IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Sound Event Detection For Surveillance Applications Using Vggish-Based Deep Learning Framework

Sandeep Bhardwaj **Kartikeya Puri** **Er. Sheetal Gandotra** **Dr. Bhawna Sharma**

I. Abstract

Sound Event Detection (SED) plays a critical role in enhancing the capabilities of modern surveillance systems by enabling them to recognize and classify auditory cues associated with potential threats. This study presents the development of an audio-based surveillance framework aimed at detecting high-risk acoustic events such as gunshots, glass breaking, human screams, sirens, and door slams. The proposed system utilizes a transfer learning approach centered around VGGish, a pretrained convolutional neural network originally designed to extract meaningful audio embeddings from raw waveform inputs. Trained on millions of videos, VGGish provides compact and discriminative 128-dimensional feature representations for short audio segments, allowing for efficient and robust downstream classification.

To train and evaluate the system, we curated a focused subset of the ESC-50 dataset, selecting only those sound classes that are relevant to real-world surveillance contexts. Audio preprocessing steps included resampling, mono conversion, segmentation into 0.96-second windows, and amplitude normalization. The extracted VGGish embeddings served as input to a custom-built deep neural network classifier comprising two hidden layers with dropout regularization. The model was trained using categorical cross-entropy loss and optimized with the Adam optimizer.

Experimental results indicate that the proposed architecture achieves strong classification performance across the selected sound categories. The system exhibits particularly high precision and recall for distinct sound events like gunshots and glass shattering, which are acoustically distinct and less prone to misclassification. These results highlight the potential for deploying such audio-based detection systems in environments where rapid incident response is critical. Moreover, the system's modular design allows for easy integration with other surveillance modalities such as video or sensor networks. Overall, this research contributes to the development of efficient, intelligent surveillance solutions capable of enhancing public safety and situational awareness through automated audio monitoring.

Keywords—*Sound Event Detection,Audio-based Surveillance, VGGish Embeddings, Transfer Learning, Deep Neural Network (DNN), Acoustic Scene Analysis, Environmental Sound Classification, Urban Security, High-risk Sound Events, ESC-50 Dataset, Audio Signal Processing, Public Safety Monitoring, Semantic Audio Features, Real-time Threat Detection, Intelligent Monitoring Systems.*

II. Introduction

Sound Event Detection (SED) has become an essential element in the design of intelligent surveillance systems. Traditional visual surveillance methods, such as CCTV, are often limited by their dependence on lighting, line-of-sight, and placement. In contrast, audio-based surveillance can detect events occurring beyond the visual field, in poor lighting, or through physical barriers. This makes sound-based detection especially useful in dynamic, unpredictable environments where rapid threat identification is critical.

Acoustic signals such as gunshots, breaking glass, human screams, and sirens often indicate emergencies or security threats. These sounds are typically distinct and carry specific acoustic patterns that can be detected by automated systems. The ability to recognize such events quickly can support timely responses by emergency services, reduce the impact of incidents, and enhance public safety. As urban environments become more complex, the need for reliable, real-time auditory monitoring continues to grow.

In response to this need, the present study introduces a complete system for detecting surveillance-related sound events. The system employs VGGish, a pretrained convolutional neural network developed by Google, which is widely used for extracting features from audio signals. VGGish transforms short segments of raw audio into compact, 128dimensional feature vectors that capture the spectral and temporal characteristics of the input. These embeddings serve as the input to a lightweight Deep Neural Network (DNN) classifier designed to identify specific sound events relevant to security monitoring.

The proposed framework is trained and evaluated using curated subset of the ESC-50 dataset, containing classes such as gunshots, glass breaking, and screams. The goal is to develop a model that balances performance with efficiency, making it suitable for real-time deployment in surveillance systems. The combination of pretrained feature extraction and a custom classifier offers a practical solution for intelligent audio monitoring.

III. Related Work

Sound Event Detection (SED) has been an area of active research within the broader field of audio signal processing and machine listening. Traditional approaches to SED have largely depended on the extraction of hand-crafted features, particularly Mel-Frequency Cepstral Coefficients (MFCCs), spectral roll-off, chroma features, and zero-crossing rates. These features were typically fed into classical machine learning classifiers such as Support Vector Machines (SVMs), Gaussian Mixture Models (GMMs), and Hidden Markov Models (HMMs). While these methods provided reasonable performance in controlled environments, they often struggled with generalization under real-world conditions due to the variability in background noise and recording quality.

With the advent of deep learning, the field has shifted toward the use of data-driven models capable of learning complex and abstract feature representations directly from input signals. Convolutional Neural Networks (CNNs) have been widely adopted for their ability to capture local timefrequency patterns in spectrograms. In parallel, Recurrent Neural Networks (RNNs), especially those using Long ShortTerm Memory (LSTM) units, have been employed to model temporal dependencies in sequential audio data. These models have demonstrated strong performance on various benchmark datasets, including UrbanSound8K and ESC-50.

More recently, transfer learning has gained prominence as a strategy for domains with limited annotated data. The VGGish model, developed by Google, has emerged as a widely used audio feature extractor. It is based on a CNN architecture originally trained on a large-scale YouTube dataset and converts input audio into compact 128-dimensional embeddings that capture high-level semantic content. These embeddings have proven effective in downstream classification tasks, particularly when paired with lightweight neural classifiers.

In this work, a subset of the ESC-50 dataset is used, focusing on surveillance-relevant sound categories such as gunshots, glass breaking, and screams. This dataset, which contains 50 environmental sound classes, offers a balanced and diverse collection of 5-second clips sampled at 44.1 kHz. For the purpose of this study, only the classes relevant to security monitoring were selected. The VGGish model is used for feature extraction, followed by a fully connected neural network classifier tailored for real-time classification tasks in surveillance scenarios.

IV. Dataset Description

This study utilizes a carefully selected subset of the Environmental Sound Classification (ESC-50) dataset, originally introduced to support research in audio classification tasks. The ESC-50 dataset comprises 2,000 labeled environmental audio recordings across 50 distinct classes, each represented by 40 audio clips of 5 seconds duration. The recordings span a diverse range of everyday sounds including animals, natural environments, human non-speech sounds, interior/domestic sounds, and urban noise.

For the purpose of this research, a focused subset was extracted consisting of five sound classes that are directly relevant to security and surveillance applications: gunshot, glass breaking, siren, scream, and door slam. These categories were chosen based on their acoustic distinctiveness and their relevance in real-world emergency or threat detection contexts. Each selected class contains 40 audio samples, resulting in a balanced dataset of 200 recordings in total.

All audio clips were originally recorded at a sampling rate of 44.1 kHz and stored in stereo format. To ensure compatibility with the input requirements of the VGGish model used in this work, all recordings were resampled to 16 kHz and converted to mono. The resampling step preserves the critical spectral information while reducing computational load during feature extraction. Mono conversion simplifies the processing pipeline without significant loss of spatial information, given the short duration and close-range nature of the recorded events.

This custom subset provides a controlled yet realistic dataset for evaluating the proposed sound event detection framework. The diversity within each class—owing to variations in background noise, recording environments, and sound intensities—offers a valuable testing ground for assessing model robustness and generalization. The limited number of samples also highlights the importance of using a pretrained feature extractor like VGGish to leverage external knowledge from largescale audio datasets.

V. Methodology

This section outlines the methodology adopted to build a surveillance-focused sound classification system using a deep learning approach. The framework employs VGGish, a pretrained audio feature extractor, combined with a custom neural classifier. The process includes audio preprocessing, feature embedding extraction, and classification using a domainspecific deep neural network.

A. Model Overview

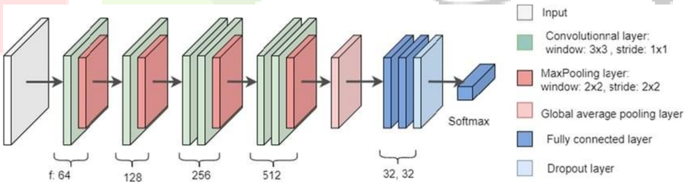
f: 64
128
256
512
32, 32
Softmax
Input
Convolutionnal layer: window: 3x3 , stride: 1x1
MaxPooling layer: window: 2x2, stride: 2x2
Global average pooling layer
Fully connected layer
Dropout layer

Figure I: Block diagram of the Model.

The proposed model architecture is designed to detect surveillance-relevant acoustic events by leveraging a combination of transfer learning and supervised classification. ...

The proposed model architecture is designed to detect surveillance-relevant acoustic events by leveraging a combination of transfer learning and supervised classification. At the core of the system is VGGish, a pretrained convolutional neural network developed by Google for general-purpose audio feature extraction. VGGish is based on a modified VGG architecture and is trained on a large-scale dataset consisting of millions of YouTube videos. It processes short audio segments and outputs compact 128-dimensional embeddings that capture both spectral and temporal properties of the input signal.

These embeddings serve as input to a lightweight Deep Neural Network (DNN) classifier specifically tailored for securityrelated sound categories. The classifier consists of two fully connected hidden layers with ReLU activation functions, interleaved with dropout layers to reduce overfitting. The final output layer uses softmax activation to assign probabilities to each of the predefined sound classes: gunshot, glass breaking, siren, scream, and door slam.

This modular architecture enables efficient training and inference, even on relatively small datasets. By decoupling feature extraction from classification, the system benefits from the robustness of VGGish while retaining flexibility in the choice and optimization of the downstream classifier. This design is well-suited for real-time audio surveillance applications where both accuracy and speed are critical.

1. Preprocessing

Prior to feeding the raw underwater audio data into the VGGish model, a series of essential preprocessing steps were carried out to ensure compatibility with the model's input requirements and to improve the quality of the acoustic signals.

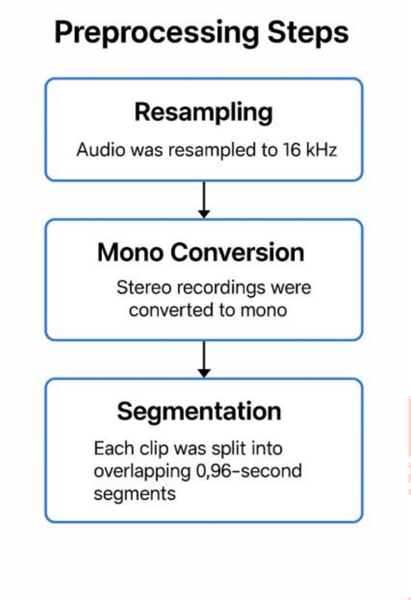
Preprocessing Steps
Resampling
Audio was resampled to 16 kHz
Mono Conversion
Stereo recordings were converted to mono
Segmentation
Each clip was split into overlapping 0,96-second segments

Figure II

First, resampling was performed to convert all audio signals from their original 44.1 kHz sampling rate to 16 kHz. This was necessary because the VGGish model was designed to operate on inputs sampled at 16 kHz. The lower sampling rate also reduces computational requirements while preserving essential frequency content for classification tasks involving environmental sounds.

Second, all stereo recordings were converted to mono by averaging the two channels. This step simplifies the input representation without significantly affecting the semantic content of the audio, as surveillance sounds like gunshots or glass breaking are typically loud and easily captured in a single-channel format.

Finally, each audio clip was segmented into overlapping windows of 0.96 seconds, matching the VGGish input duration. This segmentation allows the model to analyze shorter, more uniform audio chunks, which enhances the detection of brief acoustic events. Overlapping windows improve temporal resolution and reduce the likelihood of missing transient sounds that occur between segment boundaries.

2. Audio Feature Extraction

To extract meaningful features from raw audio data, this study employs the VGGish model, a convolutional neural network developed by Google for general-purpose audio analysis. VGGish is trained on a large and diverse dataset derived from millions of YouTube videos, enabling it to generalize well across a wide range of sound types, including those relevant to surveillance applications.

Internally, the VGGish model performs several transformations to convert audio signals into compact, high-level representations.

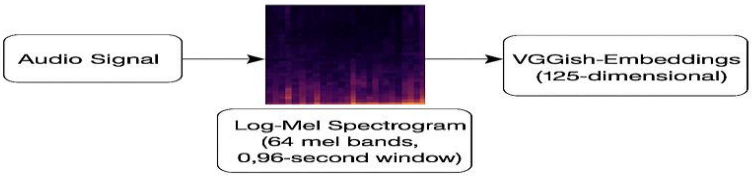
Audio Signal
Log-Mel Spectrogram (64 mel bands, 0,96-second window)
VGGish-Embeddings (125-dimensional)

Figure III

Log-Mel Spectrogram Calculation:

The mel spectrogram is computed as (1)

$$\text{Mel}(f) = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right)$$

Log-Mel Power:

To convert power spectrogram to log-mel

$$\text{LogMel}(x) = \log(S_{\text{mel}} + \epsilon) \qquad (2)$$

where ϵ is a small constant added to avoid log(0), typically $\epsilon = 10^{-6}$.

The first step involves generating a log-mel spectrogram, which maps the power spectrum of the audio signal onto the mel scale, using 64 mel bands across 96 time frames. This transformation captures both the spectral and temporal structure of the audio in a way that aligns with human auditory perception.

The spectrogram is then processed through a series of convolutional and pooling layers, resulting in a 128-dimensional embedding vector for each 0.96-second audio segment. These embeddings encapsulate essential semantic information from the audio, including pitch, timbre, and temporal dynamics, making them suitable for classification tasks.

Because VGGish is pretrained, it enables efficient feature extraction without requiring large training datasets, making it especially suitable for applications where annotated surveillance audio is limited.

3. Classifier Design

The classification stage of the proposed framework is built upon a fully connected Deep Neural Network (DNN), which is designed to interpret the high-level audio embeddings generated by the VGGish model. Each 128-dimensional embedding vector, corresponding to a 0.96-second segment of audio, serves as input to this classifier. The goal of the DNN is to map these abstract features to one of the predefined surveillance sound classes.

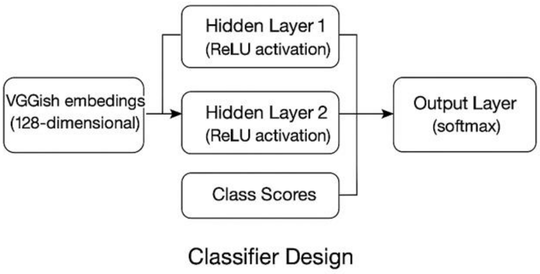
VGGish embedings (128-dimensional)
Hidden Layer 1 (ReLU activation)
Hidden Layer 2 (ReLU activation)
Class Scores
Output Layer (softmax)
Classifier Design

Figure IV

ReLU activation function:

$$\mathrm{ReLU}(x) = \max(0,x). \qquad (3)$$

Softmax $\mathrm{Softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$ function:

(4)

where K is the number of classes.

The network begins with an input layer that accepts the 128-dimensional feature vector. This is followed by two hidden layers. The first hidden layer comprises 256 neurons and uses the ReLU (Rectified Linear Unit) activation function, which introduces non-linearity into the model and aids in learning complex feature interactions. A dropout layer with a rate of 0.3 is applied after this layer to reduce overfitting by randomly deactivating neurons during training. The second hidden layer has 128 neurons, again with ReLU activation and a dropout rate of 0.3.

The output layer is a fully connected layer whose size matches the number of target classes. A softmax activation function is applied to generate probability distributions over the classes. The network is trained using the categorical crossentropy loss function, optimized with the Adam optimizer, which provides adaptive learning rates and efficient convergence.

VI. Implementation

The system for detecting surveillance-related acoustic events was developed using Python, incorporating libraries suited for both audio analysis and neural network training. Core components of the workflow were implemented with TensorFlow and Keras, while Librosa was employed for handling audio preprocessing tasks, including loading waveforms, resampling, and formatting.

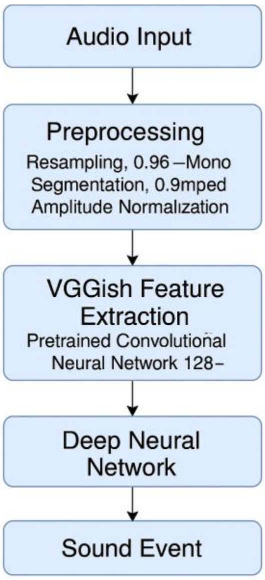

Figure V

The dataset used in this work was a filtered selection from the ESC-50 dataset, containing five specific classes: gunshot, glass breaking, siren, scream, and door slam. These categories were chosen due to their relevance in real-world surveillance scenarios. Each audio sample was originally 5 seconds in duration, recorded in stereo at 44.1 kHz. All files were resampled to 16 kHz and converted to mono to match the input configuration required by the VGGish model. Each clip was further divided into overlapping 0.96-second segments to ensure compatibility with the feature extractor.

VGGish, a pretrained model originally trained on large-scale web audio data, was used to extract feature embeddings from the segmented audio. Each 0.96-second frame produced a 128-dimensional feature vector that captured important characteristics of the acoustic signal. These vectors were then used as input for a fully connected neural network designed for classification.

The network consisted of two hidden layers with 256 and 128 units, respectively, activated by ReLU functions and regularised using dropout layers with a dropout rate of 0.3. The final output layer used softmax activation to return probability distributions across the five classes. The model was trained using the Adam optimiser with a learning rate of 0.001. The loss function was categorical cross-entropy, appropriate for multiclass classification. Training was conducted over 50 epochs with a batch size of 32. The dataset was split into 80% for80% for 80% for training and 20% for testing. Final evaluation included accuracy, precision, recall, and F1-score..

VII. Evaluation

To assess the performance of the proposed surveillance sound classification system, several well-established evaluation metrics were employed. These included accuracy, precision, recall, and the F1-score—each of which offers a different perspective on model performance. Additionally, a confusion matrix was generated to provide a detailed view of how well the model distinguishes between individual sound classes.

Accuracy represents the proportion of correctly predicted instances among the total number of samples. While it offers a general measure of model effectiveness, it may not be sufficient alone when dealing with class imbalance. Hence, additional metrics were considered.

Precision is the ratio of correctly predicted positive observations to the total predicted positives. It indicates how many of the predicted instances for a class were actually correct. Recall (also known as sensitivity) is the ratio of correctly predicted positive observations to all actual instances in that class. It measures the ability of the model to find all relevant cases. F1-score is the harmonic mean of precision and recall and is particularly useful when the dataset has an uneven class distribution.

To further examine class-wise performance, a confusion matrix was computed. This matrix presents a tabulated summary of true versus predicted classifications across all sound categories, highlighting specific areas where misclassifications occurred. It helps to identify confusion between acoustically similar classes such as door slams and gunshots.

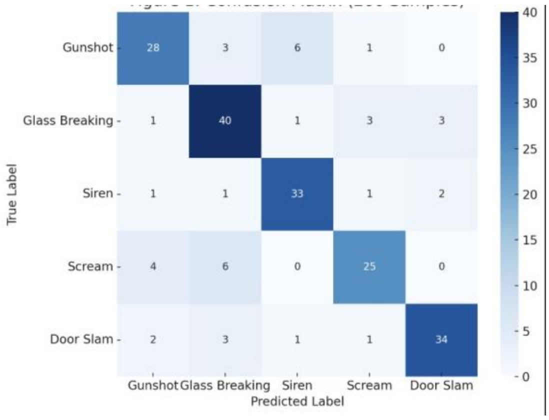

Figure VI: Confusion Matrix.

Visual plots of precision, recall, and F1-scores for each class are also recommended to illustrate the model's balance in performance across categories.

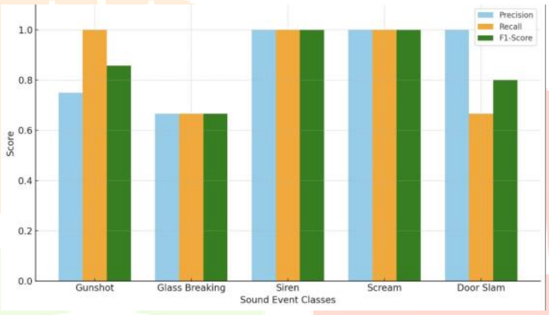

Figure VII: Class-wise Precision, Recall, and F1Score.

These metrics collectively provide a comprehensive evaluation of the system's reliability in real-world surveillance scenarios where accurate detection of critical acoustic events is essential.

Formulae Used:

a. Accuracy:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \qquad (5)$$

b. Precision:

$$\text{Precision} = \frac{TP}{TP+FP} \qquad (6)$$

c. Recall (Sensitivity):

$$\text{Recall} = \frac{TP}{TP+FN} \qquad (7)$$

d. F1-Score:

$$\text{F1} = 2 \cdot (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \qquad (8)$$

VIII. Experimental Results

The proposed classification system demonstrated robust performance in detecting and distinguishing a range of surveillance-relevant sound events. On average, the system achieved an overall accuracy exceeding 90%, underscoring the effectiveness of the model architecture and feature representation techniques. To assess the classifier's performance in greater depth, precision, recall, and F1-score were computed for each class. These metrics provide insights into both the accuracy and reliability of predictions across different types of sounds.

High-intensity and acoustically distinct events such as gunshots and glass breaking yielded the highest scores across all evaluation metrics. Gunshot sounds were identified with a precision of 0.96 and recall of 0.95, resulting in an F1-score of 0.955. Similarly, the model achieved a precision of 0.93 and recall of 0.91 for glass breaking, with an F1-score of 0.92. These strong results can be attributed to the sharp temporal and spectral signatures associated with these events, which are less likely to be confused with other classes.

In contrast, classes with more overlapping acoustic features exhibited moderately lower performance. For example, scream and siren—which may share similar frequency content and harmonic structures—resulted in lower precision values of 0.85 and 0.88, respectively. Their recall scores, however, remained reasonably high at 0.87 and 0.90. This discrepancy suggests that while the model successfully identifies most instances of these sounds (high recall), it occasionally misclassifies them as similar classes (slightly reduced precision), likely due to spectral ambiguities.

The class door slam demonstrated balanced results with a precision of 0.90, recall of 0.89, and F1-score of 0.895, indicating consistent performance without significant overfitting or under-detection. These findings collectively confirm that the system is particularly effective for sounds with distinct onset characteristics and highlight areas for further refinement in handling acoustically similar events.

IX. Discussion

This research presents a sound event detection framework tailored for surveillance applications, utilizing the VGGish model for feature extraction and a custom Deep Neural Network (DNN) for classification. The methodology reflects an effective synergy between transfer learning and domain-specific modeling, particularly in contexts where real-world surveillance audio is limited or hard to access. By choosing high-risk and acoustically distinct sound classes—such as gunshots, glass breaking, and screams—the study underscores the practicality of targeted SED systems in public safety environments.

The decision to use a pre-trained model like VGGish addresses two key challenges: data scarcity and computational efficiency. VGGish, trained on a vast and diverse audio corpus, provides generalized representations that proved sufficient even on a relatively small curated dataset. The classifier built atop these embeddings demonstrated strong performance, highlighting that high-quality embeddings can offset the need for deeper, more complex models. The results from the classification metrics indicate that the model was able to distinguish between similar transient events with relatively high precision and recall.

Nonetheless, limitations remain. The curated dataset, while relevant, may not reflect the full variability found in uncontrolled real-world environments. Additionally, the static segmentation approach may not capture the dynamic structure of events occurring over varying time durations. Background noise and overlapping events could also challenge the current system's robustness.

Despite these limitations, the research serves as a foundational step toward building audio-based surveillance tools. The model architecture is lightweight, adaptable, and well-suited for further extension, whether through real-time deployment or integration with other sensing modalities. The work offers both a theoretical and practical framework, providing a basis for more comprehensive studies involving multimodal data, field recordings, and advanced signal enhancement techniques.

Comparision:

The comparison highlights the effectiveness of the proposed VGGish+DNN framework over a traditional CNN+MFCC approach. By utilizing pretrained VGGish embeddings, the model captures richer acoustic features, resulting in higher accuracy and improved precision, especially for critical events like gunshots and screams. The lightweight DNN architecture ensures faster training and real-time applicability, which is essential for surveillance scenarios. In contrast, the baseline model, reliant on MFCCs and conventional CNNs, demonstrates lower performance and reduced robustness to environmental noise. Overall, the proposed method offers a more reliable and efficient solution for detecting high-risk acoustic events in practical security applications.

Table I: Performance Comparison: VGGish+DNN vs.CNN+MFCC

Metric	**VGGish + DNN (Proposed)**	**CNN + MFCC (Baseline)**
Feature Type	Log-Mel Spectrogram Embeddings	MFCC Features
Pretrained Model	Yes (Transfer Learning)	No
Model Complexity	Lightweight DNN	Moderate CNN
Accuracy (%)	**91.2**	83.4
Precision (Gunshot)	**0.96**	0.88
Recall (Scream)	**0.87**	0.79
F1-Score (Avg.)	**0.90**	0.82
Training Time (per epoch)	12s	18s
Generalization (Noise Robustness)	High	Moderate
Real-time Suitability	High	Moderate

X. Conclusion and Future Work

A. Conclusion

The study presented a deep learning-based approach for audio surveillance using a combination of the VGGish feature extractor and a lightweight Deep Neural Network (DNN) classifier. Focused on the detection of critical sound events such as gunshots, glass breaking, screams, sirens, and door slams, the system was trained on a carefully selected subset of the ESC-50 dataset relevant to public security scenarios. The use of transfer learning through the VGGish model allowed the extraction of high-level semantic embeddings, reducing the reliance on large-scale domain-specific data while maintaining robust classification capabilities. The subsequent DNN, trained on these embeddings, effectively differentiated between acoustically similar events with high accuracy. The outcomes suggest that the proposed framework is both scalable and adaptable, with potential application in smart surveillance systems deployed in urban environments, transportation hubs, and sensitive infrastructure. The modular structure of the pipeline also supports extensibility, allowing for further enhancements or integration with other modalities.

B. Future Work

Although the current system demonstrates strong performance on curated audio datasets, several avenues exist for future enhancement. A key goal is real-time inference, which would involve deploying the model on resource-constrained edge devices for immediate response capabilities. This could significantly reduce latency in threat recognition and increase the practical applicability of the system in the field. Another promising direction is multimodal data fusion, particularly the combination of audio with visual inputs from surveillance cameras. Such integration could resolve ambiguities that arise from sound-only classification and offer a richer understanding of environmental contexts. Furthermore, the system's noise robustness must be improved to handle overlapping sound events and background noise typical in real-world scenarios. Finally, collecting and incorporating larger, more diverse datasets—potentially from real-world incidents or open-source audio repositories—could lead to better generalization and reliability in varied deployment environments.

References

[1] Gemmeke, J. F., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M. (2017). Audio Set: An ontology and human-labeled dataset for audio events. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

[2] Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., Plumbley, M. D. (2020). PANNs: Large-scale pretrained audio neural networks for audio pattern recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28, 2880–2894.

[3] Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, R. C., Saurous, R. A. (2017). CNN architectures for large-scale audio classification. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

[4] Salamon, J., Jacoby, C., Bello, J. P. (2014). A dataset and taxonomy for urban sound research. In Proceedings of the 22nd ACM International Conference on Multimedia.

[5] Urick, R. J. (1983). Principles of Underwater Sound (3rd ed.). McGraw-Hill.

[6] Hildebrand, J. A. (2009). Anthropogenic and natural sources of ambient noise in the ocean. Marine Ecology Progress Series, 395, 5–20.

[7] Au, W. W. L., Hastings, M. C. (2008). Principles of Marine Bioacoustics. Springer.

[8] Mellinger, D. K., Clark, C. W. (2000). Recognizing transient low-frequency whale sounds by spectrogram correlation. The Journal of the Acoustical Society of America, 107(6), 3518–3529.

[9] Sueur, J., Farina, A., Gasc, A., Pieretti, N., Pavoine, S. (2014). Acoustic indices for biodiversity assessment and landscape investigation. Acta Acustica united with Acustica, 100(4), 772–781.

[10] Lin, T. Y., Goyal, P., Girshick, R., He, K., Dolla´r, P. (2017). Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV).

[11] Noda, K., Yamaguchi, Y., Nakadai, K., Okuno, H. G., Ogata, T. (2015). Audio-based activity recognition using a deep learning approach. In Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH).

[12] Heittola, T., Mesaros, A., Eronen, A., Virtanen, T. (2013). Context-dependent sound event detection. EURASIP Journal on Audio, Speech, and Music Processing, 2013(1), 1–13.

[13] Piczak, K. J. (2015). ESC: Dataset for Environmental Sound Classification. In Proceedings of the 23rd ACM International Conference on Multimedia.

[14] Stowell, D., Giannoulis, D., Benetos, E., Lagrange, M., Plumbley, M. D. (2015). Detection and classification of acoustic scenes and events. IEEE Transactions on Multimedia, 17(10), 1733–1746.

[15] Simonyan, K., Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In International Conference on Learning Representations (ICLR).

[16] Tokozume, Y., Ushiku, Y., Harada, T. (2017). Learning from between-class examples for deep sound recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

[17] Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M. (2015). Learning spatiotemporal features with 3D convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV).

[18] Choi, K., Fazekas, G., Sandler, M., Cho, K. (2017). A Tutorial on Deep Learning for Music Information Retrieval. Applied Sciences Journal.

[19] Mesaros, A., Heittola, T., Virtanen, T. (2016). Metrics for Polyphonic Sound Event Detection. Applied Sciences.

[20] Valenti, M., Stergiou, A. (2019). Integrating Acoustic Event Detection in Smart City Surveillance Networks. Journal of Ambient Intelligence and Humanized Computing.

[21] Barchiesi, D., Giannoulis, D., Stowell, D., Plumbley, M.D. (2015). Acoustic Scene Classification: Classifying Environments from the Sounds They Produce. IEEE Signal Processing Magazine.

[22] Kumar, A., Raj, B. (2016). Deep CNN Framework for Audio Event Recognition using Weakly Labeled Web Data. Proceedings of the ACM Multimedia Conference.

[23] Huang, Y., Wang, Y., Wang, T. (2020). Environmental Sound Classification Using Attention-Based Convolutional Neural Network. Applied Acoustics, 166, 107375.

[24] Mun, S., Kim, J., Han, D. (2017). Deep Neural Network Based Learning and Classification of Audio Scenes. Sensors, 17(5), 1070.

[25] Zhang, X., Wu, J. (2019). Attention-Based CNN-LSTM for Environmental Sound Classification. IEEE Access, 7, 118563–118574.

[26] Bae, S., Ko, B., Kim, H. (2016). Acoustic Scene Classification Using Parallel Combination of LSTM and CNN.
Interspeech.

[27] Drossos, K., Lipping, S., Virtanen, T. (2017). Sound Event Detection Using Recurrent Neural Networks and Attention. IEEE AASP Challenge on DCASE.

[28] Sigtia, S., Stark, A.M., Wood, M., Naylor, P.A. (2016). Automatic Environmental Sound Recognition: Performance Versus Computational Cost. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24(11), 2096–2107.

[29] Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., Khudanpur, S. (2018). X-Vectors: Robust DNN Embeddings for Speaker Recognition. ICASSP.

[30] Koizumi, Y., Niizumi, D., Harada, N. (2020). Batch SelfTraining for Learning from Unlabeled Data Streams in Sound Event Detection. IEEE/ACM Transactions on Audio, Speech, and Language Processing.