GUN SOUND RECOGNITION USING NLP AND YAMNET MODEL

[1]G. Sailikith, [2]G. Yashwanth, [3]J. Pavan, [4] Mr V. Devasekhar
[1,2,3]UG Scholars,[4]Associate Professor
[1,2,3,4]Department of Computer Science and Engineering
[1,2,3,4]Guru Nanak Institutions Technical Campus, Hyderabad, Telangana, India
*Corresponding author E-mail: gundetisailikith@gmail.com

ABSTRACT

This approach introduces a hybrid methodology for detecting gunshot sounds by combining traditional and deep learning techniques. Initially, audio signals are analyzed using Mel-Frequency Cepstral Coefficients (MFCC), which effectively capture the unique spectral features of the sound. These extracted features are then fed into a Support Vector Machine (SVM), which classifies the sounds into gunshots or other types. To boost the system's accuracy and reliability, YAMNet—a pre-trained deep neural network model designed for sound classification—is also incorporated. YAMNet categorizes the audio across a wide range of classes, offering an additional perspective to support or refine the SVM's output. By integrating the focused classification strength of SVM with the comprehensive audio recognition capabilities of YAMNet, the system achieves improved precision in identifying gunshots. This dual-layered technique is particularly useful for real-time audio analysis in environments where accurate detection is critical. The synergy between handcrafted audio features and deep learning models results in a powerful and adaptable detection framework suitable for diverse real-world scenarios.

Keywords: Gunshot Detection, MFCC Features, Audio Classification, Support Vector Machine (SVM), YAMNet Model, Hybrid Approach, Real-Time Audio Processing, Sound Recognition, Machine Learning Integration , Deep Learning for Audio

I.INTRODUCTION:

In recent years, the need for accurate and efficient sound recognition systems has grown substantially, particularly in security and surveillance domains. Among various acoustic signals, gunshot detection holds critical importance due to its implications for public safety and rapid emergency response. This project proposes a novel hybrid approach for detecting gunshot sounds by combining both traditional machine learning techniques and modern deep learning models. The core idea is to leverage the strengths of Mel-Frequency Cepstral Coefficients (MFCC), Support Vector Machines (SVM), and YAMNet, a powerful pre-trained neural network designed for general audio classification.

The process begins with the extraction of MFCC features from audio inputs. MFCCs are widely recognized for their ability to represent the short-term power spectrum of a sound, effectively capturing key auditory information such as pitch, tone, and intensity. These features are particularly suitable for distinguishing gunshots from other environmental sounds. After feature extraction, the data is passed to an SVM classifier, which excels in separating data points using a high-dimensional feature space. The SVM model is trained to differentiate between gunshot sounds and non-gunshot

noises, providing a highly reliable first-level classification.

To enhance the overall detection accuracy and robustness, the system incorporates YAMNet, a deep convolutional neural network pre-trained on a large dataset of audio events. YAMNet adds a secondary layer of analysis by classifying audio clips into a broad array of sound categories. This additional classification step helps validate and support the initial decision made by the SVM, minimizing false positives and improving confidence in the final output.

By integrating MFCC, SVM, and YAMNet, this study presents a comprehensive and effective solution for gunshot detection. The synergy between handcrafted features and deep learning-based sound categorization ensures high performance across diverse acoustic environments. This hybrid system demonstrates significant potential for real-time deployment in safety-critical scenarios, offering timely alerts and aiding in the rapid response to incidents involving firearms.

II. LITERATURE SURVEY:

In 2021, S. Dogan presented a novel approach for identifying different gun models using the acoustic patterns found in gunshot audios. The method is based on the use of a fractal H-tree structure, a geometric model known for its recursive and self-similar properties. This fractal pattern is applied to the audio signals to extract detailed features that capture the unique characteristics of each gunshot. These extracted features serve as input to a machine learning classifier trained to distinguish between various firearm types. The system was evaluated using a dataset containing gunshot recordings from different guns. The results revealed that the proposed approach performed effectively in identifying specific gun models. It also demonstrated competitive accuracy when compared with conventional methods. The use of fractal geometry adds a new dimension to audio signal representation, enabling better recognition performance. By leveraging the complex, repeating nature of the H-tree pattern, the model captures intricate sound details. This enhances the classifier's ability to make precise distinctions between similar audio samples. The technique offers an innovative way to approach firearm identification using sound alone. It has potential applications in forensic analysis, surveillance, and law enforcement systems. Overall, the study shows how combining fractal features with machine learning can improve gunshot audio classification. It opens up new possibilities in the field of acoustic-based firearm recognition.

III. EXISTING SYSTEM:

Convolutional Neural Networks (CNNs) have played a vital role in medical image analysis, particularly in the detection of strokes using CT scans. Their strength lies in their ability to automatically identify and extract key features from complex image data, eliminating the need for manual intervention in feature design. These networks utilize a layered structure—including convolutional, pooling, and fully connected layers—that work in unison to capture various levels of image detail. Through this hierarchical learning process, CNNs can effectively identify patterns associated with stroke indicators. Despite their success and accuracy in numerous studies, CNN models often face practical limitations. One such limitation is their dependence on large labeled datasets, which are often difficult to obtain in medical domains. Moreover, their depth and complexity can lead to increased computational demands, making real-time processing more difficult in clinical environments. These early CNN-based systems, while foundational, require improvements to ensure greater efficiency and adaptability. To build upon these strengths while minimizing weaknesses, the proposed system introduces the ResNet architecture. ResNet enhances traditional CNNs by introducing residual connections, which help

mitigate issues like the degradation of accuracy in deeper networks, thus enabling better performance and stability in complex stroke detection tasks.

One major drawback of CNN-based models is their high computational cost, which can hinder deployment in real-time medical systems. Additionally, they are susceptible to overfitting, especially when trained on small or limited datasets. Another concern is the lack of interpretability, as understanding how these models arrive at specific decisions can be challenging, posing a barrier in sensitive fields like healthcare.

IV. PROPOSED SYSTEM:

The proposed system offers several notable advantages that contribute to its effectiveness in audio classification tasks. By using MFCCs, it provides a compact yet highly informative representation of audio signals, capturing essential spectral features that are crucial for accurate recognition. Since MFCCs are designed to reflect the way humans perceive sound, they enhance the system's ability to differentiate between various audio patterns, leading to improved recognition accuracy. The integration of SVM further strengthens the model by ensuring a clear and well-defined margin between different classes. This precise separation contributes to better classification performance, especially in complex audio environments. Moreover, SVM's reliance on support vectors makes the system resilient to outliers and noise in the data, ensuring stable and reliable predictions even when working with challenging or imperfect inputs.

The proposed system efficiently represents audio signals using MFCCs, closely aligning with how humans perceive sound. This leads to improved accuracy in recognizing and classifying audio events. With SVM's clear separation of classes and resistance to outliers, the system remains both precise and robust.

METHODOLOGIES:

Feature Extraction using MFCC

The system begins with audio preprocessing where unwanted noise is filtered, and volume levels are normalized to maintain consistency. The clean audio is segmented into short frames to capture variations over time. Each frame is transformed to the frequency domain using Fourier Transform. A Mel-scale filter bank is then applied to focus on frequency ranges sensitive to the human ear. The output is passed through a logarithmic function followed by a Discrete Cosine Transform (DCT) to generate Mel-Frequency Cepstral Coefficients (MFCCs). These coefficients form a compact and effective representation of the audio's spectral content and are particularly useful in distinguishing the unique patterns of gunshot sounds from other noises.

Classification with SVM

After feature extraction, the MFCCs are fed into a Support Vector Machine (SVM) for classification. This model is trained on labeled audio samples where gunshot and non-gunshot sounds are clearly marked. The SVM learns to separate the two classes by identifying an optimal boundary in the feature space. Feature selection may be applied to reduce complexity and improve training performance. To validate the model's accuracy and reliability, evaluation metrics such as precision, recall, and F1-score are used. Cross-validation helps ensure the model generalizes well to new data, minimizing errors and overfitting.

Integration of YAMNet

YAMNet, a pre-trained deep learning model designed for audio classification, is used to complement the SVM. It can classify a wide range of environmental and human-made sounds, including gunshots. By processing the same audio input, YAMNet offers an independent prediction that supports or refines the SVM's output. This dual-model structure

increases the confidence and accuracy of the final detection, especially in noisy or unclear audio conditions.

Real-Time Detection

Both models are combined into a single real-time detection system. A continuous audio stream is analyzed on-the-go, with MFCC extraction and classification occurring instantly. The predictions from both the SVM and YAMNet are fused to generate a final output. To ensure fast and responsive detection, the system is optimized through techniques like parallel processing or hardware acceleration. This allows it to function effectively in time-critical scenarios such as surveillance or emergency alerts.

Testing and Validation

A well-prepared dataset containing both gunshot and non-gunshot samples is used to test the system. It includes diverse audio environments and firearm types to ensure broad applicability. The system is evaluated on how well it identifies gunshots while minimizing false positives. Standard metrics and cross-validation techniques are used to confirm its accuracy and consistency. This ensures the model is ready for real-world deployment where dependable detection is essential.

EXISTING TECHNIQUE USED OR ALGORITHM USED:

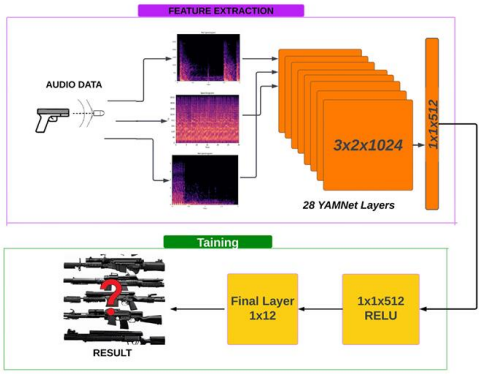
Convolutional Neural Networks (CNNs) have played a key role in medical image analysis, particularly in detecting strokes due to their strong ability to extract meaningful patterns from images. By applying convolutional filters across input data, CNNs can recognize essential visual features such as edges, textures, and structural patterns that signal abnormalities like infarcts or hemorrhages. These networks are structured to progressively learn from basic visual cues in the early layers to more complex patterns in deeper layers, making them highly suitable for identifying subtle medical conditions. Typically, CNN architectures used for stroke detection include multiple convolutional and pooling layers to extract and compress image data, followed by fully connected layers for classification. While CNNs have delivered promising results in this domain, they are not without limitations. One of the primary drawbacks is their dependence on large amounts of labeled data, which can be difficult to obtain in medical scenarios. Limited datasets increase the risk of overfitting, reducing the model's ability to generalize. Moreover, training deep CNNs can be computationally demanding and time-consuming. Another critical issue arises as networks grow deeper: the problem of vanishing gradients, which can hinder effective learning and lead to reduced model performance. These challenges highlight the need for improvements or alternatives in medical imaging techniques.

PROPOSED TECHNIQUE USED OR ALGORITHM USED:

In real-world audio processing applications, MFCC, SVM, and YAMNet are often integrated to create a more accurate and resilient classification system. MFCCs serve as a reliable method for extracting critical features from audio signals by capturing patterns that represent the sound's core characteristics. These extracted features can then be input into an SVM, a machine learning algorithm known for its effectiveness in handling complex, high-dimensional data, to perform classification tasks. Alternatively, MFCCs may be used as inputs for deep learning models like YAMNet, which adds further refinement through its pre-trained neural architecture. YAMNet enhances the system's capability by detecting a wide range of audio events with greater detail. This hybrid approach combines the strengths of traditional machine learning with modern deep learning techniques, leading to improved classification accuracy. The combined use of MFCC, SVM, and YAMNet supports robust audio analysis across applications such as

speech recognition, environmental sound monitoring, and music classification. The complementarity of these tools allows for precise detection in noisy or variable environments. Moreover, the system remains adaptable, capable of generalizing well to different types of audio data. This integration ensures a high-performing, versatile audio classification framework suited for dynamic, real-time scenarios.

SYSTEM ARCHITECTURE:



V.RESULT AND IMPLEMENTATION

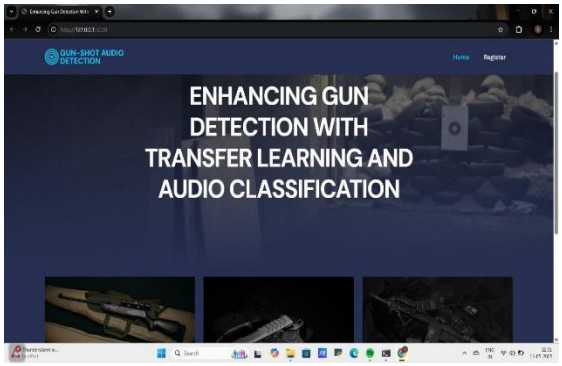


Figure 1: Home Page

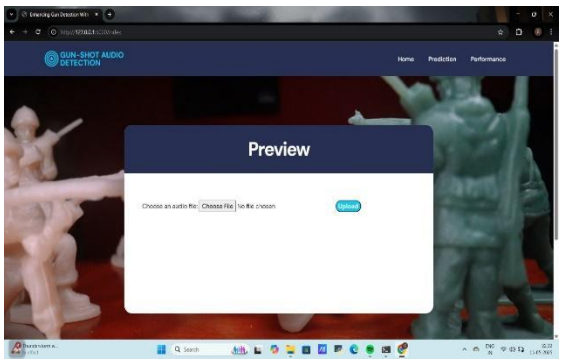


Figure 2: Select a File from this window

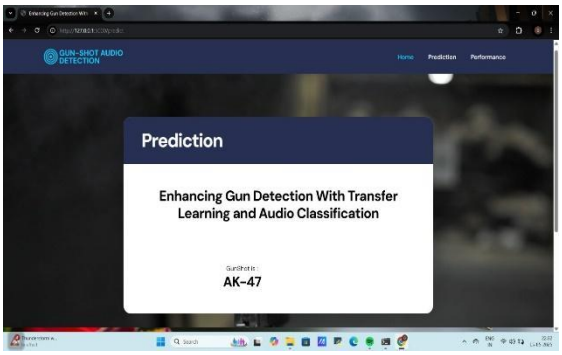


Figure 3: Gun sound detected

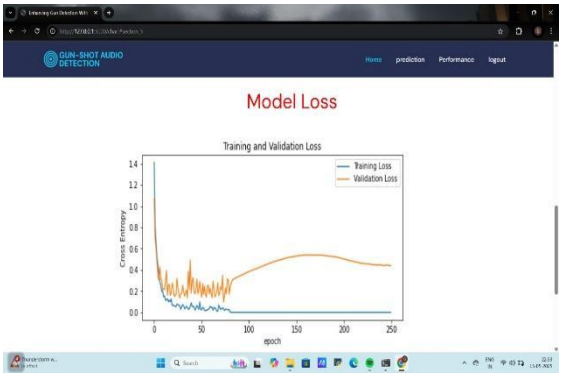


Figure 4: Performance of gun sound

VI. FUTURE ENHANCEMENT:

Future improvements to this audio classification system could focus on integrating advanced deep learning models like Transformer architectures, which are known for effectively capturing complex temporal and contextual patterns in audio data. These models may significantly boost detection precision and contextual understanding. To reduce dependency on large annotated datasets, incorporating unsupervised or semi-supervised learning methods could allow the system to learn from limited or unlabeled data. Enhancing scalability and real-time performance is also essential; optimizing model efficiency through algorithmic upgrades or hardware acceleration would support faster, more responsive processing. This would be especially valuable for time-sensitive applications like surveillance and interactive audio systems.

VII. CONCLUSION:

In summary, the combination of MFCC, SVM, and YAMNet forms a robust and efficient framework for audio classification, particularly in tasks like gunshot detection. This hybrid approach leverages the precise feature extraction of MFCC, the strong classification capabilities of SVM, and the deep learning power of YAMNet to deliver accurate and reliable results. The integration allows the system to perform well across varied and complex audio environments. Future improvements may include incorporating advanced models like Transformers for better pattern recognition and context handling. Employing semi-supervised learning techniques can also help reduce the dependency on large labeled datasets. Additionally, enhancing real-time performance and exploring multi-modal data integration could further increase the system's adaptability and effectiveness in real-world applications.

VIII. REFERENCES:

[1] E. Tsalera, A. Papadakis, and M. Samarakou, ''Comparison of pre-trained CNNsfor audio classification using transfer learning,'' J. Sensor Actuator Netw., vol. 10, no. 4, p. 72, Dec. 2021.

[2] S. Patil and K. Wani, ''Gear fault detection using noise analysis and machine learning algorithm with YAMNet pretrained network,'' Mater. Today, Proc., vol. 72, pp. 1322–1327, 2023.

[3] J. Bajzik et al., ''Independent channel residual convolutional network for gunshot detection,'' Int. J. Adv. Comput. Sci. Appl., vol. 13, no. 4, pp. 950–958, 2022.

[4] R. Baliram Singh, H. Zhuang, and J. K. Pawani, ''Data collection, modeling, and classification for gunshot and gunshot-like audio events: Acase study,'' Sensors, vol. 21, no. 21, p. 7320, Nov. 2021.

[5] A. K. Sharma, G. Aggarwal, S. Bhardwaj, P. Chakrabarti, T. Chakrabarti, J. H. Abawajy, S. Bhattacharyya, R. Mishra, A. Das, and H. Mahdin, ''Classification of Indian classical music with time-series matching deep learning approach,'' IEEE Access, vol. 9, pp. 102041–102052, 2021.