Multimodal Deep Learning for Audio Classification in Diverse Listening Environments

[1]Sunilkumar M. Hattaraki, [2]Shankarayya G. Kambalimath

[1]Department of Electronics and Communication Engineering,
B.L.D.E.A's V. P. Dr. P. G. Halakatti College of Engineering and Technology,
Vijayapura-586103, Karnataka, India.
Visvesveraya Technological University, Belagavi-590018.
Karnataka, India.

[2]Department of Electronics and Communication Engineering,
Basaveshwar Engineering College,
Bagalkote, Karnataka, India.
Visvesveraya Technological University, Belagavi-590018.
Karnataka, India.

Abstract

This study aims to enhance audio classification in diverse environments by leveraging a multimodal deep learning model. The primary objective is to address the limitations of existing models that often rely on a single type of audio feature, which may not fully capture the complexities of different audio environments. Our approach combines Mel-Frequency Cepstral Coefficients (MFCC) and spectrogram features, providing a more comprehensive representation of audio data. The process involves loading audio data, extracting MFCC and spectrogram features, and splitting the dataset into training and testing sets. Each feature type is processed through a dedicated Convolutional Neural Network (CNN) branch: one for MFCCs and another for spectrograms. The outputs from these branches are then concatenated and passed through fully connected dense layers to refine the learned features, capturing complex interactions between the two feature sets. The final classification results are produced using a SoftMax activation function, predicting the most likely audio environment. Our multimodal approach significantly improves classification accuracy, achieving a training accuracy of 99.96% and a test accuracy of 97.23%. This research not only addresses the gap in utilizing multiple audio features for classification but also has practical applications in various societal contexts, such as environmental monitoring, security systems, and smart city infrastructures, where accurate audio classification can lead to enhanced safety and efficiency.

Keywords: Audio Classification, Multimodal Deep Learning, Convolutional Neural Networks, Mel-Frequency Cepstral Coefficients, Mel Spectrogram, and Environmental Sound Classification.

1. Introduction

In recent years, the field of audio classification has experienced significant advancements, primarily driven by the rapid development of deep learning technologies. Audio classification involves identifying and categorizing audio signals into predefined classes, such as speech, music, and environmental sounds. This capability has a wide range of applications, including speech recognition, music genre classification, environmental sound recognition, and audio-based monitoring systems. Accurate classification of audio signals is crucial across various domains, enhancing user experience in consumer electronics and ensuring safety in security systems.

Traditional audio classification methods often rely on a single type of audio feature, such as Mel-Frequency Cepstral Coefficients (MFCCs), which capture the short-term power spectrum of sound. While these features have proven effective in many scenarios, they may not fully capture the complexities and variations present in different audio environments. For instance, environmental sounds can vary significantly based on factors such as background noise, recording conditions, and the presence of overlapping sounds. This variability necessitates a more robust approach to audio feature extraction and classification.

The primary motivation behind this study is to address the limitations of existing audio classification models that rely solely on single-feature types. By incorporating multiple types of audio features, we aim to create a more comprehensive representation of audio data, which can lead to improved classification performance. The integration of multimodal features, specifically MFCCs and spectrograms, provides a richer and more diverse set of information that can better capture the nuances of different audio environments.

The choice of MFCCs and spectrograms is driven by their complementary nature. MFCCs are widely used in speech and audio processing due to their ability to represent the power spectrum of sound in a compact form. Spectrograms, on the other hand, provide a visual representation of the frequency spectrum over time, capturing both temporal and spectral characteristics of the audio signal. By combining these features, we can leverage their strengths and mitigate their individual weaknesses.

The main objectives of this study are as follows:

- To develop a multimodal deep learning model for audio classification: This model will integrate MFCC and spectrogram features to provide a more comprehensive representation of audio data.

- To enhance classification accuracy in diverse audio environments: By utilizing a multimodal approach, we aim to improve the robustness and accuracy of audio classification, especially in challenging and variable conditions.
- To demonstrate the practical applications of the proposed model: We will highlight how the enhanced classification performance can be applied in real-world scenarios, such as environmental monitoring, security systems, and smart city infrastructures.

The proposed multimodal deep learning model follows a systematic approach to audio classification:

- Data Acquisition: Audio data is collected from diverse environments, encompassing a wide range of sound sources and conditions.
- Feature Extraction: Both MFCC and spectrogram features are extracted from the audio data. MFCCs capture the short-term power spectrum, while spectrograms provide a time-frequency representation of the audio signal.
- Data Splitting: The dataset is divided into training and testing sets to evaluate the model's performance.
- Model Architecture: The multimodal model consists of two dedicated Convolutional Neural Network (CNN) branches: one for processing MFCC features and another for spectrogram features. The outputs from these branches are concatenated and passed through fully connected dense layers to refine the learned features.
- Classification: The final classification is performed using a SoftMax activation function, which predicts the most likely audio environment based on the combined features.
- Performance Evaluation: The model's performance is evaluated based on accuracy metrics for both the training and testing sets. The results demonstrate the effectiveness of the multimodal approach in improving classification accuracy.

Significance and Contributions

This research makes several significant contributions to the field of audio classification:

- Multimodal Feature Integration: By combining MFCC and spectrogram features, we provide a more comprehensive representation of audio data, leading to improved classification performance.

- Robustness in Diverse Environments: The proposed model is designed to handle the variability and complexity of different audio environments, making it more robust and reliable.

- Practical Applications: The enhanced classification accuracy has practical implications for various societal contexts, such as environmental monitoring, security systems, and smart city infrastructures. Accurate audio classification can lead to improved safety, efficiency, and user experience in these applications.

- Benchmark Performance: The model achieves a training accuracy of 99.96% and a test accuracy of 97.23%, setting a new benchmark for audio classification in diverse environments.

2. Related Work

Jabeen et al. [1] examine a range of modalities such as images, videos, text, audio, body gestures, facial expressions, and physiological signals. Their study offers an in-depth analysis of both baseline approaches and recent advancements in multimodal deep learning from 2017 to 2021. They propose a comprehensive taxonomy of multimodal deep learning methods and discuss various applications in detail. Additionally, the paper highlights key challenges within each domain and suggests potential future research directions.

Kakub et al. [2] explore trends and challenges in bimodal Speech Emotion Recognition (SER), focusing on natural environment deployment. They introduce the DBMER model, which integrates CNNs, RNNs, and multi-head attention mechanisms, and identify optimal acoustic feature combinations and the importance of attention mechanisms. The study underscores the vital role of attention mechanisms in bimodal dyadic SER systems. Despite the advantages of combining these deep learning techniques, challenges such as limited datasets, difficulties in data acquisition, and issues in cross-corpus and multilingual studies persist. Their experiments demonstrate that combining these techniques with multi-level fusion approaches results in more accurate and robust outcomes.

Jeon et al. [3] introduce an AVSR model that mimics human dialogue recognition and remains robust in noisy environments. It transforms word embeddings and log-Mel spectrograms into feature vectors using a dense spatial-temporal CNN, enhancing auditory and visual recognition. Tested in nine noisy environments, the model achieves a 1.711% error rate with a three-feature multi-fusion method, compared to the general rate of 3.939%, showcasing its effectiveness and stability.

Utebayeva et al. [4] explored a deep learning method using Gated Recurrent Unit (GRU) to classify drone sounds, particularly at varying distances. The study aimed to determine if the sound classification method could recognize drones flying at different distances. In the experiments, drones were launched at five-meter intervals from ground level. The results showed that drone sounds could be accurately recognized from distances of 10 meters or more, with an average accuracy of 94-98 percent. This system could be integrated into complex recognition systems as a functional component of bimodal and multi-modal systems.

Kushwaha et al. [5] propose a multimodal prototypical approach that uses local audio-text embeddings to enhance the relevance of answers to audio queries, improving sound detection adaptability in diverse environments. The method first uses text to query a community of audio embeddings, selecting group centroids as prototypes. Then, it compares unseen audio to these prototypes for classification. Multiple ablation studies were conducted to assess the impact of embedding models and prompts. This unsupervised approach outperforms the zero-shot state-of-the-art by an average of 12% across three sound recognition benchmarks.

Shaqra et al. [6] introduce a multimodal emotion detection system using an Arabic dataset. Their model combines audio and visual data, showing that gender identification improves emotion recognition. The multimodal system achieves 75% accuracy for emotion detection and 60.11% for emotion recognition, outperforming individual audio (70% for anger) and visual (56.2% for surprise) models, and is the first to focus on Arabic content.

Ding et al. [7] summarize resources for Acoustic Scene Classification (ASC) research and analyze ASC tasks from DCASE challenges. They discuss current algorithm limitations and future challenges for practical ASC applications.

Chelali et al. [8] enhance recognition system robustness against environmental noise through audiovisual data fusion. Their approach involves two steps: extracting low-level features—LPC and MFCC for acoustic data, and ZM and HOG for visual data—and then fusing these descriptors to improve each modality's efficiency, compared to score-based late integration. Using a multilayer perceptron (MLP) for classification, results show that the visual modality outperforms the acoustic one in noisy environments, and the fusion technique significantly boosts performance.

Qu et al. [9] introduce a multi-branch 3D CNN model for precise classification, featuring frequency-domain signal representations, a 3D-SE-ResNet for capturing sound correlations, and an auxiliary supervised branch to reduce overfitting. Tested on the DCASE 2019 dataset, this model significantly outperforms existing methods.

Tripathi et al. [10] present an attention-based model that highlights key frames in spectrograms and learns spatio-temporal relationships. Tested on ESC-10 and DCASE 2019 Task-1(A) datasets, it achieves 11.50% and 19.50% accuracy improvements over baseline models, respectively, and effectively focuses on relevant spectrogram regions.

3. PROPOSED METHODOLOGY

The dataset for this study is compiled from the NOIZEUS, AURORA databases, and a Kaggle audio dataset, featuring recordings from ten distinct environments [20]. These environments are meticulously categorized into directories as follows: Quiet Environment, Car Noise Environment, Cocktail Noise Environment, Restaurant Noise Environment, Street Noise Environment, Airport Noise Environment, Train Station Noise Environment, Group Setting Environment, Reverberant Spaces Environment, and Telephone Conversations Environment [11-16][21].

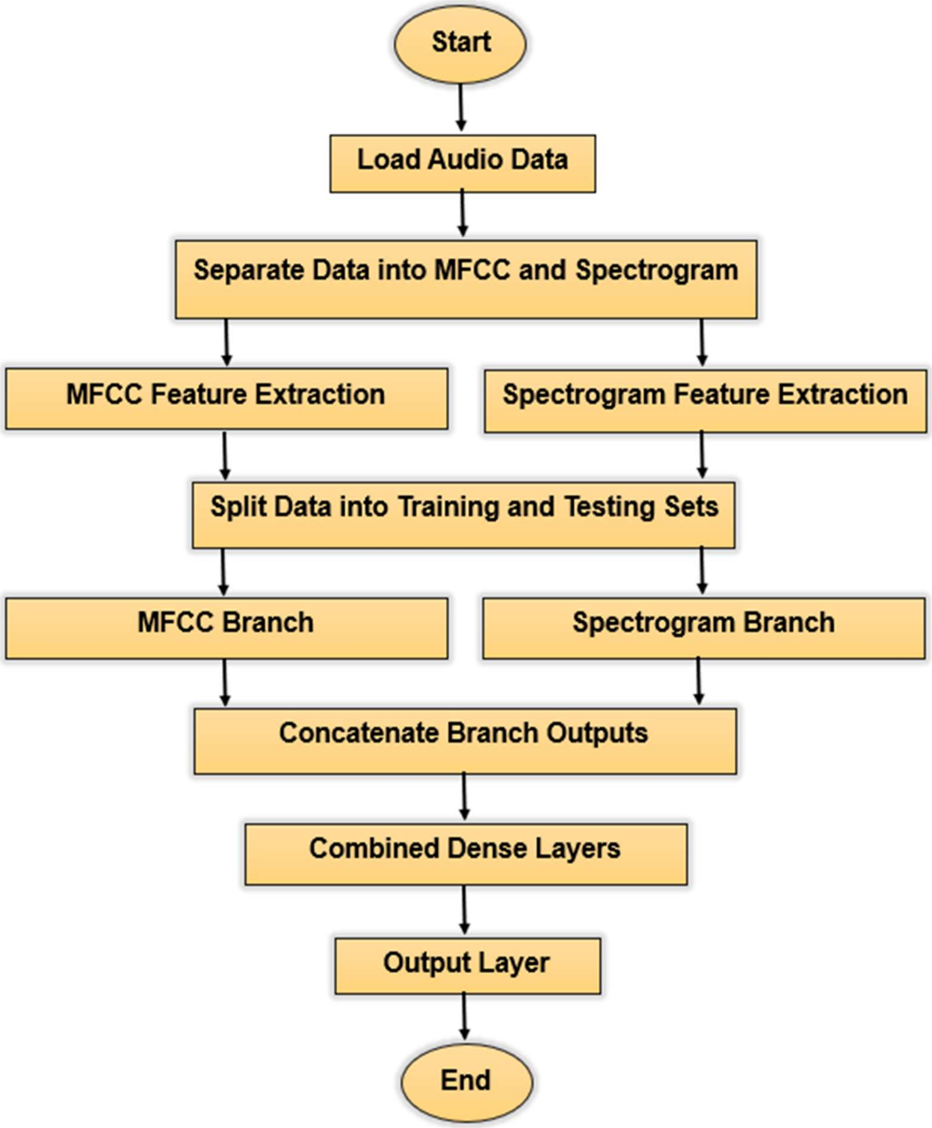
Start
Load Audio Data
Separate Data into MFCC and Spectrogram
MFCC Feature Extraction
Spectrogram Feature Extraction
Split Data into Training and Testing Sets
MFCC Branch
Spectrogram Branch
Concatenate Branch Outputs
Combined Dense Layers
Output Layer
End

Fig.1. Flow Chart of Proposed Method.

The flowchart provides a comprehensive overview of the process for training and evaluating a multimodal CNN model using MFCC and spectrogram features extracted from audio data. The process begins with loading the audio data, which is then separated into MFCC and spectrogram features. These features undergo extraction, after which the dataset is split into training and testing sets. The MFCC features are processed through a specific convolutional neural network (CNN) branch designed to learn patterns from these coefficients, while the spectrogram features are processed through a separate CNN branch tailored to interpret the visual representation of audio frequencies over time [17-19].

Following feature extraction and processing through their respective CNN branches, the outputs from the MFCC and spectrogram branches are concatenated. This combined output is then passed through fully connected dense layers that refine the learned features, capturing complex interactions between the MFCC and spectrogram data. The final output layer produces classification results, typically using a softmax activation function to predict the most likely audio environment. This detailed yet streamlined workflow ensures the model can effectively utilize both types of audio features to achieve high classification accuracy.

4. Results and Discussions

The following section presents a detailed analysis of the results obtained from the study, highlighting the key findings and their implications.

TABLE1: ACCURACY OF MULTIMODAL MODEL IN TRAINING AND TEST PHASES

Multimodal Model	**Accuracy**
Training	99.96%
Testing	97.23%

The multimodal model demonstrated exceptional performance, achieving an accuracy of 99.96% during training and maintaining a high accuracy of 97.23% on the test set, indicating strong generalization capabilities as shown in Table 1.

TABLE2: PERFORMANCE METRICS FOR AUDIO CLASSIFICATION ACROSS DIFFERENT ENVIRONMENTS

Environment	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
Quiet Environment	100	100	100
Car Noise Environment	89	100	94
Cocktail Noise Environment	93	100	97
Restaurant Noise Environment	80	100	89
Street Noise Environment	100	90	95
Airport Noise Environment	83	95	89
Train Station Noise Environment	100	67	80
Group Setting Environment	100	89	94
Reverberant Spaces Environment	83	94	88
Telephone Conversations Environment	94	89	91

The table 2 presents the performance metrics—Precision, Recall, and F1-Score—across various audio environments. Precision measures the accuracy of the model's positive predictions, Recall indicates the model's ability to identify all relevant instances, and F1-Score is the harmonic mean of Precision and Recall, providing a balanced measure of performance.

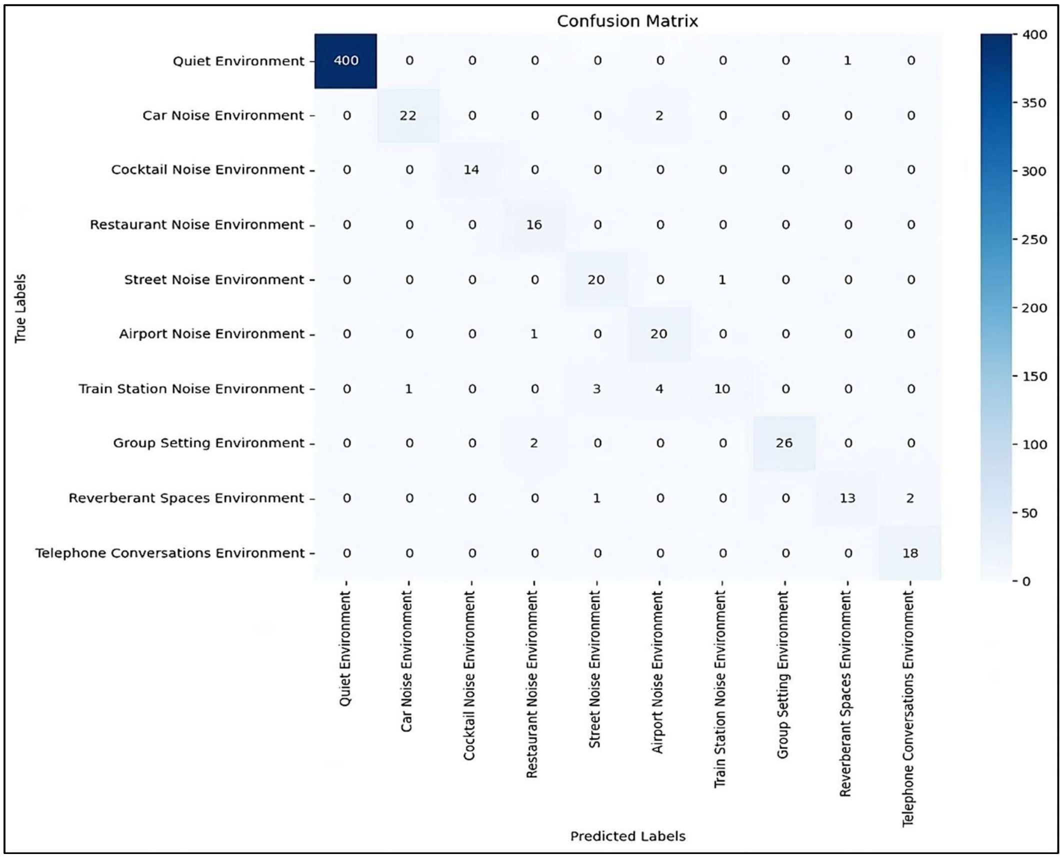

Fig. 2. Confusion Matrix for Multimodal Model

The provided figure 2 is a confusion matrix illustrating the performance of the audio classification model on the test dataset. The vertical axis represents the actual classes of the test samples, while the horizontal axis represents the predicted classes. Diagonal elements indicate correct predictions, with the "Quiet Environment" class showing an impressive 400 correct predictions and only 1 misclassification. Other classes like "Car Noise Environment," "Cocktail Noise Environment," and "Restaurant Noise Environment" also show strong performance with minimal misclassifications. Off-diagonal elements reveal misclassifications, such as "Street Noise Environment" being incorrectly classified as "Car Noise Environment" and "Cocktail Noise Environment." The color gradient, ranging from white to dark blue, visually represents the number of samples, with darker shades indicating higher numbers. Overall, the model performs well on most classes but shows areas for improvement, particularly in distinguishing between similar noise environments, highlighting the effectiveness and limitations of the model.

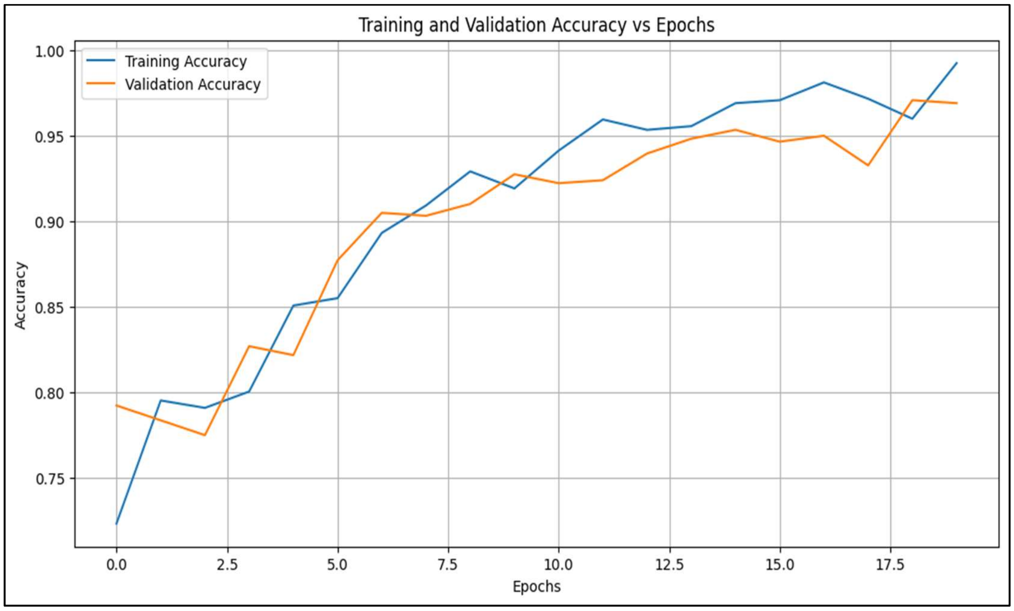

Fig. 3. Training and Validation Accuracy vs Epochs

The graph in Figure 3 illustrates the training and validation accuracy of the multimodal CNN model over 20 epochs. Initially, both training and validation accuracies improve rapidly, reflecting the model's learning process. Around the 10th epoch, the training accuracy reaches above 95%, while the validation accuracy closely follows, indicating the model's ability to generalize well to unseen data. Post the 10th epoch, training accuracy continues to rise steadily, ultimately nearing 100%, whereas validation accuracy shows minor fluctuations yet remains above 95%. These trends suggest that the model effectively learns from the training data and performs consistently on the validation set, demonstrating its robustness and reliability in classifying audio environments.

Figure 4 displays the training and validation loss of the multimodal CNN model over 20 epochs. At the start, the training loss decreases sharply, indicating that the model quickly learns the underlying patterns in the training data. This rapid drop is mirrored by the validation loss, though it follows a slightly more gradual decline. After the initial epochs, both losses stabilize, with the training loss steadily decreasing towards zero, showcasing the model's improved performance. Meanwhile, the validation loss, although fluctuating, maintains a downward trend and remains consistently low. These observations suggest that the model not only fits the training data well but also generalizes effectively to the validation data, demonstrating minimal overfitting and high predictive accuracy.

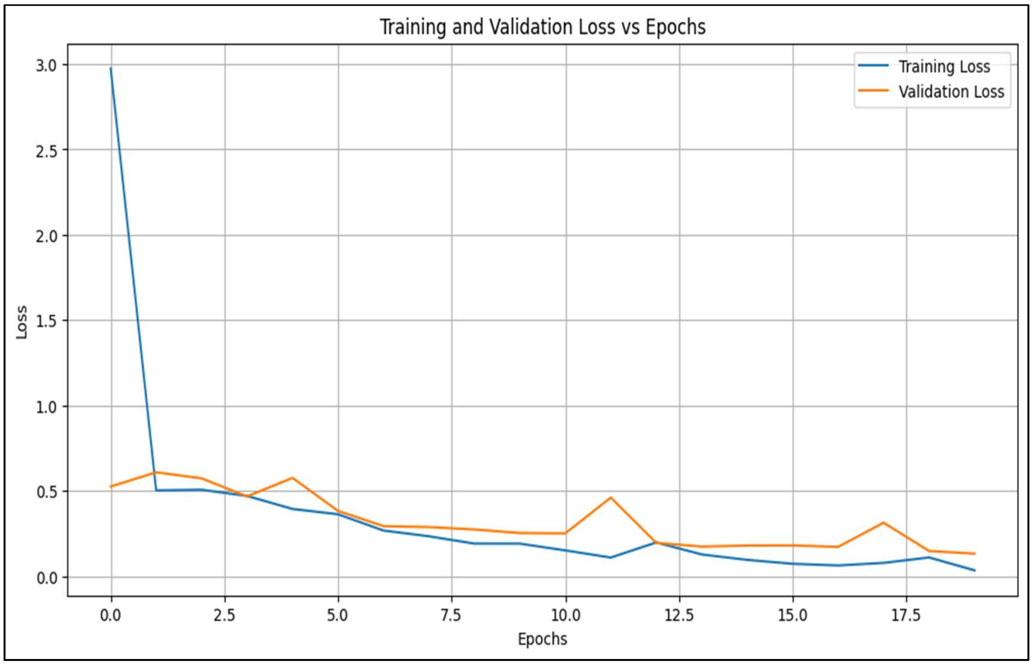

Fig. 4. Training and Validation Loss vs Epochs

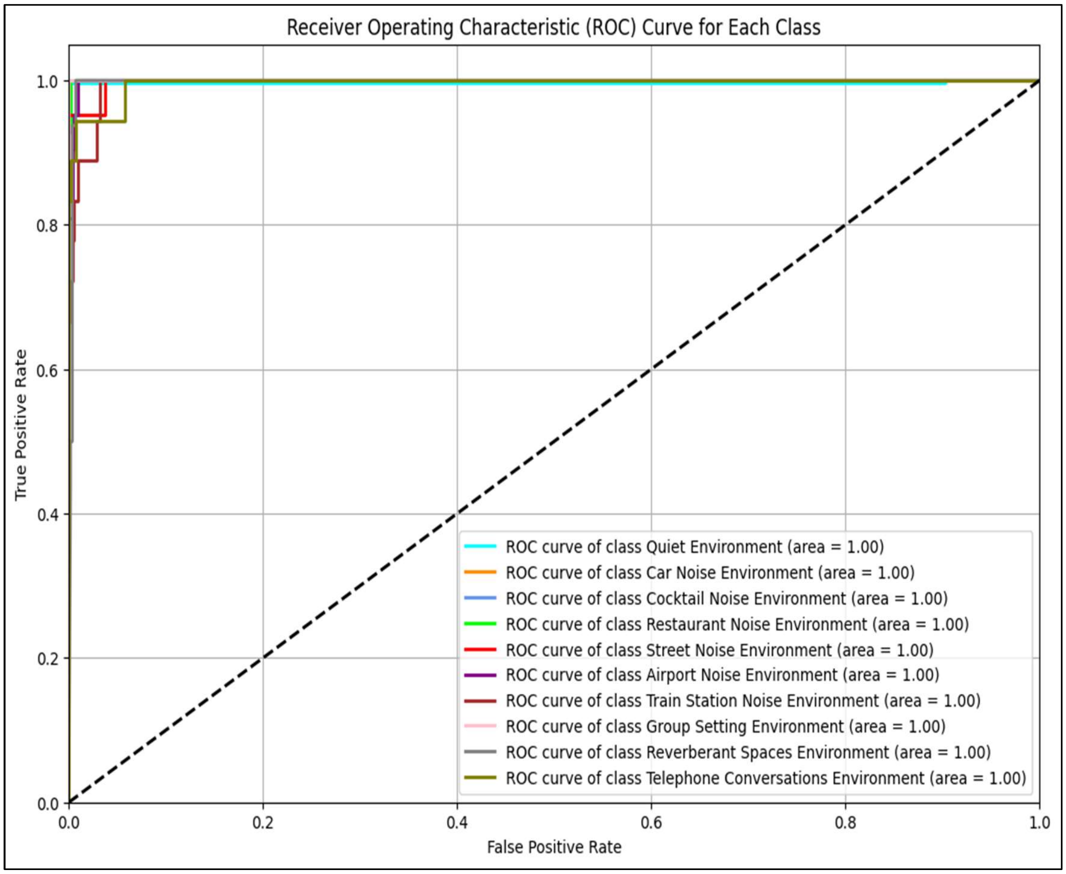

Fig. 5. Receiver Operating Characteristic (ROC) Curve for Each Class

This figure 5 displays the ROC curves for a classification model tested across various noise environments, each represented by a different color. The classes include Quiet, Car Noise, Cocktail Noise, Restaurant Noise, Street Noise, Airport Noise, Train Station Noise, Group Setting, Reverberant Spaces, and Telephone Conversations. With an area under the curve (AUC) of 1.00 for all classes, the model demonstrates perfect classification performance, achieving ideal discrimination between the positive and negative classes in each environment.

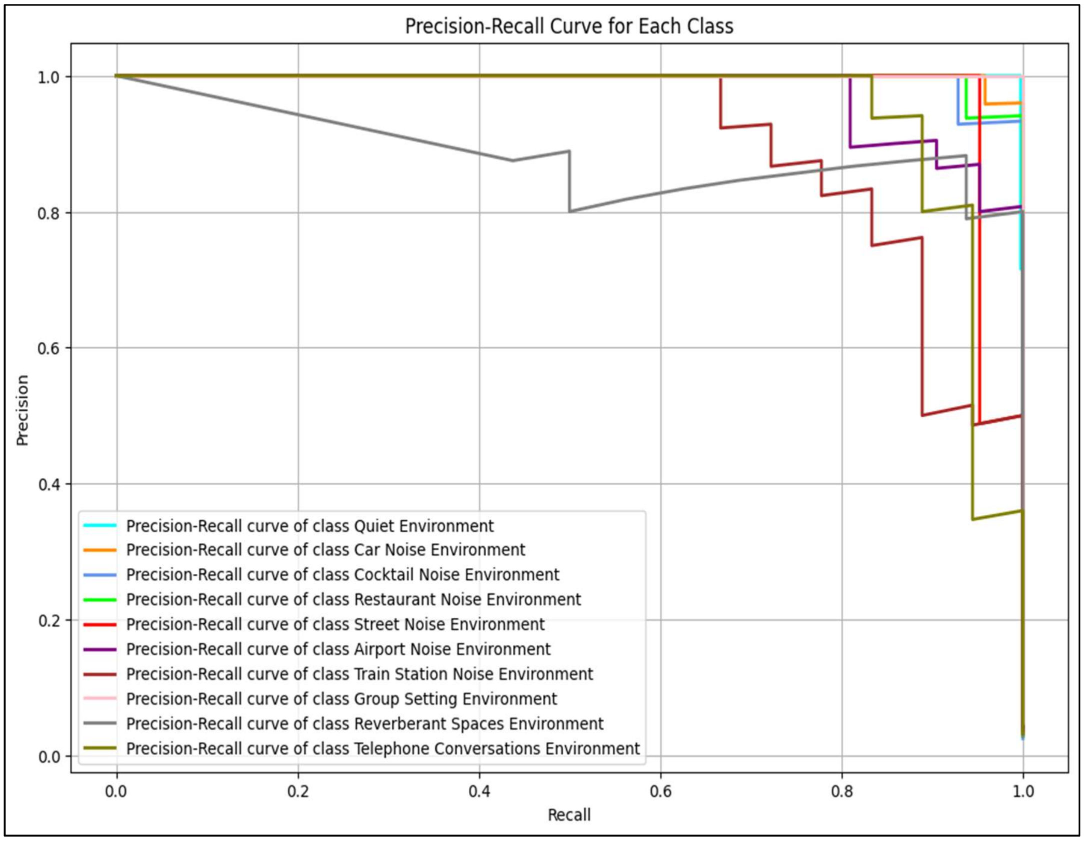

Fig. 6. Precision- Recall Curve for Each Class

This figure 6 illustrates the Precision-Recall (PR) curves for a classification model tested on various noise environments, with each class represented by a distinct color. The classes include Quiet, Car Noise, Cocktail Noise, Restaurant Noise, Street Noise, Airport Noise, Train Station Noise, Group Setting, Reverberant Spaces, and Telephone Conversations. The PR curves show how the model's precision and recall trade-off across different thresholds. Despite some variation, the curves indicate high precision and recall for most classes, demonstrating the model's strong performance in distinguishing between relevant and irrelevant instances across diverse noise environments.

5. Conclusion and Future Scope

The previous work utilizing an RNN model with a FIR filter achieved a training accuracy of 98.50% and a test accuracy of 94.97% [21]. Building on this, the current study improved performance through a multimodal CNN approach that integrates both MFCC and spectrogram features. This enhanced method achieved a 99.96% training accuracy and a 97.23% test accuracy, reflecting a 1.46% and 2.26% improvement, respectively. These results highlight the effectiveness of combining multiple audio features, resulting in more accurate classification of listening conditions, with practical applications in environmental monitoring and smart city systems.

Future work can improve the multimodal model's robustness by incorporating more data sources and exploring new feature extraction techniques or additional modalities like visual or textual data. Advanced methods like transfer learning and domain adaptation could enhance performance in challenging environments. Real-time implementation in applications such as noise monitoring, smart devices, and assistive technologies should be explored. Continuous refinement with user feedback will be crucial for maintaining and advancing the model's performance and relevance.

Declarations:

Availability of data and materials

The datasets used in this study were sourced from the NOIZEUS database and the AURORA-2 corpus dataset. A noisy speech corpus (NOIZEUS) was developed to facilitate comparison of speech enhancement algorithms among research groups. The noisy database contains 30 IEEE sentences (produced by three male and three female speakers) corrupted by eight different real-world noises at different SNRs. The noise was taken from the AURORA database and includes suburban train noise, babble, car, exhibition hall, restaurant, street, airport and train-station noise. This dataset is available on the following URL.

https://ecs.utdallas.edu/loizou/speech/noizeus/

https://www.kaggle.com/datasets.

Conflicts of Interest

The authors declare that they have no Conflicts of Interest.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Authors' contributions

Sunilkumar M. Hattaraki, collected the data, analyzed the data, implemented the proposed work and drafted the complete manuscript. Shankarayya G. Kambalimath defined the problem statement and provided critical reviews. All authors read and approved the final manuscript.

Acknowledgements

I, Sunilkumar M. Hattaraki, would like to express my gratitude to the Research Centre, Department of Electronics and Communication Engineering at Basaveshwar Engineering College, Bagalkote, Karnataka, India.

References

[1]. S. Jabeen, et al., "A review on methods and applications in multimodal deep learning," ACM Transactions on Multimedia Computing, Communications and Applications, vol. 19, no. 2s, pp. 1-41, 2023.

[2]. S. Kakuba, A. Poulose, and D. S. Han, "Deep learning approaches for bimodal speech emotion recognition: Advancements, challenges, and a multi-learning model," IEEE Access, 2023.

[3]. S. Jeon, et al., "Multimodal audiovisual speech recognition architecture using a three-feature multi-fusion method for noise-robust systems," ETRI Journal, vol. 46, no. 1, pp. 22-34, 2024.

[4]. D. Utebayeva and A. Yembergenova, "Study a deep learning-based audio classification for detecting the distance of UAV," in 2024 IEEE International Conference on Evolving and Adaptive Intelligent Systems (EAIS), IEEE, 2024.

[5]. S. S. Kushwaha and M. Fuentes, "A multimodal prototypical approach for unsupervised sound classification," arXiv preprint arXiv:2306.12300, 2023.

[6]. F. A. Shaqra, R. Duwairi, and M. Al-Ayyoub, "A multi-modal deep learning system for Arabic emotion recognition," International Journal of Speech Technology, vol. 26, no. 1, pp. 123-139, 2023.

[7]. B. Ding, et al., "Acoustic scene classification: a comprehensive survey," Expert Systems with Applications, p. 121902, 2023.

[8]. F. Z. Chelali, "Bimodal fusion of visual and speech data for audiovisual speaker recognition in noisy environment," International Journal of Information Technology, vol. 15, no. 6, pp. 3135-3145, 2023.

[9]. Y. Qu, et al., "Acoustic scene classification based on three-dimensional multi-channel feature-correlated deep learning networks," Scientific Reports, vol. 12, no. 1, p. 13730, 2022.

[10]. A. M. Tripathi and A. Mishra, "Environment sound classification using an attention-based residual neural network," Neurocomputing, vol. 460, pp. 409-423, 2021.

[11]. Y. Hu and P. Loizou, "Subjective evaluation and comparison of speech enhancement algorithms," Speech Communication, vol. 49, pp. 588-601, 2007.

[12]. Y. Hu and P. Loizou, "Evaluation of objective quality measures for speech enhancement," IEEE Transactions on Speech and Audio Processing, vol. 16, no. 1, pp. 229-238, 2008.

[13]. J. Ma, Y. Hu, and P. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," Journal of the Acoustical Society of America, vol. 125, no. 5, pp. 3387-3405, 2009.

[14]. ITU-T Recommendation P. 862, "Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," ITU, 2000.

[15]. H. Hirsch and D. Pearce, "The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions," ISCA ITRW ASR2000, Paris, France, September 18-20, 2000.

[16]. UrbanSound8K dataset, [Online]. Available: https://www.kaggle.com/datasets/chrisfilo/urbansound8k.

[17]. Kundur, N.C. and Mallikarjuna, P.B. 2022. Deep Convolutional Neural Network Architecture for Plant Seedling Classification. Engineering, Technology & Applied Science Research. 12, 6 (Dec. 2022), 9464–9470. DOI:https://doi.org/10.48084/etasr.5282.

[18]. S Nuanmeesri, S. 2021. A Hybrid Deep Learning and Optimized Machine Learning Approach for Rose Leaf Disease Classification. Engineering, Technology & Applied Science Research. 11, 5 (Oct. 2021), 7678–7683. DOI:https://doi.org/10.48084/etasr.4455.

[19]. Owida, H.A., Al-Ghraibah, A. and Altayeb, M. 2021. Classification of Chest X-Ray Images using Wavelet and MFCC Features and Support Vector Machine Classifier. Engineering, Technology & Applied Science Research. 11, 4 (Aug. 2021), 7296–7301. DOI:https://doi.org/10.48084/etasr.4123.

[20]. S. Loizou, "NOIZEUS: A noisy speech corpus for evaluation of speech enhancement algorithms," University of Texas at Dallas, [Online]. Available: https://ecs.utdallas.edu/loizou/speech/noizeus/. [Accessed: 09-Sep-2024].

[21]. S. M. Hattaraki and S. G. Kambalimath, "Detection and Classification of Diverse Listening Conditions for Hearing-Impaired Individuals using RNN Model and FIR Filter,"Journal of Basic Science and Engineering, vol. 21, no. 01, pp. 592-612, ISSN: 1005-0930, 2024,doi: 10.0802/JBSE.2024984437.