IJETIR

ISSN (Online) 2583-0554

International Journal of Emerging Technologies and Innovative Research (IJETIR)

Volume 5, Issue 6, June 2025

Impact Factor: 5.731

Robust Threat Identification through Real-Time Audio Event Detection with 1D CNNs

P. Anusha[1], K. Nikhil Saradhi[2], K. Sravan Kumar[3], M. Tejesh[4]

Assistant Professor, Department of Computer Science & Engineering[1]

Students, Department of Computer Science & Engineering[2,3,4]

Guru Nanak Institute of Technology, Hyderabad

Abstract: *In today's world, ensuring personal safety especially in remote or isolated areas where people often work alone is more important than ever. In such environments, threats like robbery or assault often come with certain sounds that can act as early warning signs. Unfortunately, traditional security systems aren't always equipped to recognize or react to these sounds accurately or in real time. This paper tackles that challenge by developing an intelligent system that can listen to and classify sounds in the surrounding environment. Whether it's footsteps approaching, a car engine revving, or background chatter, the system aims to detect these audio cues and identify what they might mean. At the heart of this system is a 1D Convolutional Neural Network (1D-CNN), a deep learning model that's particularly good at understanding time-based signals like audio. By training the model on a labeled dataset of real-world sounds, it learns to pick up on patterns and features that distinguish one sound from another. Once trained, the model can process live audio, categorizing it into different sound types with high accuracy. This real-time sound classification offers a smarter way to understand what's happening in the environment whether in a busy urban street, a quiet workplace. Ultimately, this project isn't just about recognizing sounds it's about enhancing awareness, improving safety, and laying the groundwork for smarter.*

Keywords: Sound Classification, 1D Convolutional Neural Network (1D-CNN), Real-time Audio Processing, Personal Safety Monitoring, Environmental Sound Recognition

I. INTRODUCTION

In today's fast-paced and unpredictable world, ensuring personal safety especially for people working alone in remote or high-risk areas. Whether it's night-shift workers, solo travelers, or individuals in isolated locations, many face the risk of criminal activities like theft, assault, or even violence. Often, these threats are accompanied by specific sounds footsteps, shouting, breaking glass that could serve as early warning signs. But despite their importance, these audio cues are frequently overlooked or missed by traditional security systems, which tend to struggle with detecting and interpreting sounds in real-time and with enough accuracy. The real challenge is that everyday environments are filled with a mix of harmless background noises like traffic, street chatter, or machinery that make it hard to distinguish threatening sounds from normal ones. Existing systems like CCTV cameras and alarm setups often fail to pick up on these audio signals quickly or correctly, which can delay crucial responses. This project sets out to bridge that gap by developing a smart audio classification system that can detect and make sense of different environmental sounds in real time. By using a deep learning model specifically a 1D Convolutional Neural Network (1D-CNN) the system learns to recognize patterns in audio data, helping it tell the difference between a friendly voice and a potential danger. The goal is to create a tool that improves situational awareness by offering real-time insight into what's happening around an individual. Whether it's for public surveillance, workplace safety, or managing urban environments, this sound-based monitoring approach could play a key role in keeping people safer. In an age where timely information can make all the difference, being able to analyze and understand environmental sounds isn't just helpful it's essential.

IJETIR
ISSN (Online) 2583-0554
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Volume 5, Issue 6, June 2025
Impact Factor: 5.731

II. LITERATURE REVIEW

S. Dong et. al (2023) : The research paper titled *"Environmental Sound Classification Based on Improved Compact Bilinear Attention Network"* introduces a fresh and effective approach to classifying environmental sounds. The authors tackle the common limitations of traditional sound classification methods by proposing an enhanced deep learning model that uses a compact bilinear attention mechanism. This attention-based approach helps the model better focus on the most relevant parts of an audio signal, improving its ability to understand both spatial and temporal patterns within the sound data. By refining how the model processes audio features, the proposed method significantly boosts classification accuracy while maintaining computational efficiency. The paper also compares this improved network with other existing models and demonstrates its superior performance in handling complex environmental sound datasets.Overall, the study offers a promising direction for more accurate and efficient sound classification systems, which could be especially valuable for real-time audio analysis in safety, surveillance, and monitoring applications.

Mohaimenuzzaman et al (2023): The paper titled *"Environmental Sound Classification on the Edge: A Pipeline for Deep Acoustic Networks on Extremely Resource-Constrained Devices"*. In many real-world situations like remote monitoring or portable safety devices hardware is often constrained by minimal memory, low computational power, and no GPU support. This makes it difficult to run typical deep neural networks for tasks like sound classification. To address this challenge, the authors propose a universal pipeline that transforms large, complex convolutional networks into lightweight, efficient models through compression and quantization techniques. Their solution, a model called **ACDNet**, is specially designed for edge devices like microcontrollers. Despite the dramatic reduction in size shrinking the network by an impressive **97.22%**the model still delivers excellent performance. On the popular ESC-10 benchmark dataset, ACDNet achieves a **state-of-the-art accuracy of 97.28%**, proving that high performance can be maintained even on minimal hardware. This work stands out for making environmental sound classification more accessible and deployable in real-time applications, especially in scenarios where devices need to be compact, efficient, and capable of functioning independently in the field.

K.Presannakumar and A.Mohamed (2023): The paper titled *"Deep Learning Based Source Identification of Environmental Audio Signals Using Optimized Convolutional Neural Networks"*. Environmental sound classification is a complex task due to factors like limited data availability and background noise. This study tackles those challenges by fine-tuning CNN architectures, enhancing their performance in real-world scenarios. By applying advanced optimization techniques, the authors improved the CNNs' ability to generalize across different environments, making the model more robust and reliable.The proposed approach demonstrates strong capabilities in distinguishing between a variety of sound sources, such as animal calls or mechanical noises, even under noisy or variable conditions. These improvements are particularly valuable for real-time monitoring applications, like wildlife tracking or urban sound surveillance.Importantly, the optimized models also show promise for deployment in low-resource settings, including edge devices, which means they can be effectively used in the field without needing heavy computing power.Overall, the study showcases how tailored deep learning solutions can advance sound-based monitoring systems, enabling smarter and more efficient audio classification in diverse environments.

R.Viveros-Muñoz et al (2023): The paper titled "Dataset for polyphonic sound event detection tasks in urban soundscapes: The Synthetic Polyphonic Ambient Sound Source (SPASS) dataset" is designed to facilitate research in polyphonic sound event detection (PSED) in urban environments. It was created to help train deep neural networks for detecting multiple overlapping sound events within complex urban soundscapes. The dataset features synthetic recordings that simulate real-world environments like parks, streets, markets, squares, and waterfronts. These recordings were curated using the RAVEN software library, which allows for the creation of virtual soundscapes. The dataset is part of the FuSA System project, aimed at advancing detection and classification of environmental sound sources using deep learning models.The SPASS dataset is a valuable resource for researchers looking to improve sound event detection algorithms, particularly in environments where sounds overlap and are not easily isolated. It also provides metadata associated with each recording, offering valuable context for training AI model.

A. B. Shrestha et al (2024): The paper titled *"Navigating the Role of Smartwatches in Cardiac Fitness Monitoring: Insights from Physicians and the Evolving Landscape"* by A. B. Shrestha et al., published in *Current Problems in*

Cardiology(January 2024), takes a close look at how smartwatches are reshaping the way we monitor heart health. The study gathers perspectives from physicians to understand how wearable devices like smartwatchesare being used in the field of cardiovascular care. The authors highlight the benefits of smartwatches in tracking key health indicators such as heart rate, ECG, and other vital signs in real time. This kind of continuous monitoring can play a critical role in early detection of potential heart issues, making it a valuable tool in preventive cardiology. At the same time, the paper acknowledges some important challenges. These include concerns over the accuracy of the data collected, patient privacy, and how to effectively integrate wearable data into traditional healthcare systems. Despite these hurdles, the study emphasizes the growing potential of wearable technologies to transform cardiovascular care. As smartwatches become more advanced and widely used, they may pave the way for a more proactive, tech-driven approach to managing heart health.

Existing System

The current system is all about spotting threats in real-time by picking up on audio signals from a person's environment. It primarily tackles the growing danger faced by individuals, particularly those who find themselves working alone at night in remote or isolated spots think women who are at risk of things like robbery, assault, or even worse.This system takes advantage of the fact that these threats usually come with specific sounds or noises that can be detected and categorized to give early warnings. The aim is to build an automated system that can analyze surrounding audio signals, pinpoint potential threats, and send timely alerts to registered contacts through their smartphones, all without needing extra hardware.To make this happen, the system taps into audio data from the Kaggle dataset, which features a variety of environmental and human sounds. It employs Exploratory Data Analysis (EDA) techniques to visualize and grasp the audio features. These EDA techniques are essential for spotting patterns, anomalies, and key characteristics in the audio data, which are vital for crafting an accurate sound classification model. The results from this analysis are then fed into sophisticated deep learning models, mainly Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNN), for deeper analysis and classification of the audio events. The LSTM model is particularly effective at capturing long-term dependencies in sequential data, making it great for processing time-series audio signals where the timing of sounds matters for classification.

Existing System Disadvantages:

- Dependence on Dataset Quality.
- Enviranmental Noise Impact.
- Real-Time Processing Delays**.**
- Misclassification Risk.

Proposed System

The 1D Convolutional Neural Network (1D CNN) is a unique deep learning model designed specifically for sequence data, such as time-series or audio signals. Unlike the more commonly known 2D CNNs that are typically used for images, 1D CNNs concentrate on data that has just one spatial dimension imagine it as a series. When it comes to audio classification, 1D CNNs really excel at analyzing sequential features like Mel-Frequency Cepstral Coefficients (MFCCs) or other time-series representations of sound. In the model we're talking about, the 1D CNN is structured with a series of convolutional layers followed by pooling layers. Each convolutional layer applies a set of filters (or kernels) to the input sequence. These filters slide across the data (convolving), looking for specific patterns such as peaks, edges, or other distinctive features. The main aim of the convolution is to pinpoint local patterns in the input data that are essential for classification.

Proposed System Advantages:

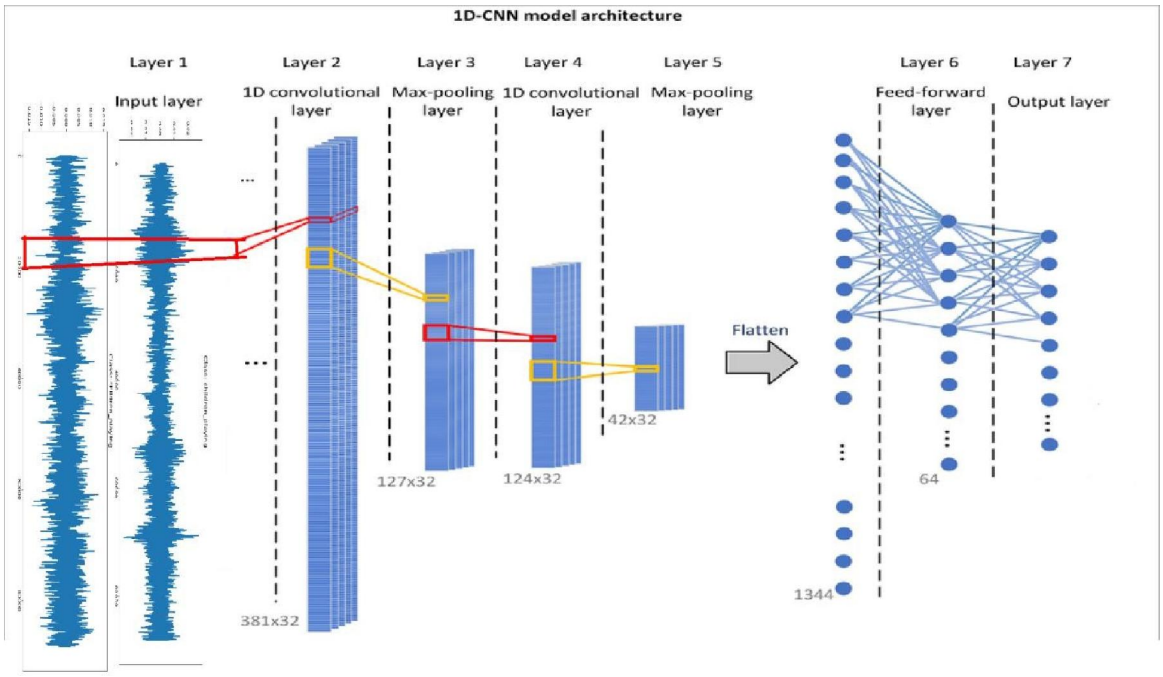
- Efficient for sequence data.
- Automatic feature extraction.

IJETIR
ISSN (Online) 2583-0554
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Volume 5, Issue 6, June 2025
Impact Factor: 5.731

- Local Feature detection.

III. SYSTEM ARCHITECTURE

The figure demonstrates the architecture of a 1D Convolutional Neural Network (1D-CNN) specialized in dealing with sequential data such as time series, audio, or sensor signals. It starts with an input layer that takes in data with a shape of 381×32, where 381 stands for time steps and 32 is the number of features or channels. The model architecture contains a sequential flow that includes time series data as the first feature. The input is sent to the first 1D convolutional layer where filters are applied along the time axis to extract important local patterns. This is also known as the convolution step in the model. Then, there is a max-pooling layer that downsamples (reduces the size of) the output to reduce dimensionality and help generalization within the model. A set of more complicated features are fetched from the pooled data through a second convolutional layer. Other dominating features are retained while non-masked sized portions are further reduced via a second max-pooling layer. The final output of 42 x 32 is changed to 1344 units which all together make a single vector that can be directed to the densely connected subsequent layers. Through a fully connected layer containing 64 neurons, the flattened vector is directed and high-level representation commences. After all these steps are executed, an output layer is predicted enabled to do final probabilistic prediction/classification. Each of these neural parts are implemented parallel through the edges coloured in the diagram which represent the flow of feature extraction at each layer.



IV. METHODOLOGY

MODULES:

Data Collection:

The first step in the project is to gather an appropriate dataset of audio files. Since this is a sound event classification task, the dataset should include a variety of labeled sound events relevant to the problem domain. Publicly available datasets, such as those from Kaggle, are commonly used, containing various sound categories. Each audio file in the dataset must be clearly labeled with its corresponding class(e.g.,dog barking, street music, etc.).The quality and variety of the audio samples are essential for training a robust model capable of generalizing to real-world situations.

IJETIR
ISSN (Online) 2583-0554
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Volume 5, Issue 6, June 2025
Impact Factor: 5.731

Preprocessing:
Preprocessing is a crucial step before feeding the audio data into the model. The raw audio filesneed to be converted into a format that can be processed by the neural network. In this step, the audio is loaded and converted into a numerical format using the librosa library. The audio is then analyzed, and Mel-frequency cepstral coefficients (MFCCs) are extracted, as these features represent the spectral properties of the sound. The MFCCs are averaged across time frames to generate a fixed-size feature vector for each audio clip, ensuring uniform input to the model. Normalization may also be applied to scale the data for better model performance.

Feature Extraction:
Feature extraction involves transforming the raw audio into a more meaningful representation that can be processed by a machine learning model. MFCCs are widely used in speech and audio classification tasks as they capture important information about the timbre and texture of the sound. The features are then averaged over time to create a compact representation of each audio clip, which significantly reduces the complexity of the input data while retaining essential information. These features serve as input to the CNN model.

Model Design:
The next step is designing the deep learning model. For this project, a 1D Convolutional Neural Network(1D-CNN)is chosen, as it is particularly effective for sequential data like audio signals. The model consists of multiple layers: the input layer accepts the MFCC features, followed by Conv1D layers that act as filters to detect patterns in the audio data. MaxPooling1D layers are added to reduce the dimensionality of the data while retaining key features. The final fully connected (dense) layers are used for classification.

Model Compilation:
Once the model structure is ready, it is compiled by defining the optimizer, loss function, and evaluation metrics. The Adam optimizer is chosen for its efficiency and adaptability. Since the task involves classifying data into multiple categories, categorical cross-entropy is used as the loss function.

Model Training:
Training the model involves feeding the preprocessed and feature-extracted data into the model and allowing it to learn from the labeled examples. The model will adjust its internal weights using backpropagation to minimize the loss function. The dataset is typically divided into training and validation sets to assess the model's performance on unseen data. Monitoring the loss and accuracy during training helps ensure the model is learning effectively and not overfitting.

Model Evaluation:
After training, the model is evaluated on a separate test set to assess its generalization ability. The test data should consist of audio samples that the model has never seen before. This helps determine how well the model performs on unseen data. To measure how well the model is performing, we use evaluation metrics like accuracy, precision, recall, and the F1-score.

Model Saving:
Once the model has been trained and evaluated successfully, it is essential to save the trained model. Saving the model allows it to be reused in future predictions without the need to retrain it each time. In Tensor Flow/Keras, this can be done using the model.save() function, which saves the model architecture, weights, and optimizer state to a file. This saved model can then be easily loaded for inference on new data.

IJETIR
ISSN (Online) 2583-0554
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Volume 5, Issue 6, June 2025
Impact Factor: 5.731

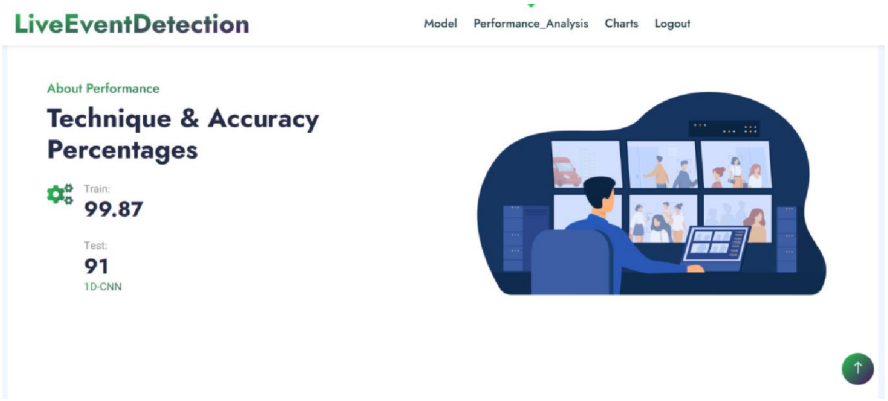
User Interface Development:

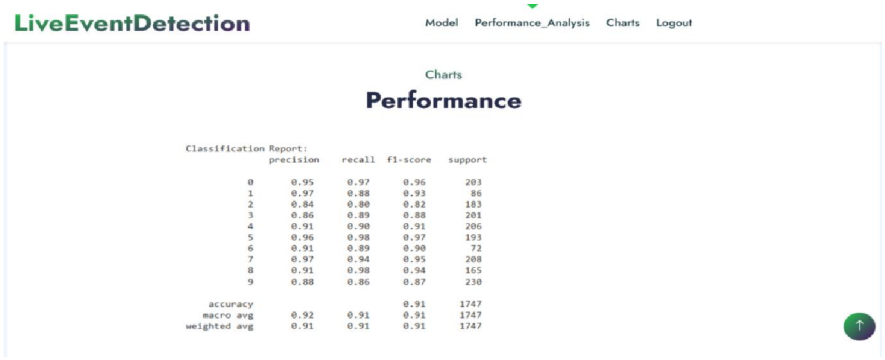
In this step, a user interface (UI) is developed to allow users to interact with the trained model. The UI should allow users to upload audio files, which are then processed and classified by the model. The UI could be web-based, created using frameworks like Flask or Django, where users can simply drag and drop audio files for classification. The model will return the predicted class of the audio event (e.g., dog barking, street music, etc.), which can be displayed to the user on the UI.

Connecting Model to Frontend for Prediction:

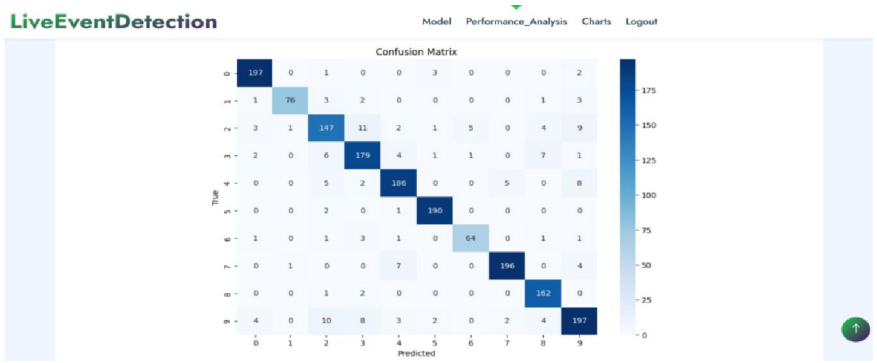
The final step involves integrating the trained model with the frontend for real-time prediction. The frontend collects the audio file uploaded by the user and sends it to the backend, where the model processes the file. The backend should include the necessary code to load the saved model and perform inference on the input audio. After the model makes a prediction, the result is sent back to the frontend, where it is displayed to the user. This allows the model to make predictions on new audio samples in real-time.

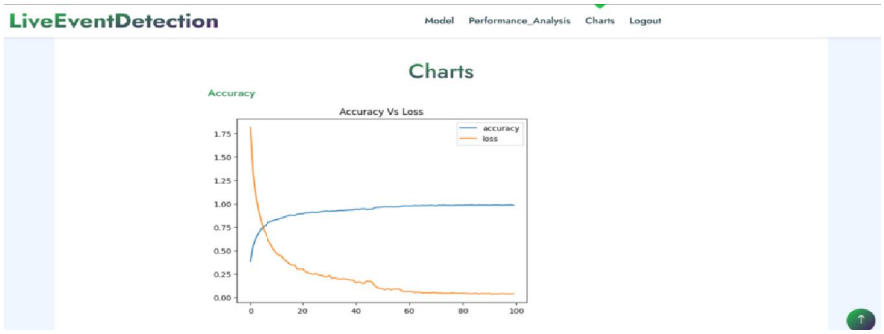
V. EXPERIMENTAL RESULTS

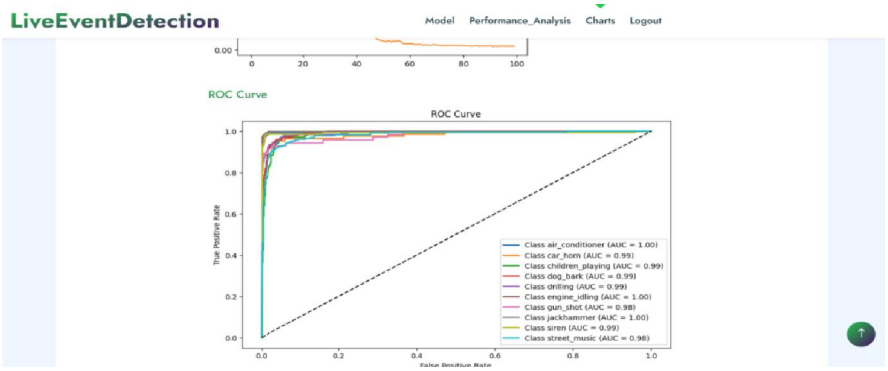


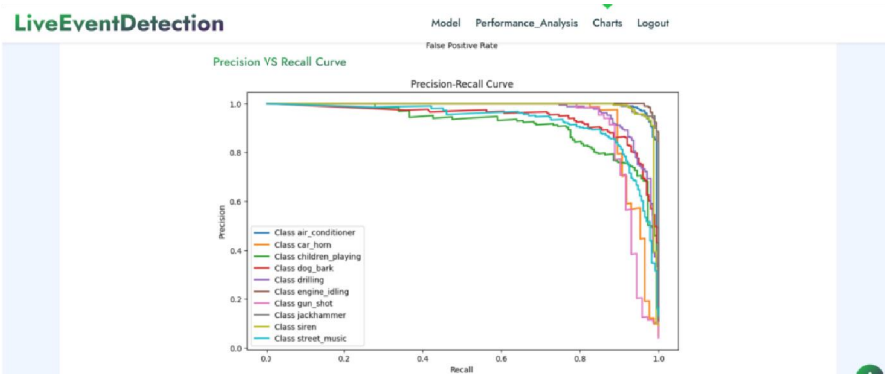


IJETIR
ISSN (Online) 2583-0554
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Volume 5, Issue 6, June 2025
Impact Factor: 5.731



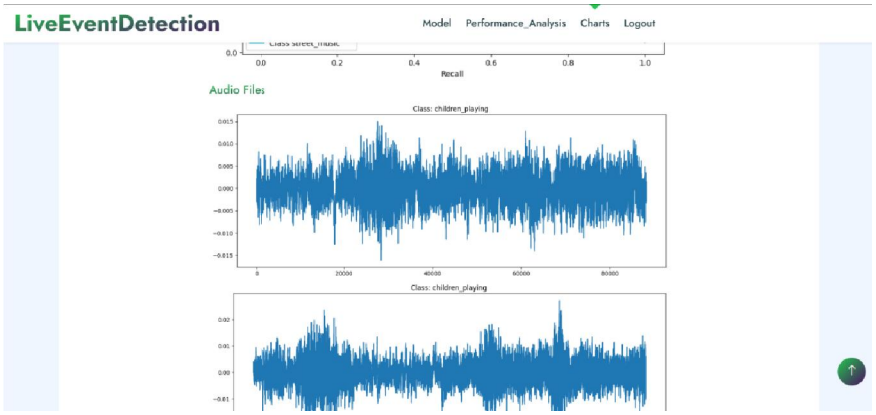






DOI: 10.48175/IJETIR-6002

IJETIR
ISSN (Online) 2583-0554
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Volume 5, Issue 6, June 2025
Impact Factor: 5.731



VI. CONCLUSION

A 1D Convolutional Neural Network (CNN) model is a powerful way to detect and classify different audio events by analyzing features like Mel-frequency cepstral coefficients (MFCCs). By effectively processing and learning from sound signals, this model can distinguish between various sounds. This makes it a valuable tool for applications such as security, surveillance, and monitoring the environment. Although the current model works well with its original architecture, potential future developments like data augmentation, transfer learning, multimodal integration of input, and real-time processing could significantly enhance its accuracy, robustness, and applicability. In general, this model is a promising solution to applying deep learning to audio classification problems, and further development could result in more advanced systems able to tackle realistic challenges in sound-based event detection.

REFERENCES

[1]. T. P. Suma and G. Rekha, ‘‘Study on IoT based women safety devices with screaming detectionandvideocapturing,’’Int.J.Eng.Appl.Sci.Technol.,vol.6,no.7,pp.257–262,2021.

[2]. P. Zinemanas, M. Rocamora, M. Miron, F. Font, and X. Serra, ‘‘An interpretable deep learning model for automatic sound classification,’’ Electronics, vol. 10, no. 7, p. 850, Apr. 2021.

[3]. D. G. Monisha, M. Monisha, G. Pavithra, and R. Subhashini, ‘‘Women safety device and application-FEMME,’’ Indian J. Sci. Technol., vol. 9, no. 10, pp. 1–6, Mar. 2016.

[4]. G. Ciaburro and G. Iannace, ‘‘Improving smart cities safety using sound events detection based on deep neural network algorithms,’’ Informatics, vol. 7, no. 3, p. 23, Jul. 2020.

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

[6]. A.Triantafyllopoulos, G. Keren, J.Wagner, I. Steiner, and B.W. Schuller, ‘‘Towards robust speech emotion recognition using deep residual networks for speech enhancement,’’ in Proc. INTERSPEECH, Graz, Austria, Sep. 2019.

[7]. M. T. García-Ordás, H. Alaiz-Moretón, J. A. Benítez-Andrades, I. García-Rodríguez, O. García-Olalla, and C. Benavides, ‘‘Sentiment analysis in non-fixed length audios using a fully convolutional neural network,’’ Biomed. Signal Process. Control, vol. 69, Aug. 2021, Art. no. 102946.

[8]. A Bhardwaj, P. Khanna, S. Kumar, and Pragya, ‘‘Generative model for NLP applications based on component extraction,’’ Proc. Comput. Sci., vol. 167, pp. 918–931, Jan. 2020.

[9]. S. Malova and D. V. Tikhomirova, ‘‘Recognition of emotions in verbal messages based on neural networks,’’ Proc. Comput. Sci., vol. 190, pp. 560–563, Jan. 2021.

[10]. Hodorog, I. Petri, and Y. Rezgui, ‘‘Machine learning and natural language processing of social media data for event detection in smart cities,’’ Sustain. Cities Soc., vol. 85, Oct. 2022, Art. no. 104026.

[11]. D. D. Nguyen, M. S. Dao, and T. V. T. Nguyen, ''Natural language processing for social event classification,'' in Knowledge and Systems Engineering. Cham, Switzerland: Springer, 2015, pp. 79–91. 6470 VOLUME 12, 2024 A. Sen et al.: Live Event Detection for People's Safety Using NLP and Deep Learning.

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

[13]. D. Tsiktsiris, A. Vafeiadis, A. Lalas, M. Dasygenis, K. Votis, and D. Tzovaras, ''A novel image and audio-based artificial intelligence service for security applications in autonomous vehicles,'' Transp. Res. Proc., vol. 62, pp. 294–301, Jan. 2022.

[14]. Q.Kong,Y.Cao,T.Iqbal,Y.Wang,W.Wang,andM.D.Plumbley,''PANNs:Large-scale pretrained audio neural networks for audio pattern recognition,'' IEEE/ACM Trans. Audio, Speech, Language Process., vol. 28, pp. 2880–2894, 2020.

[15]. L. Yang and H. Zhao, ''Sound classification based on multihead attention and support vector machine,'' Math. Problems Eng., vol. 2021, pp. 1–11, May 2021