detection," *World Wide Web-Internet and Web Information Systems*, vol. 22, no. 2, pp. 571–601, 2019.

[8] S. U. Amin, M. S. Hossain, G. Muhammad, M. Alhussein and M. A. Rahman, "Cognitive smart healthcare for pathology detection and monitoring," *IEEE Access*, vol. 7, no. 1, pp. 10745–10753, 2019.

[9] H. Zogan, I. Razzak, X. Wang, S. Jameel and G. Xu, "Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media," *World Wide Web-Internet and Web Information Systems*, vol. 25, no. 1, pp. 281–304, 2022.

[10] C. Heipke and F. Rottensteiner, "Deep learning for geometric and semantic tasks in photogrammetry and remote sensing," *Geo-Spatial Information Science*, vol. 23, no. 1, pp. 10–19, 2020.

[11] B. Omarov, N. Saparkhojayev, S. Shekerbekova, O. Akhmetova, M. Sakypbekova *et al.,* "Artificial intelligence in medicine: Real time electronic stethoscope for heart diseases detection," *Computers, Materials & Continua*, vol. 70, no. 2, pp. 2815–2833, 2022.
[12] A. Rajbanshi, D. Das, V. Udutalapally and R. Mahapatra, "DLeak: An IoT-based gas leak detection framework for smart factory," *SN Computer Science*, vol. 3, no. 4, pp. 1–12, 2022.
[13] Y. Arslan and H. Canbolat, "Sound based alarming based video surveillance system design," *Multimedia Tools and Applications*, vol. 81, no. 6, pp. 7969–7991, 2022.
[14] R. Sun, Q. Cheng, F. Xie, W. Zhang, T. Lin *et al.,* "Combining machine learning and dynamic time wrapping for vehicle driving event detection using smartphones," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 1, pp. 194–207, 2019.
[15] G. Chen, F. Wang, S. Qu, K. Chen, J. Yu *et al.,* "Pseudo-image and sparse points: Vehicle detection with 2D LiDAR revisited by deep learning-based methods," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 12, pp. 7699–7711, 2020.
[16] B. Omarov, A. Tursynova, O. Postolache, K. Gamry, A. Batyrbekov *et al.,* "Modified unet model for brain stroke lesion segmentation on computed tomography images," *Computers, Materials & Continua*, vol. 71, no. 3, pp. 4701–4717, 2022.
[17] V. Osipov, N. Zhukova, A. Subbotin, P. Glebovskiy and E. Evnevich, "Intelligent escalator passenger safety management," *Scientific Reports*, vol. 12, no. 1, pp. 1–16, 2022.
[18] I. H. Peng, P. C. Lee, C. K. Tien and J. S. Tong, "Development of a cycling safety services system and its deep learning bicycle crash model," *Journal of Communications and Networks*, vol. 24, no. 2, pp. 246–263, 2022.
[19] L. Kou, "A review of research on detection and evaluation of the rail surface defects," *Acta Polytechnica Hungarica*, vol. 19, no. 3, pp. 167–186, 2022.
[20] L. M. Bine, A. Boukerche, L. B. Ruiz and A. A. Loureiro, "Leveraging urban computing with the internet of drones," *IEEE Internet of Things Magazine*, vol. 5, no. 1, pp. 160–165, 2022.
[21] S. Khan, L. Alarabi and S. Basalamah, "Toward smart lockdown: A novel approach for COVID-19 hotspots prediction using a deep hybrid neural network," *Computers*, vol. 9, no. 4, pp. 1–16, 2020.
[22] M. Dua, D. Makhija, P. Manasa and P. Mishra, "A CNN-RNN–LSTM based amalgamation for Alzheimer's disease detection," *Journal of Medical and Biological Engineering*, vol. 40, no. 5, pp. 688–706, 2020.
[23] H. Gill, O. Khalaf, Y. Alotaibi, S. Alghamdi and F. Alassery, "Multi-model CNN-RNN-LSTM based fruit recognition and classification," *Intelligent Automation & Soft Computing*, vol. 33, no. 1, pp. 637–650, 2022.
[24] K. Chandriah and R. Naraganahalli, "RNN/LSTM with modified Adam optimizer in deep learning approach for automobile spare parts demand forecasting," *Multimedia Tools and Applications*, vol. 80, no. 17, pp. 26145–26159, 2021.
[25] S. Hansun and J. Young, "Predicting LQ45 financial sector indices using RNN-LSTM," *Journal of Big Data*, vol. 8, no. 1, pp. 1–13, 2021.
[26] Y. Xue, P. Shi, F. Jia and H. Huang, "3D reconstruction and automatic leakage defect quantification of metro tunnel based on SfM-deep learning method," *Underground Space*, vol. 7, no. 3, pp. 311–323, 2022.
[27] L. Zhang, L. Yan, Y. Fang, X. Fang and X. Huang, "A machine learning-based defensive alerting system against reckless driving in vehicular networks," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 12, pp. 12227–12238, 2019.
[28] A. M. Youssef, B. Pradhan, A. Dikshit, M. M. Al-Katheri, S. S. Matar *et al.,* "Landslide susceptibility mapping using CNN-1D and 2D deep learning algorithms: Comparison of their performance at Asir Region, KSA," *Bulletin of Engineering Geology and the Environment*, vol. 81, no. 4, pp. 1–22, 2022.
[29] S. Asadianfam, M. Shamsi and A. Rasouli Kenari, "Hadoop deep neural network for offending drivers," *Journal of Ambient Intelligence and Humanized Computing*, vol. 13, no. 1, pp. 659–671, 2022.
[30] L. M. Koerner, M. A. Chadwick and E. J. Tebbs, "Mapping invasive strawberry guava (Psidium cattleianum) in tropical forests of Mauritius with Sentinel-2 and machine learning," *International Journal of Remote Sensing*, vol. 43, no. 3, pp. 841–872, 2022.

[31] D. K. Dewangan and S. P. Sahu, "Deep learning-based speed bump detection model for intelligent vehicle system using raspberry Pi," *IEEE Sensors Journal*, vol. 21, no. 3, pp. 3570–3578, 2020.
[32] Z. Fang, B. Yin, Z. Du and X. Huang, "Fast environmental sound classification based on resource adaptive convolutional neural network," *Scientific Reports*, vol. 12, no. 1, pp. 1–18, 2022.
[33] V. Gugnani and R. K. Singh, "Analysis of deep learning approaches for air pollution prediction," *Multimedia Tools and Applications*, vol. 81, no. 4, pp. 6031–6049, 2022.
[34] X. Yang, L. Shu, Y. Liu, G. P. Hancke, M. A. Ferrag *et al.,* "Physical security and safety of IoT equipment: A survey of recent advances and opportunities," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 7, pp. 4319–4330, 2022.
[35] H. Kyle, P. Agarwal and J. Zhuang, "Monitoring misinformation on Twitter during crisis events: A machine learning approach," *Risk Analysis*, vol. 42, no. 8, pp. 1728–1748, 2022.
[36] M. Esmail Karar, O. Reyad, A. Abdel-Aty, S. Owyed and M. F. Hassan, "Intelligent IoT-aided early sound detection of red palm weevils," *Computers, Materials & Continua*, vol. 69, no. 3, pp. 4095–4111, 2021.
[37] T. Thomas Leonid and R. Jayaparvathy, "Classification of elephant sounds using parallel convolutional neural network," *Intelligent Automation & Soft Computing*, vol. 32, no. 3, pp. 1415–1426, 2022.
[38] Z. Ma, G. Mei and F. Piccialli, "Machine learning for landslides prevention: A survey," *Neural Computing and Applications*, vol. 33, no. 17, pp. 10881–10907, 2021.
[39] X. Zhao, L. Zhou, Y. Tong, Y. Qi and J. Shi, "Robust sound source localization using convolutional neural network based on microphone array," *Intelligent Automation & Soft Computing*, vol. 30, no. 1, pp. 361–371, 2021.
[40] B. Omarov, A. Batyrbekov, K. Dalbekova, G. Abdulkarimova, S. Berkimbaeva *et al.,* "Electronic stethoscope for heartbeat abnormality detection," in *5th Int. Conf. on Smart Computing and Communication (SmartCom 2020)*, Paris, France, pp. 248–258, 2020.
[41] A. Altayeva, B. Omarov and Y. Cho, "Towards smart city platform intelligence: PI decoupling math model for temperature and humidity control," in *2018 IEEE Int. Conf. on Big Data and Smart Computing (BigComp)*, Shanghai, China, pp. 693–696, 2018.
[42] F. Abid, "A survey of machine learning algorithms based forest fires prediction and detection systems," *Fire Technology*, vol. 57, no. 2, pp. 559–590, 2021.
[43] X. Yan, B. Cui, Y. Xu, P. Shi and Z. Wang, "A method of information protection for collaborative deep learning under GAN model attack," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 18, no. 3, pp. 871–881, 2019.
[44] D. Li, D. Zhao, Q. Zhang and Y. Chen, "Reinforcement learning and deep learning based lateral control for autonomous driving," *IEEE Computational Intelligence Magazine*, vol. 14, no. 2, pp. 83–98, 2019.
[45] Y. Sun, J. Liu, J. Wang, Y. Cao and N. Kato, "When machine learning meets privacy in 6G: A survey," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 4, pp. 2694–2724, 2020.
[46] M. Carminati, O. Kanoun, S. L. Ullo and S. Marcuccio, "Prospects of distributed wireless sensor networks for urban environmental monitoring," *IEEE Aerospace and Electronic Systems Magazine*, vol. 34, no. 6, pp. 44–52, 2019.
[47] N. Anantrasirichai, J. Biggs, K. Kelevitz, Z. Sadeghi, T. Wright *et al.,* "Detecting ground deformation in the built environment using sparse satellite InSAR data with a convolutional neural network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 4, pp. 2940–2950, 2020.
[48] Q. Chen, W. Wang, F. Wu, S. De, R. Wang *et al.,* "A survey on an emerging area: Deep learning for smart city data," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 3, no. 5, pp. 392–410, 2019.
[49] M. Kabir, J. Mir, C. Rascon, M. Shahid and F. Shaukat, "Machine learning inspired efficient acoustic gunshot detection and localization system," *University of Wah Journal of Computer Science*, vol. 4, no. 1, pp. 1–11, 2022.
[50] S. Rahman, A. Khan, S. Abbas, F. Alam and N. Rashid, "Hybrid system for automatic detection of gunshots in indoor environment," *Multimedia Tools and Applications*, vol. 80, no. 3, pp. 1–11, 2021.

[51] M. Sigmund and M. Hrabina, "Efficient feature set developed for acoustic gunshot detection in open space," *Elektronika Ir Elektrotechnika*, vol. 27, no. 4, pp. 62–68, 2021.

[52] N. Gupta, P. Deshpande, J. Diaz, S. Jangam and A. Shirke, "F-alert: Early fire detection using machine learning techniques," *International Journal of Electronics Engineering and Applications*, vol. 9, no. 3, pp. 34–43, 2021.