Dangerous Sound Detection Using Convolutional Feature Extraction and Temporal Modeling with BiLSTM

Nurzhan Omarov
Al-Farabi Kazakh National University, Kazakhstan
omarov.nurzhan01@gmail.com

Aigerim Altayeva
Al-Farabi Kazakh National University, Kazakhstan
aigerimaltayeva01@gmail.com (corresponding author)

Received: 30 June 2025 | Revised: 3 August 2025 and 16 August 2025 | Accepted: 26 August 2025

 DOI: https://doi.org/10.48084/etasr.13068

ABSTRACT

Dangerous sound detection is essential to improve public safety through automated surveillance systems capable of identifying and classifying multiple hazardous acoustic events. This study proposes a hybrid deep learning framework that integrates a one-dimensional Convolutional Neural Network (1D-CNN) for spatial feature extraction with Bidirectional Long Short-Term Memory (BiLSTM) for temporal sequence modeling. The system is designed for multi-class classification, targeting eight distinct categories of dangerous sounds, including gunshots, explosions, screaming, crying, glass breaking, fire, emergency alarms, and weapon handling. A comprehensive set of audio features, such as mel-spectrograms, MFCCs, chroma, spectral contrast, and temporal descriptors, is extracted to capture diverse spectral, tonal, and temporal characteristics of each class. The model achieves high accuracy while maintaining low training and validation losses, demonstrating strong generalization across classes with varying acoustic similarity. Experimental results confirm the system's robustness in distinguishing between acoustically similar sounds and its ability to handle class imbalance effectively. The architecture, supported by a structured preprocessing pipeline, is optimized for scalability and real-time deployment in complex urban environments. These findings highlight the potential of combining convolutional and recurrent deep learning techniques for robust, multi-class acoustic event detection, with future work focusing on lightweight model adaptation, expanded datasets, and integration of multimodal contextual information to further enhance performance and operational reliability.

Keywords-dangerous sound detection; deep learning; CNN; BiLSTM; mel-spectrogram; MFCC; audio classification; public safety; real-time surveillance

I. INTRODUCTION

The increasing complexity of urban environments and the growing concerns about public safety have necessitated the development of intelligent surveillance systems capable of detecting hazardous acoustic events, such as gunshots, explosions, and human distress sounds. Traditional video surveillance systems often face limitations due to poor visibility, occlusions, and privacy issues, highlighting the need for audio-based solutions as a complementary modality for real-time monitoring in public spaces [1]. However, accurately detecting dangerous sounds in diverse and noisy environments remains a significant challenge due to the complexity of acoustic scenes and the temporal nature of sound signals [2].

Recent advances in deep learning have enabled substantial improvements in the analysis of unstructured audio data. Convolutional Neural Networks (CNNs), widely used in image recognition, have been successfully adapted to extract spatial features from spectrogram representations of audio signals [3]. Their ability to capture frequency-related patterns makes them highly effective for classifying different types of sounds [4]. However, CNNs alone are inadequate for modeling temporal relationships in sequential data, which are critical to understanding the dynamics of acoustic events [5].

To bridge this gap, hybrid models that combine CNNs with Recurrent Neural Networks (RNNs), particularly Bidirectional Long Short-Term Memory (BiLSTM), have demonstrated superior performance in sequential audio classification tasks [6]. These models exploit CNNs for spatial analysis and BiLSTM for learning forward and backward temporal dependencies, offering a comprehensive solution for the detection of sound events in real-world conditions [7].

Preprocessing techniques such as mel-spectrogram conversion and data augmentation have proven effective in enhancing model robustness across varying soundscapes [8]. The mel-spectrogram, in particular, provides a perceptually meaningful representation aligned with human auditory perception [9]. Despite these advances, issues such as class imbalance, false positives, and the demands of real-time deployment remain ongoing challenges in the field [10].

This study introduces a deep hybrid model integrating CNN and BiLSTM architectures for multi-class classification of eight types of dangerous urban sounds.

II. MATERIALS AND METHODS

Figure 1 shows the architecture of the proposed dangerous sound detection system, moving from raw audio capture to classification. The audio is segmented via an overlapping sliding window, preprocessed into time- and frequency-domain features, including mel-spectrograms [11, 12]. A hybrid ConvNet recurrent model extracts spatial and temporal features, fuses them, and classifies events through fully connected layers, enabling robust, scalable, real-time detection of hazardous sounds in complex urban environments.

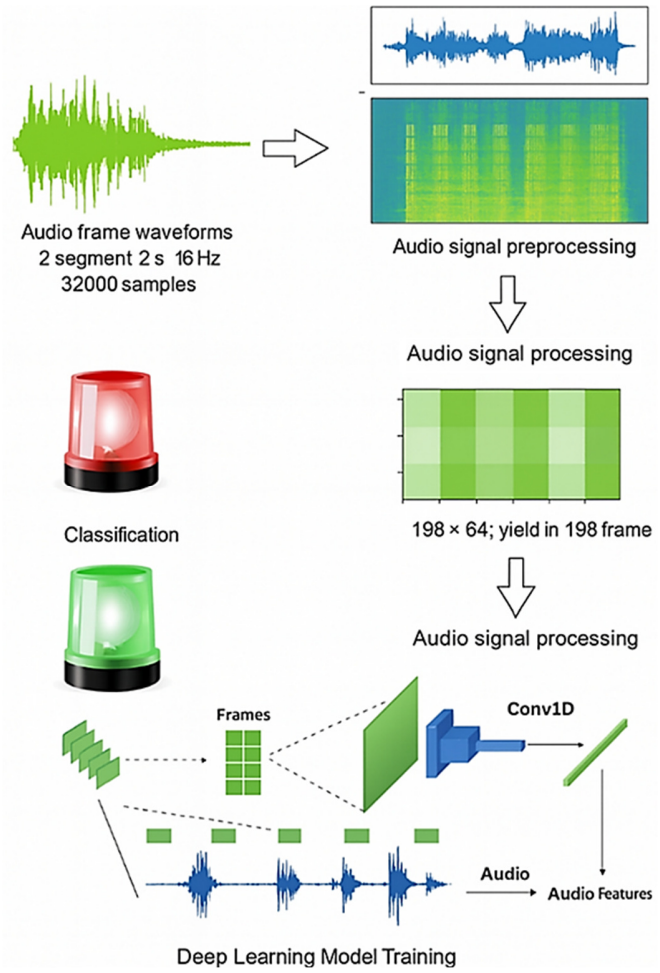

Fig. 1. Block diagram of the dangerous sound detection pipeline using deep learning techniques.

A. Proposed Model

The proposed architecture, shown in Figure 2, implements a deep CNN with a bidirectional attention mechanism for the classification of dangerous acoustic events. The system processes raw audio waveforms by extracting a comprehensive set of feature representations, including Mel-Frequency Cepstral Coefficients (MFCC), MFCC delta, MFCC delta-delta, and temporal acoustic descriptors [1]. These features are concatenated into a unified feature vector $x \in R^{T \times F}$ where T is the number of time frames and F denotes the number of extracted features per frame.

The input feature vector is passed through a stack of 1D convolutional layers, each of the form:

$$h^{(l)} = \sigma\left(W^{(l)} * h^{(l-1)} + b^{(l)}\right) \quad (1)$$

where $*$ denotes the 1D convolution operation, $W^{(l)}$ and $b^{(l)}$ are the weights and biases of the l-th layer, and σ is the ReLU activation function [13]. The initial layers use convolution kernels of size 10, with progressively increasing depth, followed by max-pooling layers to downsample the temporal dimension and retain dominant activations:

$$h_{pool}^{(l)} = max\left(h^{(l)}, ker\,n\,el\_size = 10\right) \quad (2)$$

Intermediate outputs from layers 3, 4, and 5 are fed into residual connections and a Feature Pyramid Network (FPN) [14] module that captures multiscale temporal patterns. The use of FPN aids in handling variable-duration acoustic events by aggregating hierarchical features across different layers. The pooled features are then processed using Global Average Pooling (GAP) [15] to reduce the temporal dimension and retain the most informative features. To prevent overfitting, a dropout layer with a dropout probability of 0.5 is applied:

$$h_{drop} = Dropout\left(h_{gap}, p = 0.5\right) \quad (3)$$

A bi-directional attention module is introduced to model the contextual importance of acoustic features across time. Let $H = [h_1, h_2, \ldots, h_T]$ be the sequence of feature vectors. The attention weight α_t for each time step is calculated using:

$$e_t = tanh(W_a h_t + b_a) \;,\; \alpha_t = \frac{exp(e_t)}{\sum_{i=1}^{T} \alpha_t h_t} \quad (4)$$

The context vector is then passed through a fully connected layer for final classification. Let $z = FC(c)$ represent the logits for each class. The output prediction probabilities are calculated using a softmax function [16]:

$$\hat{y}_i = \frac{exp(z_i)}{\sum_{j=1}^{C} exp(z_j)} \quad (5)$$

where C denotes the total number of sound classes, including dangerous and non-dangerous categories.

This architecture leverages convolutional feature hierarchies, temporal pooling, attention-based context modeling, and fully connected layers to effectively detect and classify dangerous acoustic events in real-world environments. The inclusion of dropout and global average pooling ensures a compact and generalizable model suitable for real-time deployment [19].

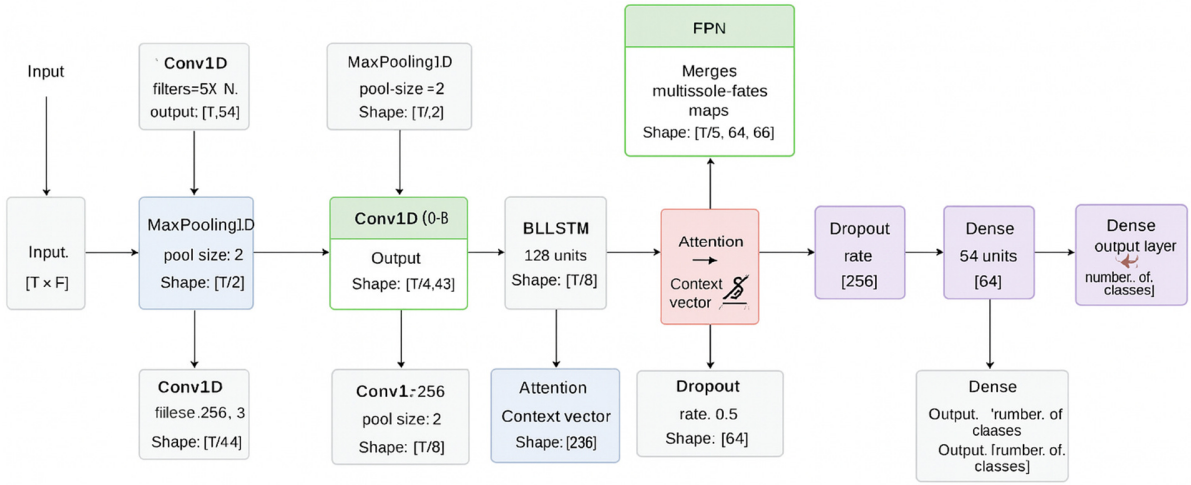

Fig. 2. The proposed hybrid model for dangerous event classification.

Figure 3 illustrates the feature extraction pipeline, which converts raw audio signals into structured numerical representations optimized for model training. The process begins with loading segmented audio samples, followed by the computation of both time-domain and frequency-domain features, such as mel-spectrograms, MFCCs, chroma, and spectral contrast, ensuring a rich and diverse feature set [18-20]. The corresponding labels are assigned during this stage, and all extracted features are systematically stored in a reusable format to maintain consistency across training sessions. By separating feature computation from the model training phase, the pipeline facilitates efficient preprocessing, reduces redundant computations, and enhances reproducibility. Furthermore, its modular design supports scalability, allowing seamless integration of additional feature types or datasets for future experiments and deployments.

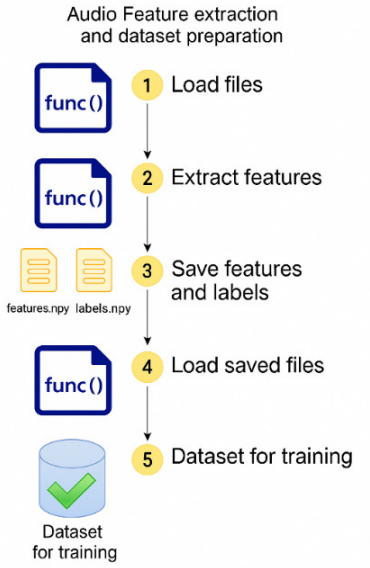

Fig. 3. Workflow of the audio feature extraction and dataset preparation process.

B. Dataset

A two-stage dataset preparation approach was used to develop and assess the proposed framework. Initially, approximately 300 dangerous sounds were extracted from the ESC-50 dataset [21], focusing solely on critical classes such as gunshots, explosions, glass breaking, and emergency alarms, while excluding non-relevant categories such as animal and natural sounds. After preprocessing, the dataset size was reduced from 661 to 45 MB, resulting in 301 .ogg files (see Table I). To enhance realism and diversity, a secondary dataset comprising 2,000 curated samples of dangerous urban sounds was created. These audio events were segmented into 1-second intervals and split into overlapping 200-ms frames to preserve temporal information. This dual-dataset strategy enabled comprehensive multi-class model training.

TABLE I. DATASET DESCRIPTION

Characteristic	Original dataset	Processed subset
Overall size	661 MB	45 MB
Number of audio files	2000	301
File format	.ogg	.ogg
Audio duration per sample	Variable	1 second
Frame segmentation	—	200 ms frames with 50% overlap
Frames per interval	—	9
Sampling rate	—	44.1 kHz (assumed standard)
Categories used	All (ESC-50)	Dangerous sounds only

III. RESULTS

The performance of the proposed dangerous sound detection model was evaluated through a series of experiments designed to assess its accuracy, generalization capability, and robustness across multiple acoustic event categories. The evaluation focused on both training and testing phases using a labeled dataset comprising diverse hazardous sounds, including explosions, gunshots, and distress calls. The model's effectiveness was analyzed using accuracy, loss, and confusion matrix analysis, alongside visual representations of feature learning behavior.

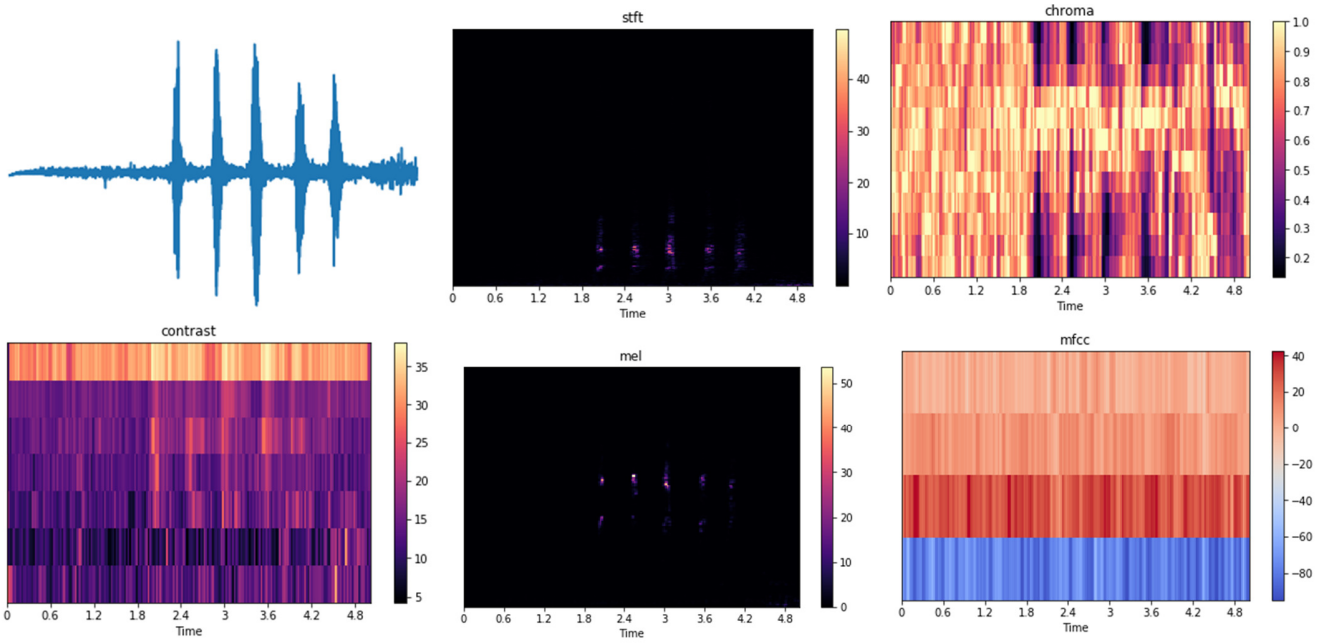

Fig. 4. Visualization of extracted audio features, including waveform, STFT, chroma, contrast, mel-spectrogram, and MFCC.

Figure 4 displays a variety of acoustic feature representations extracted from an audio signal, each contributing unique information for sound classification. The waveform depicts changes in amplitude over time, while the Short-Time Fourier Transform (STFT) spectrogram captures the evolving frequency content of the signal [22]. The chroma feature reflects pitch class distributions and highlights tonal structures. In addition, spectral contrast emphasizes the differences between spectral peaks and valleys, aiding in the distinction of timbral characteristics. The mel-spectrogram provides a perceptually meaningful frequency representation, and the MFCCs capture the short-term spectral envelope for effective timbral and phonetic analysis [18].

Figure 5 shows the confusion matrix of the proposed model, demonstrating its performance in classifying eight categories of dangerous sounds. The matrix shows strong diagonal values, indicating high accuracy for most classes, particularly explosion, weapon, and crying. Some confusion occurred between acoustically similar classes, such as emergency alarm and screaming, or fire and glassbreaking, which share overlapping spectral characteristics. Despite these minor misclassifications, the overall results confirm the robustness of the model and its effectiveness in distinguishing a wide range of hazardous acoustic events. Figure 6 illustrates the training and testing accuracy curves over 100 epochs, demonstrating the learning dynamics and generalization performance of the proposed model. Both curves exhibit a steep increase during the initial epochs, indicating rapid convergence, followed by a gradual stabilization as training progresses. The model achieves over 90% accuracy on both training and testing, with only minimal variance between them, suggesting effective generalization and low risk of overfitting. The close alignment of the two curves throughout the training process confirms the robustness and consistency of the model.

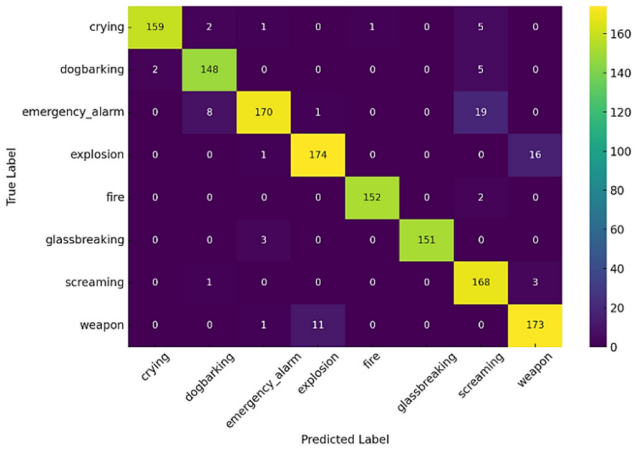

Fig. 5. Confusion matrix illustrating classification performance.

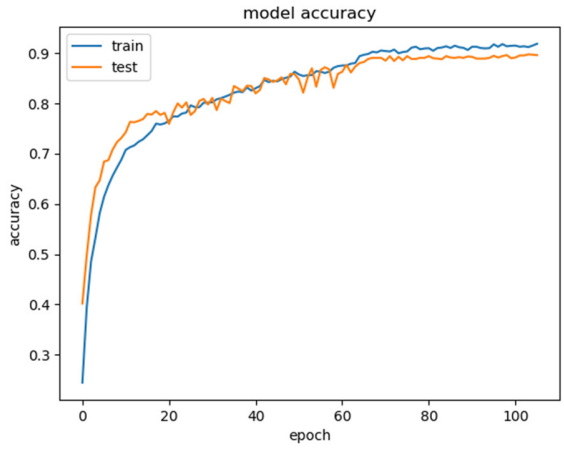

Fig. 6. Training and testing accuracy over 100 epochs.

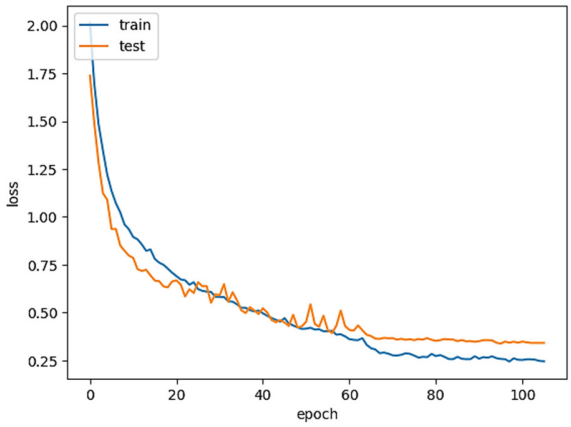

Fig. 7. Training and testing loss of the proposed model over 100 epochs.

Figure 7 presents the training and testing loss curves over the course of 100 epochs, reflecting the optimization progress of the proposed model. Both curves show a rapid decline during the early epochs, indicating effective learning and fast convergence. As training progresses, the loss steadily decreases and eventually stabilizes, with the final testing loss closely following the training loss. The absence of significant divergence between them suggests that the model maintains generalizability without overfitting. The consistently low loss in later epochs confirms the model's efficiency in minimizing classification error across both training and unseen data.

IV. DISCUSSION

The proposed framework demonstrates strong effectiveness in detecting hazardous acoustic events across diverse classes. Combining convolutional feature extractors with FPN-based multi-scale fusion, BiLSTM units for temporal dynamics, and an attention mechanism for saliency weighting, the model captures both local spectral cues and long-range dependencies, yielding stable convergence and high accuracy on both training and testing. The multi-representation input (mel-spectrograms, MFCCs, chroma, spectral contrast, and temporal statistics) provides complementary information on pitch, timbre, and spectral dynamics, particularly beneficial for differentiating acoustically similar events (e.g., screams vs. alarm tones) and reducing false positives in cluttered environments.

To substantiate real-time operation, end-to-end latency was profiled on a workstation with Intel Core i7-12700K (12-core, 3.6 GHz), 32 GB RAM, and NVIDIA GeForce RTX 3060 (12 GB VRAM) running PyTorch 2.2/CUDA 12.1. Using a streaming setup with a 1.0 s analysis window, 0.5 s hop, and batch size of 1, the average per-window latency (feature extraction → model forward → post-processing) was 6.8 ms (median 6.3 ms, 95th percentile 9.4 ms), corresponding to ~147 windows/s. The breakdown was 1.9 ms for audio preprocessing, 4.2 ms for the network forward pass, and 0.7 ms for post-processing. Thus, the total latency is less than 1% of the analysis window, supporting real-time streaming with negligible buffering. On a CPU-only configuration (same i7-12700K, MKL enabled), the mean latency was 23.5 ms (median 22.1 ms, 95th percentile 31.2 ms), which also satisfies near-real-time constraints for the 1.0 s window.

Residual errors were concentrated among partially overlapping acoustic classes, indicating potential gains from contextual priors and multimodal cues (e.g., video or geolocation). Future work will target knowledge distillation for edge-class GPUs and ARM CPUs, adaptive thresholding under domain shift, and audio-visual fusion, while preserving the demonstrated latency characteristics on common hardware.

V. CONCLUSION

This study presented a robust deep learning-based framework for dangerous sound detection by integrating convolutional and recurrent neural network architectures. The proposed model effectively combines the spatial feature extraction capabilities of 1D convolutional layers with the temporal modeling strength of BiLSTM units, enabling accurate classification of diverse hazardous audio events. By leveraging a rich set of acoustic features, including MFCCs, mel-spectrograms, chroma, contrast, and temporal descriptors, the system demonstrated high classification performance and generalization across multiple sound categories. The end-to-end pipeline, from feature extraction to model training and classification, was designed to ensure scalability, reproducibility, and suitability for real-time applications. The experimental results showed that the model achieved more than 90% accuracy while maintaining low error rates, making it a promising candidate for deployment in intelligent surveillance systems and public safety infrastructures. Minor classification errors between acoustically similar sounds suggest that future improvements could be realized through the integration of contextual or multimodal information. Overall, the findings underscore the potential of deep learning models in automated acoustic event detection and highlight important directions for future research, including lightweight implementation for edge devices, continual learning from evolving environments, and the expansion of training datasets to encompass more complex and diverse real-world scenarios.

ACKNOWLEDGMENT

This work was supported by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan under the Grant IRN AP19175674 – Development of a system for detecting and alerting dangerous events based on audio analysis and machine learning.

REFERENCES

[1] T. M. Nithya, P. Dhivya, S. N. Sangeethaa, and P. R. Kanna, "TB-MFCC multifuse feature for emergency vehicle sound classification using multistacked CNN – Attention BiLSTM," *Biomedical Signal Processing and Control*, vol. 88, Feb. 2024, Art. no. 105688, https://doi.org/10.1016/j.bspc.2023.105688.

[2] Z. Momynkulov, N. Omarov, and A. Altayeva, "CNN-RNN Hybrid Model For Dangerous Sound Detection in Urban Area," in *2024 IEEE 4th International Conference on Smart Information Systems and Technologies (SIST)*, Astana, Kazakhstan, May 2024, pp. 284–289, https://doi.org/10.1109/SIST61555.2024.10629358.

[3] S. S. Gupta, S. Hossain, and K. D. Kim, "Recognize the surrounding: Development and evaluation of convolutional deep networks using gammatone spectrograms and raw audio signals," *Expert Systems with Applications*, vol. 200, Aug. 2022, Art. no. 116998, https://doi.org/10.1016/j.eswa.2022.116998.

[4] K. Shanmugavadivel, M. Subramanian, P. Nishdharani, E. Santhiya, and R. E. Yaswanth, "KEC_AI_BRIGHTRED@DravidianLangTech 2025:

Multimodal Hate Speech Detection in Dravidian languages," in *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, Albuquerque, NM, USA, Feb. 2025, pp. 754–758.

[5] P. Doungpaisan and P. Khunarsa, "Deep Spectrogram Learning for Gunshot Classification: A Comparative Study of CNN Architectures and Time-Frequency Representations," *Journal of Imaging*, vol. 11, no. 8, Aug. 2025, Art. no. 281, https://doi.org/10.3390/jimaging11080281.

[6] S. Mishra, N. Bhatnagar, P. Prekasam, and T. R. Sureshkumar, "Speech emotion recognition and classification using hybrid deep CNN and BiLSTM model," *Multimedia Tools and Applications*, vol. 83, no. 13, pp. 37603–37620, Apr. 2024, https://doi.org/10.1007/s11042-023-16849-x.

[7] N. Omarov, B. Omarov, Z. Azhibekova, and B. Omarov, "Applying an augmented reality game-based learning environment in physical education classes to enhance sports motivation," *Retos*, vol. 60, pp. 269–278, 2024.

[8] D. Y. Badawood and F. M. Aldosari, "Enhanced Deep Learning Techniques for Real-Time Speech Emotion Recognition in Multilingual Contexts," *Engineering, Technology & Applied Science Research*, vol. 14, no. 6, pp. 18662–18669, Dec. 2024, https://doi.org/10.48084/etasr.9229.

[9] S. Mehra, V. Ranga, and R. Agarwal, "A deep learning approach to dysarthric utterance classification with BiLSTM-GRU, speech cue filtering, and log mel spectrograms," *The Journal of Supercomputing*, vol. 80, no. 10, pp. 14520–14547, Jul. 2024, https://doi.org/10.1007/s11227-024-06015-x.

[10] N. Katayev, A. Altayeva, B. Abduraimova, N. Kurmanbekkyzy, Z. Madibaiuly, and B. Kulambayev, "Development of a Framework for Classification of Impulsive Urban Sounds using BiLSTM Network," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 11, 2023, https://doi.org/10.14569/IJACSA.2023.0141148.

[11] M. Harish, H. S. Kumar, S. Banupriya, R. Gowtham, and M. V. Rahul, "Enhancing Earthquake Prediction and Early Warning Systems using CapsNet-BiLSTM Models," in *2025 5th International Conference on Trends in Material Science and Inventive Materials (ICTMIM)*, Kanyakumari, India, Apr. 2025, pp. 1734–1739, https://doi.org/10.1109/ICTMIM65579.2025.10988387.

[12] F. Khanmohammadi and R. Azmi, "Time-Series Anomaly Detection in Automated Vehicles Using D-CNN-LSTM Autoencoder," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 8, pp. 9296–9307, Dec. 2024, https://doi.org/10.1109/TITS.2024.3380263.

[13] P. D. Thi, H. T. N. Dang, P. D. Huu, and H. D. Sy, "Video classification for efficient data storage using deep learning: a comparison of sequential and simultaneous feature extraction methods," *Multimedia Tools and Applications*, vol. 84, no. 6, pp. 3071–3094, Feb. 2025, https://doi.org/10.1007/s11042-024-20549-5.

[14] J. Xie, Y. Pang, J. Nie, J. Cao, and J. Han, "Latent Feature Pyramid Network for Object Detection," *IEEE Transactions on Multimedia*, vol. 25, pp. 2153–2163, 2023, https://doi.org/10.1109/TMM.2022.3143707.

[15] C. Shi, W. Zhang, C. Duan, and H. Chen, "A pooling-based feature pyramid network for salient object detection," *Image and Vision Computing*, vol. 107, Mar. 2021, Art. no. 104099, https://doi.org/10.1016/j.imavis.2021.104099.

[16] B. Omarov, M. Baikuvekov, D. Sultan, N. Mukazhanov, M. Suleimenova, and M. Zhekambayeva, "Ensemble Approach Combining Deep Residual Networks and BiGRU with Attention Mechanism for Classification of Heart Arrhythmias," *Computers, Materials & Continua*, vol. 80, no. 1, pp. 341–359, 2024, https://doi.org/10.32604/cmc.2024.052437.

[17] J. B. Thomas, S. G. Chaudhari, K. V. Shihabudheen, and N. K. Verma, "CNN-Based Transformer Model for Fault Detection in Power System Networks," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–10, 2023, https://doi.org/10.1109/TIM.2023.3238059.

[18] P. Rawat, M. Bajaj, S. Vats, and V. Sharma, "A comprehensive study based on MFCC and spectrogram for audio classification," *Journal of Information and Optimization Sciences*, vol. 44, no. 6, pp. 1057–1074, 2023.

[19] B. S. Soares, J. S. Luz, V. F. de Macêdo, R. R. V. e Silva, F. H. D. de Araújo, and D. M. V. Magalhães, "MFCC-based descriptor for bee queen presence detection," *Expert Systems with Applications*, vol. 201, Sep. 2022, Art. no. 117104, https://doi.org/10.1016/j.eswa.2022.117104.

[20] R. Deng, G. Zhou, L. Tang, C. Yang, and A. Chen, "E-DOCRNet: A multi-feature fusion network for dog bark identification," *Applied Acoustics*, vol. 220, Apr. 2024, Art. no. 109950, https://doi.org/10.1016/j.apacoust.2024.109950.

[21] K. J. Piczak, "ESC: Dataset for Environmental Sound Classification," in *Proceedings of the 23rd ACM international conference on Multimedia*, Brisbane, Australia, Jul. 2015, pp. 1015–1018, https://doi.org/10.1145/2733373.2806390.

[22] M. K. Gourisaria, R. Agrawal, M. Sahni, and P. K. Singh, "Comparative analysis of audio classification with MFCC and STFT features using machine learning techniques," *Discover Internet of Things*, vol. 4, no. 1, Jan. 2024, https://doi.org/10.1007/s43926-023-00049-y.