frontiers | Frontiers in Computer Science

TYPE Original Research
PUBLISHED 28 April 2025
DOI 10.3389/fcomp.2025.1517346

OPEN ACCESS

EDITED BY
Eduard Babulak,
National Science Foundation (NSF),
United States

REVIEWED BY
Caleb Rascon,
National Autonomous University of Mexico,
Mexico
Mamadou Dia,
Université Paris-Est Créteil Val de Marne,
France

*CORRESPONDENCE
Ahmad Almadhor
✉ aaalmadhor@ju.edu.sa
Michal Gregus
✉ michal.gregus@fm.uniba.sk

RECEIVED 25 October 2024
ACCEPTED 04 April 2025
PUBLISHED 28 April 2025

CITATION
Iqbal F, Abbasi A, Almadhor A, Alsubai S and Gregus M (2025) Real-time active-learning method for audio-based anomalous event identification and rare events classification for audio events detection.
Front. Comput. Sci. 7:1517346.
doi: 10.3389/fcomp.2025.1517346

Real-time active-learning method for audio-based anomalous event identification and rare events classification for audio events detection

Farkhund Iqbal[1], Ahmed Abbasi[2], Ahmad Almadhor[3]*, Shtwai Alsubai[4] and Michal Gregus[5]*

[1]College of Technological Innovation, Zayed University, Abu Dhabi, United Arab Emirates, [2]Faculty of Computing and AI, Air University, Islamabad, Pakistan, [3]Department of Computer Engineering and Networks, College of Computer and Information Sciences, Jouf University, Sakaka, Saudi Arabia, [4]College of Computer Engineering and Sciences, Prince Sattam bin Abdulaziz University, AlKharj, Saudi Arabia, [5]Faculty of Management, Comenius University in Bratislava, Bratislava, Slovakia

Introduction: Audio event detection, the application of scientific methods to analyze audio recordings, can be helpful in examining and analyzing audio recordings to preserve, analyze, and interpret sound evidence. Furthermore, it can be helpful in safety and compliance, security, surveillance, maintenance, and predictive analysis. Audio event detection aims to recover meaningful information from audio recordings, such as determining the authenticity of the recording, identifying the speakers, and reconstructing conversations. However, filtering out noise for better accuracy in audio event detection is a major challenge. A greater sense of public security can be achieved by developing automated event detection systems that are both cost-effective and real-time.

Methods: In response to these challenges, this study presented a method for identifying anomalous events based on noisy audio evidence and a real-time scenario to help the audio event detection investigator during the investigation. This study created a large audio dataset containing both noisy and original audio. The dataset includes diverse environmental background settings (e.g., office, restaurant, and park) and some abnormal events (e.g., explosion, car crash, and human attack). This study used an ensemble learning model to conduct experiments in an active learning environment. Nine methods are employed to create the feature vector.

Results: The experiments show that the proposed ensemble learning model using the active learning settings obtained an accuracy score of 99.26%, while the deep learning model obtained an accuracy of 95.35%. The proposed model was tested using noisy audio evidence and a real-time scenario.

Discussion: The experiment results show that the proposed approach can efficiently detect abnormal events from noisy audio evidence and a real-time scenario in real-time.

KEYWORDS

audio event detection, forensics investigation, deep learning, machine learning, noisy data

1 Introduction

Audio event detection is essential in assuring effectiveness, safety, and compliance in the industrial sector because it uses scientific methods to analyze audio recordings (Ross et al., 2020). Detecting anomalies in production line operations or machine sounds offers real-time quality control, ensuring that the products satisfy the necessary criteria

(Abbasi et al., 2022). Additionally, continuously tracking audio data makes predictive maintenance possible, enabling businesses to plan maintenance of pricey failures. Audio event detection is used in various fields, including law enforcement, legal investigations, and security (Abbasi et al., 2022). Audio recordings may provide crucial evidence in criminal investigations, such as eyewitness accounts, confessions, or incriminating Statements (Hamza et al., 2022; Iqbal et al., 2022). In civil and family court cases, audio recordings can be evidence for harassment, custody, or property disputes. Audio event detection may monitor and analyze communication to detect and prevent security threats.

The emergence of numerous uses for audio event detection (AED) in areas like audio-based forensics investigation and audio anomaly detection has made it an essential topic in recent years. Many academics have looked into different AED approaches to raise the accuracy of detection rates. One method for enhancing detection performance is using deep learning to detect auditory events (Greco et al., 2020; Zhang et al., 2021). An accurate AED model can be used for various audio events but is challenging to construct (Mesaros et al., 2021). So, if one wants your AED to function better in a particular situation, it is smart to build a model tailored to that circumstance. The lack of clear separation between audio events in feature space hinders the detection accuracy of traditional machine learning methods. To do this, deep learning methods such as convolutional neural networks (CNNs) trained on spectral images (Sharan and Moir, 2019), deep neural networks (DNNs) trained on feature extraction (Zhang et al., 2021), and RNNs trained on sequence capture (Mesaros et al., 2021). Moreover, better audio event recognition depends on a carefully chosen feature set. In particular, when choosing features, it is crucial to consider factors like the selection of temporal events, noise robustness of features, and minimal computational complexity.

Time-frequency (T-F) cochleagram features, which use a model of human auditory perception to determine when events occurred, have recently gained traction among audio researchers (Mondal and Barman, 2020). Cochleagram is an example of a filter bank method that makes use of gammatone filters (Wang and Brown, 2006). The usage of a mel-scale filter bank is also mentioned in a couple of additional research papers (Vafeiadis et al., 2020). For weak signal-to-noise ratio (SNR) environments, gammatone features outperform mel-scale features regarding robustness to noise (Mondal and Barman, 2022). As opposed to spectrogram-based methods, filter bank approaches are shown to be vastly superior (Mulimani and Koolagudi, 2019). In the presence of noise, the significant spectral-domain granularity in the spectrogram might occasionally mislead the system, but this can be remedied by employing the filter bank technique. In addition, spectrogram characteristics call for greater storage space due to the increased computational load they impose. Therefore, the features of a gammatone filter bank are well-suited for use in real-time applications due to their minimal complexity and time-frequency smoothness.

However, AED remains challenging because of the complexity of real-world noises in audio and the challenging real-time task involved in deploying various AED models (Mesaros et al., 2019). In this paper, we propose a novel approach using an active learning method to identify abnormal events from noisy audio and detect abnormal events in a real-time environment.

This paper makes the following contributions.

- Create a large-benchmark dataset composed of noisy and original audio signals to detect abnormal events from the audio signal from the real world (i.e., explosion, sounds of a machine gun, an assault by humans, a police siren, a car crash, explosions, screams, footsteps, broken glass, a gunshot, and a baby crying).
- Proposed an efficient active learning-based approach to train the ensemble model, combining machine and deep learning approaches to quickly assist the audio event detection investigator in identifying anomalous events. Furthermore, nine different feature extraction approaches are used to create a single feature vector to assist the model during classification.
- A real-time automated AED system is developed to identify abnormal events from different environments (noisy or original audio). This system reduces event detection time while requiring no human intervention.

The rest of the article is organized as follows. Section 2 goes over the existing work. Section 3 contains a detailed overview of the dataset. Section 4 outlines the proposed methodology in detail. Section 5 provides the experimental results and commentary. Section 6 shows the study's conclusion, future work, and limitations.

2 Literature review

This section discusses AED methods and analyzes the importance of AED characteristics used in various research studies.

2.1 Methods for audio event detection

Research within Automatic Event Detection employs deep learning technology at an increasing rate. Wealth or scarcity of deep learning methods does not affect how AED research utilizes audio-derived features as represented in Mesaros et al. (2016); Gemmeke et al. (2017). The production of these feature representations follows three primary methodologies, which include manual feature extraction techniques together with deep learning-generated feature representations and integrated methods that unite these two strategies.

The audio signal is typically converted into the frequency domain, and manual procedures obtain approximate coefficients. The mel-domain and the log-mel-domain are popular representations of an audio signal derived from its spectrogram (Eutizi and Benedetto, 2021; Kothinti et al., 2019; Wang et al., 2021). On the other hand, the learning-based approaches aim to acquire the optimal features for the AED purpose. Two common approaches are: (i) training a feature set using self-supervised or unsupervised learning techniques like auto-encoders (Çakir and

Virtanen, 2018), and (ii) training a separate network specifically for the AED task using the extracted features (Lee et al., 2017).

Hybrid approaches combine features developed by hand with those discovered through machine learning to achieve the best of both worlds (Lee et al., 2017). Context-aware features are learned through a training technique and applied to a specific task, whereas handcrafted features are taken from an algorithm and are founded on expert knowledge. Early or late fusion is preferable in the feature extraction phase because learned features are hard to interpret, and handcrafted features are limited in scope. The classification accuracy greatly increases when the normal handcrafted features are merged with the learned feature representations.

Authors in Zaman et al. (2023) investigated multiple deep learning architectures that include CNNs, RNNs, transformers and hybrid models to perform audio classification functions. The research findings establish deep learning as a superior approach compared to traditional classification methods because it produces better accuracy alongside scalability and adaptability benefits. Large dataset requirements for maximum performance act as a central focus within this study while the researchers identify overfitting problems along with computational resource needs. The author of the paper (Rehman et al., 2021) establishes a cutting-edge method for surveillance detection through the joint processing of audio and visual information. Visual analysis through optical flow techniques, particle swarm optimization and social force model, together with audio features based on MFCCs and spectral features, and RF classification, produce this method. The research shows that applying different data sources together enhances system reliability, particularly when one source fails or experiences performance degradation. A new reference dataset has been introduced for the dual purpose of anomaly detection and uncommon audio event classification by the authors in the paper (Abbasi et al., 2022). The system performs feature extraction through MFCC while using principal component analysis for feature selection and various machine-learning classification models. Testing different machine learning systems across various environments verifies the dataset through research and achieves superior anomaly detection capabilities.

2.2 Researches analyzing features for AED

Research on AED advancement remains superficial about specific approaches to represent these features, whereas the AED problem deserves greater attention to effective representations. These analyses typically only cover a specific proportion of the literature's widely accepted feature extraction methods. Furthermore, the existing studies do not thoroughly study the approaches' hyper-parameters (see Table 1). For instance, the study's author (Piczak, 2015) employed 60 mel-band-log-melspectrogram features with a 25 ms window size and 10 ms hop size. Log-mel-spectrogram characteristics with a 40 ms window size, a 20 ms hop size, and 40 mel bands were used in this study (Cakır et al., 2017). Additionally, they utilized MFCC characteristics with a 40 ms window, 20 ms hop, and 40 mel bands.

2.3 Literature review and our contributions summary

Most previous research has focused on a single feature extraction method for AED, optimizing its settings for maximum efficiency. One can see how they evaluate performance by selecting an extraction technique and focusing on their architecture in Table 1. No study compares popular feature extraction techniques using various parameter settings with the same dataset, as shown in the Table. In contrast to the previous studies, we focused on AED from original and noisy audio evidence and created a feature vector using multiple feature extraction methods.

3 Dataset creation

An automated surveillance application requires the proposed system to identify the abnormality in the observed environment. Sounds of a machine gun, an assault by humans, a police siren, a car crash, explosions, screams, footsteps, broken glass, a gunshot, and a baby crying may all be heard. Monitoring for unusual sounds in real time helps protect people and property from hazards like explosions and car accidents. In light of this, this study developed a database for detecting anomalous audio occurrences called abnormal event detection in audio forensics (AEDAF). Forensics experts can use the AEDAF, a collection of audio based on unexpected abnormal events and diverse background environmental scenes, to better spot these anomalies. We compile the dataset by combining 15 separate background-sound files and 10 important abnormal events. The entire dataset contains ten abnormal events. The Python script integrated these abnormal events at different positions with 15 different background environments. A normal distribution is used in this research to include random noise in the input data. This method ensures that the noise level remains proportionate to the maximum value of the input data, resulting in realistic noise of varying levels. Adding events at varying times and decibel (db) levels complicates the data. In total, there are 2,790 audio samples of background scenes (beach, bus, cafe/restaurant, car, city center, forest path, grocery store, home, library metro station, office, park, residential area, train, and tram); a python script artificially augments this with a variety of abnormal audio events, creating a new dataset of 10,960 audios. Firstly, this dataset contains the original audio; however, the audio noise in the dataset is added using the Python script to make it a more challenging and complex dataset and to obtain the effectiveness of the active learning approach. The technique employs Equations 1, 2 to add the synthetic noise to the dataset. Here, NV represents noise value, 0.015 is the scaling factor, r represents random value, max($data$) represents the maximum value in the dataset and $\mathcal{N}(0,1)$ represents a standard normal distribution (Gaussian noise).

$$NV = 0.015 \times r \times \max(data) \tag{1}$$

$$data' = data + NV \times \mathcal{N}(0,1) \tag{2}$$

The size of the dataset has also increased (21,920 audio), and the final dataset includes both original and noisy audio. The dataset

TABLE 1 A comparison of feature representation in audio-based event detection studies.

Ref.	Spectrogram	Mel-Spectrogram	Log-Mel Spectrogram	MFCC
Cakır et al. (2017)			✓	
Li et al. (2017)			✓	✓
Çakir and Virtanen (2018)	✓		✓	
Shah et al. (2018)	✓			
Becker et al. (2018)		✓		
Dinkel et al. (2018)				✓
Kothinti et al. (2019)	✓			
Zhang et al. (2019)		✓		
Turpault and Serizel (2020)		✓		
Sun and Ghaffarzadegan (2020)			✓	
Kong et al. (2020)			✓	
Kwak and Chung (2020)			✓	
Maria and Jeyaseelan (2021)				✓
Eutizi and Benedetto (2021)		✓		
Wang et al. (2021)			✓	
This paper	✓	✓	✓	✓

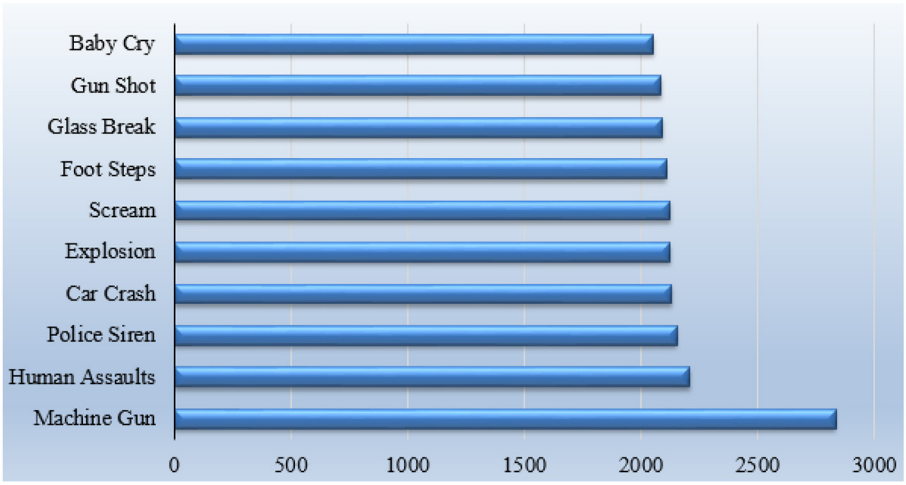


FIGURE 1
Dataset distribution.

is not publicly available online but can be provided upon request. The distribution of the final dataset is shown in Figure 1.

4 Proposed approach

The audio event detection investigator takes a long time to determine the anomalous events from audio evidence based on the noisy surroundings. The researchers are still having trouble identifying anomalous events in noisy audio. This study suggested a technique that will assist the audio event detection investigator in quickly distinguishing these events from a noisy environment.

This section provides a complete description of the processes and procedures that form the basis of our suggested strategy. The proposed approach is presented in Figure 2. The proposed approach mainly consists of three primary stages: Data generation, feature extraction and active ensemble learning model development for real-time prediction. The dataset used in this study contains a mixture of both noisy audio and original audio. The abnormal event is presented in each audio for the final model training. Various preprocessing steps are performed to get insights from

the audio. Feature extraction converts raw input data into a more meaningful representation for model training. Feature extraction seeks to isolate a dataset's most salient and instructive characteristics that can be used to predict a target variable. This study proposes the feature vector by combining the features extracted from nine different audio feature extraction methods. Finally, the active learning-based approach is applied. The ensemble learning model proposed in this study uses the active learning approach during the experiments. The proposed Active Ensemble Learning Model (AELM) combines machine learning and deep learning models. The model is developed to detect abnormal events from the testing set and a real-time environment. In our application, the ensemble-based technique is optimized for real-time performance, allowing quick decision-making with minimal latency. This efficiency is achieved in various ways, including methodology selection based on effectiveness, streamlined implementation, and computational resource optimization. By using these efficiencies, the ensemble may swiftly integrate outputs from several approaches and offer a timely response in seconds, making it ideal for real-time systems that require speedy decision-making.

4.1 Data pre-processing

The performance output of AI models highly relies on effective pre-processing methods. Data framing represents the first procedure in audio processing, which transforms acoustic signals into numeric data that a machine can understand. The audio must be divided into separate frames, and users should keep all sampling rates consistent at 44.1 kHz for our system. A compilation of sample amplitudes from different time points serves as fundamental data for obtaining features. The calculation of audio file frame quantity depends on the sampling rate into the length of the audio as shown in Equation 3. The AEDAF database includes audio samples with a 30-second duration, which frame rates can be determined through Equation 4, the sampling rate is 44.1kHz, and the length of the audio is 30 seconds. The data framing method maintains a uniform sample rate during the whole processing period.

$$Frames = Sampling\ rate \times Time\ of\ an\ audio \tag{3}$$

$$AEDAF(file_1) = 44.1kHz \times 30 = 1,323,000 \tag{4}$$

The audio signal moves from time-based representations to frequency-based representations during the extraction process. A spectrogram displays frequency content evolution through time to detect significant acoustic components. This spectrogram shows a single event, such as footsteps in Figure 3, while representing time through the horizontal axis and frequencies from 0 to 10 kHz through the vertical axis. The recorded sound level uses purple color variations to show sound intensity. When specific data points should be reconstructed, we implement linear interpolation (Equation 5) for deriving estimates from current measured values, here $\{x_1, x_2\}$ and $\{y_1, y_2\}$ are two known data points, x is an intermediate value between $\{x_1, x_2\}$ and y is the estimated value. Standardization plays a vital role in the process of implementing uniform feature scaling. We implement Standard Scaler to standardize features by normalizing both their mean value to zero and standard deviation to one according to Equation 6, where X is the original feature value, X_mean is the mean of the value X, X_std is the standard deviation of X and $X^{'}$ is the standardized value. The encoding process for categorical data uses one-hot encoding techniques, which results in numerical representations. The one hot encoding creates a binary (0,1) column for each category that prepares data for the Machine learning models (Choong and Lee, 2017).

$$y = \frac{y_1 + (y_2 - y_1) \times (x - x_1)}{(x_2 - x_1)} \tag{5}$$

$$X^{'} = \frac{(X - X\_mean)}{X\_std} \tag{6}$$

We validate the dataset through these steps: spectrogram and waveform analysis in addition to quality evaluation using data distribution analysis and correlation structure evaluation alongside outlier detection. The dataset's integrity is confirmed by domain expert validation. The team takes ethical and privacy aspects into account during the dataset utilization process to guarantee responsible usage. As per their evaluation, Experts were provided with a subset of annotated audio samples from different classes. They independently verified the accuracy of labels and sound categories using a blind evaluation method, where the original labels were hidden. The model's output was compared to real samples using MFCCs. The dataset was analyzed for temporal patterns (event duration, frequency distribution over time) and spatial diversity (background noise variations). We ensured that the dataset complies with GDPR and ethical guidelines for audio data collection. Audio samples containing personal identifiers or sensitive content were manually reviewed and excluded from the dataset.

4.2 Feature extraction

In many circumstances, feature extraction approaches noise reduction in some way, particularly when focused on spectral centroid bandwidth, spectral bandwidth, and zero-crossing rate. These features typically capture signal aspects that are less influenced by noise. Furthermore, approaches such as spectral contrast and chroma STFT can create robust audio signal representations by focusing on frequency bands or harmonic content. This helps in limiting the influence of noise.

Every sound stream has a wide variety of features and qualities. However, feature extraction from the event is required before we can do this detection. Firstly, we employ Mel-Frequency Cepstral Coefficients (MFCC) to extract suitable features from the audio signal. This study draws on 40 MFCC-extracted features. In the realm of audio processing, MFCC features are most often employed in the context of speech recognition. In this investigation, MFCC is employed to aid in identifying abnormal sound events. Along with the MFCC features, the feature vector includes (zero crossing rate features, spectral roll-off features, spectral centroid features, spectral contrast features, spectral bandwidth features, chroma stft

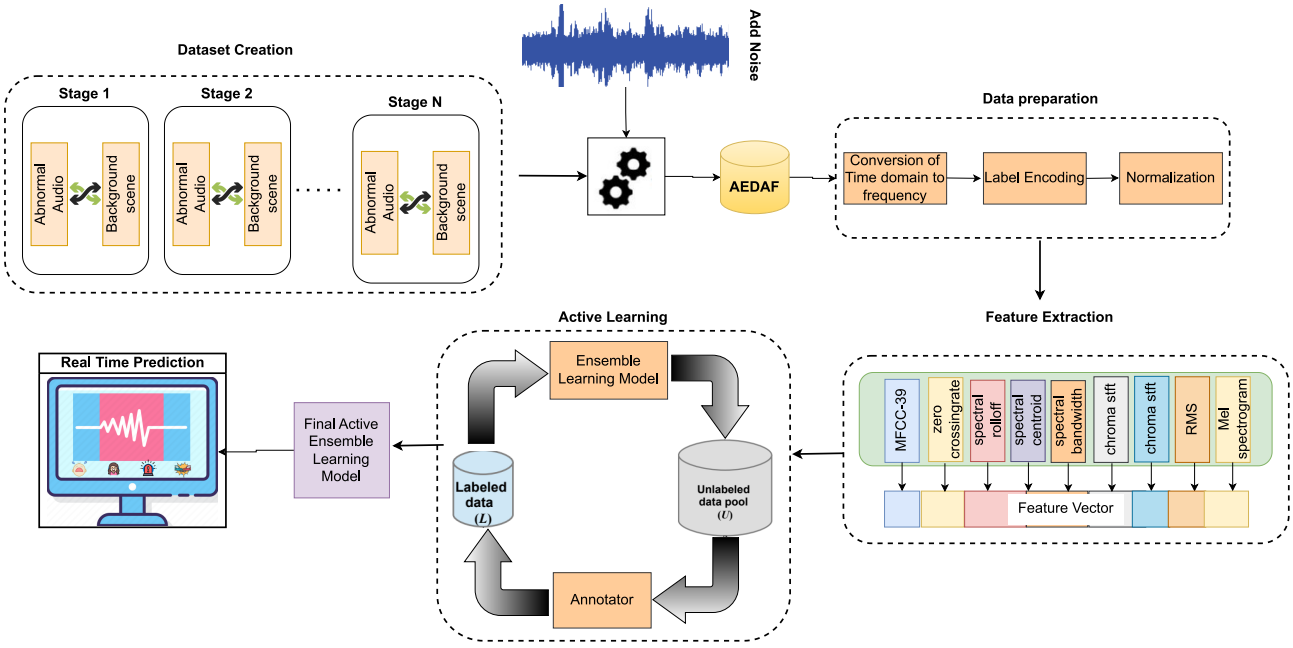


FIGURE 2
Proposed approach for AEDAF creation and abnormal event detection.

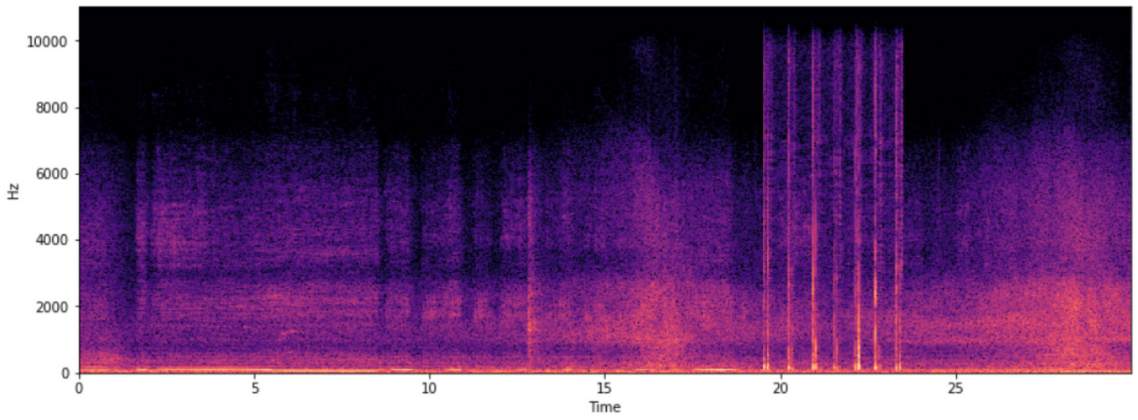


FIGURE 3
Audio spectrogram representation.

features, RMS features, and Mel spectrogram features). The feature vector extracted 288 features from individual audio. These features are further passed to the AELM classifier. The primary purpose of this feature vector is to extract the most optimal features from the noisy audio signal and reduce the processing time for model training and prediction.

4.3 Comparison with other datasets

The ESC-50, alongside DCASE, provides datasets that consist mainly of environmental sounds but do not contain forensic-specific classes like natural events or diverse acoustic scenes. The large database of AudioSet contains numerous sounds, but its insufficient forensic labeling hinders its usefulness in applications that need distinct categorization of relevant audio events. The proposed dataset has been designed to include specific classes of forensic sounds, which include police sirens together with explosions and baby cries, for applications in forensic investigations, emergency systems, and security monitoring. The proposed dataset provides feature vectors already extracted from audio samples to users instead of the raw files users would need for the ESC-50 and AudioSet datasets. The proposed feature representation technique lowers computational requirements while keeping training costs down. Thus, it works well for deep learning and machine learning models, especially when resources

are limited. The structured feature data in the dataset combines complete fidelity with the elimination of extensive preprocessing requirements.

The proposed dataset stands out because it contains more sample data than ESC-50. Due to its 21,920 samples, the proposed dataset exceeds ESC-50 by a factor of more than ten, which enhances the robustness of the deep learning model training. The millions of samples in AudioSet remain unreliable for forensic applications since its poor annotation quality combines with inherent noise to diminish its accuracy in precise task assignments. The proposed dataset stands out as the best choice for security and law enforcement research because its curated selection process creates higher-quality annotations with forensic importance. The quality of data remains a primary issue for large-scale datasets based on YouTube metadata, including AudioSet. The labeling approach generates unreliable annotations because it creates inconsistent data and noisy labels, which decreases its fitness for use in forensic work. The proposed dataset consists of manually checked and curated data points that guarantee higher precision since forensic models depend on highly accurate annotation for their function. The proposed dataset does not have an equivalent in the DCASE challenges since they concentrate on environmental sound classification and acoustic scene analysis. The DCASE effort includes diverse real-world soundscapes but lacks specific content related to important forensic sounds, including gunshots, alarms, emergency sirens, and explosions. The proposed dataset solves this problem by delivering specialized, high-quality forensic audio classification data that specifically fulfills the needs of applications centered on exact forensic sound recognition processes.

4.4 Active learning approach

For evaluation of the proposed method's ability to detect abnormal events from a raw signal and report our findings. The ensemble model based on machine learning (ML) and deep learning (DL) is used in this study to identify abnormal events in noisy audio. Using the entropy-based method described in this article (see Algorithm 1, lines 1–4), data distribution is chosen from a set of unlabeled data. This model employs an evolutionary algorithm-based optimization technique to discover generalizations from limited data. From Algorithm 1's line 6, we can see that the initial training set consists of a limited dataset and an entropy-based approach. The goal is to choose data points for the training set based on their entropy. Based on the pool size, the model decides how many instances are given out, as shown in lines 6 and 7 of the Algorithm 1. This score is based on the uncertainty sample. At the start of each cycle, the pool values are reset. They are incorporated into the training set so an alternative model can be trained using this data. This procedure is repeated until all the unlabeled data has been transformed. All the data used to train the final model is listed on line 9 of Algorithm 1. The proposed method could help reduce the amount of time spent annotating data and identify anomalous audio in a short amount of time.

The active learning approach, which generates new occurrences by selecting from the same distribution as the training set, can be used to generate new examples following the uncertainty of a model. We are training a second model with data generated by an existing model. This process will continue until all the data is used to train the model. On the other hand, active learning does not make any assumptions about what is being studied. One problem with active learning is that model training takes longer because the model has to be retrained every time a new example is made. To mitigate the increased training time in active learning, the study most likely used approaches like batch mode active learning and efficient feature selection to reduce computing complexity. These approaches seek to reduce the impact of retraining on model training time while ensuring that the advantages of adding new examples are realized. A model combines two data sets: training and test sets. Each item in the training set has a label, but none in the test set does. A model successfully separates the training data from the test data using the two datasets. A genetic algorithm was used to tune the parameters of the proposed semi-supervised learning method to improve the model's performance.

1: **Input:**
2: $X_{\text{labeled}}, y_{\text{labeled}}$ {Initially labeled dataset}
3: $X_{\text{pool}}, y_{\text{pool}}$ {Unlabeled dataset (pool)}
4: $X_{\text{test}}, y_{\text{test}}$ {Test dataset}
5: clf {Classifier model}
6: max_iterations {Maximum number of iterations}
7: min_accuracy {Desired accuracy threshold}
8: **for** $i = 1$ to max_iterations **do**
9: Train classifier on labeled data: clf.fit($X_{\text{labeled}}, y_{\text{labeled}}$)
10: Predict labels for pool set: y_{pred} = clf.predict(X_{pool})
11: Compute accuracy: accuracy = $\frac{\sum(y_{\text{pred}} == y_{\text{pool}})}{\text{len}(y_{\text{pool}})}$
12: **if** accuracy $\geq$ min_accuracy **then**
13: **Break** {Stop if accuracy threshold is met}
14: **end if**
15: Compute prediction probabilities: P = clf.predict_proba(X_{pool})
16: Compute entropy for each sample: $H(x) = -\sum P(x) \log P(x)$
17: Select most uncertain sample: $x^* = \arg\max H(x)$
18: Add uncertain sample to labelled set:
19: $X_{\text{labeled}} = X_{\text{labeled}} \cup \{X_{\text{pool}}[x^*]\}$
20: $y_{\text{labeled}} = y_{\text{labeled}} \cup \{y_{\text{pool}}[x^*]\}$
21: Remove selected sample from pool:
22: $X_{\text{pool}} = X_{\text{pool}} \setminus \{X_{\text{pool}}[x^*]\}$
23: $y_{\text{pool}} = y_{\text{pool}} \setminus \{y_{\text{pool}}[x^*]\}$
24: **end for**
25: Predict on test set: $y_{\text{test\_pred}}$ = clf.predict(X_{test})
26: Compute confusion matrix: cm = confusion_matrix($y_{\text{test}}, y_{\text{test\_pred}}$)
27: Compute classification report: cr = classification_report($y_{\text{test}}, y_{\text{test\_pred}}$)
28: **Output:** Trained classifier, confusion matrix, classification report

Algorithm 1. Selection of uncertainty pool using active learning.

TABLE 2 Model architecture with layer details.

Layer (type)	Output shape	Param #
Dense (dense)	(None, 1,000)	289,000
Dense (dense_1)	(None, 750)	750,750
Dense (dense_2)	(None, 500)	375,500
Dense (dense_3)	(None, 250)	125,250
Dense (dense_4)	(None, 100)	25,100
Dense (dense_5)	(None, 50)	5,050
Dense (dense_6)	(None, 10)	510
Total parameters:		1,571,160
Trainable parameters:		1,571,160
Non-trainable parameters:		0

4.5 Machine/deep ensemble learning approach

This study deploys ensemble learning to construct a new model through the integration of machine and deep learning systems. This research applied the voting classifier to select a method from the sklearn library. The research with random forest among machine learning systems along with deep learning models named multilayer feedforward. The selection of ML and DL models for audio event classification relied on different criteria, such as the data characteristics together with task complexity and available computational resources, as well as prior research results. Random forest was selected due to their straightforward nature, good accuracy on datasets of small to medium scale, and easily interpretable capabilities. The multilayer feedforward neural network was selected among DL methods because it efficiently detects complex patterns in audio data and frequently achieves great results in similar classification scenarios. Performance alongside interpretability and computational efficiency led to our decision to incorporate multiple ML and DL techniques despite numerous available options because of research boundaries and objectives. A composite model system helps evaluate the performance strengths between both ML and DL methods within a unified network scheme. Models proceed with their established default settings. The deep learning component applies 500 batch size and executes 100 epochs with Adam optimizer and categorical cross-entropy loss. Table 2 describes the architecture used for the proposed approach.

Aspect-based decision-making involves entropy measurements to evaluate model uncertainty for selecting optimal choices during the decision-making stage. Model prediction confidence levels get evaluated through entropy measurement tools since entropy delivers information theory uncertainty evaluations. It is computed as follows in Equation 7:

$$H(x) = -\sum_{i} P(y_i|x) \log P(y_i|x) \tag{7}$$

The calculated prediction probability for class yi takes the form of $P(y_i|x)$ from an input data point x. A prediction becomes less certain when entropy values grow higher, but lower values indicate more confident predictions. The incorporation of entropy in ensemble systems helps voting strategies lower the effects of imprecise predictions when generating final recommendations. The implementation of entropy produces better classification results when dealing with situations under uncertainty. Following is a representation of the voting strategy shown in the Equation 8:

$$y = \arg\max_{i} \sum_{i=1}^{n} Z_i A_x(V_i(x) = i) \tag{8}$$

In the Equation above, y is the final predicted label determined by selecting the label i that receives the highest weighted vote, denoted by $\arg\max_i$. The process has n number of voting classifiers; each is assigned a weight of Z_i. To check the validity of the vote, a function A_x is used, and V_i stands for the voting classifier.

4.6 Model justification

RF delivered an ideal selection due to its robustness in handling high-dimensional data and its ability to capture complex patterns effectively through ensemble learning. RF provides an outstanding trade-off between interpretability and strong classification performance, making it a reliable choice compared to other tree-based methods. The few characteristics of the dataset support RF selection since it demonstrates strong resistance to overfitting and effective management of various features. Gradient Boosting or Extra Trees are ensemble-based methods, but Random Forest was selected over other options because of its ability to handle large datasets at high speed and its excellent scalability attributes. RF achieves superior performance compared to boosting methods in different situations because it reduces overfitting risks by combining multiple trees for aggregation. RF exhibits exceptional suitability for practical applications because it eliminates the need for time-consuming hyperparameter parameterization. Dense models necessitate large datasets for their effectiveness, but Bayesian models successfully apply prior information to handle limited training data. Risks of overfitting are minimized in Bayesian models because they generate probabilistic outputs that enhance predictive confidence in classification tasks. The implemented selection criteria established an equilibrium between accuracy, interpretability, and computational efficiency, aligning with our research objectives.

5 Experimental results and discussion

The results and methods of the experiments are explained in this section. As can be seen in Figure 4, there are two phases to the AED algorithm: feature extraction and classification. The proposed AELM receives a feature vector constructed with several distinct feature extraction techniques. Using AELM, a voting strategy-based model, we trained the feature subset with the optimal classification model and hyperparameters for the proposed AEDAF benchmark dataset, which included 21,920 audios. 60% (14,028 audios) data is used for training the model, 20% (3,508) data is utilized throughout the active learning process during model

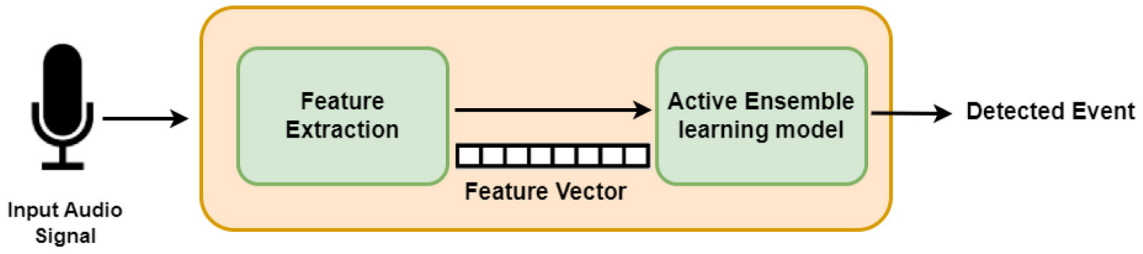


FIGURE 4
Audio event detector.

TABLE 3 Computing environmental setup.

Specifications	Values
Framework	Jupyter notebook
Operating system	Windows 10 professional
CPU	Core(TM) i5-8300H CPU 2.30GHz
GPU	NVIDIA GetForce GTX 1050
RAM	16GB
Programming language & Version	Python 3.8.8

training, and the remaining twenty per cent is used once the AELM model has been fully trained and is ready to test on a new dataset. The experiments were performed in two phases: (i) In the first phase, we evaluated the trained AELM on an unseen dataset, and (ii) In the second phase, the proposed AELM was evaluated using the real-time scenario.

The five essential evaluation metrics were employed in this study. These important evaluation metrics (accuracy, precision, recall, F1-score, and confusion matrix) were employed to determine the ability of the proposed model to test data. The precision, recall and F1-score can be measured using the Equations 9–11.

$$Precision = \frac{True\ Positive}{False\ Positive\ +\ True\ Positive} \tag{9}$$

$$Recall = \frac{True\ Positive}{False\ Negative\ +True\ Positive} \tag{10}$$

$$F1 - measure = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)} \tag{11}$$

Table 3 displays the setup used in the experiment. Several major programming libraries (Pandas, Matplotlib, Numpy, and Sklearn) were also used to aid the studies.

5.1 Results

The experimental results were measured using the testing and new data for the proposed AELM after completing the training using an active learning approach. The proposed Active Ensemble Learning Model results are shown in Table 4 and of Deep Learning are shown in Table 5. The experiments show each class's precision, recall, and F1 score results. The proposed AELM obtained a 99.26% accuracy score using the unseen dataset. Meanwhile, the deep learning model could only reach up to 95%. The support shows the number of audio samples used in the experiments. The model obtained 99.26% accuracy by predicting 4,384 audio samples. The precision, recall, and F1-score outcomes were calculated using the additional two evaluation measures (Macro average and weighted average). When deciding between a macro average and a weighted average, it is important to consider the weights assigned to each model. The weighted average considers the relative relevance of each model, while the macro average treats them all the same. The impressive precision, recall, and F1 score, all reaching 99% for AELM, indicate a robust overall performance. However, the presence of misclassifications suggests the existence of limitations in certain situations or classes. Examining these misclassifications can yield valuable insights into the model's limitations and opportunities for enhancement.

The precision, recall and F1-score of the proposed model is 99.26%. The confusion matrix of the proposed model is shown in Figure 5. Elements on the diagonal of a confusion matrix represent the proportion of correct predictions made by the proposed model, while those off the diagonal represent the proportion of incorrect predictions. The proposed method worked well, as shown by a confusion matrix value indicating that the audio samples were correctly classified. It is critical to assess the supplied data in light of general and specific patterns and provide justification for achieving these results. Discussing general trends entails looking for overarching patterns or tendencies across multiple metrics (e.g., accuracy, recall, and F1-score) and classes. This debate may highlight consistent strengths or drawbacks in the proposed Active Ensemble Learning Model (AELM) across several elements of categorization performance. Moving on to specific trends, digging further into individual class performance and detecting significant differences or inequalities is critical. According to the confusion matrix, this research reveals the classes that the model excels at accurately identifying and which provide more difficulty. The

TABLE 4 Proposed AELM results.

Parameters	Precision	Recall	F1-score	Support
Human assault	0.99	1.00	0.99	450
Baby cry	0.96	1.00	0.98	422
Car crash	1.00	1.00	1.00	420
Explosion	0.99	0.99	0.99	455
Footsteps	0.98	1.00	0.99	424
Glass break	1.00	0.96	0.97	426
Gunshot	0.98	0.91	0.99	394
Machine gun	1.00	0.99	1.99	569
Police siren	1.00	1.00	1.00	413
Scream	1.00	0.98	1.00	411
AELM accuracy	-	-	0.99	4384
Macro average	0.99	0.99	0.99	4384
Weighted average	0.99	0.99	0.99	4384

TABLE 5 Classification report of DL model.

Parameter	Precision	Recall	F1-Score	Support
Human assault	0.92	0.98	0.95	450
Baby cry	0.96	0.98	0.97	422
Car crash	0.96	0.90	0.93	420
Explosion	0.96	0.83	0.89	455
Foot steps	0.91	0.95	0.93	424
Glass break	0.94	0.92	0.93	426
Gunshot	0.89	0.91	0.90	394
Machine gun	0.94	0.98	0.96	569
Police siren	0.99	0.98	0.99	413
Scream	0.96	0.98	0.97	411
Accuracy	-	-	0.94	4384
Macro average	0.94	0.94	0.94	4384
Weighted average	0.94	0.94	0.94	4384

confusion matrix can also help identify class imbalance or improper data distribution by analyzing the classification results.

5.2 Discussion

The proposed AELM model outperforms previous studies based on accuracy levels, which appear in Table 6. The assessed accuracy of AELM reached 0.99.26 while surpassing the findings of earlier research studies. The accuracy scores reported by Wang et al. (2021), Kothinti et al. (2019), and Çakir and Virtanen (2018) were 0.70, 0.75, and 0.60 respectively. AELM shows superior ability in the identification of abnormal events that occur within audio data. The proposed approach demonstrates significant accuracy improvement because it successfully identifies relevant features while delivering good generalization to new data, thus establishing itself as a stronger solution than existing approaches.

AELM shows outstanding results against state-of-the-art audio event detection approaches due to its superior accuracy alongside its robustness to noise and efficient operation. The performance metrics of AELM reached 99.26% accuracy, which surpassed the Convolutional Recurrent Neural Network (CRNN) proposed by Cakir et al. since their best F1 score marked only 66.4%. The CRNN model attempted to extract time-frequency representations directly from waveforms but failed to achieve results better than spectrogram features that experts handcrafted. The analysis by Kothinti et al. for the DCASE 2019 Challenge used supervised along with unsupervised learning approaches to boost the F1-score by 11% against baseline measures yet had less precision compared to AELM. AELM delivers its main strength from an active learning procedure that continuously selects the most doubtful examples to improve the model performance while operating on low amounts of labeled data.

The deep learning models implemented different methods to boost their ability to perform under noise conditions. According to Cakir et al., the log-mel spectrogram features demonstrated better resistance to noise compared to end-to-end feature learning in the 0–3 kHz frequency range, which held the most meaningful data. The approach of Kothinti et al. incorporated event detection based on salience while using Kalman filters to monitor audio signal changes, which enabled their system to adapt to noisy conditions. AELM achieves better acoustic environment generalization by using iterative refinement and uncertainty sampling to choose informative samples regardless of whether they include dedicated noise-mitigation approaches. Additional research should investigate the integration of spectrograms or hybrid approaches to noise reduction methods, which would improve AELM's performance quality in actual operational environments.

The strong computational capability of AELM provides practical benefits for real-world use that surpass deep learning models. The CRNN-based method from Cakir et al. needed extensive computational power because of its deep feature extraction process along with its recurrent layers. The DCASE 2019 Challenge hybrid approach combined supervised and unsupervised learning methods through pseudo-labeling and consistency losses, which increased model complexity and execution time. AELM implements Random Forest with active learning as its classifier while maintaining lower computational costs when compared to other models. By employing the query-by-uncertainty method, AELM minimizes labeling tasks by concentrating on uncertain samples that provide maximum information, thus offering an efficient solution to performance quality and resource management needs. AELM proves better than traditional and deep learning-based audio event detection techniques through its superior accuracy, practical noise resistance, and efficient computational processing. Modern deep learners demonstrate significant advancement, yet they continue to need major datasets and significant computing power. AELM functions as an effective practical model which performs better in situations when labeled data availability is limited along with efficiency requirements.

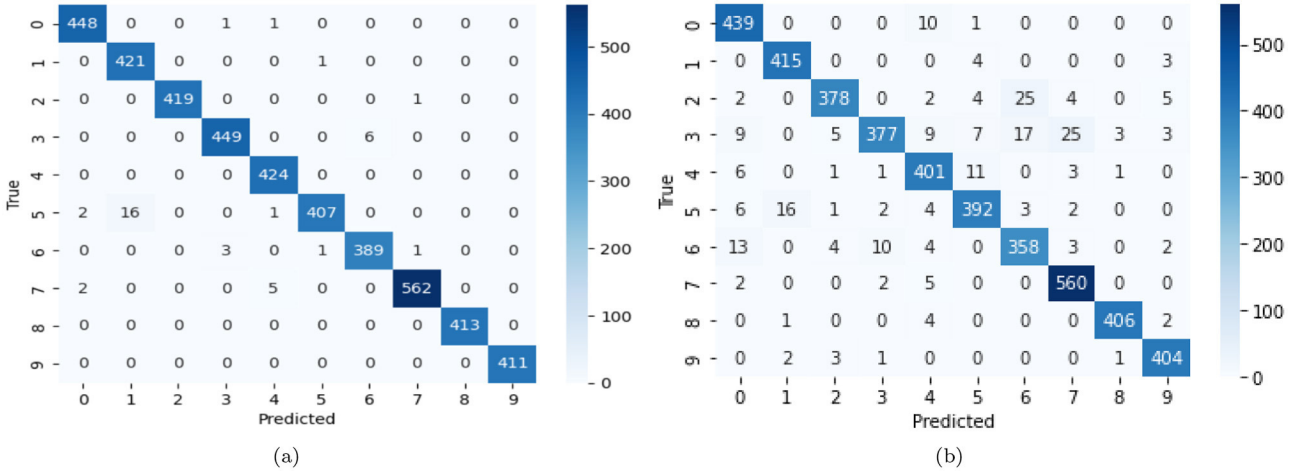


FIGURE 5
Confusion matrix of the proposed AELM and DL Model. **(a)** AELM. **(b)** DL.

TABLE 6 Comparison of proposed AELM model with previous papers.

Papers models	Accuracy
Wang et al. (2021)	0.70
Kothinti et al. (2019)	0.75
Çakir and Virtanen (2018)	0.60
AELM	99.26

Future development should incorporate noise-resistant processing techniques together with hybrid feature extraction systems to enhance performance in noisy acoustic conditions.

5.3 Ablation study

The evaluation of the Active Ensemble Learning Model (AELM) employed Random Forest as its classifier through an active learning technique which enhanced its classification precision in successive iterations. The research examined performance changes and adaptive capabilities of the model through modifications in ensemble tree numbers. A single set of active learning parameters was retained as researchers executed three experimental trials by selecting a distinct number of estimators for each trial. A maximum of ten active learning cycles and five selected samples at each cycle determined the operation of the model, which performed selections based on sample uncertainty to acquire labels before retraining until reaching the defined accuracy threshold of 95%.

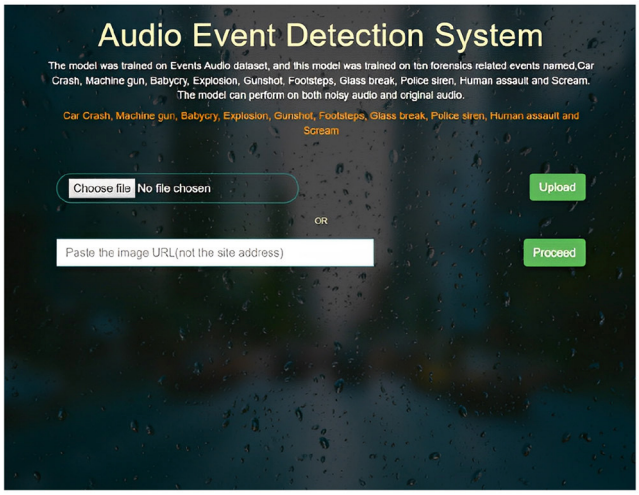
The Random Forest model established a 99.26% accuracy with its 100 estimators because its large ensemble enabled an excellent capture of complex patterns as well as strong decision boundaries. When the number of estimators decreased to 10, the model accuracy dropped to 97.86%, thus showing diminished performance, though the model continued functioning effectively. A model using five estimators reached 95.44% accuracy during the third trial, demonstrating that decreased tree numbers impact the data representation stability and produce unpredictable predictions. We set the Random Forest classifier to test split quality using Gini impurity to achieve both strong performance and efficient processing. The state 42 randomized settings served as our standard for duplicate tests, and the bootstrap sampling method made our model more stable. An algorithm splits up to the root number of total features to stop tree growth before overfitting. The algorithm remained at its basic setup for minimum samples per split, which gave the model flexibility in its decision tree expansion process.

A larger ensemble size produces better results in future applications but requires more computing power to operate. After decreasing the number of estimators, the system processed more easily but achieved less precise results. Using active learning techniques helped the model succeed because it picked up only the best training samples across all estimator counts. Data point selection through active learning became the key factor for preserving high-accuracy performance across all experimental conditions. The model utilized Shannon entropy metrics for uncertainty evaluation by choosing the five least certain samples per execution to boost its training process. Probabilities generated by the model were employed to determine prediction confidence levels, which ensured the selection of unclear samples.

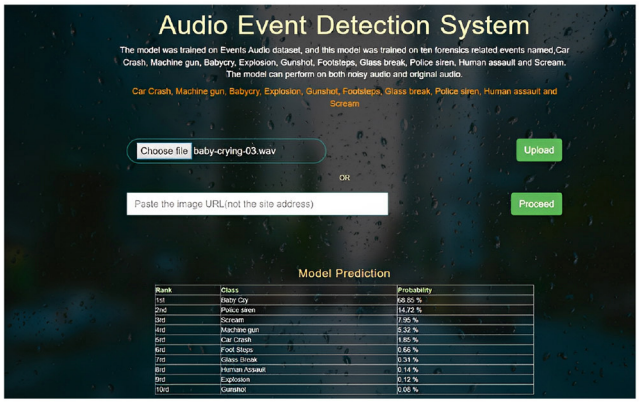
5.4 Real time prediction

We used the Flask framework to build the web-based application for real-time prediction. The trained model was saved using the joblib python library. We created a Python application programming interface (API) in the Flask framework. Figure 6 shows the developed application interface and real-time AED. Figure 6a depicts the interface of the application. The user can either use the online audio link to submit the audio or select it from the local machine for analysis (Figures 6b, c). The proposed

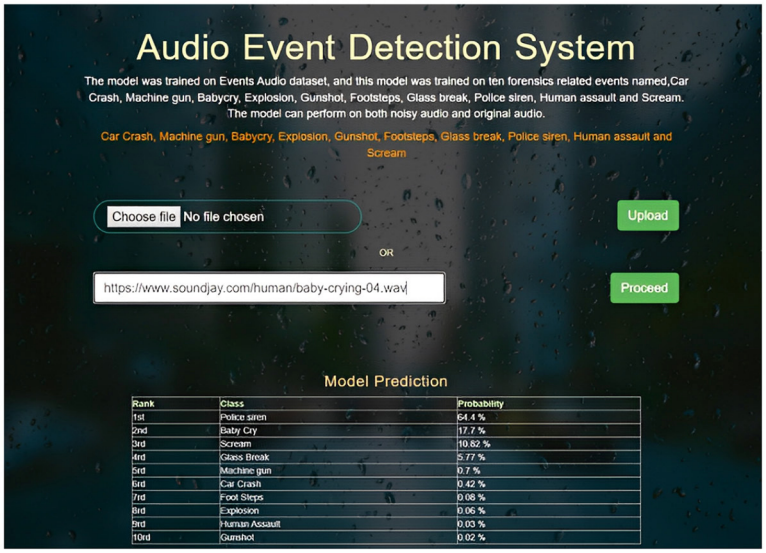
10.3389/fcomp.2025.1517346



(a)



(b)



(c)

FIGURE 6
Real-time AED. **(a)** Interface of the application. **(b)** Event detection from audio in local machine. **(c)** Event detection from audio link.

approach can detect an event from noisy or original audio quickly. The model takes audio evidence as input. The Flask API starts the process by cleaning the audio and extracting the features. Then, the efficient model identifies the abnormal events from the audio. The proposed AED System can identify multiple audio events from audio simultaneously. A higher probability is assigned to an event if it occurs more frequently in the audio. Existing forensics toolkits analyze audio manually, listing it with high and low volumes to discover abnormal events. Human beings require a lot of time and effort to follow this procedure. The primary benefit of AEDs is that they can immediately identify abnormal occurrences in audio, which aids the forensics team and saves time, human effort, and cost.

In a real-time system, genetic evolution generally requires numerous generations to optimize, but we used the Flask framework to create a web-based application for real-time prediction. We optimized the event detection procedure from audio by saving the trained model with the joblib Python package and implementing a Python API in Flask. Our algorithm can quickly recognize aberrant occurrences in noisy and original audio inputs, saving time and effort compared to manual analysis. This automated approach helps forensic teams by quickly recognizing odd occurrences, saving time, human work, and money.

6 Conclusion

The practice of audio event detection has expanded rapidly in recent times because audio technology improvements coincide with the escalating adoption of digital audio devices like smartphones and digital recorders. The rising demand for audio event detection experts created a need for specialized and advanced approaches in the field. Enhancing public safety depends on creating low-cost systems which automatically detect events in real-time. The researchers devised a method to identify abnormal events from imperfect sound recordings in actual operational settings. The proposed technique serves forensic auditors who need to recognize unusual events in noisy environments. For the research, a mass audio dataset was created that included both modified recordings and unmodified audio files. All of the data is split into 15 different locations and contains 10 abnormal incidents. A model based on ensemble learning operated in an active learning framework to execute experiments through nine different approaches for creating feature vectors. The proposed ensemble learning model succeeded in reaching 99.26% accuracy as an active learning solution. The testing phase involved real-world data as well as noisy audio evidence, which proved that the model effectively detected abnormal events when used in real-time situations. The study has one main issue because it does not establish exact times for abnormal event occurrences. The research team plans to improve the model precision for event localization so new dependable automatic event detection systems can be developed.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/supplementary material.

Author contributions

FI: Formal analysis, Investigation, Methodology, Validation, Writing – original draft, Writing – review & editing, Conceptualization, Data curation, Project administration, Supervision. AAb: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Validation, Writing – original draft, Writing – review & editing, Software, Visualization. AAl: Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. SA: Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review & editing, Project administration. MG: Methodology, Project administration, Writing – original draft, Writing – review & editing, Funding acquisition, Resources, Supervision.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This study was supported via funding from Prince Sattam bin Abdulaziz University, project number (PSAU/2024/R/1446).

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Abbasi, A., Javed, A. R. R., Yasin, A., Jalil, Z., Kryvinska, N., and Tariq, U. (2022). A large-scale benchmark dataset for anomaly detection and rare event classification for audio forensics. *IEEE Access* 10, 38885–38894. doi: 10.1109/ACCESS.2022.3166602

Becker, S., Ackermann, M., Lapuschkin, S., Müller, K.-R., and Samek, W. (2018). Interpreting and explaining deep neural networks for classification of audio signals. *arXiv preprint arXiv:1807.03418*.

Cakır, E., Parascandolo, G., Heittola, T., Huttunen, H., and Virtanen, T. (2017). Convolutional recurrent neural networks for polyphonic sound event detection. *IEEE/ACM Trans. Audio Speech Lang. Proc.* 25, 1291–1303. doi: 10.1109/TASLP.2017.2690575

Çakir, E., and Virtanen, T. (2018). "End-to-end polyphonic sound event detection using convolutional recurrent neural networks with learned time-frequency representation input," in *2018 International Joint Conference on Neural Networks (IJCNN)* (Rio de Janeiro: IEEE), 1–7. doi: 10.1109/IJCNN.2018.8489470

Choong, A. C. H., and Lee, N. K. (2017). "Evaluation of convolutionary neural networks modeling of DNA sequences using ordinal versus one-hot encoding method," in *2017 International Conference on Computer and Drone Applications (IConDA)* (IEEE), 60–65. doi: 10.1109/ICONDA.2017.8270400

Dinkel, H., Qiand, Y., and Yu, K. (2018). *A hybrid asr model approach on weakly labeled scene classification*. Technical report, Tech. Rep., DCASE2018 Challenge.

Eutizi, C., and Benedetto, F. (2021). "On the performance improvements of deep learning methods for audio event detection and classification," in *2021 44th International Conference on Telecommunications and Signal Processing (TSP)* (IEEE), 141–145. doi: 10.1109/TSP52935.2021.9522625

Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., et al. (2017). "Audio set: an ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE), 776–780. doi: 10.1109/ICASSP.2017.7952261

Greco, A., Petkov, N., Saggese, A., and Vento, M. (2020). Aren: a deep learning approach for sound event recognition using a brain inspired representation. *IEEE Trans. Inf. Foren. Secur.* 15, 3610–3624. doi: 10.1109/TIFS.2020.2994740

Hamza, A., Javed, A. R. R., Iqbal, F., Kryvinska, N., Almadhor, A. S., Jalil, Z., et al. (2022). Deepfake audio detection via MFCC features using machine learning. *IEEE Access* 10, 134018–134028. doi: 10.1109/ACCESS.2022.3231480

Iqbal, F., Abbasi, A., Javed, A. R., Jalil, Z., and Al-Karaki, J. (2022). "Deepfake audio detection via feature engineering and machine learning," in *Proceedings of the CIKM 2022 Workshops co-located with 31st ACM International Conference on Information and Knowledge Management (CIKM 2022), Atlanta, USA, October 17–21, 2022, volume 3318 of CEUR Workshop Proceedings. CEUR-WS.org*.

Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., and Plumbley, M. D. (2020). Panns: large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Trans. Audio Speech Lang. Proc.* 28, 2880–2894. doi: 10.1109/TASLP.2020.3030497

Kothinti, S., Sell, G., Watanabe, S., and Elhilali, M. (2019). *Integrated bottom-up and top-down inference for sound event detection technical report*. Detection and Classification of Acoustic Scenes and Events.

Kwak, J.-Y., and Chung, Y.-J. (2020). Sound event detection using derivative features in deep neural networks. *Appl. Sci.* 10:4911. doi: 10.3390/app10144911

Lee, J., Kim, T., Park, J., and Nam, J. (2017). Raw waveform-based audio classification using sample-level cnn architectures. *arXiv preprint arXiv:1712.00866*.

Li, J., Dai, W., Metze, F., Qu, S., and Das, S. (2017). "A comparison of deep learning methods for environmental sound detection," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE), 126–130. doi: 10.1109/ICASSP.2017.7952131

Maria, A., and Jeyaseelan, A. S. (2021). Development of optimal feature selection and deep learning toward hungry stomach detection using audio signals. *J. Control Autom. Electr. Syst.* 32, 853–874. doi: 10.1007/s40313-021-00727-8

Mesaros, A., Diment, A., Elizalde, B., Heittola, T., Vincent, E., Raj, B., et al. (2019). Sound event detection in the dcase 2017 challenge. *IEEE/ACM Trans. Audio Speech Lang. Proc.* 27, 992–1006. doi: 10.1109/TASLP.2019.2907016

Mesaros, A., Heittola, T., and Virtanen, T. (2016). "Tut database for acoustic scene classification and sound event detection," in *2016 24th European Signal Processing Conference (EUSIPCO)* (IEEE), 1128–1132. doi: 10.1109/EUSIPCO.2016.7760424

Mesaros, A., Heittola, T., Virtanen, T., and Plumbley, M. D. (2021). Sound event detection: a tutorial. *IEEE Signal Process. Mag.* 38, 67–83. doi: 10.1109/MSP.2021.3090678

Mondal, S., and Barman, A. D. (2020). Speech activity detection using time-frequency auditory spectral pattern. *Appl. Acoust.* 167:107403. doi: 10.1016/j.apacoust.2020.107403

Mondal, S., and Barman, A. D. (2022). Human auditory model based real-time smart home acoustic event monitoring. *Multimed. Tools Appl.* 81, 887–906. doi: 10.1007/s11042-021-11455-1

Mulimani, M., and Koolagudi, S. G. (2019). Segmentation and characterization of acoustic event spectrograms using singular value decomposition. *Expert Syst. Appl.* 120, 413–425. doi: 10.1016/j.eswa.2018.12.004

Piczak, K. J. (2015). "Environmental sound classification with convolutional neural networks," in *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)* 1–6. IEEE. doi: 10.1109/MLSP.2015.7324337

Rehman, A.-U., Ullah, H. S., Farooq, H., Khan, M. S., Mahmood, T., and Khan, H. O. A. (2021). Multi-modal anomaly detection by using audio and visual cues. *IEEE Access* 9, 30587–30603. doi: 10.1109/ACCESS.2021.3059519

Ross, A., Banerjee, S., and Chowdhury, A. (2020). Security in smart cities: a brief review of digital forensic schemes for biometric data. *Pattern Recognit. Lett.* 138, 346–354. doi: 10.1016/j.patrec.2020.07.009

Shah, A., Kumar, A., Hauptmann, A. G., and Raj, B. (2018). A closer look at weak label learning for audio events. *arXiv preprint arXiv:1804.09288*.

Sharan, R. V., and Moir, T. J. (2019). Acoustic event recognition using cochleagram image and convolutional neural networks. *Appl. Acoust.* 148, 62–66. doi: 10.1016/j.apacoust.2018.12.006

Sun, Y., and Ghaffarzadegan, S. (2020). "An ontology-aware framework for audio event classification," in *ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE), 321–325. doi: 10.1109/ICASSP40776.2020.9053389

Turpault, N., and Serizel, R. (2020). Training sound event detection on a heterogeneous dataset. *arXiv preprint arXiv:2007.03931*.

Vafeiadis, A., Votis, K., Giakoumis, D., Tzovaras, D., Chen, L., and Hamzaoui, R. (2020). Audio content analysis for unobtrusive event detection in smart homes. *Eng. Appl. Artif. Intell.* 89:103226. doi: 10.1016/j.engappai.2019.08.020

Wang, D., and Brown, G. J. (2006). *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. New York: Wiley-IEEE press.

Wang, Z., Casebeer, J., Clemmitt, A., Tzinis, E., and Smaragdis, P. (2021). "Sound event detection with adaptive frequency selection," in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* (IEEE), 41–45. doi: 10.1109/WASPAA52581.2021.9632798

Zaman, K., Sah, M., Direkoglu, C., and Unoki, M. (2023). A survey of audio classification using deep learning. *IEEE Access.* 11, 106620–106649. doi: 10.1109/ACCESS.2023.3318015

Zhang, Z., Xu, S., Zhang, S., Qiao, T., and Cao, S. (2021). Attention based convolutional recurrent neural network for environmental sound classification. *Neurocomputing* 453, 896–903. doi: 10.1016/j.neucom.2020.08.069

Zhang, Z. Z., Yang, M., and Liu, L. (2019). *An improved system for dcase 2019 challenge task 4*. Technical Report.