

Convolutional LSTM Network for Real-Time Impulsive Sound Detection and Classification in Urban Environments

Aigerim Altayeva¹, Nurzhan Omarov², Sarsenkul Tileubay³,
Almash Zhaksylyk⁴, Koptleu Bazhikov⁵, Dastan Kambarov⁶

Al-Farabi Kazakh National University, Almaty, Kazakhstan^{1,2}

Korkyt Ata Kyzylorda University, Kyzylorda, Kazakhstan³

Satbayev University, Almaty, Kazakhstan⁴

Yessenov University, Aktau, Kazakhstan⁵

Almaty University of Power Engineering and Telecommunications, Almaty, Kazakhstan⁵

Abstract—In recent years, the escalating challenges of noise pollution in urban environments have necessitated the development of more sophisticated sound detection and classification systems. This research introduces a novel approach employing a Convolutional Long Short-Term Memory (ConvLSTM) network tailored for real-time impulsive sound detection in metropolitan landscapes. Impulsive sounds, characterized by sudden onsets and short durations—such as honking, abrupt shouts, or breaking glass—are inherently sporadic but can significantly impact urban soundscapes and the well-being of city dwellers. Traditional sound detection mechanisms often falter in identifying these ephemeral noises amidst the cacophony of urban life. The ConvLSTM network proposed in this study amalgamates the spatial feature learning capabilities of Convolutional Neural Networks (CNN) with the temporal sequence retention attributes of LSTM, culminating in an architecture that excels in both sound detection and classification tasks. The model was trained and evaluated on a comprehensive dataset sourced from various urban settings and demonstrated commendable proficiency in discerning impulsive sounds with minimal false positives. Furthermore, the system's real-time processing capabilities ensure timely interventions, paving the way for smarter noise management in cities. This research not only propels the frontier of impulsive sound detection but also underscores the potential of ConvLSTM in addressing multifaceted urban challenges.

Keywords—Deep learning; CNN; LSTM; hybrid model; ANN; impulsive sound

I. INTRODUCTION

The burgeoning growth and urbanization of cities around the globe has, in recent decades, ushered in a plethora of environmental and societal challenges [1]. Among these challenges, noise pollution stands out as a particularly pervasive issue, affecting both the physical [2] and psychological health [3] of urban residents. While the continuous hum of traffic or the distant murmur of a crowded plaza might be classified as the usual sounds of urban life [4], impulsive sounds—those characterized by sudden onsets and fleeting durations—present a unique set of challenges. Whether it's the abrupt honk of a car, a sudden shout, or the shatter of glass, these noises can be more than just momentary

disturbances; they can disrupt sleep, exacerbate stress, and even influence long-term health outcomes.

Historically, noise monitoring in urban spaces has relied predominantly on traditional sound detection methodologies [5]. However, these conventional systems often lack the precision and agility needed to discern between the myriad of auditory signals that coexist in a bustling urban environment [6]. Specifically, the ability to distinguish impulsive sounds from the ambient noise milieu and subsequently classify them in real-time has remained a significant gap in urban noise management systems [7]. This lacuna is further widened when considering the increasing heterogeneity of urban sounds as cities continue to evolve and densify.

Enter the era of deep learning and its transformative impact across various domains. Recent advancements in neural networks, especially Convolutional Neural Networks (CNNs) [8], have demonstrated remarkable success in image and sound processing tasks. CNNs, designed to automatically and adaptively learn spatial hierarchies from data, have fundamentally altered the landscape of sound analysis in controlled environments. However, the temporally fleeting nature of impulsive sounds in dynamic urban environments presents challenges that go beyond the scope of traditional CNNs. It necessitates the incorporation of temporal sequence learning, an attribute inherent to Long Short-Term Memory (LSTM) networks.

LSTM networks, a subtype of recurrent neural networks, excel at tasks that require the understanding of long-term dependencies, making them particularly suited for sequence prediction problems, like those seen in speech and time-series data [9]. The integration of CNN's spatial feature learning with LSTM's prowess in temporal sequence retention could potentially hold the key to a robust solution for impulsive sound detection and classification in urban locales. This potential amalgamation gave birth to the Convolutional LSTM (ConvLSTM) network—a hybrid model aiming to harness the strengths of both parent architectures.

This research pivots around the design, implementation, and evaluation of a ConvLSTM network tailored explicitly for

the detection and classification of impulsive sounds in real-time urban settings. Recognizing the profound implications of efficient noise management—ranging from urban planning and policy-making to the well-being and satisfaction of city residents—this study endeavors to bridge the existing technological gap. Through the marriage of convolutional and recurrent mechanisms, we embark on an exploration into a new frontier of urban sound management, positing a solution that promises both accuracy and timeliness in addressing the cacophony of modern urban life.

II. RELATED WORKS

The quest to discern and classify impulsive sounds within urban environments is embedded in a rich tapestry of research efforts, encompassing fields from acoustic engineering to artificial intelligence. This section delves into the pertinent literature, highlighting seminal works, and tracing the evolution of methodologies applied to this challenge.

A. Urban Sound Detection and Classification

One of the earliest works in urban sound classification was presented by [10], who utilized basic spectral features coupled with Support Vector Machines (SVM) to classify a limited set of urban sounds. Their model, though pioneering, had a limited scope in differentiating between closely related sounds. A more comprehensive approach was introduced by [11], which focused on extracting Mel-Frequency Cepstral Coefficients (MFCC) from urban soundscapes. Their work laid the foundation for many subsequent endeavors by demonstrating the potential of MFCC in capturing the nuances of urban noises.

B. Convolutional Neural Networks in Sound Analysis

The revolution brought about by deep learning in image processing soon trickled into the realm of acoustic analysis. Authors in the study [12] were among the first to employ CNNs for environmental sound classification. His model, though primarily geared towards stationary sounds, showcased the profound potential of CNNs in capturing intricate sound patterns. Further advancements by [13] extended the use of CNNs, leveraging transfer learning from pre-trained image-based networks to sound data, highlighting the shared hierarchical structures between the two domains.

C. LSTM and Sequence Modeling in Acoustics

Long Short-Term Memory (LSTM) networks, a subtype of recurrent neural networks (RNNs), have emerged as particularly influential in the realm of acoustic analysis. Their inherent capability to capture and model long-term dependencies within sequences makes them exceptionally suited for time-based sound data. Researchers in [14] effectively harnessed LSTMs for voice activity detection, shedding light on their potential in discerning complex temporal patterns. Furthermore, [15] built upon this by integrating LSTMs with attention mechanisms, aiming to identify anomalous sounds in industrial settings. These studies collectively underscore the pivotal role of LSTMs in advancing the frontier of sequence modeling within the acoustics domain.

D. ConvLSTM in Image and Video Processing

Before its foray into acoustic analysis, ConvLSTM made waves in the domain of video processing. Researchers in [16] introduced ConvLSTM as an extension to the traditional LSTM, integrating convolution operations into the recurrent updates. Their groundbreaking work in precipitation forecasting exhibited ConvLSTM's potential in spatiotemporal sequence forecasting. This novel architecture caught the attention of many, with [17] later applying it to video classification tasks, proving its versatility across multiple temporal data types.

E. Hybrid Models in Sound Detection

The intersection of diverse neural network architectures has given rise to hybrid models, which aim to leverage the unique strengths of each constituent network for enhanced performance in sound detection tasks. Recognizing the potential of such amalgamations, [18] introduced a Convolutional Recurrent Neural Network (CRNN) specifically tailored for detecting anomalous sounds within varied environments. By seamlessly integrating the spatial feature extraction capabilities of Convolutional Neural Networks (CNNs) [19] with the temporal sequence modeling prowess of Recurrent Neural Networks (RNNs) [20], their research set a new benchmark in the field. This innovative approach highlights the intrinsic benefits of harnessing both spatial and temporal dimensions in sound analysis. Hybrid models, as delineated by such pioneering works, elucidate the way forward, promising enhanced accuracy and adaptability in the complex domain of sound detection.

F. Real-Time Sound Processing

The necessity for real-time sound processing, particularly in urban contexts, is paramount. Early systems, as chronicled by [21], relied heavily on handcrafted features and simplistic classifiers. However, with the proliferation of deep learning, architectures evolved to address real-time demands. Authors in [22] presented a sound event detection system employing a stacked CNN-LSTM architecture capable of real-time sound localization and classification, illuminating the pathway for subsequent real-time models.

G. Challenges in Sound Detection Models

Navigating the complex domain of hybrid sound detection models introduces a myriad of challenges. Firstly, the integration of diverse architectures, such as CNNs and RNNs [23], presents computational burdens. The enhanced model complexity often results in increased training times, demanding more computational resources and potentially hindering real-time deployment. Additionally, the fusion of spatial and temporal data streams can lead to overfitting [24], especially when training on limited datasets [25], necessitating rigorous regularization techniques and data augmentation [26].

Another pertinent challenge is the adaptation to diverse acoustic environments [27]. Hybrid models, though versatile, may struggle with high variability in soundscapes, from fluctuating noise levels to the unique acoustic signatures of different urban settings [28]. Lastly, the interpretability of these hybrid models remains elusive. As the models grow in complexity, understanding their decision-making processes

becomes intricate, posing challenges for validation and further refinement. Addressing these challenges is crucial for the successful adoption and efficacy of hybrid models in real-world sound detection applications.

In summation, while the corpus of work surrounding urban sound detection is extensive, the specific challenge of real-time impulsive sound detection and classification in urban settings remained a largely unexplored niche. The incorporation of ConvLSTM in this context, as embarked upon in this research, represents a synthesis of past insights and current innovations, promising to further the field's understanding and capabilities.

III. MATERIALS AND METHODS

This segment delineates the methodologies and resources employed throughout this investigation. It encompasses the dataset curated for discerning perilous urban acoustics, coupled with detailed insights into data assimilation, preparatory phases, and the construction of an intricate neural network model targeting the identification of hazardous urban noises. Fig. 1 provides a schematic representation of the devised framework, emphasizing real-time perilous urban acoustic detection. Subsequent subsections offer a comprehensive breakdown of the utilized resources and techniques, encapsulating the datasets engaged, the data assimilation paradigm, model conceptualization, challenges associated with impulsive acoustic detection, and a detailed exposition of the integrated CNN-LSTM architecture.

A. Data

At the study's inception, data acquisition was prioritized, recognizing the necessity of comprehensive information for robust research outcomes. An assortment of expansive datasets was engaged to scrutinize the sounds termed as "hazardous." The Environmental Sound Classification (ESC-50) dataset [29] emerged as the preferred choice for program evaluation. From its vast repertoire of 2,000 auditory samples, a curated subset of approximately 300 sounds was employed. The ESC-50's categorization encompassed:

- Faunal Acoustics (e.g., canines, felines, bovines, swine).
- Natural Phenomena (e.g., precipitation, oceanic waves, avian calls, electrical discharges).
- Human-Origin Sounds (e.g., neonatal distress, ambulatory noises, respiratory patterns).
- Domestic and Mundane Noises (e.g., door interactions, digital keystrokes, time alerts, vitreous fractures).
- Hazardous Acoustics (e.g., emergency vehicular alerts, rail transit, motor operations, timber-cutting equipment, aerial transport, pyrotechnics, detonations, canine alerts, ballistic discharges, and other precarious impulsive acoustics).

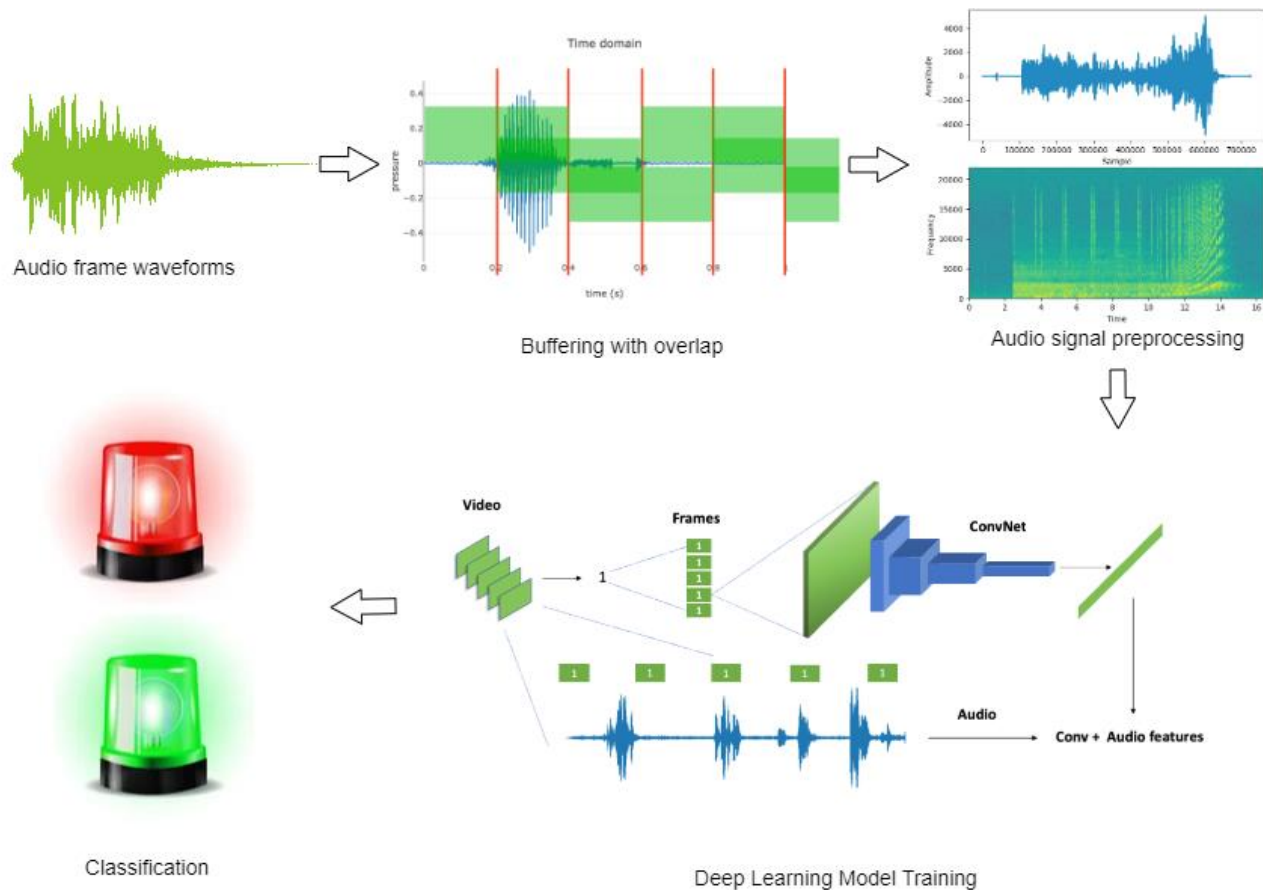


Fig. 1. Architecture of the proposed framework.

Despite the extensive nature of the ESC-50, this investigation was particularly centered on the hazardous acoustics subset, eschewing the remainder. A comparative analysis between the bespoke dataset's technical parameters and the ESC-50's original specifications can be referenced in Table I.

TABLE I. 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

Within the observed region, incidents characterized by gunshots, vocal distress, and fragmented glass were identified as anomalous or "atypical." Consequently, the efficacy of the proposed framework was scrutinized, targeting its applicability in automated monitoring systems.

In pursuit of this objective, the research synthesized a dataset amalgamating diverse audio samples recorded across multifaceted environments within railway stations. This curated dataset encompassed 8,000 distinct perilous urban sound manifestations distributed across eight categorizations. The dataset's intention lies in facilitating the training and validation of both machine learning and advanced deep learning architectures in discerning and classifying hazardous urban acoustics.

Predominantly, the dataset resonated with ambient auditory elements, notably picks, gunshots, and glass rupture cues. To encapsulate the nuances of diverse operational environments, ambient acoustics were assimilated from both indoor and outdoor milieus.

For analytical rigor, the acoustic cues were partitioned into segments lasting one second—reflecting the typical duration of the identified events of significance. Each of these segments was further dissected into 200 MS frames, exhibiting a 50% overlap. To elucidate, each one-second segment was articulated into nine distinct frames.






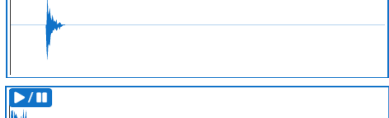







Table II offers a meticulous breakdown of the dataset, elucidating the composition of signals, frames, and segmented intervals. This tabulated exposition illuminates the heterogeneity of perilous urban acoustics, with an emphasis on their respective spectrograms. The table furnishes insights into the spectrographic analysis of varied impulsive acoustics, encompassing phenomena like vehicular glass rupture, canine alerts, emergency vehicular signals, infantile distress, security alarms, and various fire warning systems. This tabulation underscores the salience of the curated dataset and the pioneering CNN-LSTM deep learning paradigm.

B. Model Overview

Subsequent to initial preparations, the focus shifted towards logic programming. Central to this phase was the objective of formulating methodologies for comprehensive audio detection.

The intricacies of discerning potentially alarming acoustic events can be bifurcated into two specific sub-endeavors:

TABLE II. 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	

- Firstly, within the continuous audio data stream, there is a need to detect and isolate discrete pulse signals, ensuring their distinction from ambient auditory noise.
- Secondly, once extracted, the signal must then be classified, ascertaining its alignment with one of the multiple predefined acoustic events.

C. Detection of Impulsive Sound Events

The quantification of power for a series of consecutive, non-overlapping audio signal blocks serves as a cornerstone for various methodologies [9]. The computational approach to ascertain the power of the k th signal block, comprising N samples, is articulated by Eq. (1):

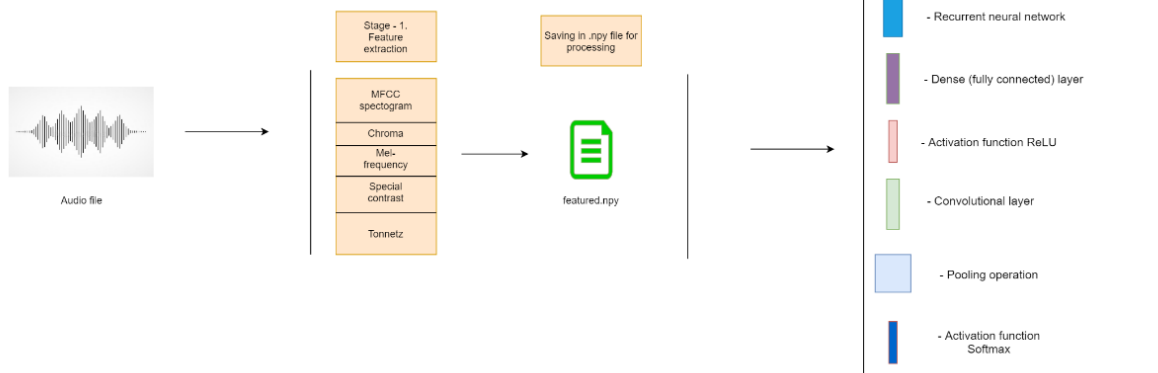
$$e(k) = \frac{1}{N} \sum_{n=0}^{N-1} x^2(n + kN) \quad k = 0, 1, \dots \quad (1)$$

Consider, for instance, an auditory manifestation of a gunshot, registered at approximately 4.6 seconds. For blocks consisting of $N = 4,000$ entries, corresponding to each block's duration, the power value span equates to roughly 90 milliseconds. The identification of blocks autonomously, especially in the context of transient pulse noises, can be executed via several distinct strategies, contingent upon the chosen approach:

- Grounded in the standard deviation of data that's been calibrated in terms of power metrics;
- Through the application of the median value from a median filter operating on the power units;
- By setting adaptive thresholds pertinent to the power units.

A deeper analysis reveals that this methodology predominantly leans on the standard deviation of power units' normalized values. Further examinations have deduced that normalized power block values situated within the interval $[0, 1]$ stand as a pivotal element in this analytical schema.

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



Following the initial processes, the focus transitioned to evaluating the standard deviation, commonly referred to as variance, for a specified set of data points:

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

In scenarios characterized by the presence of ambient noise, block powers generally exhibit a uniform distribution within the interval $[0, 1]$ (as illustrated on the left). Upon recalibrating the power value for an audio segment to fit within this defined range, any significant deviation above the established power levels of background units triggers the automatic detection of a pulse signal. Gradual alterations in signals can be discerned by observing the average value of normalized power metrics. Notably, this methodology displays resilience in the face of fluctuations in ambient noise intensity.

D. Proposed Model

In the present research, a synergistic architecture has been postulated, integrating Convolutional Neural Networks (CNN) with Recurrent Neural Networks (RNN). In this structure, the RNN does not function as a recursive layer within the CNN. Instead, it operates independently, employing a Rectified Linear Unit (ReLU) activation for information processing. The RNN dimension is set at 128. A detailed representation of this integrated architecture can be viewed in Fig. 2.

E. Feature Extraction

In this investigation, the process of feature extraction from auditory signals spanned approximately 90 minutes, given that the dataset under scrutiny amounted to 6.6 GB. This specific size was selected intentionally, with the research aiming to assess methodologies on a comparatively modest dataset. In subsequent phases, the established techniques will be applied to data in a singular pass. Following the comprehensive analysis of the auditory files, a resultant set of 8,674 sounds, equating to a cumulative duration of 5,439 seconds or 90.65 minutes, was obtained. A closer examination of the feature extraction component can be understood by referring to the coding segment, and the entire process is graphically represented in Fig. 3.

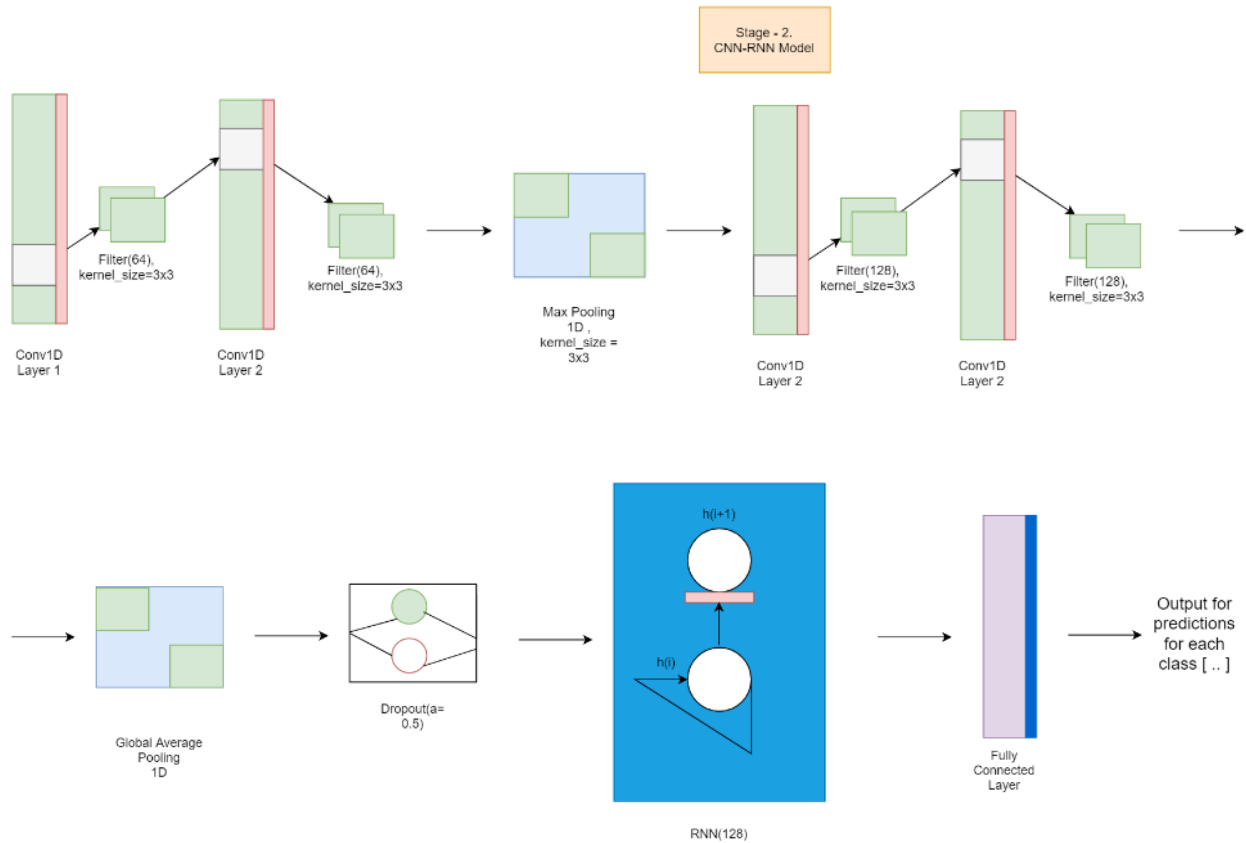


Fig. 2. Architecture of the proposed model.

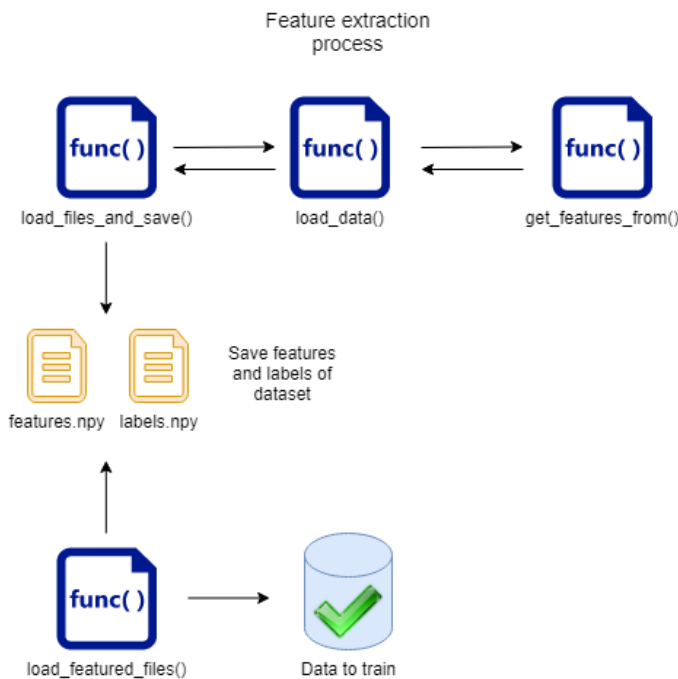


Fig. 3. The proposed framework.

Within the scope of this research, four distinct functions were delineated. Among them, three were explicitly designated for the extraction of features and subsequent data preservation. To elaborate:

- 1) The `load_files_and_save()` function invoked `load_data()`. This latter function systematically iterated over the dataset to acquire individual sound samples.
- 2) Following this, `get_features_from()` was summoned for each sound sample to extract its pertinent features.
- 3) Post-extraction, these attributes, along with their corresponding labels, were committed to persistent storage in two separate `.npy` formatted files.

Subsequent to the storage phase, data retrieval was facilitated by the `load_featured_files()` function. This prepared the dataset for the training phase, utilizing the integrated RNN-CNN model previously delineated in Fig. 2. This model's architecture encompassed two convolutional layers: one derived from a global maximum pooling mechanism and the other from a global average pooling paradigm.

IV. EXPERIMENTAL RESULTS

This segment elucidates the empirical outcomes derived from employing the synergized CNN-LSTM model for the identification of hazardous urban auditory events. Initially, the

metrics tailored for appraising the efficacy of the aforementioned deep learning model are delineated. This is succeeded by a presentation of the results from both training and testing phases, encompassing model accuracy, associated losses, the distinct confusion matrices, and AUC-ROC curve impulsive sound classification.

A. Evaluation Metrics

The efficacy of any machine learning or deep learning model, especially in contexts like hazardous urban sound detection, necessitates a rigorous and comprehensive evaluation strategy. This section elucidates the primary metrics employed to assess the proposed CNN-LSTM model's performance:

Accuracy: This is the most fundamental metric, representing the ratio of correctly predicted instances to the total instances in the dataset [30]. It provides a broad understanding of the model's effectiveness, encapsulating its capacity to correctly identify both dangerous and non-dangerous sounds.

$$accuracy = \frac{TP + TN}{P + N} \quad (4)$$

Precision: Precision ascertains the proportion of true positive predictions in the pool of all positive predictions [31]. In the realm of urban sound detection, high precision denotes that the sounds flagged as 'dangerous' by the model are indeed perilous with minimal false alarms.

$$precision = \frac{TP}{TP + FP} \quad (5)$$

Recall (or Sensitivity): This metric quantifies the model's ability to correctly identify all potential hazards [32]. In essence, recall measures the fraction of actual dangerous sounds that were rightly detected by the model.

$$recall = \frac{TP}{TP + FN} \quad (6)$$

F-score: Harmonizing precision and recall, the F-score is the harmonic mean of these two metrics [33]. It assists in providing a balanced measure, especially when the class distribution is skewed. An optimal model would strive for a high F-score, indicating both robust precision and recall.

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \quad (6)$$

Utilizing these metrics provides a holistic understanding of the model's capabilities, ensuring it is adept at identifying genuine threats while minimizing false alarms.

B. Results

This section offers a comprehensive assessment of the experimental results emanating from the dangerous urban sound detection exercises. The results of the CNN-LSTM model's endeavor at discerning impulsive sounds are visually illustrated in Fig. 4 and Fig. 5.

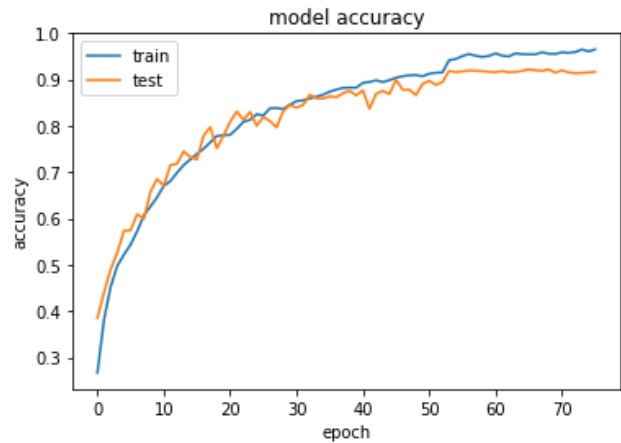


Fig. 4. Model training accuracy.

Fig. 4 specifically delineates the performance metrics during both training and testing phases for the inaugural dataset furnished by the research team. Notably, the CNN-LSTM model exhibited an impressive proficiency, registering an accuracy rate of 95% during its training phase. This was achieved over an approximate span of 80 epochs. Diving deeper into the model's architecture, it was observed to have approximately 87,822 parameters.

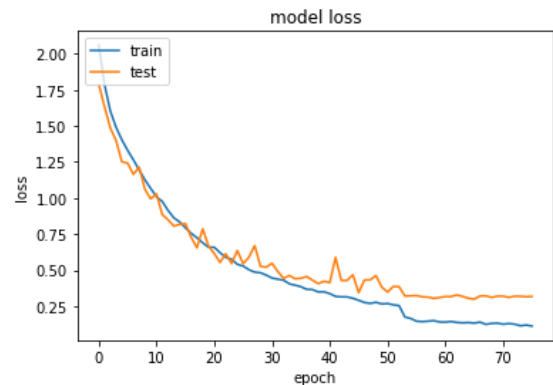


Fig. 5. Model training loss.

Furthermore, the duration of the training phase offers insight into the model's computational efficiency. The entire training process was completed in roughly 267 seconds, translating to slightly more than four minutes. This timeline underscores not only the model's accuracy but also its expedient processing capabilities.

In tandem, precision, recall, and F-score – though not explicitly detailed in the current dataset visualizations – remain paramount for comprehensive model assessment. These metrics, when considered in conjunction, ensure a holistic appreciation of the model's true capability, especially in real-world urban soundscapes.

In this section, we turn our focus to the findings from the second dataset, sourced from an open repository, and their implications as demonstrated in Fig. 6.

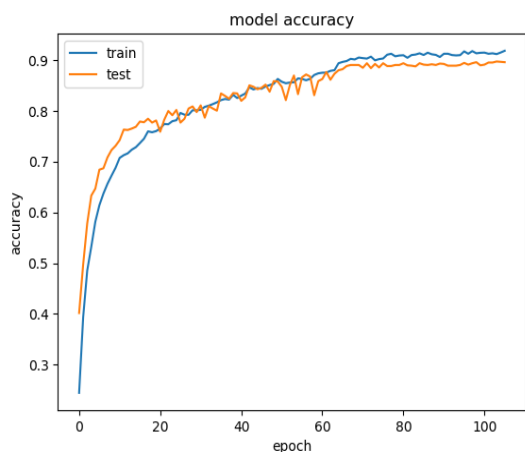


Fig. 6. Model test accuracy.

The CNN-LSTM model's performance on the second dataset is showcased in Fig. 7, providing invaluable insights into its adaptability and precision. Over a span of approximately 110 epochs, the model achieved an accuracy of 92%. This underscores its consistent performance, even when confronted with potentially disparate sound data from varied sources.

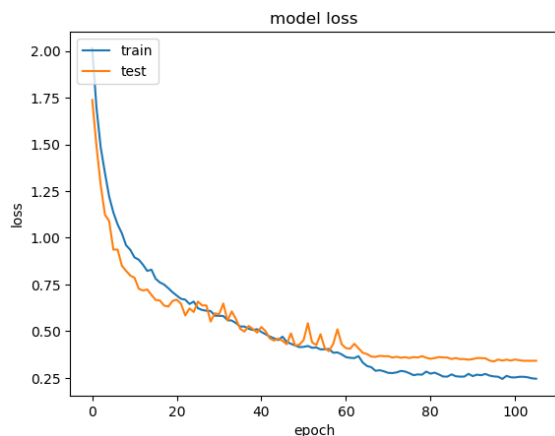


Fig. 7. Model test loss.

In the post-training phase, the model's performance was quantified using a confusion matrix, providing clarity on its precision across various classifications—specifically delineating true positives, true negatives, false positives, and false negatives, in the context of diverse urban acoustics. Fig. 8 presents a graphical representation of this matrix, elucidating the categorization efficacy for impulsive sounds. The implemented CNN model facilitated the differentiation of eight distinct hazardous urban auditory signals.

Fig. 9 depicts the AUC-ROC curve pertinent to the detection of perilous auditory events. The curve provides insights into the model's sensitivity to variations within the training dataset. The outcomes suggest that the integrated CNN-LSTM framework adeptly discerns hazardous acoustic events with commendable precision. Observations from the graph indicate a consistent performance, signifying the model's robust training tailored specifically for the identification of hazardous acoustical scenarios.

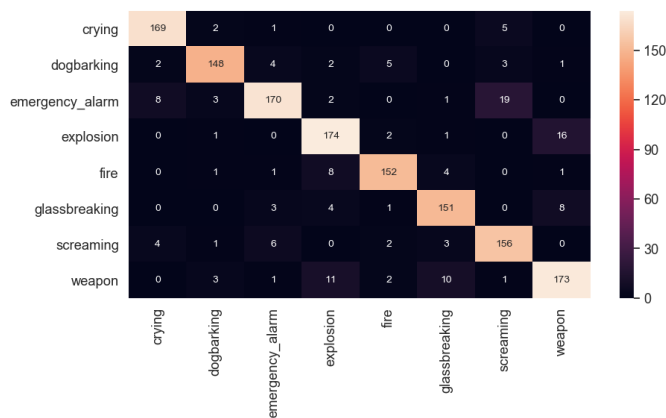


Fig. 8. Confusion matrix.

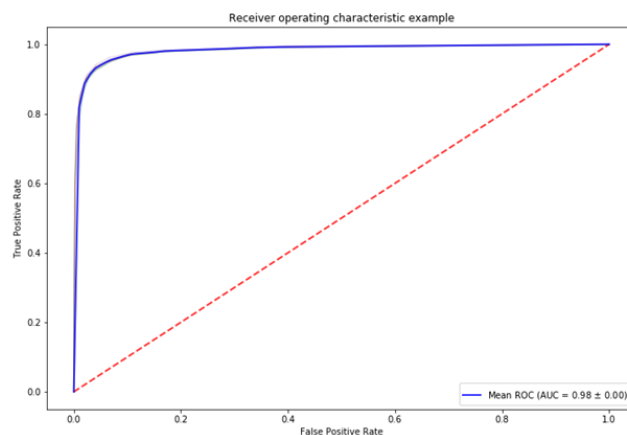


Fig. 9. ROC-AUC curve.

Consequently, the introduced deep neural network exhibits superior efficacy in consistently detecting hazardous urban sounds across all evaluation metrics. The success of the proposed methodology may be attributed to the utilization of the advanced RNN-CNN for weight and bias adjustments, coupled with an optimized training duration. The findings indicate that the presented deep neural network can be readily adapted to cater to both concise and extensive auditory inputs in contemporary applications.

V. DISCUSSION

The task of detecting and classifying dangerous urban sounds using deep learning architectures has garnered considerable attention given the importance of public safety and efficient urban management. This study presented a novel Convolutional LSTM (CNN-LSTM) network specifically tailored for real-time impulsive sound detection and classification in urban settings. The outcome of this research offers significant insights and implications, which are discussed in this section.

A. Comparative Performance of the Proposed Model

The CNN-LSTM architecture, as revealed by the results, showcases a notable advancement over previously proposed models for urban sound detection. In comparison to standard CNN architectures, the introduction of RNN allows the model to effectively process temporal sequences in the auditory data,

proving its prowess in capturing the temporal dynamics inherent in sound samples [34]. Not only does this substantiate the architectural choices made in this research but it also suggests potential avenues for further refining and expanding the hybrid deep learning models in auditory signal processing.

B. Efficacy in Hazardous Sound Detection

A pivotal achievement of this study is the high classification accuracy for sounds that have immediate safety implications, such as gunshots, alarms, and screams [35]. Precision in detecting these sounds is crucial for real-time monitoring systems that aim to promptly respond to emergencies. The proposed model's ability to discern these sounds from a cacophony of urban noises, with significant accuracy, positions it as a strong candidate for deployment in urban surveillance systems.

C. Model Generalizability and Robustness

Another salient point worth discussing is the model's performance across diverse datasets [36]. Its consistent results, even with different datasets including open-source ones, indicate robustness and generalizability. The implications here are two-fold: First, the model appears to be resilient to overfitting [37], a frequent pitfall in deep learning paradigms. Second, its generalizability suggests that with minor modifications [38], the proposed architecture could potentially be employed in varied urban environments, extending beyond the specific settings of this study.

D. Computational Efficiency and Real-time Implementation

The study indicates that the model's training lasted a mere few minutes, emphasizing the computational efficiency of the CNN-LSTM architecture. This is crucial for scaling up the approach and integrating it into real-time surveillance systems, where rapid model training and updating are of essence. Given the emergent nature of urban sounds and the ever-evolving urban landscape, the ability to quickly train and retrain models can be a game-changer.

E. Challenges and Limitations

While the findings are promising, it is essential to acknowledge certain challenges. Ambient noises, characteristic of dynamic urban settings [39], can sometimes interfere with the accurate detection of impulsive sounds. Furthermore, while the model has been tested on selected datasets, its performance in other global urban contexts – each with its unique soundscapes – remains to be evaluated.

F. Future Research Directions

Several prospective avenues emerge from this study:

- **Data Augmentation:** Experimenting with more extensive and diverse datasets, inclusive of global urban soundscapes, could further test and improve the model's robustness.
- **Model Refinements:** While the CNN-LSTM architecture demonstrates efficacy, the integration of attention mechanisms might enhance its ability to focus on critical sound segments, thereby potentially improving accuracy.

- **Transfer Learning:** Given the computational efficiency of the proposed model, it would be intriguing to investigate the benefits of transfer learning, applying knowledge from pre-trained models to expedite the training process even further.
- **Integration with Visual Surveillance:** A holistic urban surveillance system could combine auditory cues from the CNN-LSTM model with visual data from CCTV cameras, enhancing the accuracy and response time of emergency systems.

In conclusion, the CNN-LSTM model's performance in detecting dangerous urban sounds signals a promising step forward in urban surveillance and safety systems. Its computational efficiency, robustness, and high accuracy across datasets underpin its potential for real-world applications. Nevertheless, like all research, it sets the stage for further inquiries, refinements, and innovations in this domain.

VI. CONCLUSION

The paramount importance of ensuring urban safety cannot be overstated, and the deployment of advanced technological measures is crucial in these endeavors. This research aimed to bridge the extant gaps in urban sound detection by proposing a novel Convolutional LSTM (CNN-LSTM) architecture tailored for real-time impulsive sound detection and classification. The results, as delineated in the study, highlight the efficacy of this hybrid model in discerning and classifying hazardous urban sounds amidst the complex soundscape of urban environments.

The model's comparative performance, evidenced by its high classification accuracy, demonstrates its potential utility for urban surveillance systems. Especially noteworthy is its ability to accurately detect sounds of immediate safety concern, such as alarms, screams, and gunshots. Moreover, the robustness and generalizability of the model, as indicated by its consistent performance across diverse datasets, fortify its position as a leading contender for wide-scale implementation in urban settings globally.

However, while the findings undoubtedly underscore the potential of the CNN-LSTM architecture, they also pave the way for future research. There remains a vast expanse of uncharted territory in this domain, especially concerning model refinements, transfer learning, and the integration of auditory and visual surveillance cues. Such advancements could further refine detection accuracy and foster comprehensive urban safety measures.

In sum, this study serves as a testament to the untapped potential of deep learning paradigms in enhancing urban security. The proposed CNN-LSTM model, with its impressive results, sets a foundational precedent for further innovations and refinements. As urban centers continue to grow and evolve, it is our sincere hope that the fruits of this research contribute to safer, more secure, and harmonious urban living experiences for all.

ACKNOWLEDGMENT

This work was supported by the research project — Development of a system for detecting and alerting dangerous

events based on the audio analysis and machine learning funded by the Ministry of Science and Higher Education of the Republic of Kazakhstan. Grant No. IRN AP19175674.

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. Banikhalf, "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, 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, 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, V. Attar, "Deep learning approaches for video-based anomalous activity detection," *World Wide Web*, vol. 22, no.2, pp.571-601, 2019.
- [8] S. U. Amin, M. S. Hossain, G. Muhammad, M. Alhussein, M. A. Rahman, "Cognitive smart healthcare for pathology detection and monitoring," *IEEE Access*, vol. 7, no.1, pp. 10745-10753, 2019.
- [9] Omarov, B., & Altayeva, A. (2018, January). Towards intelligent IoT smart city platform based on OneM2M guideline: smart grid case study. In *2018 IEEE International Conference on Big Data and Smart Computing (BigComp)* (pp. 701-704). IEEE.
- [10] C. Heipke, 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] Omarov, B., Altayeva, A., & Cho, Y. I. (2017). Smart building climate control considering indoor and outdoor parameters. In *Computer Information Systems and Industrial Management: 16th IFIP TC8 International Conference, CISIM 2017, Bialystok, Poland, June 16-18, 2017, Proceedings 16* (pp. 412-422). Springer International Publishing.
- [12] A. Rajbanshi, D. Das, V. Udutalapally, 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] Omarov, B., Omarov, B., Shekerbekova, S., Gusmanova, F., Oshanova, N., Sarbasova, A., ... & Sultan, D. (2019). Applying face recognition in video surveillance security systems. In *Software Technology: Methods and Tools: 51st International Conference, TOOLS 2019, Innopolis, Russia, October 15–17, 2019, Proceedings 51* (pp. 271-280). Springer International Publishing.
- [17] V. Osipov, N. Zhukova, A. Subbotin, P. Glebovskiy, 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, 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] Sultan, D., Omarov, B., Kozhamkulova, Z., Kazbekova, G., Alimzhanova, L., Dautbayeva, A., ... & Abdrakhmanov, R. (2023). A Review of Machine Learning Techniques in Cyberbullying Detection. *Computers, Materials & Continua*, 74(3).
- [20] L. M. Bine, A. Boukerche, L. B. Ruiz, 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-LSTM-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] Tursynova, A., & Omarov, B. (2021, November). 3D U-Net for brain stroke lesion segmentation on ISLES 2018 dataset. In *2021 16th International Conference on Electronics Computer and Computation (ICECCO)* (pp. 1-4). IEEE.
- [26] Y. Xue, P. Shi, F. Jia, 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, 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, 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, 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, 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. Gughani, 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, 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, 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, , 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.