2024 4th International Conference on Intelligent Technologies (CONIT)
Karnataka, India. Jun 21-23, 2024

Enhanced Gunshot Sound Detection using AlexNet and XGBoost from Fourier Spectrograms

M.Yagnasri Priya, Sanjivani P. Shendre, Peeta Basa Pati*
Department of Computer Science and Engineering
Amrita School of Computing, Bengaluru
Amrita Vishwa Vidyapeetham, India
*Corresponding Author: bp_peeta@blr.amrita.edu, ORC ID: 0000-0003-2376-4591

2024 4th International Conference on Intelligent Technologies (CONIT) | 979-8-3503-4990-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/CONIT61985.2024.10626951

***Abstract*—Gunshot detection and classification are pivotal for ensuring public safety, law enforcement, defense, and forensic investigations. This paper presents a comparitive study into the effectiveness of deep learning and traditional machine learning models in classifying gunshot sounds through Short-Term Fourier Transform (STFT) spectograms is presented. The study involves integrating Convolutional Neural Networks (CNNs) with standard machine learning techniques to extract and analyze features, particularly focusing on the penultimate layer of CNNs. By leveraging these features, this paper aims to enhance the system's ability to accurately identify complex sound patterns associated with gunshots. This analysis provides insights into the strengths and weaknesses of deep learning and machine learning approaches in gunshot classification, particularly achieving a perfect accuracy score with AlexNet and XGBoost. These findings pave the way for more effective gunshot detection systems in a variety of operational scenarios, thereby significantly contributing to public safety measures.**

***Index Terms*—Alexnet, Convolutional Neural Networks (CNNs), Gunshot detection, Public safety, Spectograms, XGBoost**

I. Introduction

Since the 1990s, gunshot detection systems have been specially created to detect gunfire, supporting emergency response and law enforcement efforts [1]. Acoustic detection systems are crucial for recognizing and timing a variety of sound occurrences [2]. Sufficient training data continues to be a barrier to the accuracy of gunshot detection systems, leading to mistakes, including false positives, despite breakthroughs in a variety of applications, such as surveillance and environmental monitoring. The results of comparative signal analysis indicate that techniques like short-term Fourier transform (STFT) perform better than wavelets, Smoothed pseudo-Wigner-Ville distribution (SPWVD), and Matching Pursuit (MP), suggesting that there may be room for improvement in terms of accuracy and versatility when it comes to gunshot detection [3].

In gunshot detection, Convolutional Neural Networks (CNNs) and traditional machine learning models are combined to improve feature extraction and classification accuracy. Features extracted from the CNNs final layer, just before the fully connected layer, are used as inputs for another machine learning model. This integration improves the system's ability to recognize complex sound patterns caused by gunshots. This approach maximizes its efficacy by combining the interpretability and generalization strengths of traditional machine learning algorithms with CNNs automatic learning of hierarchical features from raw data. This integrated approach significantly improves the system's accuracy and applicability.

The major contributions of this work include :

- Gunshot audio classification and localization via CNN and deep feature extraction.
- Utilization of ensemble methods with diverse machine learning algorithms for classification.

The proposed research contributes to several key objectives that are aligned with the United Nations' Sustainable Development Goals. It contributes to Goal 16 by promoting peace, justice, and strong institutions through gunshot detection capabilities, Goal 11 by contributing to sustainable cities and communities through urban safety enhancement, and Goal 17 by fostering partnerships and interdisciplinary collaboration to address societal challenges related to violence prevention and public safety. By these SDGs, the proposed project aims to use technology and innovation to address pressing social issues, promote peace and security, and contribute to the well-being and prosperity of communities around the world.

II. Literature Survey

Xia *et al.* [2] developed a Convolutional Recurrent Neural Network (CRNN) for multi-channel audio recognition, improving performance in speech-interference settings by utilizing perceptual and spatial data from multichannel audio. They used deep neural networks for sound event localization, detection, and tracking, resulting in higher accuracy in overlapping sound event recognition. Yang *et al.* [4] used a decentralized technique to extract features in automatic voice recognition with a quantum convolutional neural network (QCNN) and recurrent neural network (RNN), resulting in competitive accuracy and improved privacy protection. Hosseini *et al.* [5] presented state-of-the-art multimodal emotion detection models that integrate auditory, visual, and textual information. However, they emphasized the importance of accurate datasets and tackling modality heterogeneity. Grinstein *et al.* [6] developed an audio style transfer method that captures speech style from a reference signal and applies it to target audio to defend against audio adversarial examples.

Demir *et al.* [7] developed a deep CNN model for ambient sound classification. They used spectrogram pictures and deep meta-model aggregation to get promising results

in sound recognition. This highlights the need for efficiency improvements in traditional sound classification systems. Liu *et al.* [8] used CNNs for food recognition and found that transfer learning and data augmentation strategies improved classification accuracy. Zhou *et al.* [9] improved acoustic scene classification accuracy with bone-conducted sound data by incorporating prosodic characteristics and transfer learning, especially with limited datasets.

Kabir *et al.* [10] propose a novel gunshot detection method combines adapted standard sound features with hand-crafted novel features, achieving high accuracy. Utilizing unique mel-frequency cepstral coefficients and features from significant points in the raw waveform, it employs a simple neural network for remarkable accuracy. However, there's a gap in evaluating the overall accuracy of a prior study using an anti-poaching system based on elephant collars.

Tsalera *et al.* [11] evaluated pre-trained CNN models for audio classification and proposed a joint training framework for audio-visual scene classification. Achieved a high accuracy on TAU Urban Audio-Visual Scenes 2021 dataset, surpassing traditional methods. Emphasized further research on model effectiveness across scenarios and the challenge of interpreting extracted features.

Nijhawan *et al.* [12] present a multimodal model for gunshot detection that combines audio and visual analysis, emphasizing the importance of improving component synergy, addressing ethical and legal concerns, and optimizing the model for real-world use. Katsis *et al.* [13] address the challenge of monitoring unsustainable hunting in tropical forests using acoustic detection but emphasize the importance of additional validation in various tropical forest environments to assess the robustness of their proposed approach.

Furthermore, Roshan *et al.* [14] investigate the vulnerability of logic locking schemes in ICs to machine learning-based attacks, proposing methods and discussing trade-offs for security enhancement, while raising critical questions for future research on scalability and key generation optimization under specific constraints. Ramya *et al.* [15] propose a weapon detection system based on SSD methodology, but the study lacks comprehensive dataset details and real-world test scenarios.

Tanmayi *et al.* [16] found that CNNs like AlexNet can accurately detect fires in real time, leading to improved fire safety. Furthermore, energy-efficient models like SqueezeNet achieve comparable accuracy, highlighting the trade-off between detection accuracy and computational efficiency in fire detection systems.

Kihal and Hamza [17] developed a powerful spam filtering system that combines visual, textual, and audio-deep characteristics with the Random Forest algorithm, resulting in excellent accuracy in multimedia message analysis. Zhao and Liu [18] underlined the effectiveness of Convolutional Neural Networks (CNNs) for handwritten digit recognition. They demonstrated better validation accuracy and reduced loss through parameter modifications. Barbhuiya *et al.* [19] used CNNs and bidirectional long short-term memory networks for sign language classification, attaining good accuracy but confronting issues with model simplicity and training efficiency.

Raghuram *et al.* [20] present a simulation modeling and genetic algorithm-based approach to warehouse management optimization. Adapting the model to dynamic environments and integrating real-world data are examples of open problems. Sridhar *et al.* [21] present an optimized GTCNN for fire detection without explicitly discussing its limitations. Vignesh Raj *et al.* [22] use the Single Shot Detector (SSD) algorithm for helmet detection but do not provide information on dataset diversity or real-world applicability.

The reviewed studies highlight advancements in audio signal processing and machine learning, with applications ranging from gunshot detection to emotion recognition. There are still areas for improvement, including real-world validation, feature extraction interpretation, dataset accuracy, and comparative analysis. The lack of comprehensive evaluations emphasizes the importance of strong validation frameworks.

III. Methodology

A. Data Description

The dataset for this study was obtained from Kaggle and consisted of audio recordings divided into gunshot and non-gunshot categories. Specifically, the dataset contains 851 gunshot audio samples and 2000 non-gunshot audio samples. The gunshot audios have been taken from the dataset on Kaggle [23]. The non gunshot audios have taken from the dataset on Kaggle [24].

B. Proposed Methodology

The proposed methodology combines acoustic signal processing and machine learning approaches to develop a gunshot detection system. A dataset consisting of 2000 non-gunshot audio samples and 850 gunshot audio samples allows for robust model training and evaluation. Python is the primary programming language used to complete the process. The Short-Term Fourier Transform (STFT) is first used to preprocess the data, extracting relevant features and converting them to spectrograms. These spectrograms are classified after being normalized for wavelength uniformity. The flowchart for the proposed methodology is shown in Fig.1.

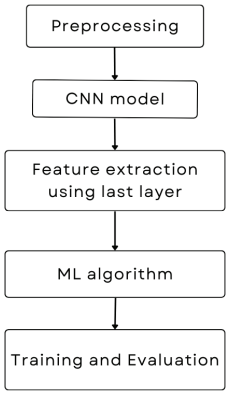


Fig. 1: Flowchart of proposed methodology

The preprocessed data is then fed into a Convolutional Neural Networks (CNNs) module to extract the features. The CNN architecture captures significant features at each layer, making it easier to represent the complex audio patterns that occur during gunshots. Feature vectors from the final layer, preceding the fully connected layer, are then extracted for further processing. These vectors are fed into a machine learning (ML) algorithm for classification after feature extraction. Based on the labeled dataset, the machine learning system can distinguish between audio samples containing gunshots and those without. This classification is investigated using several conventional machine learning techniques, including Random Forest, Naive Bayes, XGBoost, CatBoost, AdaBoost, KNN, Decision Trees, and Support Vector Machines (SVM).

C. Preprocessing

Preprocessing involves loading the audio dataset, which contains both gunshot and non-gunshot sounds, and creating TensorFlow datasets with corresponding labels. To address the class imbalance, the minority class is oversampled to ensure balance. Then, a preprocessing function extracts features from the audio files. This function resamples the audio to 16kHz and normalizes the length. The spectrogram, which depicts the magnitude of frequencies over time, is then computed using the Short-Time Fourier Transform (STFT). The spectrogram is then transformed into absolute values and expanded to include a channel dimension. This process ensures that each audio sample is represented as a consistent input shape to convolutional neural networks. The processed datasets are then cached, shuffled, batched, and prefetched to ensure consistent input shapes suitable for later classification.

D. CNN Models

1) Custom CNN Model: The custom CNN architecture is optimized for audio classification tasks, with a focus on extracting features from spectral data. It starts with two convolutional layers, each with 32 filters of size (3, 3), activated by the Rectified Linear Unit (ReLU) function. To avoid overfitting and improve generalization, dropout layers with a rate of 0.5 are strategically placed after the first convolutional layer and the flattening layer. The model then progresses to a flattened layer, which reshapes the output into a single vector. This is followed by a dense layer of 16 ReLU-activated neurons, which further abstracts the features. Another dropout layer with a rate of 0.5 is used for regularisation. The architecture sequentially integrates convolutional layers for feature extraction, dropout layers for regularisation, and dense layers for additional abstraction. This tailored design effectively captures spatial features from audio spectrograms, providing a solid foundation for subsequent classification tasks with traditional machine learning models.

2) AlexNet Model: The AlexNet model utilized in the methodology serves as a feature extractor for audio classification tasks. The architecture consists of several convolutional and max-pooling layers, followed by fully connected layers. The model's architecture includes two convolutional layers with large receptive fields to help extract features from input audio spectrograms. Subsequent max-pooling layers aid in capturing key features while minimizing spatial dimensions. Additional convolutional layers improve feature extraction, whereas fully connected layers enable abstraction and classification. The model's trainable parameters are then frozen to preserve previously learned feature representations, yielding consistent and meaningful feature extraction. Using this feature extractor, the model efficiently converts audio spectrograms into high-level feature representations, making classification tasks easier for traditional machine learning models.

3) VGG16 Model: The VGG16 model is used for feature extraction for the audio classification tasks. The VGG16 architecture, pre-trained on ImageNet for image classification, consists of multiple convolutional layers with small receptive fields that are interleaved with max-pooling layers before being fully connected. Using transfer learning, the pre-trained VGG16 model is adapted for audio data by changing its input shape and adding a new classification layer tailored to the task at hand. The model's convolutional layers efficiently extract hierarchical feature representations from the input audio spectrograms, while the dense layers aid in classification. The trainable parameters are fine-tuned to extract meaningful audio features, allowing for precise classification using conventional machine learning models.

E. Feature Extraction

Feature extraction involves utilizing custom Convolutional Neural Network (CNN) architectures to extract features from audio spectrograms. The CNN architecture is based on image classification architectures. A pre-trained VGG16 model is also used for feature extraction, leveraging its ImageNet-learned representations, while the AlexNet architecture is used for audio classification, capitalizing on its convolutional layers despite the lack of pre-trained weights. Despite their different approaches, all three methods effectively extract relevant features from spectrogram data, thereby improving their performance in audio classification tasks.

F. Classifiers

Gunshot detection systems use a variety of classification algorithms, including K-nearest neighbors (KNN), Support Vector Machines (SVM), Random Forest (RF), Naive Bayes (NB), MultiNomial NB (MNB), Decision Trees (DT), boosting algorithms such as XG Boost (XGB), CatBoost (CB), and AdaBoost (AB), as well as the Voting Classifier (VC). These algorithms efficiently analyse audio data and distinguish gunshots from other auditory events. KNN finds the most similar instances in the training data to the input data point, whereas SVM finds the best hyperplane for class separation. Random Forest generates decision trees and handles noise effectively, whereas Naive Bayes classifiers are computationally efficient. Decision trees make interpretation easier while boosting algorithms to improve accuracy iteratively. The Voting Classifier combines multiple classifiers to improve performance.

Together, these algorithms help to develop accurate gunshot detection systems, which are critical to public safety. The classifiers are rated according to their test and training accuracies, precision, recall, and F1-Score.

The hyperparameters were chosen based on their relevance to the specific algorithms and potential impact on classification performance. Furthermore, we used RandomizedSearchCV in scikit-learn to experiment and fine-tune the hyperparameters for each classifier. This approach produced robust and effective model training, improving the overall performance of the classifiers. Table I shows the hyperparameters for different classifiers.

TABLE I: Hyperparameters for Different Classifiers

Classifier	Hyperparameters
KNN	n_neighbors: 2 & 3 weights: uniform p: 2
SVM	kernel: rbf C: 1.0 gamma: scale
RF	n_estimators: 100 max_depth: None min_samples_split: 2 min_samples_leaf: 1
NB	No hyperparameters to tune.
MNB	alpha : 1.0 Fit_prior: True Class_prior: None
DT	criterion: gini max_depth: None min_samples_split: 2 max_features: None
CB	learning_rate: 0.03 depth: 6 model_size_reg: 0.5 verbose: False
XGB	learning_rate: 0.1 max$_depth$: 6 use_label_encoder: False verbosity: 0
AB	n_estimators: 50 learning_rate: 1.0 algorithm: SAMME.R

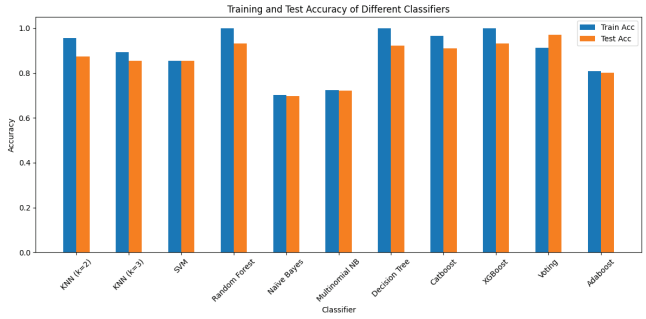
IV. Results

The results presented show the performance metrics of various classifiers used in a gunshot detection system that employs both traditional machine learning approaches and deep learning models such as CNN, AlexNet, and VGG16 for feature extraction and classification. The classifiers are evaluated based on metrics such as training accuracy, test accuracy, precision, recall, and F1 score.

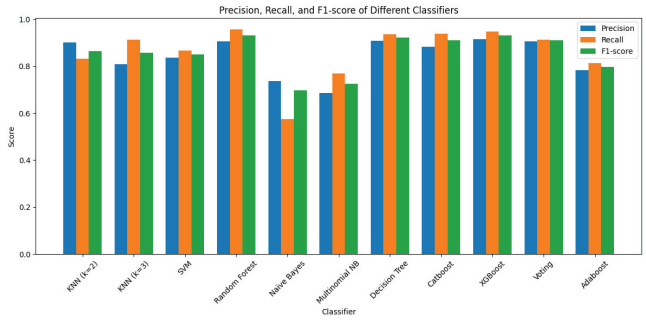
Several key parameters were used to evaluate the performance of the gunshot detection system. Accuracy was a basic measure of the model's overall correctness, but its usefulness decreased in datasets with imbalanced classes. Precision emerged as an important metric, indicating the proportion of correctly identified gunshot sounds among all cases classified as gunshots. Its emphasis on false alarm minimization is critical in real-world scenarios to reduce false positive detections. In contrast, recall was critical in determining the system's ability to correctly detect actual gunshot sounds. Maximizing recall is critical in situations where false negatives can have serious consequences. However, the F1-Score emerged as the most important evaluation parameter. It provided a thorough evaluation that balanced precision and recall, which was critical for improving the system's performance. In the context of gunshot detection, where false positives and false negatives have serious consequences, maximizing the F1-Score ensures a delicate balance between reducing false alarms and missed detections, thereby improving the system's reliability and effectiveness.

The results from the CNN model show that different classifiers perform at varying levels. For example, KNN with k=2 had a high training accuracy (0.9553) but a slightly lower test accuracy (0.875), indicating possible overfitting. RF and XGB achieved perfect training accuracy (1), indicating strong learning from the training data. However, their test accuracies (0.9313 and 0.9333, respectively) are slightly lower, indicating some generalization issues. Naïve Bayes classifiers had lower accuracies than others, including MNB and NB.

The results obtained with the AlexNet model show a significant improvement in performance across most classifiers when compared to CNN. Several classifiers achieved perfect or near-perfect training and test accuracies, indicating that the AlexNet model's extracted features are more generalizable and discriminative. Notably, KNN with k=2, RF, DT, CB, XGB, and the ensemble method Voting all received perfect scores across all metrics, demonstrating the utility of AlexNet features for classification tasks.

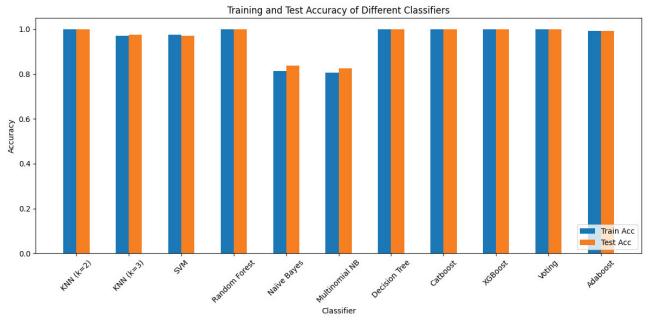


(a)

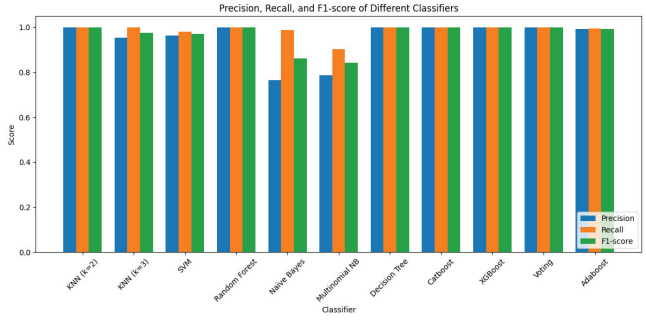


(b)

Fig. 2: (a) Accuracies of Classifiers using CNN, (b) Precision, Recall and F1-Score of Classifiers using CNN.

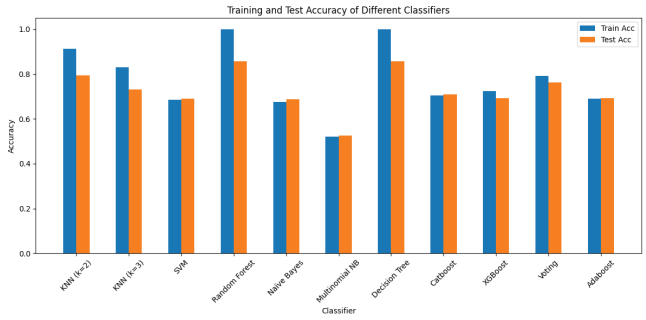


(a)

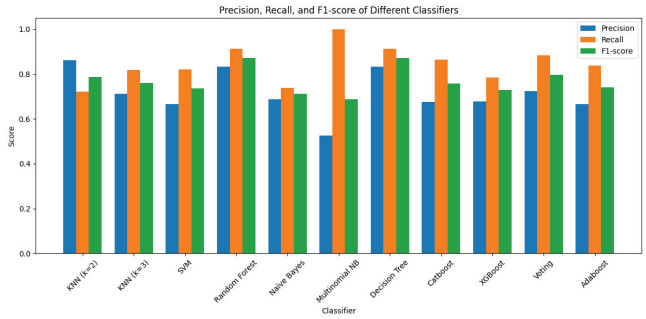


(b)

Fig. 3: (a) Accuracies of Classifiers usinf AlexNet, (b)Precision, Recall and F1-Score of Classifiers suing AlexNet.



(a)



(b)

Fig. 4: (a) Accuracies of Classifiers using VGG16, (b) Precision, Recall and F1-Score of Classifiers using VGG16.

When comparing the results of the features extracted with VGG16 to those of AlexNet and a custom CNN grayscale approach, it is clear that VGG16 does not consistently outperform the other two methods across all metrics. While VGG16 performs well in terms of accuracy, precision, recall, and F1-score, it may not always produce the highest results when compared to other approaches. In some cases, AlexNet or the custom CNN grayscale approach may achieve greater accuracy, precision, or recall. As a result, the choice between these methods should be based on the classification task's specific requirements and objectives, as well as additional experimentation to determine the best approach for achieving the desired performance.

Finally, the comparison table of all deep learning models shows information about the performance of only CNN, AlexNet, and pre-trained VGG16 models. Notably, CNN alone achieved perfect training and test accuracies, demonstrating its efficacy as a stand-alone model for this task. However, both AlexNet and VGG16 models, which are trained on large datasets, perform admirably across all metrics, indicating their potential in classification tasks with minimal fine-tuning.

The obtained results of the proposed method have been compared with the methods used in previous studies. The proposed method outperforms previous studies in automatically detecting gunshots, with 100% accuracy, precision, and recall rates. This method outperforms previous studies with accuracies ranging from 90% to 97.3%, particularly when compared to pre-trained CNNs for audio classification. The approach, which employs a combination of signal processing and advanced neural networks, ensures robust detection across a wide range of scenarios, outperforming machine learning and acoustic feature extraction techniques on their own. The best results were obtained from the combination of AlexNet and XGBoost Algorithm.

TABLE II: Best Performance of custom CNN, AlexNet & VGG16 Models

	CNN	AlexNet	VGG16
Train Acc	0.9995	0.9725	0.9950
Test Acc	1.0000	0.9958	1.0000
Precision	1.0000	0.9959	0.5078
Recall	1.0000	0.9959	0.5536
F1-Score	1.0000	0.9959	0.5297

Fig. 2 depicts the accuracies and other performance metrics using the CNN Model for feature extraction. Fig. 3 depicts the accuracies and other performance metrics using the AlexNet Model for feature extraction. Fig. 4 depicts the accuracies and other performance metrics using the VGG16 Model for feature extraction. Table II shows the best performance of the standalone Deep Learning Models used for feature extraction and classification. Table III shows the comparison of the results obtained from different methodologies used for the classification of audios.

TABLE III: Comparison of results of different methodologies used for audio detection

Work reference	Methodology	Accuracy (%)	Precision (%)	Recall (%)
[12]	• Extracting MFCCs and Mel-spectrograms from audio • Fine-tuning a pre-trained vision transformer model (L32) with Adam optimization.	90	NA	NA
[13]	• Trained ResNet18 CNN on balanced Belize datasets • Optimized decision thresholds for maximal recall • Assessed precision on separate test data from another site	NA	85	95
[10]	• Extracting features: MFCC, GTCC, LPC, SC array • Employing bagged tree ensemble and SVM for detection • Utilizing GCC-PHAT transform for localization with a fivemicrophone	97.3	100	97.8
[11]	• Using transfer learning with GoogLeNet, SqueezeNet, ShuffleNet, VGGish, and YAMNet • Retraining on UrbanSound8K, ESC-10, and Air Compressor datasets	96.7	NA	NA
Proposed Work	• Preprocessing via Short-Term Fourier Transform • Extracting features using CNN, AlexNet, and VGG16	100	100	100

V. Discussions

This paper presents a comprehensive comparison of deep learning and traditional machine learning models for classifying gunshot sounds. The study used Convolutional Neural Networks (CNNs) and standard machine learning techniques to improve the accuracy of detecting complex sound patterns associated with gunshots. The findings demonstrate that feature extraction and classification methodologies can be successfully implemented, with a focus on CNN's penultimate layer. This method yielded outstanding performance metrics, including nearly perfect accuracy, precision, recall, and F1 score, demonstrating the system's effectiveness in gunshot detection.

The study's findings indicate that both deep learning models, such as CNNs, and traditional machine learning algorithms are capable of extracting meaningful features and accurately classifying gunshot sounds. The CNN model, as well as pre-trained architectures like AlexNet and VGG16, excelled at extracting relevant features from spectral data. These features enabled precise classification, resulting in excellent performance across a wide range of metrics. Notably, the CNN model achieved perfect training and test accuracies, demonstrating its viability as a standalone model for gunshot detection.

Additionally, the comparative analysis shows that the proposed methodology, which uses AlexNet for its superior performance due to its simple architecture and XGBoost for its robustness, outperforms previous approaches [16]. It outperforms previous methodologies with perfect accuracy, precision, and recall, ranging from 90% to 97.3% accuracy, indicating a significant improvement in gunshot detection efficacy. This significant improvement demonstrates the effectiveness and utility of combining advanced signal processing techniques with cutting-edge deep learning and machine learning algorithms.

Furthermore, the comparative analysis highlights the complementary nature of deep learning and traditional machine learning approaches for dealing with complex classification tasks like gunshot detection. While deep learning models excel at automatically extracting features from raw data, traditional machine learning algorithms offer interpretability and generalization benefits that boost overall system performance. By combining the strengths of both approaches, the integrated methodology presented in this study improves the system's ability to detect gunshots in real-world scenarios, making a significant contribution to public safety.

VI. Conclusion

The results of both standalone CNN models and approaches that incorporate features from the CNN's final layer show promising advances in gunshot detection systems. When it comes to identifying gunshot events from audio spectra, both methods achieve high levels of accuracy, precision, recall, and F1-scores. Feature extraction, particularly when using CNN architectures, has been shown to improve classification performance, highlighting its potential in this field.

Future advancements in gunshot detection systems may focus on improving accuracy by combining advanced signal processing and deep learning techniques. This could include improving algorithms for distinguishing gunshot sounds from background noise and implementing them in real-time across a variety of scenarios. Furthermore, integrating gunshot detection with other sensor modalities shows promise for developing autonomous platforms that could transform emergency response strategies.

Acknowledgement

We thank Amrita Vishwa Vidyapeetham for the infrastructure and support received during this work. We gratefully acknowledge the assistance of ChatGPT, Gemini, QuillBot, and Grammarly in refining this paper through paraphrasing and editing.

REFERENCES

[1] J. H. Ratcliffe, M. Lattanzio, G. Kikuchi, and K. Thomas, "A partially randomized field experiment on the effect of an acoustic gunshot detection system on police incident reports," *Journal of Experimental Criminology*, vol. 15, pp. 67–76, 2019.

[2] X. Xia, R. Togneri, F. Sohel, Y. Zhao, and D. Huang, "A survey: neural network-based deep learning for acoustic event detection," *Circuits, Systems, and Signal Processing*, vol. 38, pp. 3433–3453, 2019.

[3] R. Álvarez, E. Borbor, and F. Grijalva, "Comparison of methods for signal analysis in the time-frequency domain," in *2019 IEEE Fourth Ecuador Technical Chapters Meeting (ETCM)*, pp. 1–6, IEEE, 2019.

[4] C.-H. H. Yang, J. Qi, S. Y.-C. Chen, P.-Y. Chen, S. M. Siniscalchi, X. Ma, and C.-H. Lee, "Decentralizing feature extraction with quantum convolutional neural network for automatic speech recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6523–6527, IEEE, 2021.

[5] S. S. Hosseini, M. R. Yamaghani, and S. Poorzaker Arabani, "Multimodal modelling of human emotion using sound, image and text fusion," *Signal, Image and Video Processing*, vol. 18, no. 1, pp. 71–79, 2024.

[6] E. Grinstein, N. Q. Duong, A. Ozerov, and P. Pérez, "Audio style transfer," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 586–590, IEEE, 2018.

[7] F. Demir, D. A. Abdullah, and A. Sengur, "A new deep cnn model for environmental sound classification," *IEEE Access*, vol. 8, pp. 66529–66537, 2020.

[8] Y. Liu, H. Pu, and D.-W. Sun, "Efficient extraction of deep image features using convolutional neural network (cnn) for applications in detecting and analysing complex food matrices," *Trends in Food Science & Technology*, vol. 113, pp. 193–204, 2021.

[9] H. Zhou, X. Bai, and J. Du, "An investigation of transfer learning mechanism for acoustic scene classification," in *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 404–408, IEEE, 2018.

[10] M. S. Kabir, J. Mir, C. Rascon, M. L. U. R. Shahid, and F. Shaukat, "Machine learning inspired efficient acoustic gunshot detection and localization system," *University of Wah Journal of Computer Science*, vol. 3, no. 1, 2021.

[11] E. Tsalera, A. Papadakis, and M. Samarakou, "Comparison of pre-trained cnns for audio classification using transfer learning," *Journal of Sensor and Actuator Networks*, vol. 10, no. 4, p. 72, 2021.

[12] R. Nijhawan, S. A. Ansari, S. Kumar, F. Alassery, and S. M. El-Kenawy, "Gun identification from gunshot audios for secure public places using transformer learning," *Scientific reports*, vol. 12, no. 1, p. 13300, 2022.

[13] L. K. Katsis, A. P. Hill, E. Pina-Covarrubias, P. Prince, A. Rogers, C. P. Doncaster, and J. L. Snaddon, "Automated detection of gunshots in tropical forests using convolutional neural networks," *Ecological Indicators*, vol. 141, p. 109128, 2022.

[14] E. Roshan, V. Nishanth, and N. Mohankumar, "Developing attack resilience in y86 processor using logic locking," in *2021 IEEE 12th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pp. 0737–0743, IEEE, 2021.

[15] R. Ramya, C. Lasya, N. M. Sai, and S. Paneerselvam, "An intelligent surveillance system for weapon detection based on efficientdet algorithm," in *2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA)*, pp. 453–459, IEEE, 2023.

[16] D. Tanmayi, E. Udaya, and P. B. Pati, "Detecting fire in color images using convolutional neural network architectures," in *2023 IEEE 8th International Conference for Convergence in Technology (I2CT)*, pp. 1–6, IEEE, 2023.

[17] M. Kihal and L. Hamza, "Robust multimedia spam filtering based on visual, textual, and audio deep features and random forest," *Multimedia Tools and Applications*, vol. 82, no. 26, pp. 40819–40837, 2023.

[18] H.-h. Zhao and H. Liu, "Multiple classifiers fusion and cnn feature extraction for handwritten digits recognition," *Granular Computing*, vol. 5, no. 3, pp. 411–418, 2020.

[19] A. A. Barbhuiya, R. K. Karsh, and R. Jain, "Cnn based feature extraction and classification for sign language," *Multimedia Tools and Applications*, vol. 80, no. 2, pp. 3051–3069, 2021.

[20] P. Raghuram, S. A. Jain, and S. Mazumder, "Order picking and storage optimisation in a warehouse using agent-based modelling," *International Journal of Operational Research*, vol. 48, no. 1, pp. 47–62, 2023.

[21] P. Sridhar, S. K. Thangavel, L. Parameswaran, and V. R. M. Oruganti, "Fire sensor and surveillance camera-based gtcnn for fire detection system," *IEEE Sensors Journal*, vol. 23, no. 7, pp. 7626–7633, 2023.

[22] A. V. Raj, N. Manohar, and G. Dhyanjith, "Helmet detection using single shot detector (ssd)," in *2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pp. 1241–1244, IEEE, 2021.

[23] T. Tuncer, S. DOGAN, E. Akbal, and E. Aydemir, "An automated gunshot audio classification method based on finger pattern feature generator and iterative relieff feature selector," *Adıyaman Üniversitesi Mühendislik Bilimleri Dergisi*, vol. 8, no. 14, pp. 225–243, 2021.

[24] M. Moreaux, "Environmental sound classification 50." https://www.kaggle.com/datasets/mmoreaux/environmental-sound-classification-50?resource=download.

APPENDIX

Table IV shows the performance of all the classifiers using CNN Model for feature extraction. Table V shows the performance of all the classifiers using AlexNet Model for feature extraction. Table VI shows the performance of all the classifiers using VGG16 Model for feature extraction.

TABLE IV: Performance of various classifiers with the features obtained from custom CNN.

	Train Acc	Test Acc	Precision	Recall	F1-Score
KNN(k=2)	0.9553	0.8750	0.9009	0.8304	0.8643
KNN(k=3)	0.8933	0.8542	0.8076	0.9130	0.8571
SVM	0.8542	0.8548	0.8361	0.8652	0.8504
RF	1.0000	0.9313	0.9053	0.9565	0.9302
NB	0.7028	0.6979	0.7374	0.5739	0.6979
MNB	0.7231	0.7208	0.6860	0.7696	0.7254
DT	1.0000	0.9229	0.9072	0.9348	0.9208
CB	0.9653	0.9104	0.8816	0.9391	0.9095
XGB	1.0000	0.9333	0.9159	0.9478	0.9316
Voting	0.9125	0.9721	0.9052	0.9131	0.9091
AB	0.8076	0.8021	0.7824	0.8131	0.7974

TABLE V: Performance of various classifiers with the features obtained from AlexNet Model

	Train Acc	Test Acc	Precision	Recall	F1-Score
KNN(k=2)	1.0000	1.0000	1.0000	1.0000	1.0000
KNN(k=3)	0.9698	0.9750	0.9536	1.0000	0.9762
SVM	0.9755	0.9708	0.9641	0.9797	0.9718
RF	1.0000	1.0000	1.0000	1.0000	1.0000
NB	0.8145	0.8375	0.7648	0.9878	0.8621
MNB	0.8058	0.8250	0.7879	0.9028	0.8415
DT	1.0000	1.0000	1.0000	1.0000	1.0000
CB	1.0000	1.0000	1.0000	1.0000	1.0000
XGB	1.0000	1.0000	1.0000	1.0000	1.0000
Voting	1.0000	1.0000	1.0000	1.0000	1.0000
AB	0.9935	0.9937	0.9919	0.9959	0.9939

TABLE VI: Performance of various classifiers with the features obtained from VGG16 Model

	Train Acc	Test Acc	Precision	Recall	F1-Score
KNN(k=2)	0.9125	0.7937	0.8626	0.7222	0.7862
KNN(k=3)	0.8317	0.7313	0.7128	0.8175	0.7616
SVM	0.6856	0.6896	0.6656	0.8214	0.7353
RF	1.0000	0.8583	0.8333	0.9127	0.8712
NB	0.6769	0.6875	0.6889	0.7381	0.7126
MNB	0.5202	0.5250	0.5250	1.0000	0.6885
DT	1.0000	0.8583	0.8333	0.9127	0.8712
CB	0.7057	0.7104	0.6749	0.8650	0.7583
XGB	0.7245	0.6916	0.6781	0.7857	0.7279
Voting	0.7913	0.7625	0.7240	0.8849	0.7964
AB	0.6904	0.6938	0.6656	0.8373	0.7417