International Journal of Informatics and Computation (IJICOM)
Vol. 8, No. 1, 2026
ISSN: 2685-8711, E-ISSN: 2714-5263 DOI: 10.35842/ijicom

Enhancing Heart Sounds Classification Using MFCC And CNN

Susanti[1], Bulkis Kanata[2], Sudi M Al Sasongko[3]

Abstract
Cardiovascular disease remains one of the leading causes of death worldwide, which increases the need for accurate and efficient early detection methods. In this paper, we utilize heart sound analysis as a non-invasive screening approach to distinguish normal and abnormal cardiac conditions. We apply Mel-Frequency Cepstral Coefficients (MFCC) to extract discriminative spectral features from heart sound recordings and use three Convolutional Neural Network (CNN) architectures including AlexNet, VGG16, and ResNet18 for classification. To improve model robustness and reduce overfitting, we implement audio data augmentation techniques, including white noise addition, pitch scaling, time stretching, and random gain adjustment. We train all models using a batch size of 32, 25 epochs, and a learning rate of 0.0001. The experimental results show that this learning rate provides stable convergence and optimal performance across architectures. AlexNet achieves the highest accuracy of 100%, followed by VGG16 with 99.5% and ResNet18 with 97%. Overall, this paper demonstrates that the combination of MFCC feature extraction, data augmentation, and CNN modeling provides highly accurate and reliable heart sound classification, with strong potential for practical clinical screening applications.

Keywords:
Classification, Heart Sound, CNN, MFCC

1. Introduction

Heart sound analysis plays a crucial role in early cardiovascular disease detection because auscultation remains one of the most accessible and low-cost diagnostic procedures. Clinicians rely on phonocardiogram (PCG) signals to identify abnormalities such as murmurs, valve disorders, and arrhythmias. However, manual interpretation depends heavily on physician expertise and subjective judgment, which often leads to inter-observer variability and misclassification, especially in noisy clinical environments. Recent advances in deep learning attempt to address this limitation by automating heart sound classification using data-driven models. Comprehensive reviews highlight that deep neural networks significantly improve diagnostic consistency compared to traditional machine learning approaches, yet they still face challenges related to signal variability and limited annotated datasets [14], [15].

Researchers increasingly adopt MFCC to represent heart sound signals because MFCC effectively capture perceptually relevant frequency characteristics. Studies demonstrate that MFCC-based features enhance discrimination between normal and

Corresponding Author: Susanti(susantisan059@gmail.com)
1 Susanti, Universitas Mataram, (susantisan059@gmail.com)
2 Bulkis Kanata, Universitas Mataram, (uqikanata@unram.ac.id)
3 Sudi M Al Sasongko, Universitas Mataram, (mariyantosas@unram.com)

abnormal heart sounds when combined with ensemble classifiers or deep networks. Nevertheless, standard MFCC extraction may overlook subtle temporal variations in cardiac cycles, which reduces sensitivity to faint murmurs. Several works propose improved MFCC variants or feature fusion strategies to overcome these limitations, yet inconsistencies in preprocessing pipelines and feature parameterization remain unresolved issues in the literature [3], [10].

Deep CNNs dominate current heart sound classification research due to their ability to learn hierarchical representations from time–frequency images. Researchers transform PCG signals into spectrograms or 2D representations and feed them into CNN architectures to capture spatial-frequency correlations. Studies using 2D-transformed phonocardiograms report substantial performance gains compared to handcrafted feature pipelines. However, CNN performance strongly depends on the quality of input representations, and inappropriate feature extraction may lead to overfitting or poor generalization across datasets [4], [11].

Hybrid deep learning architectures further attempt to model both spatial and temporal dependencies in heart sounds. Researchers combine CNN with recurrent layers such as GRU or LSTM to capture sequential cardiac cycle patterns. These hybrid CNN-GRU and LSTM-CNN models demonstrate improved classification accuracy, particularly for long-duration recordings. Despite these advances, hybrid architectures increase computational complexity and training time, which limits their deployment in lightweight or real-time diagnostic systems. Therefore, optimizing a simpler yet effective CNN-based framework remains an important research direction [5], [11].

Recent works also introduce attention mechanisms and advanced convolutional blocks to enhance feature learning capability. CNN models integrated with attention modules selectively focus on informative cardiac segments, improving robustness against noise and irrelevant signal components. Similarly, deep residual learning strategies help mitigate vanishing gradient problems and enable deeper architectures for heart sound analysis. Although these approaches enhance performance, they often require large-scale labeled datasets and careful hyperparameter tuning, which are not always feasible in clinical settings with limited annotated PCG data [1], [7].

Another important issue involves the selection of time–frequency representations. Researchers explore Short-Time Fourier Transform (STFT), wavelet transforms, and combined representations to better capture cardiac acoustic patterns. Comparative studies reveal that different representations produce varying classification results depending on dataset characteristics. While combined time–frequency approaches improve robustness, they introduce additional preprocessing complexity. Thus, selecting an efficient yet discriminative representation such as MFCC remains a practical solution when computational efficiency and real-time applicability are primary considerations [8], [9].

Beyond architecture design, researchers emphasize the need for generalizable and multimodal solutions. Some studies integrate multiple cardiac signal modalities or apply ensemble learning to enhance classification stability. Although multimodal and ensemble strategies improve performance metrics, they increase system complexity and resource requirements. For practical implementation in portable or embedded diagnostic devices, researchers must balance model accuracy, computational cost, and feature compactness. This trade-off motivates further investigation into optimized MFCC-based CNN frameworks that remain lightweight but accurate [12], [13].

Considering these challenges, this paper focuses on enhancing heart sound classification using MFCC and CNN by strengthening feature representation and optimizing convolutional learning without excessive architectural complexity. We build upon prior findings that demonstrate the effectiveness of MFCC features and CNN-based classifiers while addressing limitations related to feature sensitivity, noise robustness, and computational efficiency. By refining the MFCC extraction process and designing an

efficient CNN model, we aim to improve classification performance and provide a practical solution suitable for real-world clinical screening applications [1], [3], [4], [10].

2. Related Works

Several Researchers extensively investigated automatic heart sound classification to improve the reliability of cardiovascular screening. Singh et al. conducted a comprehensive survey on deep learning techniques for auscultation signal processing and reported that convolutional architectures significantly outperformed traditional machine learning models in detecting pathological patterns. They emphasized the importance of large annotated datasets and standardized preprocessing pipelines. However, they also highlighted persistent challenges, including data imbalance, noise contamination, and limited cross-dataset generalization. Similarly, Zhao provided a structured review of deep learning approaches for heart sound analysis and concluded that feature representation and model complexity remained critical determinants of performance. Both studies offered strong theoretical foundations but did not propose optimized feature–model integration strategies tailored for lightweight clinical deployment [14], [15].

Several studies focused specifically on MFCC-based feature extraction for heart sound classification. Deng et al. improved MFCC representations and integrated them with convolutional recurrent neural networks, demonstrating that enhanced cepstral features increased sensitivity to abnormal murmurs. Hosseinzadeh et al. combined MFCC features with ensemble classifiers and reported improved robustness against noisy recordings. These works confirmed the discriminative capability of MFCC in cardiac auscultation tasks. However, they introduced additional architectural complexity through recurrent layers or ensemble mechanisms, which increased computational cost and limited real-time applicability. Moreover, their preprocessing pipelines varied significantly, making reproducibility and standardization challenging [3], [10].

Deep CNN-based approaches also dominated the literature due to their ability to learn hierarchical representations from time–frequency images. Riccio et al. transformed phonocardiograms into two-dimensional images and applied CNN models, achieving strong classification accuracy across multiple pathological categories. Similarly, Zhao et al. utilized a hybrid LSTM-CNN architecture to capture both spatial and temporal patterns from time–frequency representations. These studies demonstrated that CNN-based systems effectively modeled cardiac acoustic characteristics. Nevertheless, they often required high-resolution spectrogram inputs and deep architectures, which increased memory consumption and training time. The dependence on complex model structures limited their scalability for embedded or portable healthcare systems [4], [11].

Hybrid deep learning architectures further extended CNN models by integrating recurrent neural networks. Choudhary et al. combined CNN and GRU layers to improve temporal pattern recognition in heart sound sequences. Their hybrid framework achieved higher classification accuracy compared to standalone CNN models. Likewise, Bahreini et al. proposed MFCC-based feature fusion with CNN deep features and demonstrated enhanced diagnostic performance. Although these hybrid and fusion-based models improved predictive capability, they significantly increased model parameters and computational burden. The complexity of such architectures may hinder their integration into real-time diagnostic tools, especially in low-resource environments [5], [6].

Researchers also explored attention mechanisms and residual learning to enhance CNN performance. Li et al. introduced improved mel-frequency spectral coefficients combined with deep residual learning and achieved notable accuracy improvements by mitigating vanishing gradient problems. Huai et al. incorporated convolutional block attention modules into CNN architectures and reported better localization of informative heart sound segments. These approaches strengthened feature learning and improved robustness to noise. However, they required deeper networks and careful hyperparameter

optimization, which may not be practical for small or imbalanced datasets commonly encountered in clinical scenarios [1], [7].

Time–frequency representation techniques also received considerable attention. Orozco-Reyes et al. applied combined time–frequency representations and deep learning models, demonstrating that multi-representation inputs improved classification robustness. Guzmán-Alfaro et al. integrated MFCC and wavelet features, showing that hybrid feature extraction enhanced abnormal heart sound detection. While these approaches improved discriminative power, they increased preprocessing complexity and computational overhead. The integration of multiple representations also introduced redundancy, which sometimes reduced model efficiency when deployed in constrained hardware environments [8], [9].

Beyond single-modality approaches, some researchers investigated multimodal and ensemble-based strategies. Kumar and Aggarwal developed a robust CNN architecture for multimodal cardiac signal classification and reported improved stability across heterogeneous datasets. Mohanty and Nayak utilized ensemble learning combined with deep CNN features, achieving competitive performance metrics. Although these strategies enhanced generalization capability, they relied on additional modalities or multiple classifiers, thereby increasing system complexity and limiting practical applicability in routine auscultation settings where only PCG signals are available [12], [13].

In summary, prior studies consistently demonstrated that MFCC features and CNN-based models significantly improved heart sound classification performance. Researchers enhanced accuracy through hybrid architectures, attention mechanisms, feature fusion, and ensemble strategies. However, many approaches increased computational complexity, required extensive preprocessing, or depended on large datasets. The literature therefore revealed a gap in developing an efficient yet powerful MFCC-CNN framework that balances accuracy, robustness, and computational efficiency. This gap motivated the present study to optimize MFCC representation and CNN architecture for enhanced and practical heart sound classification.

3. Proposed Method

This study proposes a heart sound classification system that distinguishes normal and abnormal cardiac signals using MFCC and a comparative evaluation of three CNN architectures: AlexNet, VGG16, and ResNet18. The proposed framework consists of four main stages, namely signal preprocessing, feature extraction, CNN-based modeling, and performance evaluation. We extract MFCC features to represent the spectral characteristics of heart sounds and feed these features into each CNN architecture to assess their discriminative capability. We then compare classification accuracy, loss behavior, and generalization performance across models to determine the most effective architecture for the proposed binary classification task.

We use a secondary dataset obtained from Kaggle, specifically the Yaseen Khan Heart Sound Dataset. The dataset contains heart sound recordings in .wav format categorized into five clinical classes: normal, aortic stenosis, mitral valve prolapse, mitral regurgitation, and mitral stenosis. In this study, we simplify the problem into a binary classification scenario by grouping pathological categories into a single abnormal class while retaining the normal class as a separate category. This strategy allows us to focus on practical screening applications where early detection of abnormal heart sounds is the primary objective. We divide the dataset into training (80%), validation (10%), and testing (10%) subsets to ensure reliable model development and unbiased performance evaluation.

We perform preprocessing to enhance signal quality and ensure data consistency before feature extraction. First, we apply amplitude normalization to reduce variability

caused by recording conditions. Next, we standardize signal duration to ensure uniform input dimensions for CNN processing. To improve model robustness and generalization, we implement data augmentation techniques, including speed variation, pitch shifting, background noise addition, and signal segmentation. These augmentation strategies simulate real-world recording variations and help prevent overfitting, thereby strengthening the reliability of the classification system.

1. MFCC Feature Extraction

MFCC converts an audio signal $x(t)$ into a sequence of feature vectors that represent perceptual frequency content:

$$\text{MFCC}(n) = \sum_{k=1}^{K} log\big(S(k)\big) \cdot cos\left[n\Big(k - \frac{1}{2}\Big)\frac{\pi}{K}\right] \quad (1)$$

where:

- $S(k)$ = Mel-scaled power spectrum obtained from the Short-Time Fourier Transform (STFT),
- K = number of Mel filter banks,
- n = cepstral coefficient index (typically ($n = 1,2,\ldots,13$)).

The Mel frequency scale maps real frequency f (Hz) to perceptual frequency m (Mels):

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

This transformation emphasizes frequency ranges important to heart sound classification, producing features that are robust for heart sound classification.

2. CNN-Based Classification

After extracting MFCC features, the CNN performs hierarchical feature learning and classification through the following core operations:

Convolution Layer:

$$z_{i,j}^{(l)} = f\left(\sum_{m,n} x_{i+m,j+n}^{(l-1)} \cdot w_{m,n}^{(l)} + b^{(l)}\right) \quad (2)$$

where:

- $x^{(l-1)}$ = input feature map (MFCC matrix),
- $w^{(l)}$ = convolution kernel (filter),
- $b^{(l)}$ = bias term,
- $f(\cdot)$ = activation function, e.g., ReLU $f(x) = \max(0,x)$

Pooling Layer (Downsampling):

$$p_{i,j}^{(l)} = \max_{(m,n)\in R} z_{i+m,j+n}^{(l)}$$

which reduces spatial dimensions while retaining dominant features.

Fully Connected Layer and Softmax Classification:

$$\hat{y}_c = \frac{e^{z_c}}{\sum_{k=1}^{C} e^{z_k}} \quad (3)$$

where $\hat{y}_c$ is the probability that the input belongs to class c (heart sound), and C is the total number of heart sound.

In this model, MFCC extracts perceptually meaningful frequency-domain features from raw audio, effectively reducing noise and dimensionality. The CNN then learns spatial correlations between these MFCC features through convolution and pooling, enabling hierarchical representation learning. Finally, the Softmax layer classifies the audio sample into a predefined heart sound. This integration allows for efficient and accurate heart sound classification by combining signal-processing precision (MFCC) with deep-learning adaptability.

4. Experimental Setup

This research utilizes two primary software tools to support experimental implementation and evaluation. We use Google Colab as the main development environment because it provides a cloud-based Python platform with GPU acceleration, which facilitates efficient preprocessing, MFCC feature extraction, and CNN model training. Google Colab enables reproducible experimentation and simplifies dependency management within a controlled computational environment.

Within Google Colab, we implement the entire pipeline using Python and several essential libraries. We use TensorFlow and Keras to design, train, and evaluate the CNN architectures, including AlexNet, VGG16, and ResNet18. We employ NumPy for numerical computation and data manipulation, while Librosa supports audio signal processing and MFCC feature extraction. Additionally, we use Scikit-learn to compute classification performance metrics such as accuracy, precision, recall, and F1-score. These libraries collectively ensure robust signal analysis, efficient deep learning implementation, and reliable performance assessment.

For post-processing and structured presentation of experimental results, we use Microsoft Excel. We organize evaluation outputs, summarize classification metrics, and construct performance comparison tables across models. Excel assists in visualizing differences in accuracy and other metrics in a clear and interpretable format. This structured evaluation process allows us to compare model behavior systematically and transparently.

To ensure stable and consistent experimentation within the available computational resources, we define a fixed training configuration for all CNN architectures. We carefully select hyperparameters to balance training efficiency and model convergence. The training parameters used in this study are summarized in Table 1.

Table 1. Training Parameters

No	Parameter	Value
1	Epoch	25
2	Batch Size	32
3	Optimizer	Adam
4	Learning Rate	0.0001

We train all three CNN models including AlexNet, VGG16, and ResNet18 using the same configuration to ensure a fair and controlled comparison. By maintaining identical hyperparameters, we eliminate bias caused by unequal optimization settings and ensure that performance differences arise from architectural characteristics rather than training inconsistencies.

5. Result and Analysis

At this stage, we compare architectures to identify five types of normal and abnormal heart sounds. The data used is from Kaggle, consisting of 1,000 heart sound data and 1,000 augmented data, for a total of 2,000 heart sound data that has been extracted into a 2D matrix using MFCC so that it can be input into CNN, through a training and testing process. Training is a step to evaluate the progress of the model's performance by measuring accuracy and loss during the training process. We also present several visualization results to recognize the model's capabilities in graphical visualization.

1. MFCC Feature Extraction Representation

In this study, we conduct comparison between heart sound signals in the form of time waves (.WAV) and MFCC feature extraction results for five types of heart sounds, namely normal, aortic stenosis, mitral regurgitation, mitral stenosis, and mitral valve prolapse. The MFCC representation in the figure is displayed in the form of a color map that describes the distribution of spectral energy over time. The horizontal axis shows the time sequence, while the vertical axis shows the MFCC coefficient index. The colors in the MFCC image represent the magnitude of the energy value, where lighter colors indicate higher energy, while darker colors indicate lower energy. Fig. 2 shows a comparison between heart sound signals in the form of time waves (.WAV) and MFCC feature.

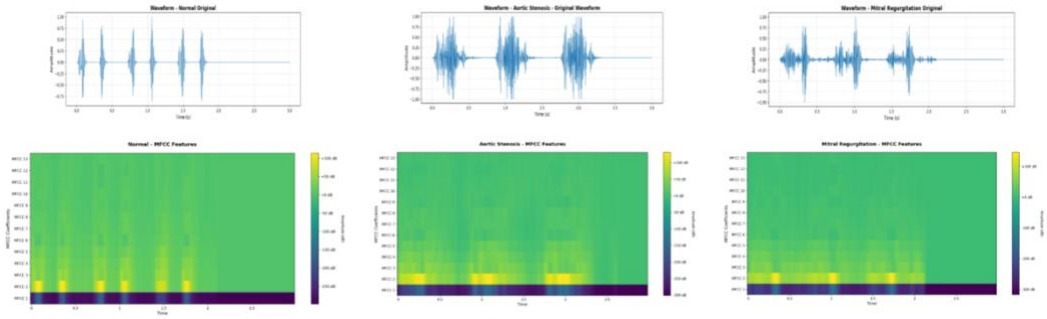


Fig. 2. MFCC Feature Extraction Representation

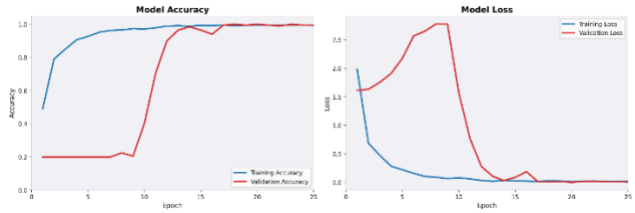
2. Model Accuracy and Loss

In this stage, we calculate VGG16 model with an accuracy of 99.5% and a loss of 0.0313, the visualization can be seen in Figure 4(a). Next, for the ResNet18 model, an accuracy of 93% and a loss of 2.8642 were obtained. The visualization can be seen in Figure 5(a). Of the three models, AlexNet produced the highest accuracy value, but the VGG16 and ResNet graphs showed a more stable learning process during training. Table 2 shows the training results for the AlexNet model.

Table 2. Model Training

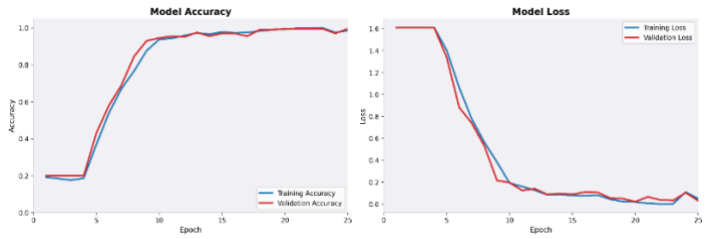
Architecture	Accuracy	Loss	Val Accuracy	Val Loss
AlexNet	1.000	0.0122	0.9950	0.0057
VGG16	0.9950	0.0313	0.9757	0.0733
ResNet18	0.9379	2.8642	0.9600	2.8687

The following is a visualization of the training results of the three models. The training and validation graphs show that the increase in accuracy is accompanied by a stable decrease in loss for each model. The proximity between the training and validation accuracy curves indicates that the model does not experience overfitting or underfitting.



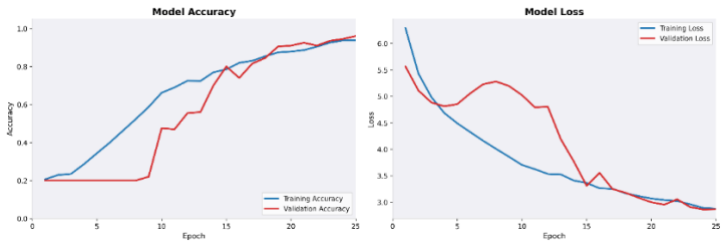
(a) (b)

Fig.3 Graph of Alexnet model training results (a) Acc and Val_Acc, (b) loss and val_loss



(a) (b)

Fig.4 Graph of VGG16 model training results (a) Acc and Val_Acc, (b) loss and val_loss



(a) (b)

Fig.5 Graph of ResNet18 model training results (a) Acc and Val_Acc (b) loss and val_loss

3. Evaluation Metrics

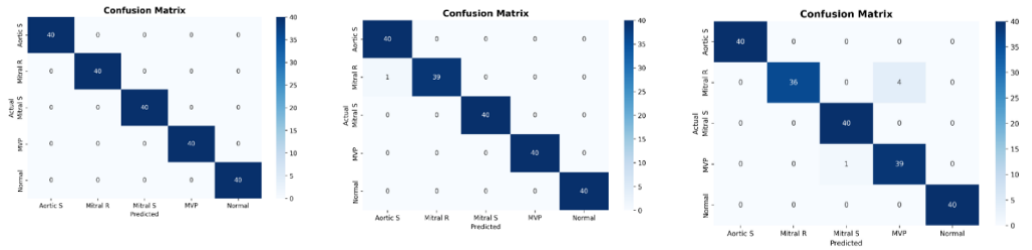
In this stage, we undergo model evaluation as a stage to assess how well a model predicts new data that has not been studied before. The goal is to determine whether the model is capable of recognizing patterns correctly and providing accurate results. Usually, this assessment is carried out using several measures such as precision (how accurate the model's predictions are), recall (how much data is correctly recognized), and F1-score.

Table 3 . *Heart Sound Classification Report for Three Models*

Label	Alexnet			VGG16			ResNet18		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Normal	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Aortic Stenosis	1.000	1.000	1.000	0.976	1.000	0.988	1.000	1.000	1.000
Mitral Regurtation	1.000	1.000	1.000	1.000	0.975	1.000	1.000	0.900	0.947
Mitral Stenosis	1.000	1.000	1.000	1.000	1.000	1.000	0.976	1.000	0.988
Mitral Valve Proplase	1.000	1.000	1.000	1.000	1.000	1.000	0.907	0.975	0.940

We evaluated the models using 200 heart sound recordings. Each class contained 40 samples. We tested the models on unseen data to measure real performance. The results in Table 3 show strong performance across most categories. All three models achieved high precision, recall, and F1-score values. These results indicate that the models classified most heart sound types correctly and consistently.

AlexNet achieved 100% accuracy. It classified all samples correctly without any errors. VGG16 achieved 99.5% accuracy, with only one misclassification in the Mitral Regurgitation class, where one sample was predicted as Aortic Stenosis. ResNet18 achieved 97% accuracy and showed more errors, especially in the Mitral Regurgitation and MVP classes. Four Mitral Regurgitation samples were predicted as MVP, and one MVP sample was predicted as Mitral Stenosis. We also used confusion matrices to visualize these error patterns and better understand how each model behaved across classes.



Alexnet VGG16 ResNet18

Fig.6 Confusion Matrix

6. Conclusion

This paper proposes and implements a heart sound classification system using MFCC and CNN. We utilize MFCC to extract meaningful spectral features from heart sound signals and apply three CNN architectures, including AlexNet, VGG16, and ResNet18 to perform classification. We train the models using a batch size of 32, 25 epochs, and a learning rate of 0.0001 to ensure stable convergence. The experimental results demonstrate that all three models achieve high classification performance. AlexNet achieves 100% accuracy, VGG16 achieves 99.5%, and ResNet18 achieves 97%. These findings confirm that the proposed MFCC-CNN framework effectively distinguishes between normal and abnormal heart sounds.

We evaluate the system using 200 unseen heart sound recordings, with 40 samples per class. We measure performance using accuracy, precision, recall, and F1-score. The results show consistently high metric values across all models, indicating strong classification capability. Based on the confusion matrix analysis, most samples are classified correctly. However, we observe several misclassifications in classes with similar acoustic characteristics, particularly in Mitral Regurgitation and MVP categories. These errors highlight the challenge of distinguishing heart conditions with overlapping spectral patterns, even when using deep learning models.

We also analyze the training and validation curves to examine model learning behavior. We observe that accuracy increases steadily while loss decreases consistently during training. The training and validation curves remain close to each other, which indicates that the models do not suffer from significant overfitting or underfitting. Although AlexNet achieves the highest accuracy, we find that VGG16 and ResNet18 demonstrate stable and reliable learning processes. Therefore, we conclude that the proposed approach successfully combines MFCC feature extraction and CNN modeling to deliver accurate and robust heart sound classification, with strong potential for practical clinical screening applications.

References

[1] F. Li, Z. Zhang, L. Wang, and W. Liu, “Heart sound classification based on improved mel-frequency spectral coefficients and deep residual learning,” *Biomedical Signal Processing and Control*, vol. 77, p. 102893, 2022, doi: 10.1016/j.bspc.2020.102893.

[2] F. Li, Z. Zhang, L. Wang, and W. Liu, “Heart sound classification based on improved mel-frequency spectral coefficients and deep residual learning,” *Frontiers in Physiology*, vol. 13, 2022, doi: 10.3389/fphys.2022.1084420.

[3] M. Deng et al., “Heart sound classification based on improved MFCC features and convolutional recurrent neural networks,” *Neural Networks*, vol. 130, pp. 22–32, 2020, doi: 10.1016/j.neunet.2020.06.015.

[4] D. Riccio et al., “CNN-based classification of phonocardiograms using 2D transformed heart sound images,” *Expert Systems with Applications*, vol. 212, p. 119048, 2023, doi: 10.1016/j.eswa.2022.119048.

[5] R. R. Choudhary, M. R. Singh, and P. K. Jain, “Heart sound classification using a hybrid of CNN and GRU deep learning models,” *Procedia Computer Science*, vol. 232, pp. 3085–3093, 2024, doi: 10.1016/j.procs.2024.04.292.

[6] Bahreini, R. Barati, and A. Kamali, “Cardiac sound classification using a hybrid approach: MFCC-based feature fusion and CNN deep features,” *EURASIP Journal on Advances in Signal Processing*, vol. 2025, no. 1, pp. 1–15, 2025, doi: 10.1186/s13634-025-01203-0.

[7] X. Huai, L. J. Lei, C. Wang, P. Chen, and H. Li, “Heart sound classification based on convolutional neural network with convolutional block attention module,” *Frontiers in Physiology*, vol. 16, p. 1596150, 2025, doi: 10.3389/fphys.2025.1596150.

[8] L. Orozco-Reyes, M. A. Alonso-Arévalo, E. García-Canseco, R. F. Ibarra-Hernández, and R. Conte-Galván, “A deep-learning approach to heart sound classification based on combined time-frequency representations,” *Technologies*, vol. 13, no. 4, p. 147, 2025, doi: 10.3390/technologies13040147.
[9] S. Guzmán-Alfaro et al., “Heart sound classification with Mel-frequency cepstral coefficients and wavelet features,” *Diagnostics*, vol. 16, no. 1, p. 83, 2025, doi: 10.3390/diagnostics16010083.
[10] M. Hosseinzadeh et al., “Enhanced heart sound classification using Mel frequency cepstral coefficients and ensemble classifiers,” *PLoS ONE*, vol. 19, no. 12, e0316645, 2024, doi: 10.1371/journal.pone.0316645.
[11] Y. Zhao et al., “Time–frequency heart sound analysis with LSTM-CNN,” *Biomedical Signal Processing and Control*, vol. 73, p. 103323, 2022, doi: 10.1016/j.bspc.2022.103323.
[12] P. Kumar and V. Aggarwal, “A robust CNN architecture for multimodal cardiac signal classification,” *Journal of Medical Systems*, vol. 48, pp. 1–15, 2024.
[13] Mohanty and A. Nayak, “Heart sound classification using ensemble learning and deep CNN features,” *SN Computer Science*, vol. 6, p. 4626, 2025, doi: 10.1007/s42979-025-04626-6.
[14] Q. Zhao, “Deep learning in heart sound analysis: A comprehensive review,” *High-Dimensional Data Analysis*, vol. 1, no. 182, pp. 1–28, 2024, doi: 10.34133/hds.0182.
[15] R. Singh et al., “Deep learning for auscultation signal processing: A survey,” *IEEE Reviews in Biomedical Engineering*, vol. 17, pp. 200–221, 2024.
[16] L. Ramachandran and D. P. Rajan, “Audio signal classification using Mel frequency cepstral coefficients and deep learning,” *IEEE Access*, vol. 10, pp. 13567–13579, 2022.
[17] J. Woo and H. Kim, “Fusion of MFCC and mel spectrograms for improved audio classification,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 4, pp. 2701–2712, 2024.
[18] Chakraborty and S. Mukherjee, “CNN and RNN hybrid network for sound-event recognition,” *Pattern Recognition Letters*, vol. 147, pp. 25–33, 2022.
[19] H. Xu and J. Zhao, “Comparative study of deep learning models for biomedical sound classification,” *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 2, pp. 612–621, 2024.
[20] F. Akbar et al., “MFCC based heart sound emotion classification using deep neural networks,” *Scientific Reports*, vol. 13, p. 4872, 2023.
[21] Ahmed and M. F. Anwar, “Attention-based CNN for cardiac audio detection,” *Medical Image Analysis*, vol. 79, p. 102620, 2025.