Classifying Wildlife Acoustic Signals using a Deep Learning Approach

Sunila Sheikh, Umer Rashid
Department of Computer Science, Quaid-i-Azam University, Islamabad
*Correspondence: umerrashid@qau.edu.pk

Citation | Sheikh. S, Rashid. U, "Classifying Wildlife Acoustic Signals using a Deep Learning Approach", IJIST, Special Issue pp 324-338, May 2026
DOI | https://doi.org/10.33411/IJIST/1781
Received | March 11, 2026 **Revised |** April 24, 2026 **Accepted |** 01 May, 2026 **Published |** May 08, 2026.

Monitoring wildlife and environmental conditions in national parks is essential for ecological research, biodiversity conservation, and public safety. This study proposes a contextual sound-based monitoring framework that addresses the limitations of vision-based systems in low-light and occluded environments commonly found in wildlife areas. The proposed approach integrates a hybrid deep learning architecture combining Convolutional Neural Networks (CNNs) for spatial feature extraction and a Bidirectional Gated Recurrent Unit (BiGRU) for temporal sequence modeling, along with a fuzzy logic decision layer for high-level contextual interpretation. To ensure diversity and robustness, multiple open-source datasets, including ESC-50, UrbanSound8K, FSC22, and Scream/Non-Scream datasets, are preprocessed, harmonized, and merged into a unified dataset comprising 15,811 audio clips across 16 low-level sound classes. The dataset includes alarming sounds, representing complex acoustic environments relevant to wildlife and park monitoring. The model employs a hierarchical classification strategy. Firstly, the CNN-BiGRU network performs low-level sound event classification, and then a fuzzy inference system maps the outputs into four high-level contextual categories: Illegal Activity, Human Distress, Natural Hazard, and Safe Activity. Experimental results demonstrate strong performance, achieving an accuracy of 95.80%, precision of 95.95%, recall of 96.14%, weighted F1-score of 95.80%, and ROC-AUC of 99.67% on the UrbanSound8K dataset. With an accuracy of 91.30%, precision of 88.11%, recall of 86.23%, weighted F1-score of 87.91%, and ROC-AUC of 99.50%, the model maintains competitive performance on the harmonized dataset with a difference of less than 5% across evaluation metrics. These findings highlight the effectiveness of contextual sound analysis in enhancing situational awareness and supporting intelligent surveillance systems for wildlife and environmental monitoring.

Keywords: CNN BiGRU Architecture; Context-Aware Sound Classification; Fuzzy Logic Decision Layer; Hybrid Deep Learning Model; Environmental Acoustic Monitoring

Introduction:

A lot of work has been done recently to evaluate and compare various approaches and techniques for sound classification, especially in applications related to security, surveillance, and monitoring systems [1]. Approaches like Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Decision Trees (DTs) are some of the traditional machine learning approaches that have been used for classification problems [1][2][3][4]. Deep learning techniques, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Deep Neural Networks (DNNs), Multi-Layer Perceptrons (MLPs), and some hybrid approaches that combine the traits of traditional and modern techniques or multiple deep learning models, have been utilized in recent years [5][6].

The first and the most important stage in classification tasks is sound detection, which involves identifying unusual or anomalous audio signals from a continuous input signal and transforming them into information that has potential to be used further [7]. The low signal-to-noise ratio and the low occurrence of target sounds make it more difficult to identify important events from environmental sounds [8]. Furthermore, because environmental sounds contain unpredictable sound patterns by nature, they frequently span a larger frequency range than other audio forms like speech or music, this trait of environmental sounds makes them difficult to classify [1][9]. Since sound detection serves as the basis for the next steps including classification, recognition, and surveillance, accurate detection of an audio event is crucial to ensure reliable system performance and high accuracy [7]. The classification stage comes next, following the detection of a sound event. This stage processes the incoming audio signal, which is rich in spectral and temporal information, to extract useful features that are then utilized to classify the signal by classification model [9]. In order to have an enhanced understanding of the events occurring, environmental sound classification focuses on recognizing and classifying sounds within a specific area [10]. Depending on the model's context, these sounds can be divided into two groups of general noises (like a dog barking or glass breaking) and audio scenes (like a forest or a market) [11]. Nowadays, the Internet of Things (IoT), deep learning-based applications, and Geographical Information Systems (GIS) address real-world problems [7]. Information and communication technologies, comprising IoT, deep learning, and real-time surveillance, have shown tremendous advancements [12]. The applications further employ various event-related attributes of geospatial entities [13]. In this context, software applications are extending to diverse domains such as medicine, agriculture, and surveillance [8]. In the following, we will discuss related surveillance systems, sound classification techniques, datasets, evaluation measures, issues, and the motivation behind this research study.

Surveillance Systems:

The surveillance systems are in demand for any environment, whether inside buildings or across urban traffic settings [8]. These systems have historically contributed to environmental safety and monitoring, even from the early methods, using human personnel to ensure safety and maintain order. Now, with the shift of surveillance systems from the early methods to the latest methods using technology, the support has become stronger. In modern intelligent surveillance, precise geospatial localization and dependable event detection are necessary for interpreting complex environments [7]. However, creating strong and precise detection or classification models has become a key part of advanced surveillance systems, whether by using videos, audio, or both [14].

While meeting quality criteria, these systems focus on accuracy, durability, and scalability to ensure high reliability and efficiency [7]. To minimize false alarms, these use deep learning algorithms that are trained on large datasets. The IoT sensors are built to work in tough environmental conditions, like extreme heat, humidity, floods, and other conditions [8]. The execution of these systems requires the use of specialized hardware components, such as high-sensitivity IoT sensors designed for acoustic monitoring, motion detection, and the

collection of environmental data [13]. Advanced low-power processors can be incorporated to optimize data processing while reducing energy consumption [8]. Furthermore, GPS modules along with GIS-compatible devices can be employed to ensure precise geospatial tracking and localization of events [7]. For making the system computationally efficient, lightweight deep learning models i.e., MobileNetV3 or ShuffleNetV2 are employed and fine-tuned in research for making the systems deployable on edge devices and mobile units while considering the accuracy of classification module of the system [15][35]. Using transfer learning along with data augmentation helped, in some cases, to overcome the problem of small training datasets and the model can generalize better to unseen data [15]. Lastly, another module, a user-friendly dashboard integrated with GIS and GPS, can help authorities to easily understand and act on sound events by geo-tagging, severity of the event, and time based insights [14]. This real-time display helps authorities to stay aware of the situation and allows them to respond quickly and make better decisions smartly [14]. The authorities can monitor activity trends across time and space, identify high-risk zones, and deploy resources more effectively based on correctly classified spatial data by mapping events onto dynamic geographical interfaces [14]. As Sathruhan et al. developed a system for emergency vehicle sirens and traffic noise using a dataset of 926 samples for both classes to train a 1D CNN model with multiple convolutional, pooling, and dense layers, which achieved 93% accuracy, demonstrating effective performance [17].

Sound Classification:

The automated interpretation of audio events in several kinds of domains, including smart homes, industrial monitoring, and environmental surveillance systems, is made possible by the sound classification, which forms the basis of intelligent surveillance systems [18]. Recent research indicates that hybrid deep learning models, which incorporate convolutional, recurrent, and transformer-based components to improve accuracy, temporal sensitivity, and noise robustness, are showing effective results.

Because of the complicated structure, patterns, and polyphonic character of audio, these modern techniques for audio classification have clearly improved sound classification. The system operates better when audio aspects are caught more effectively [15].

Convolutional Neural Networks:

As CNNs are known to be effective at extracting hierarchical features from time-frequency representations like spectrograms or log-Mel spectrograms, many researchers have used them extensively for audio classification problems. As Herranz et al. investigated the use of transfer learning with the ReaLISED database, including a lightweight CNN-based model for sound event classification in smart settings [12]. The results of their method outperformed traditional machine learning models like SVM, KNN, and MLP in the industrial area by classifying 2,479 real-life household noises into 18 categories using Google's pretrained YAMNet model, which was trained on AudioSet, proved the effective use of CNNs. By using transformed audio recordings into Mel-spectrogram images, Mushtaq et al. introduced a model for environmental sound classification (ESC) that uses deep convolutional neural networks (DCNNs) and transfer learning, showing more precise and effective classification. Two custom CNN architectures, CNN-1 and CNN-2, were developed and evaluated alongside popular pretrained models such as ResNet, DenseNet, and VGG, which were fine-tuned using discriminative and cyclic learning strategies [18]. They demonstrated their results using ESC-10, ESC-50, and UrbanSound8K for training and computing their model's test results [18]. This proposed approach showed that using spectrograms and data augmentation helps in recognizing environmental sounds accurately [18].

Recurrent Neural Networks:

Recurrent Neural Networks (RNNs) have the ability to capture temporal relationships in sequential data, which makes them suitable for analyzing temporal signals, which is why RNNs are often utilized in audio classification applications. An RNN based classifier

distinguished between snoring episodes (SE) and non-snoring episodes (NSE), using features extracted from overnight sleep recordings was presented in the study by [19]. After they manually segmented and labeled these overnight recorded audio samples, Zero-Crossing Rate (ZCR), Short-Time Fourier Transform (STFT), and MFCC were used to extract features out of the raw audio signals. Then, an RNN based model was trained, validated, and tested using the retrieved features in order to classify the episodes as either SE or NSE [19]. The model used RNNs with LSTM, a type of RNN, to take advantage of their ability to learn from sequences, capturing both short-term and long-term patterns in the audio data [19]. Huy et al. proposed a deep GRU-based recurrent neural network architecture, also to take advantage of sequence to label classification, for audio scene classification, by utilizing label tree embedding (LTE) features to capture complex patterns present in the audio. The model used a stack of GRU layers to learn time-based patterns in the audio. Instead of using a softmax layer for classification, a linear SVM was applied to the output of the RNN for better accuracy. Three types of LTE features, Gammatone, MFCC, and log-frequency, were extracted and passed through both single and multi-stream RNNs [20]. The LITIS-Rouen dataset was used for evaluation and it demonstrated that the fusion of multiple feature streams and probabilistic voting strategies improved the classification of acoustic scenes [20].

Another study of Tyagi et al. proposed an audio classification approach using a Long Short-Term Memory (LSTM) model on the UrbanSound8K dataset. The proposed architecture consists of two LSTM layers with 128 and 64 units, respectively along with a dense softmax output layer for classification. The model was trained using the Adam optimizer and sparse categorical cross-entropy loss for 50 epochs. Results evaluated that the enhanced LSTM model outperformed by achieving the highest accuracy of 86.7, demonstrating its effectiveness in capturing temporal patterns in audio data for improved classification performance. [21].

Hybrid Deep Learning Techniques:

Hybrid deep learning techniques have gained significant attention from researchers in audio classification tasks because of their ability to combine the strengths of multiple neural architectures. Dissanayaka et al. proposed a real-time acoustic anomaly detection (RTAAD) system, which can easily detect and predict failure in machinery early [12]. The study used a combination of Convolutional Autoencoders (CAE), Variational Autoencoders (VAE), and Temporal Convolutional Networks (TCN) with Log-Mel spectrograms to detect anomalies based on reconstruction errors, and the model was trained exclusively on normal data and used the MIMII dataset with AUC and pAUC scores for evaluation of the system [12].

Anum et al. worked on classifying environmental sounds using a Convolutional Recurrent Neural Network (CRNN). The proposed model was trained on the UrbanSound8K dataset [18]. The model effectively classified complex, unstructured environmental sounds by the combination of CNNs for spatial feature extraction and Long Short-Term Memory (LSTM) layers for capturing temporal dependencies. The input audio signals were transformed into Mel Frequency Cepstral Coefficients (MFCCs) and passed through convolutional, pooling, LSTM, and fully connected layers, with softmax for final classification. Hyperparameters such as the number of LSTM layers, momentum, filter count, batch size, and LSTM units were tuned, resulting in better accuracy by hyperparameter testing [18].

Xiong et al. used a CRNN architecture, a deep learning approach that integrated CNNs and BiGRUs, which addressed significant weaknesses in sound-based construction activity monitoring. Using a dual-threshold method, the model showed accurate temporal localization of activities and handled polyphonic construction sound settings. Audios were collected from real construction sites to train the model, which included four different classes [7].

Context Aware Systems:

Much of the work has been done by simply classifying the classes of sounds by learning the pattern of each of the class. However, this classification data could be more meaningful by processing it with its environmental constraints which make a system context aware [22]. Following this idea, Simone et al. proposed a speech command recognition system for industrial environments using a Conformer-based architecture, combined with Voice Activity Detection and Keyword Spotting for a context aware system [23]. The system was trained on the MIVIA-ISC dataset, using real-world recordings and synthetic samples. Results showed that the model trained with synthetic data achieved better command classification performance with an F1-score of 0.947 [23].

However, for making the system contextually aware many researchers used fuzzy logic methods, as fuzzy logic addresses real world challenges by using reasoning through linguistic variables and rule-based inference [24]. As fuzzy logic helps to mimic human like decision making, it is used across diverse domains such as accident severity prediction, robotic control, medical diagnosis, industrial optimization, and IoT-based decision systems [25]. Wu et al. proposed a model that combined quantum fuzzy logic, deep neural networks, and feature fusion for image classification. The model used DNN to extract deep features while the quantum fuzzy module captured fuzzy relationships, both merged in a fusion layer. The model was tested on five different datasets and results showed 95.6% accuracy on CIFAR-10 dataset [26].

The following is the organization of this research article. Literature and related work to this study are discussed under the section of Introduction, including the issues found in the literature, as well as the motivation and novelty of this study. Overview of datasets is discussed in the section Materials and Methods. Also, the model structure and training are elaborated in the same section. The Results and Discussion section highlights the model training and evaluation. At last, the discussion of the study is concluded in the section named Conclusion.

Issues and Motivation:

Recent research on sound classification has pointed out many problems and limitations that still need solution. One of the major challenges is the handling of polyphonic sound environments, where different sounds overlap and most models cannot separate these sounds well that lead to lower accuracy in real world cases. Another problem is that many models cannot accurately determine starting and ending points when a sound occurs. This is important for time-sensitive applications. Another issue is that many models are trained on limited or unbalanced datasets that lack fully representing real world situations. This limitation reduces model performance in unseen environments or when hearing unfamiliar input sounds and also there is still no dataset for so many specific types of environments. Also, some deep learning methods do not use transfer learning or pretraining, so they need a lot of labeled data and take a long time to train and more resources. Additionally, it is noticed that several studies lack on evaluating their models in real-time audio streaming conditions, making them less practical for deployment in real-time surveillance systems or monitoring systems. Lastly and importantly, a notable limitation in most of the reviewed models is the lack of ability to understand context, which prevents them from understanding the order and relations between events.

Our study addresses few major limitations found in existing research. Firstly, many existing models are trained on small or unbalanced datasets, which makes it difficult for them to perform well in real world environments like parks or wildlife areas. Secondly, this study addresses the lack of datasets for park and wildlife environments. Thirdly, most models lack contextual awareness, making them unable to understand the order or the relationship between events like recognizing that a gunshot followed by a scream may be connected. In order to close this gap, our proposed model uses a fuzzy logic module that enables

contextual interpretation of important temporal patterns and dependencies present in the audio stream. Our research provides a more contextually aware and broadly applicable approach for sound classification in park and wildlife areas by focusing on these fundamental issues.

Objectives:

The objective of this study is to develop a context aware environmental sound classification framework for wildlife and park monitoring applications. The proposed approach focuses on the improvement of sound event detection and the interpretability of acoustic patterns through hierarchical classification and contextual reasoning.

Novelty:

The novelty of this research lies in the development of a hybrid framework that combines deep learning and fuzzy logic for contextual environmental sound classification, unlike conventional approaches that focus solely on low-level event detection or use only fuzzy logic reasoning. The key novel contributions of this work are summarized as follows:

Integration of CNN and BiGRU with a decision layer to convert low-level sound predictions into high-level contextual categories using fuzzy logic.

Development of a harmonized multi-source dataset by combining four different datasets, including ESC-50, UrbanSound8K, FSC22, and Scream/Non-Scream datasets.

Implementation of a context-aware classification mechanism to transform raw audio detection into meaningful situational understanding.

Materials and Methods:

Proposed Framework:

The goal of this research is to make monitoring of real wildlife environments possible by using different sound types. Therefore, it is crucial to propose a deep learning method that can categorize different kinds of wildlife related sounds. Wildlife distress signals, illegal human activities like poaching or deforestation, and environmental disturbances are all included in the classified sounds. The classification of wildlife related sounds is possible by using a standard dataset and a deep learning technique. In order to find anomalies and trends that might point to environmental dangers, the presence of wildlife, and illegal activities, we will propose a deep learning system that goes through an extensive training process. We particularly focus on a novel dataset, model construction, and training in this study.

Datasets:

A diverse and well-curated dataset is essential to effectively classify acoustic events in natural environments such as parks and wildlife zones. This research integrated four widely used environmental sound datasets: ESC-50 [27], UrbanSound8K [28], FSC22 [29] and the Scream/Non-Scream [30] dataset, mentioned in Table 1. The ESC-50 dataset has 2,000 five-second audio samples from 50 environmental classes, including natural sounds like rain, birds, and animal noises. More than 8,000 tagged clips from ten urban-related classes such as dog barking and footsteps, which are frequently heard in public parks, are publicly available on UrbanSound8K. With 2,025 clips which cover 27 classes, including mechanical, environmental, human, and wildlife sounds, FSC22 concentrates exclusively on forest surroundings. The Scream and Non-Scream dataset, including 3126 samples divided in two categories. For the purpose of identifying complicated park and animal acoustic settings, combining these datasets improves model generalization and increases class variety. The unified dataset simulates real-world acoustic variations and supports deep learning models for environmental monitoring and surveillance applications.

Table 1. Comparison of ESC-50, UrbanSound8K, FSC22 and Scream/Non-scream datasets

Property	ESC-50	UrbanSound8K	FSC22	Scream/Non-scream
Total Clips	2,000	8,732	2,025	3126
Classes	50	10	27	2
Duration	5 sec	4 sec	5 sec	1-4 sec
Format	WAV, 16-bit Mono	WAV, 16-bit Mono	WAV, 16-bit Mono	WAV, 16-bit Mono
Sample Rate	44.1 kHz	44.1 kHz	44.1 kHz	44.1 kHz
Focus	General environment	Urban and indoor sounds	Forest, wildlife sounds	Screams and Non-scream

Dataset Taxonomy:

For integrating different datasets into a single dataset, it is necessary to resolve a number of obstacles, such as distinct audio formats, varied sample rates, and overlapping or ambiguous class names. An organized preprocessing pipeline was developed in order to standardize this audio data. To ensure that every audio sample had the same duration, the longer recordings were clipped and the shorter ones were padded with zeros to produce a constant length of five seconds across all inputs. In order to facilitate effective batch processing and reliable model training, this step helped in maintaining uniform temporal resolution for every audio sample.

The problem of inconsistent names was solved by combining labels with similar meanings into unified categories. For instance, sounds categorized as "engine idling," "engine," and "engine" were merged into a single class known as "engine idling." In order to prevent redundant meanings and lessen class disparity, this label unification was necessary. The name, title, duration, and source dataset of each audio file were then recorded to build a master metadata file in CSV format. After audios were standardized, data from all the datasets were combined into a single repository. This process not only expanded the size of the dataset and variety but also improved the ability of the model to generalize to unfamiliar sound distributions. At the end, the labels and audios are synthesized for fuzzy logic module to mimic real world scenarios. The final taxonomy of the harmonized dataset is shown in Fig. 1.

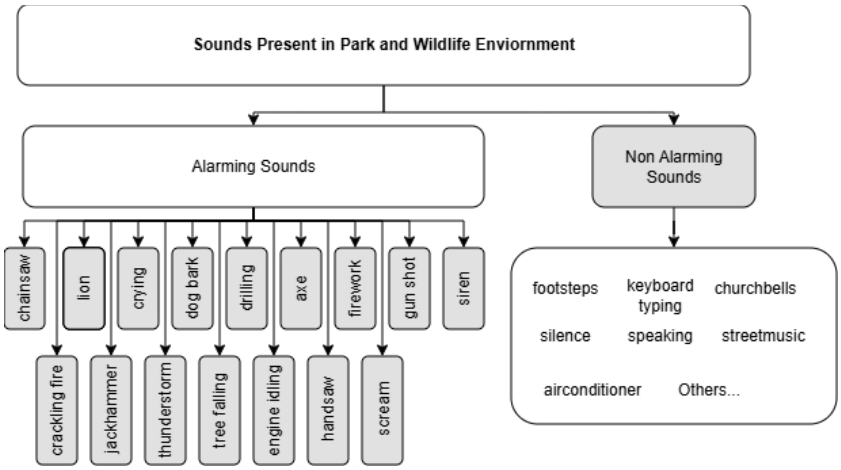


Figure 1. Separation of Alarming Sound Classes and Non-alarming Class.

Dataset Preprocessing:

Preprocessing played a critical role by improving performance of the model and ensuring robustness of the model against external noise and varying recording conditions. Initially, persistent background noises such as wind or electrical hum were minimized using spectral gating for denoising of the audio signal using Sensitivity to Noise threshold as 1.25, FFT size as 1024. To ensure temporal consistency across samples, audio normalization was performed by converting all audio clips to a fixed duration of 5 seconds and were converted

to a bit depth of 16-bit, mono WAV format with a sampling rate of 22,050 Hz. Data augmentation was applied for balancing the number of samples per class, by upsampling classes with fewer samples and down sampling classes with a higher number of samples, resulting in 7984 samples (499 in each class) out of 15,811 samples. Accurate sound classification depends on how useful and meaningful the extracted features are. In this setup, Mel-Frequency Cepstral Coefficients (MFCCs) were used as the main features. These features captured the shape of the frequency content of audio in a compact way. MFCCs were calculated from the log-Mel spectrogram using the Discrete Cosine Transform (DCT), which are also well known for reflecting how humans actually hear sound, specifically its tone and texture.

In this work, 40 MFCCs were extracted per frame from a 5-second monochrome audio clip. The first and second-order derivatives, often referred to as delta and delta-delta features, were also calculated from MFCCs in order to capture temporal dynamics with more information. The three components, MFCC, delta, and delta-delta features were stacked together to form a three-channel input tensor, similar to the RGB format used in image processing. MFCCs were used to take advantage of their smaller size and better noise resilience, even though most of the models relied on raw log Mel spectrograms as input features for the CNN. Convolutional layers were next incorporated into the tensor, ensuring efficient extraction of spatial features.

Model Structure and Training:

As shown in Fig. 2, the proposed method used a hybrid deep learning architecture to categorize wildlife and environmental sounds with contextual awareness. In order to manage the spatial and temporal complexity of acoustic inputs and understand the contextual links of alarming sounds, the model incorporated a number of components, including convolutional, recurrent, and fuzzy logic layers. The following sections provide an explanation of each stage used in the model. By integrating these processes, the model was able to recognize sounds as well as the significance of their order. Due to this contextual understanding capability", it is more practical in natural settings like national parks or woods.

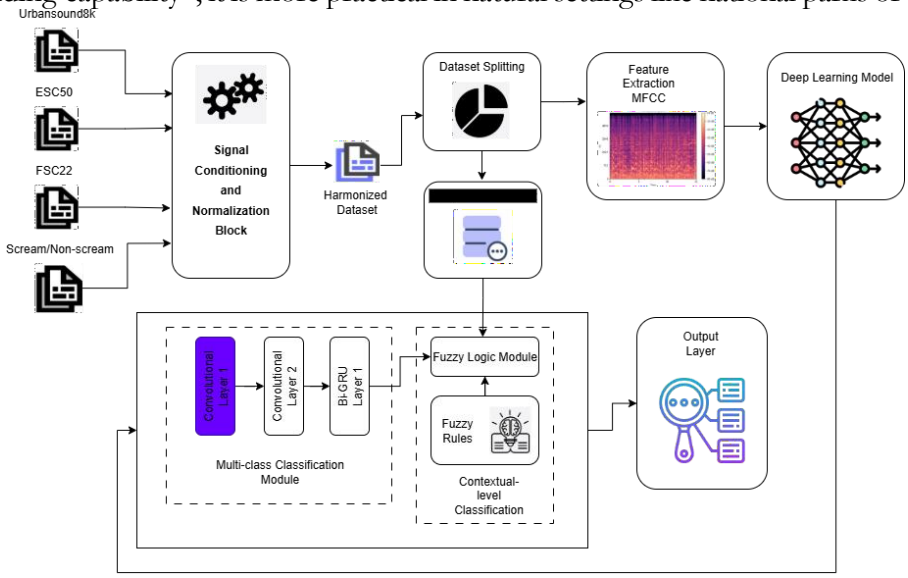


Figure 2. Architecture of the proposed hybrid deep learning

Feature Extraction and Temporal Modeling:

When dealing with inputs that resemble images, Convolutional Neural Networks have been very successful in finding spatial patterns within structured data. The MFCC-based audio representations were first processed using CNN layers in the suggested hybrid model. MFCCs with their first-order derivatives and their second-order derivatives make up each input, which is structured as a 3-channel tensor represented as $X = \text{stack}(X_1, X_2, X_3) \in R^{3\times F}$

$^{\times T}$ where X_1, X_2, X_3 are the features extracted, F is the number of 40 mel-frequency bins, and T is 216 temporal frames. By extracting localized time-frequency data, these CNN layers made it possible to identify distinctive audio signals like dog barking, gunshots, thunder, or screaming, with the nature of these sounds like sudden energy spikes and harmonic patterns.

CNN Layers:

Two convolutional blocks were used to initiate the architecture, which takes the input $X \in R^{3\times F\times T}$. In a 2D convolutional layer, the first convolutional block L_1 used 32 convolutional kernels of size 3 × 3 with a stride and padding of 1 and the second convolutional layer L_2 uses 64 convolutional filters on the previous block output followed by Batch Normalization, ReLU activation, and a max-pooling operation with a kernel size of 2 and a stride of 2 to provide spatial invariance and reduce the dimensionality of the feature maps.

BiGRU Layer:

The final output from the CNN stage was a high-dimensional feature map, $L_2 \in R^{64\times F/2\times T/2}$, rich in spatial information. However, CNNs could not model the sequential nature of sound over time. So, to handle this, the feature map was passed to the next stage of this model, the Bi-GRU, represented as H, specialized in capturing temporal dependencies. This method filtered out unnecessary noise and boosted important parts of the sound while preparing a clear compact view of how each sound changes over both time and frequency. It became very helpful for environmental sound classification because the same event, like a scream, depending on background conditions, can sound completely different. But, unlike models fed with raw audio, this CNN-based approach with MFCCs struck a good balance between the levels of complexity, accuracy, and interpretability. It made use of the ability of CNN to work with 2D format inputs of MFCCs, while also preparing the features for further sequential analysis by BiGRU layer. In this layer the sequences were processed in mini batches of 32, with hidden size of 128 for each direction, gave the output as $H \in R^{T\times 256}$ where T is the time dimension. This step enabled the model to not only detect sound events, but also understand how their acoustic patterns are changed over time.

Fuzzy Logic Module:

After the model has extracted local sound patterns using CNNs and captured temporal relationships through BiGRU, the next step is to extract contextual meaning using fuzzy logic module. This module used the temporal features (training feature vector) from BiGRU as an input and the temporal features or feature vector of the synthesized dataset and applied temporal aggregation by using attention weighted pooling for creating a fixed feature vector of $x \in R^{256}$. This synthesized contextual dataset included combinations of alarming sound classes from the harmonized dataset such as dog bark followed by scream or tree falling sound with a handsaw etc. Similarity between the feature vector and training feature vector was calculated using cosine similarity and normalized using min-max normalization. For each sound class the fuzzy sets were defined as LOW, MEDIUM and HIGH. Triangular Membership Function was used to assign degrees of membership to make a smooth representation of uncertainty between LOW, MEDIUM, and HIGH feature intensities by linearly increasing and decreasing values across a defined range for each class. A separate knowledge base was created named Fuzzy rules which includes rules such as IF scream is High AND dog_bark is High THEN Human Distress etc. which produced degrees of activation for each output category. The firing strength of each rule was computed using fuzzy operators, such as AND OR. To obtain the overall membership value for each class, the outputs were aggregated using the maximum operator. At the end, this module generates a set of fuzzy outputs, to represent the degree to which the input belongs to each category. This fuzzy output vector $\mu(x)$ consisted of four membership values corresponding to Illegal Activity, Human Distress, Natural Hazard, and Safe Activity, each ranging between 0 and 1.

Output Layer:

Next was the final step of classification using defuzzification. The defuzzification stage served as the output component of the fuzzy inference system, mapping aggregated fuzzy memberships into a single crisp class label for final decision-making specific sound categories; Illegal Activity, Human Distress, Natural Hazard and Safe Activity using a maximum membership decision rule.

In this model, two levels of classification are being used, multi-class classification and the contextual-level classification. Multi-class classification (e.g., classifying sounds like thunder, dog bark, scream, etc.) acts as a low-level classification, creating a base for the contextual reasoning by correctly classifying the specific events. However, contextual-level classification acts as a high-level classification which actually classifies the audio into meaningful categories using threshold values, which is a beneficial improvement for surveillance systems.

Training Configuration:

The model was trained using TensorFlow with GPU acceleration, which enabled efficient handling of large-scale environmental sound datasets. Before training, the input features, MFCCs, along with their corresponding delta and delta-delta coefficients, were standardized, and stacked as 3 channels.

Training was conducted using the AdamW optimizer with a learning rate of 0.001, selected for its adaptive learning capabilities and efficient convergence. A batch size of 32 was used, and training proceeded with 50 epochs.

The dataset was divided into training, validation, and testing sets in a 70:15:15 ratio. Data augmentation strategies were used to address the problem of class imbalance. Techniques like mini batch processing and data shuffling were used to improve generalization of the model. Batch normalization was used, and dropout layers with a rate of 0.5 were added after dense layers to further reduce overfitting which helped to enhance the learning stability and smoothness of the model.

Experimental Setup:

The experiments for this study were conducted on a high-performance workstation with an Intel Core i9-11900 @ 3.50 GHz processor, 64 GB RAM, and an NVIDIA RTX 300 GPU with 12 GB VRAM. For programming language, Python 3.9 was used in the computational environment, which ran on Windows 11. To guarantee a stable and completely repeatable configuration, pip was used to install all necessary libraries, including PyTorch, NumPy, OS, Pandas, Librosa, Matplotlib, Seaborn, Scikit-learn, and Scikit-Fuzzy. The GPU played a crucial role in accelerating the training of the deep learning model, especially during convolutional and recurrent operations. This hardware and software configuration enabled efficient execution of the proposed hybrid deep learning model.

Results and Discussion:

The proposed hybrid model is evaluated using common performance metrics such as Accuracy, Precision, Recall, F1-score, and ROC-AUC. For comparison, results from recent deep learning models and surveillance-focused models were also considered. The improved performance of the proposed model comes from several design choices. The performance of the proposed hybrid deep learning model is evaluated on the unified dataset created from ESC-50, UrbanSound8K, FSC22, and the Scream/Non-Scream dataset, with the stratified split of 70% for training, 15% for validation, and 15% for testing. However, for the comparison with the existing literature the model was evaluated on Urbansound8k dataset separately. The loss curves for training and validation are shown in the Fig 3 showing the stable convergence and healthy learning.

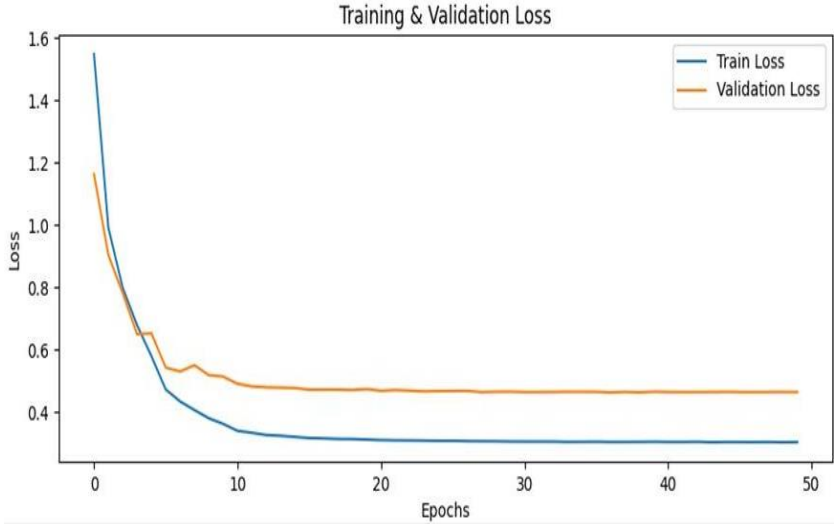


Figure 3. Graph of Training and Validation Loss on UrbanSound8k

The confusion matrix in Figure 4 shows the behavior of class wise classification of the model across 16 alarming classes. This shows that the model effectively distinguishes between most of the environmental sound classes with limited inter-class confusion. However, a slight overlap can be observed among mechanically similar classes such as jackhammer, engine idling, and drilling, as these share similar spectral and temporal patterns. In spite of this, the model maintains strong overall discriminative capability, by feature learning across diverse sound events.

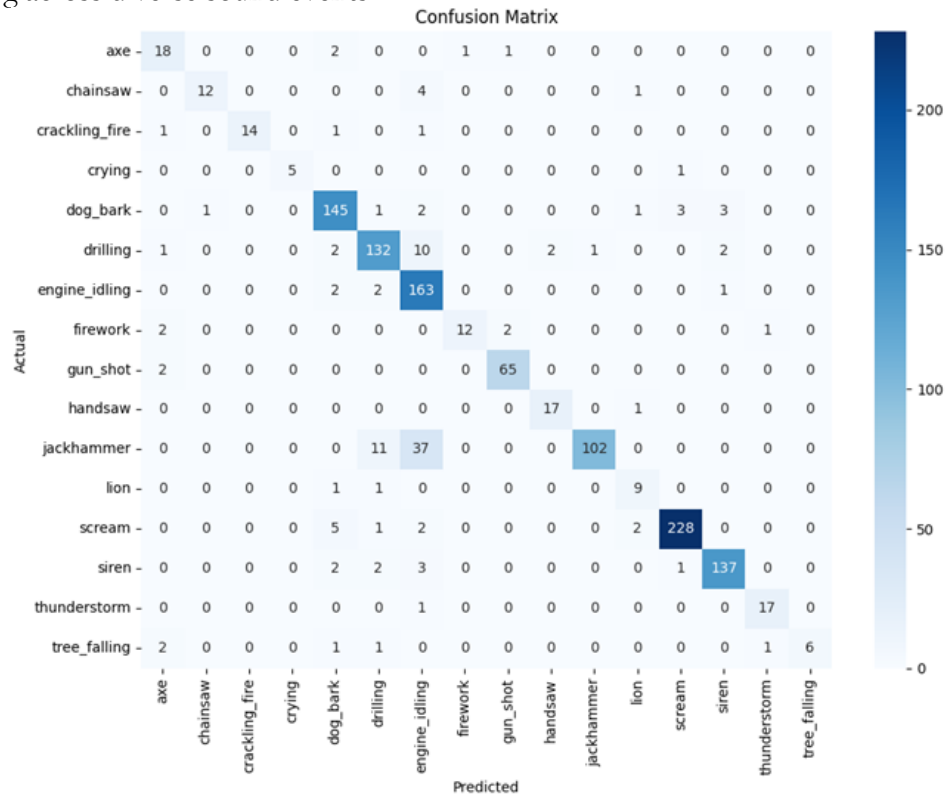


Figure 4. Confusion Matrix results representing class wise behavior

Hyperparameter testing was done to find the better hyperparameter values which also helped the model to show better performance than previous studies. The overall performance of the model is shown in the results in Table 2. using different performance metrics.

Table 2. Results and Performance Comparison

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
[14]	-	-	-	94.86%	-
[31]	-	66.43%	62.75%	63.67%	-
[32]	-	95.91%	95.81%	-	-
[33]	-	-	-	87.66%	-
[34]	92.16%	-	-	-	-
Ours (UrbanSound8k)	**95.80%**	**95.95%**	**96.14%**	**95.80% (weighted)**	**99.67%**
Ours (Harmonized Dataset)	**91.30%**	**88.11%**	**86.23%**	**87.91% (weighted)**	**99.50%**

The results reported in the Table 2 shows the model outperformed existing models on the UrbanSound8K dataset. However, despite the fact that the dataset was heterogeneous, with a variety of sounds like indoor, outdoor, distress, human-related, ambient, mechanical sounds, the model maintained competitive performance with less than a 5% reduction in accuracy with the UrbanSound8K due to its strength and the strategic offline augmentation applied on the training dataset. However, these results confirm better generalization across diverse acoustic conditions by the proposed hybrid system and offers a solution to real world environmental sound classification.

Conclusion:

This paper discusses recent studies on environmental sound classification and how it can be used in many different ways, such as for safety in parks and wildlife areas, monitoring, and surveillance. Traditional machine learning models like SVMs and KNNs have been used a lot, but newer methods use deep learning architectures like CNNs, RNNs, and their hybrid architectures to get better accuracy. In the proposed model, CNN is used for capturing spatial patterns extracted from features such as MFCCs, delta features, and delta-delta features. Next, Bi-GRU learns patterns over time and finishes the low-level classification process. Furthermore, a fuzzy logic module is incorporated after the CNN and BiGRU layers to improve situational awareness by doing high-level classification based on context awareness and helping decisions become more accurate when it comes to monitoring wildlife in national parks. As a result, the suggested hybrid outperformed numerous current studies on environmental sound classification using a comparable dataset, with an overall accuracy of 95.80%. This shows that a fuzzy logic module for context-based high-level classification combined with CNNs and BiGRU for learning detailed features may offer a more dependable and efficient solution for practical applications involving wildlife monitoring and environmental surveillance.

Future Work:

The proposed model can be extended by adding a sound source localization module for alarming classes, which can enable the system to identify the origin of detected critical sounds and activities, which can make it more suitable for large scale monitoring applications. Extension of the rules according to real world scenarios can make the model more efficient for surveillance systems. The proposed model can also be optimized for deployment on edge devices by converting it into a lightweight architecture to support real-time processing. Furthermore, the dataset can be expanded by adding more diverse and locally relevant environmental sounds, including region-specific acoustic events, which can help the model to improve generalization and robustness in real world scenarios.

Acknowledgement: We are thankful to the Lab staff of Department of Computer Science, Quaid-i-Azam University Islamabad for their support in facilitating the development and experimental evaluation of this research.

Author's Contribution: Sunila Sheikh: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing - Original Draft, Writing - Review & Editing.

Umer Rashid: Conceptualization, Formal analysis, Methodology, Project administration, Supervision, Validation, Resources.

Conflict of interest. Authors declare no conflict of interest.

References:

[1] Olusola O. Abayomi-Alli, Robertas Damaševičius, "Data Augmentation and Deep Learning Methods in Sound Classification: A Systematic Review," *Electronics*, vol. 11, no. 22, p. 3795, 2022, doi: https://doi.org/10.3390/electronics11223795.

[2] S. L. Ullo, S. K. Khare, V. Bajaj and G. R. Sinha, "Hybrid Computerized Method for Environmental Sound Classification," *IEEE Access*, vol. 8, pp. 124055–124065, 2020, doi: 10.1109/ACCESS.2020.3006082.

[3] Mahendra Kumar Gourisaria, Rakshit Agrawal, Manoj Sahni & Pradeep Kumar Singh, "Comparative analysis of audio classification with MFCC and STFT features using machine learning techniques," *Discov. Internet Things*, vol. 4, no. 1, 2024, [Online]. Available: https://link.springer.com/article/10.1007/s43926-023-00049-y

[4] Zeinel Momynkulov, Zhandos Dosbayev, "Fast Detection and Classification of Dangerous Urban Sounds Using Deep Learning," *Comput. Mater. Contin.*, vol. 75, no. 1, pp. 2191–2208, 2023, doi: https://doi.org/10.32604/cmc.2023.036205.

[5] M. Bubashait and N. Hewahi, "Urban Sound Classification Using DNN, CNN LSTM a Comparative Approach," *2021 Int. Conf. Innov. Intell. Informatics, Comput. Technol. 3ICT 2021*, pp. 46–50, Sep. 2021, doi: 10.1109/3ICT53449.2021.9581339.

[6] Jozef Kotus, Kuba Lopatka, "Detection and localization of selected acoustic events in acoustic field for smart surveillance applications," *Multimed. Tools Appl.*, vol. 68, pp. 5–21, 2024, [Online]. Available: https://link.springer.com/article/10.1007/s11042-012-1183-0

[7] Wuyue Xiong, Xuenan Xu, "Sound-Based Construction Activity Monitoring with Deep Learning," *Buildings*, vol. 12, no. 11, p. 1947, 2022, doi: https://doi.org/10.3390/buildings12111947.

[8] "(PDF) Application of CNN Models to Detect and Classify Leakages in Water Pipelines Using Magnitude Spectra of Vibration Sound." Accessed: May 05, 2026. [Online]. Available: https://www.researchgate.net/publication/368774766_Application_of_CNN_Models_to_Detect_and_Classify_Leakages_in_Water_Pipelines_Using_Magnitude_Spectra_of_Vibration_Sound

[9] Garima Sharma, Kartikeyan Umapathy, "Trends in audio signal feature extraction methods," *Appl. Acoust.*, vol. 158, p. 107020, 2020, doi: https://doi.org/10.1016/j.apacoust.2019.107020.

[10] K. Zaman, M. Sah, C. Direkoglu and M. Unoki, "A Survey of Audio Classification using Deep Learning," *IEEE Access*, vol. 11, pp. 106620–106649, 2023, doi: 10.1109/ACCESS.2023.3318015.

[11] D. Vij, Y. Yogesh, D. Srivastava, and H. Shankar, "Detection of Acoustic Scenes and Events using Audio Analysis - A Survey," *2023 3rd Int. Conf. Adv. Comput. Innov. Technol. Eng. ICACITE 2023*, pp. 316–320, 2023, doi: 10.1109/ICACITE57410.2023.10183195.

[12] Sahan Dissanayaka, Manjusri Wickramasinghe, Pasindu Marasinghe, "Temporal Convolution-based Hybrid Model Approach with Representation Learning for Real-Time Acoustic Anomaly Detection," *arXiv:2410.19722*, 2024, [Online]. Available: https://arxiv.org/abs/2410.19722

[13] Zohaib Mushtaq, Shun Feng Su, "Spectral images based environmental sound classification using CNN with meaningful data augmentation," *Appl. Acoust.*, vol. 172, p. 107581, 2021, doi: https://doi.org/10.1016/j.apacoust.2020.107581.

[14] L. Luo, "A System for the Detection of Polyphonic Sound on a University Campus Based on CapsNet-RNN," *IEEE Access*, vol. 9, pp. 147900–147913, 2021, doi: 10.1109/ACCESS.2021.3123970.

[15] Jia Wei Chang, Hao Shang Ma, "Multi-Level Transfer Learning using Incremental Granularities for environmental sound classification and detection," *Appl. Soft Comput.*, vol. 169, p. 112619, 2025, doi: https://doi.org/10.1016/j.asoc.2024.112619.

[16] H. M. Do, K. C. Welch and W. Sheng, "SoHAM: A Sound-Based Human Activity Monitoring Framework for Home Service Robots," *IEEE Trans. Autom. Sci. Eng.*, vol. 19, no. 3, pp. 2369–2383, 2022, doi: 10.1109/TASE.2021.3081406.

[17] S. Sathruhan, O. K. Herath, T. Sivakumar, and A. Thibbotuwawa, "Emergency Vehicle Detection using Vehicle Sound Classification: A Deep Learning Approach," *6th SLAAI - Int. Conf. Artif. Intell. SLAAI-ICAI-2022*, 2022, doi: 10.1109/SLAAI-ICAI56923.2022.10002605.

[18] A. Bansal and N. K. Garg, "Robust technique for environmental sound classification using convolutional recurrent neural network," *Multimed. Tools Appl. 2023 8318*, vol. 83, no. 18, pp. 54755–54772, Dec. 2023, doi: 10.1007/s11042-023-17066-2.

[19] Seung Ju Lim, Seong Jin Jang, "Classification of snoring sound based on a recurrent neural network," *Expert Syst. Appl.*, vol. 123, pp. 237–245, 2019, doi: https://doi.org/10.1016/j.eswa.2019.01.020.

[20] Huy Phan, Philipp Koch, Fabrice Katzberg, Marco Maass, Radoslaw Mazur, Alfred Mertins, "Audio Scene Classification with Deep Recurrent Neural Networks," *arXiv:1703.04770*, 2017, [Online]. Available: https://arxiv.org/abs/1703.04770

[21] "Urban Sound Classification using Long Short-Term Memory Neural Network | IEEE Conference Publication | IEEE Xplore." Accessed: May 05, 2026. [Online]. Available: https://ieeexplore.ieee.org/document/8859780

[22] Liane Marina Meßmer, Christoph Reich, "Context-aware acoustic signal processing," *Procedia Comput. Sci.*, vol. 225, pp. 1073–1082, 2023, doi: https://doi.org/10.1016/j.procs.2023.10.095.

[23] Giuseppe De Simone, Antonio Greco, Francesco Rosa, Alessia Saggese & Mario Vento, "Context-aware data augmentation for enhanced speech command recognition in industrial environments," *Sci. Rep.*, 2025, [Online]. Available: https://www.nature.com/articles/s41598-025-01886-3

[24] "Fuzzy Logic in Surveillance Big Video Data Analysis: Comprehensive Review, Challenges, and Research Directions | Request PDF." Accessed: May 05, 2026. [Online]. Available: https://www.researchgate.net/publication/351792725_Fuzzy_Logic_in_Surveillance_Big_Video_Data_Analysis_Comprehensive_Review_Challenges_and_Research_Directions

[25] Reza Saatchi, "Fuzzy Logic Concepts, Developments and Implementation," *Information*, vol. 15, no. 10, p. 656, 2024, [Online]. Available: https://shura.shu.ac.uk/34360/

[26] S. Wu, R. Li, Y. Song, S. Qin, Q. Wen, and F. Gao, "Quantum-Assisted Hierarchical Fuzzy Neural Network for Image Classification," *IEEE Trans. Fuzzy Syst.*, vol. 33, no. 1, pp. 491–502, 2025, doi: 10.1109/TFUZZ.2024.3435792.

[27] Karol J. Piczak, "ESC: Dataset for Environmental Sound Classification," *MM 2015 - Proc. 2015 ACM Multimed. Conf.*, vol. 10, 2015, [Online]. Available: https://dl.acm.org/doi/10.1145/2733373.2806390

[28] Justin Salamon, Christopher Jacoby, "A Dataset and Taxonomy for Urban Sound Research," *MM 2014 - Proc. 2014 ACM Conf. Multimed.*, vol. 11, 2014, [Online]. Available: https://dl.acm.org/doi/10.1145/2647868.2655045

[29] Meelan Bandara, Roshinie Jayasundara, "Forest Sound Classification Dataset: FSC22," *Sensors*, vol. 23, no. 4, p. 2032, 2023, doi: https://doi.org/10.3390/s23042032.

[30] "Audio Dataset of Scream and Non Scream." Accessed: Mar. 11, 2026. [Online]. Available: https://www.kaggle.com/datasets/aananehsansiam/audio-dataset-of-scream-and-non-scream

[31] Jinhua Liang, Ines Nolasco, Burooj Ghani, Huy Phan, Emmanouil Benetos, Dan Stowell, "Mind the Domain Gap: a Systematic Analysis on Bioacoustic Sound Event Detection," *Eur. Signal Process. Conf.*, 2024, [Online]. Available: https://arxiv.org/abs/2403.18638

[32] Shilpa Gupta, Varun Srivastava, "Environment Sound Classification using stacked features and convolutional neural network," *ACM Int. Conf. Proceeding Ser.*, 2024, [Online]. Available: https://dl.acm.org/doi/10.1145/3675888.3676028

[33] I. Mohino-Herranz, J. García-Gómez, "Implementing transfer learning for sound event classification using the realised audio database," *Meas. Sensors*, vol. 38, p. 101711, 2025, doi: https://doi.org/10.1016/j.measen.2024.101711.

[34] Feilong Chen, Zhenjun Zhu, "Evaluating metric and contrastive learning in pretrained models for environmental sound classification," *Appl. Acoust.*, vol. 232, p. 110593, 2025, doi: https://doi.org/10.1016/j.apacoust.2025.110593.

[35] A. Bakhshi, Joaquín García-Gómez, R. Gil-Pita, and S. Chalup, "Violence Detection in Real-Life Audio Signals Using Lightweight Deep Neural Networks," *Procedia computer science*, vol. 222, pp. 244–251, Jan. 2023, doi: https://doi.org/10.1016/j.procs.2023.08.162



Copyright © by authors and 50Sea. This work is licensed under Creative Commons Attribution 4.0 International License.