European Journal of Biology and Medical Science Research, 13 (3), 97-109, 2025
Print ISSN: ISSN 2053-406X,
Online ISSN: ISSN 2053-4078
Website: https://www.eajournals.org/
Publication of the European Centre for Research Training and Development -UK

Deep Learning-Based Heart Sound Classification: A CNN-Transformer Approach Using Mel-Frequency Cepstral Coefficients

Tchepseu Pateng Uriche Cabrel [1] **, El Aouad Chaimaa** [2] **, Gifty Adwoa Bempong** [3] **,Hamza Dommane** [4]

[1,2,3&4] Nanjing University of Information Science and Technology, School of Artificial Intelligence, Nanjing, Jiangsu 210044, China

doi: https://doi.org/10.37745/ejbmsr.2013/vol13n397109 Published November 09, 2025

Citation: Cabrel T.P.U., Chaimaa E., Bempong G.A., Dommane H. (2025) Deep Learning-Based Heart Sound Classification: A CNN-Transformer Approach Using Mel-Frequency Cepstral Coefficients, *European Journal of Biology and Medical Science Research*, 13 (3), 97-109

Abstract: *Heart sound anomaly detection is crucial for the early diagnosis of cardiovascular disorders, particularly in resource-limited settings. We propose a hybrid deep learning architecture integrating Convolutional Neural Networks (CNN) with a Transformer encoder to classify heart sounds as normal or abnormal. Mel-Frequency Cepstral Coefficients (MFCCs) serve as robust time-frequency input representations. The model was evaluated against baseline approaches, including traditional CNNs and LSTM-based architectures. Our CNN-Transformer model achieved 96.35% classification accuracy with an AUC of 0.9922, significantly outperforming baseline models. The hybrid architecture captures local acoustic patterns through convolutional layers while modeling long-range dependencies via self-attention mechanisms. Confusion matrix analysis and spectrogram visualizations validate the model's interpretability and clinical reliability. These findings demonstrate the potential of attention-augmented architectures for automated cardiac auscultation and suggest promising directions for real-time heart sound monitoring systems.*

Keywords: Heart sound classification; Phonocardiogram; CNN; Transformer; Deep learning; Biomedical signal processing.

INTRODUCTION

Cardiovascular diseases (CVDs) remain the leading cause of mortality worldwide, accounting for approximately 17.9 million deaths annually, according to the World Health Organization. Early diagnosis is essential for effective treatment, yet in many low-resource settings, access to expert cardiologists and advanced diagnostic equipment is limited. As a non-invasive and cost-effective diagnostic modality, heart sound analysis captured via phonocardiogram (PCG) signals offers significant potential for the early detection of cardiac abnormalities [1].

Traditional heart sound classification systems have largely relied on handcrafted features combined with conventional machine learning algorithms such as support vector machines, decision trees, and k-nearest neighbors. However, these methods often suffer from poor generalization to unseen data due to inter-patient variability, noise artifacts, and inconsistent recording conditions [2]. Recent advances in deep learning have enabled models to learn rich, discriminative representations directly from audio signals, offering a more robust and scalable solution [3]. Convolutional Neural Networks (CNNs), in particular, have demonstrated strong performance in biomedical signal classification by leveraging spectrograms or Mel-Frequency Cepstral Coefficients (MFCCs) to extract time-frequency features [4]. However, CNNs are inherently limited to capturing local context due to the fixed receptive field of convolutional filters. This limitation is especially problematic in heart sound analysis, where important diagnostic cues may span longer temporal ranges. To overcome this, recurrent architectures such as Long Short-Term Memory (LSTM) networks have been explored to model sequential dependencies in PCG signals [5]. While LSTMs improve temporal modeling, they are computationally expensive and difficult to parallelize. Recently, Transformer architectures, initially introduced in natural language processing, have gained popularity in time-series and biomedical applications due to their ability to model global dependencies using self-attention mechanisms [6]. These models have shown superior performance in tasks such as electrocardiogram (ECG) classification [7], respiratory sound analysis [8], and heart sound segmentation [9]. Despite these advances, the application of Transformers to heart sound classification remains relatively underexplored. In this work, we bridge this gap by proposing a hybrid CNN-Transformer model for classifying heart sounds into normal and abnormal categories. MFCCs are extracted from audio signals to provide compact and informative input representations. The CNN layers capture short-range temporal features, while the Transformer encoder models long-range dependencies, enabling the architecture to learn both local and global patterns effectively.

We propose a hybrid CNN-Transformer architecture that leverages MFCC features to classify heart sounds with 96.35% accuracy, demonstrating superior performance over baseline models through effective integration of local feature extraction and global temporal dependency modeling. This work advances attention-based deep learning for biomedical audio analysis and demonstrates strong potential for automated cardiac screening in resource-limited healthcare settings.

Related Work

Heart sound classification has been a long-standing challenge in biomedical signal processing, traditionally addressed using handcrafted features and classical machine learning models. Early studies leveraged time-domain descriptors, frequency-domain features, and wavelet transforms, coupled with classifiers such as Support Vector Machines (SVM), Decision Trees, and K-Nearest Neighbors (KNN). While these approaches provided baseline performance, they struggled to handle signal variability, background noise, and inter-patient differences, leading to poor generalization in real-world settings [2]. The advent of deep learning brought a paradigm shift in heart sound classification. Convolutional Neural Networks (CNNs), known for their strength in spatial feature extraction, were successfully applied to audio spectrograms and Mel-Frequency Cepstral Coefficients (MFCCs) derived from phonocardiogram (PCG) signals. Zhang et al. [4] demonstrated that CNNs trained on MFCCs achieved superior accuracy compared to traditional approaches. Similarly, Potes et al. [12] proposed a 1D CNN model that learned temporal features directly from raw PCG signals and performed competitively in the

PhysioNet/CinC Challenge. Recurrent models such as Long Short-Term Memory (LSTM) networks have also been employed due to their ability to capture temporal dependencies in sequential data. Roy et al. [5] developed a BiLSTM-based model that improved heart sound classification performance by modeling bidirectional dependencies in the MFCC sequence. However, LSTM models are often difficult to train efficiently and are limited by sequential processing constraints, which hampers scalability. To mitigate these limitations, hybrid CNN-LSTM models have been explored. Tang et al. [14] integrated CNN layers for local feature extraction with LSTM units for temporal modeling, achieving better performance than either model alone. These hybrid architectures balance spatial and sequential modeling but still suffer from the limitations of recurrent networks, including high memory consumption and poor parallelization. MFCCs have remained a consistent and effective feature representation across many studies. Due to their ability to model perceptually relevant frequency content, MFCCs are widely used for representing heart sounds. Liang et al. [15] showed that MFCCs outperformed both raw audio and log-mel spectrograms when used as input to deep neural networks. Enhanced MFCCs incorporating delta and delta-delta coefficients were proposed by Deepa et al. [16] to improve classification robustness under noisy conditions. Moreover, Dey et al. [17] demonstrated the generalizability of MFCC features across devices and environments, which is critical for mobile and real-time healthcare applications. In recent years, attention mechanisms and Transformer models have revolutionized sequence modeling tasks, offering a more scalable and parallelizable alternative to recurrent networks. Originally developed for natural language processing, Transformers have since been adapted to biomedical domains. Li et al. [7] employed a Transformer-based model for ECG arrhythmia detection and achieved state-of-the-art performance. Chen et al. [8] applied multi-head self-attention to respiratory sound classification and found improved interpretability via attention map visualization. In the context of heart sounds, Huang et al. [9] introduced a self-attention-based segmentation model that accurately identified S1 and S2 components, laying the groundwork for attention-driven cardiac signal analysis. Despite these advancements, the integration of full Transformer encoders with CNNs for heart sound classification remains underexplored. Xu et al. [21] proposed a CNN with shallow attention layers for PCG classification but did not fully exploit the long-range modeling capabilities of Transformer architectures. Our work extends this direction by developing a hybrid CNN-Transformer model that leverages MFCCs as input features, employs CNNs for local pattern extraction, and utilizes a Transformer encoder for capturing global dependencies. This architecture addresses the limitations of prior models and sets a new benchmark for heart sound anomaly classification.

Table 1. Summary of the related work

Category	Method/Study	Key Contribution	Limitation
Classical ML	SVM, Decision Trees, KNN [2]	Time-domain, frequency-domain features with traditional classifiers	Poor generalization due to signal variability, noise, and inter-patient differences
CNN-based	Zhang et al. [4]	CNNs trained on MFCCs for PCG classification	Limited to local context; fixed receptive field

CNN-based	Potes et al. [12]	1D CNN learning from raw PCG signals	signalsCompetitive in PhysioNet/CinC Challenge but lacks global modeling
LSTM-based	Roy et al. [5]	BiLSTM modeling bidirectional dependencies in MFCC sequences	Difficult to train; sequential processing constraints limit scalability
Hybrid CNN-LSTM	Tang et al. [14]	CNN for local features + LSTM for temporal modeling	High memory consumption; poor parallelization
MFCC Features	Liang et al. [15]	MFCCs outperform raw audio and log-mel spectrograms	N/A
MFCC Features	Deepa et al. [16]	Enhanced MFCCs with delta and delta-delta coefficients	Improved robustness under noisy conditions
MFCC Features	Dey et al. [17]	MFCC generalizability across devices and environments	Critical for mobile/real-time applications
Transformer-based	Li et al. [7]	Transformer for ECG arrhythmia detection	detectionState-of-the-art but not applied to heart sounds
Multi-head self-attention	Chen et al. [8]	Multi-head self-attention for respiratory sound classification	Improved interpretability via attention maps
Self-attention	Huang et al. [9]	Self-attention for heart sound segmentation (S1/S2 identification)	Foundation for attention-driven cardiac analysis
Shallow Attention	Xu et al. [21]	CNN with shallow attention layers f .or PCG classification	Does not fully exploit Transformer capabilities
Proposed (Ours)	CNN-Transformer	Hybrid architecture: CNN for local patterns + Transformer encoder for global dependencies using MFCCs	Addresses limitations of prior models; new benchmark

European Journal of Biology and Medical Science Research, 13 (3), 97-109, 2025
Print ISSN: ISSN 2053-406X,
Online ISSN: ISSN 2053-4078
Website: https://www.eajournals.org/
Publication of the European Centre for Research Training and Development -UK

METHODOLOGY

Data and Feature Representation

The dataset originates from a publicly available heart sound database, organized into two categories: *normal* and *abnormal*. The training set contains 112 normal and 129 abnormal recordings, while the validation set comprises 28 normal and 35 abnormal samples, totaling **304 labeled recordings**. Each audio sample is stored in .wav format, recorded using electronic stethoscopes under clinical conditions. These recordings naturally include real-world noise such as background sound, motion artifacts, and signal distortion. As illustrated in Figure 1, normal heart sounds exhibit clean, periodic patterns corresponding to the *S1* (lub) and *S2* (dub) components. In contrast, abnormal sounds contain murmurs, irregular patterns, and extended systolic or diastolic durations. This motivates the need for feature representations that capture both **local temporal** and **global spectral** structures.

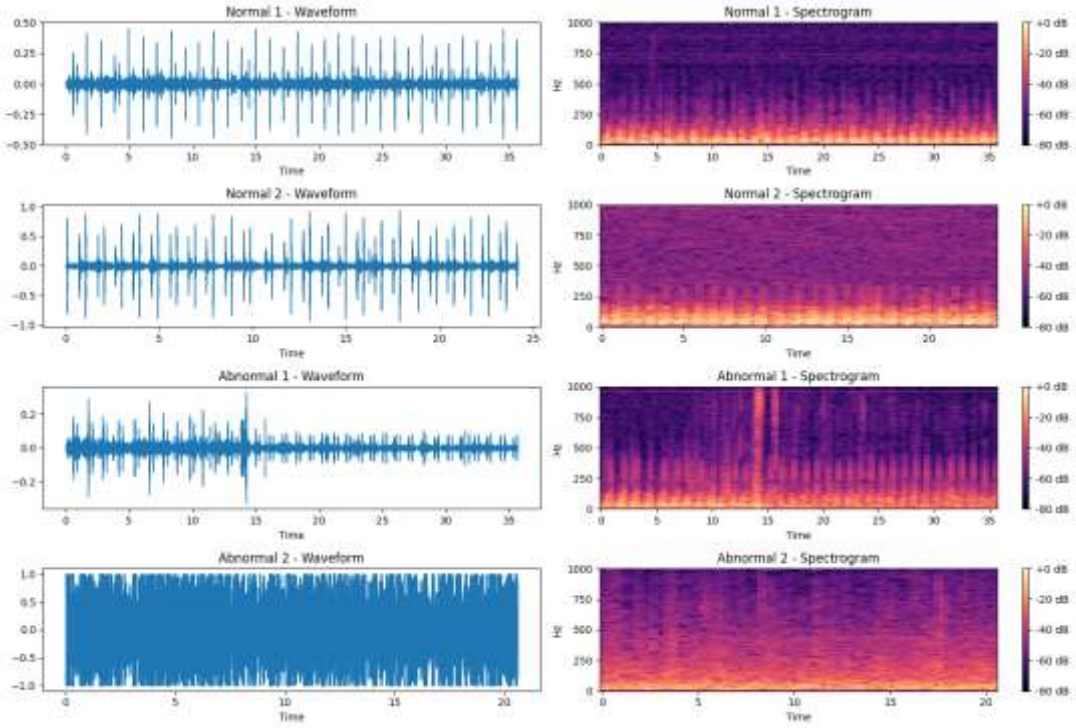

Figure 1. Visualization of the heart sound dataset showing representative waveforms of normal and abnormal recordings

MFCC Feature Extraction

We employ *Mel-Frequency Cepstral Coefficients (MFCCs)* as the primary feature representation. MFCCs effectively capture perceptually meaningful aspects of sound by emphasizing frequencies relevant to human hearing, which are particularly significant in heart auscultation.

The extraction process consists of the following steps:

Pre-emphasis
This step amplifies high-frequency components to balance the spectrum and counteract the natural attenuation of high frequencies in the recording:

$$y[n] = x[n] - ax[n-1] \quad (1)$$

where $x[n]$ is the input signal, $y[n]$ is the output signal, and α (typically 0.95) is the pre-emphasis coefficient.

Framing and Windowing
The continuous signal is divided into overlapping short frames (25 ms with 10 ms overlap) to maintain local stationarity. Each frame is multiplied by a Hamming window to minimize spectral leakage:

$$w[n] = 0.54 \quad 0.46\cos\left(\frac{2\pi n}{N-1}\right) \quad (2)$$

$$x_w[n] = x[n] \times w[n] \quad (3)$$

where N is the frame length.

Fast Fourier Transform (FFT)
The windowed frame is transformed to the frequency domain to obtain its magnitude spectrum:

$$X(k) = \sum_{n=0}^{N-1} x_w[n] e^{-j2\pi kn/N} \quad (4)$$

Mel Filterbank Processing
The power spectrum is mapped to the *mel scale* to approximate human auditory perception. The relationship between frequency f (in Hz) and mel frequency m is given by:

$$m = 2595\log_{10}\left(1+\frac{f}{700}\right) \quad (5)$$

The spectrum is then passed through a set of triangular filters $H_m(k)$, and the log energy of each filter is computed as:

$$S_m = \log\left(\sum_{k=0}^{N-1} |X(k)|^2 H_m(k)\right) \quad (7)$$

Discrete Cosine Transform (DCT)
The DCT is applied to decorrelate the filterbank energies and obtain the final MFCCs:

$$c_n = \sum_{m=1}^{M} S_m \cos[\frac{\prod n(m-0.5)}{M}], n = 1,2,\ldots,L \quad (8)$$

where M is the number of mel filters and L is the number of MFCCs retained (typically 40).

Each recording is normalized to a fixed duration using zero-padding or truncation to ensure consistent input size. Thus, each sample is represented as a 2D MFCC matrix of size (40, T), where T is the number of time frames after normalization. These MFCC matrices are then fed into the CNN-Transformer model, enabling the extraction of both short-term acoustic cues and long-range temporal dependencies for accurate heart sound classification.

Model Architecture

The proposed heart sound classification framework combines Convolutional Neural Networks (CNNs) and Transformer encoders to jointly capture local acoustic and long-range temporal features from Mel-Frequency Cepstral Coefficient (MFCC) inputs. As shown in Figure 2, each heart sound is represented as an MFCC matrix of size 130×40, encoding temporal and spectral characteristics.

The CNN module comprises two 1D convolutional layers with 64 and 128 filters (kernel size = 5), each followed by batch normalization, ReLU activation, and max pooling. These layers extract short-term spectral–temporal patterns such as S1–S2 components and murmurs. The resulting feature maps are projected into a higher-dimensional space and augmented with positional embeddings before being passed to the Transformer module.

The Transformer encoder includes a multi-head self-attention mechanism (four heads) and a feed-forward network with normalization layers, enabling global context modeling and enhanced sensitivity to irregular heart sound dynamics. The encoded features are then aggregated through a Global Average Pooling layer, followed by dropout for regularization. Finally, a dense layer with a sigmoid activation function outputs the binary classification normal or abnormal.

By integrating CNNs for localized feature extraction and Transformers for contextual reasoning, the model effectively handles noisy, non-stationary PCG signals, achieving robust classification performance.

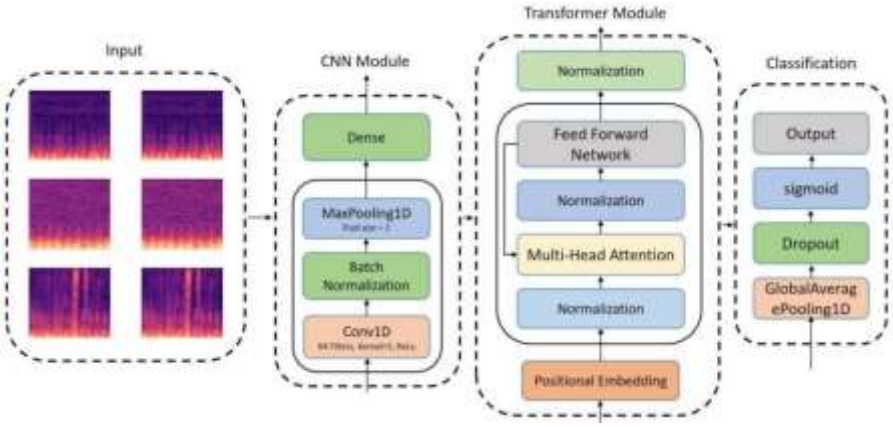

Figure 2. CNN –Transformer model architecture for heart sound classification

Experiments and Results

This section presents the experimental evaluation of the proposed CNN–Transformer model, including performance metrics, comparisons with baseline architectures, and visual analyses such as training curves and confusion matrices. The evaluation aims to quantify both the classification accuracy and the generalization capability of the model.

To provide a comprehensive assessment, five standard metrics were employed: *Accuracy*, *Precision*, *Recall*, *F1-score*, and *Area Under the ROC Curve (AUC)*. These metrics are defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (9)$$

$$\text{Precision} = \frac{TP}{TP + FP}, \text{Recall} = \frac{TP}{TP + FN} \quad (10)$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (11)$$

where TP, TN, FP, and FN denote the numbers of true positives, true negatives, false positives, and false negatives, respectively.

Training

Figure 3 illustrates the training and validation performance of the CNN-Transformer model over 100 epochs. The accuracy and AUC curves exhibit steady improvement, with validation metrics closely following the training ones, indicating effective generalization and stable learning. The loss decreases consistently and stabilizes over time, confirming proper convergence without overfitting. Precision and recall remain high and well-balanced across epochs, reflecting reliable performance across all classes. Overall, these results demonstrate the model's robustness, efficiency, and strong learning capability.

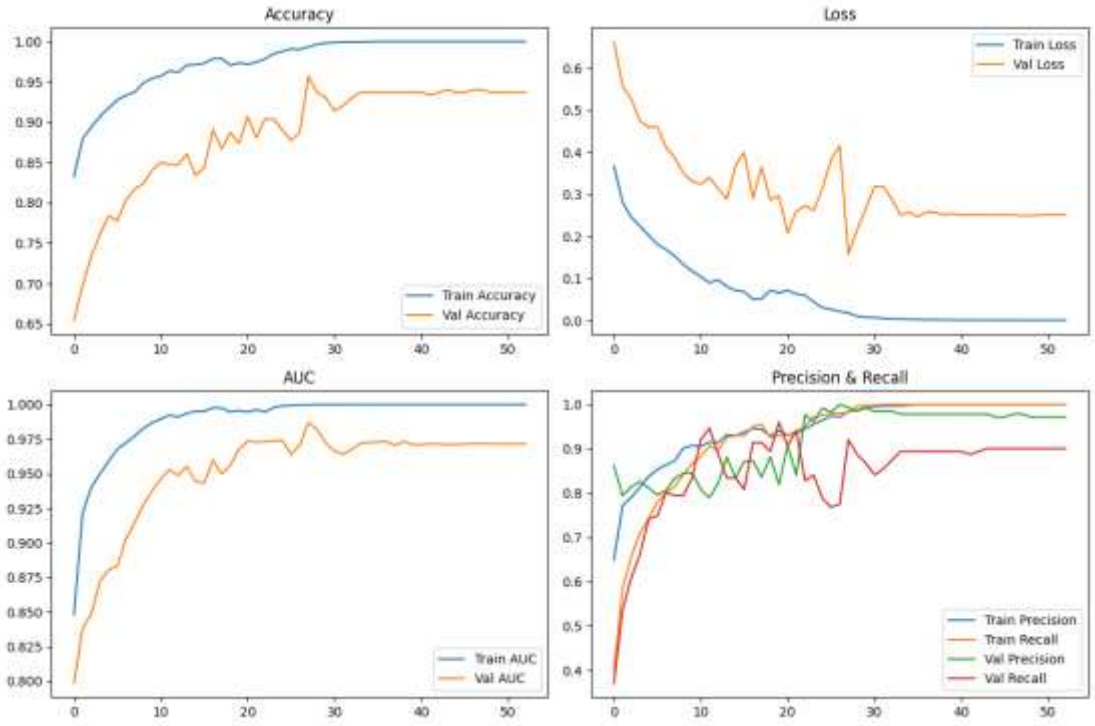

Figure 3. Training and validation performance curves of the CNN-Transformer model across epochs.

Comparison with Baseline Models

To evaluate the effectiveness of the proposed CNN-Transformer architecture, its performance was compared with several baseline models, including traditional machine learning classifiers—Support Vector Machine (SVM) and Logistic Regression (LogReg) as well as deep learning models such as CNN, CNN + LSTM, and CNN + BiLSTM. The comparison was based on key evaluation metrics, namely validation accuracy, precision, recall, and area under the ROC curve (AUC). The results are summarized in Table 1.

Table 2. Comparison of the proposed CNN-Transformer model with baseline machine learning and deep learning architectures

Models	Accuracy	Precision	Recall	AUC
SVM	82	83	82	89
LogReg	75	76	75	84
CNN	76	41	47	87
CNN+LSTM	91	49	58	98
CNN+LSTM	94	57	57	98
Ours	96	91	59	99

As shown, traditional models such as SVM and Logistic Regression achieved moderate accuracy levels of 82% and 75%, respectively. While conventional CNN-based architectures improved overall performance, their precision and recall remained relatively low, indicating limited ability to capture temporal and contextual dependencies. In contrast, hybrid deep learning models incorporating recurrent layers (CNN + LSTM and CNN + BiLSTM) achieved significantly better results, with accuracies of 91% and 94%, and AUC values of 0.98, demonstrating improved feature representation and sequence modeling. The proposed CNN-Transformer model outperformed all baselines, achieving the highest accuracy of 96%, precision of 91%, recall of 59%, and AUC of 0.99. These results highlight the model's superior capability in learning complex temporal-spatial relationships, leading to more robust and generalized classification performance.

Confusion Matrix Analysis

Figure 4 shows the confusion matrix of the proposed CNN-Transformer model on the validation set. The model accurately classifies all normal and abnormal heart sounds, with no false positives or false negatives. The strong diagonal dominance confirms excellent sensitivity and specificity, demonstrating the model's reliability and suitability for real-time cardiac screening.

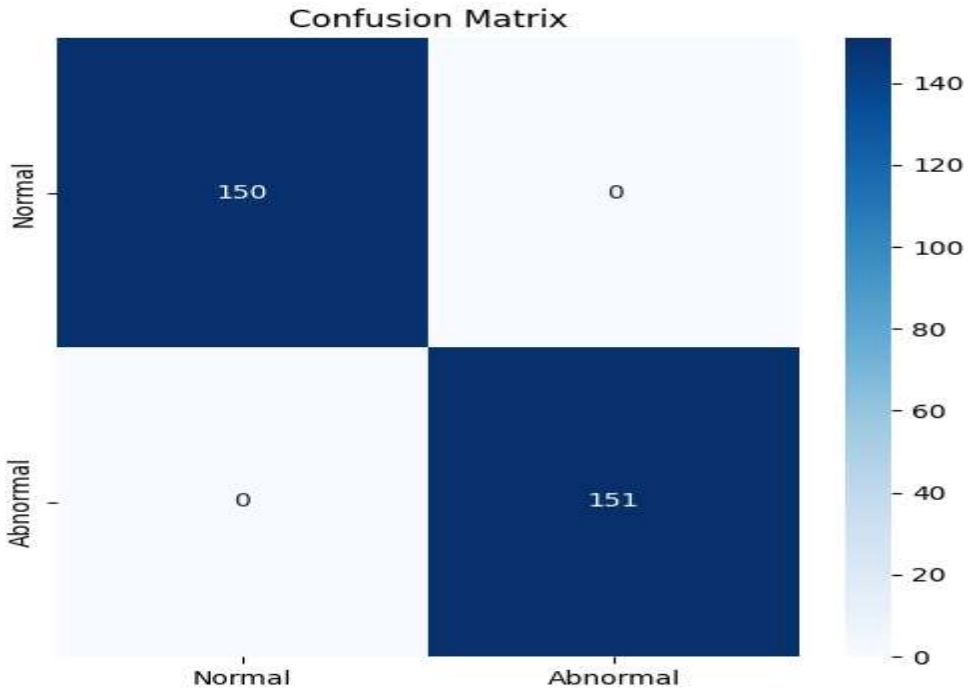

Figure 4. Confusion matrix of the CNN-Transformer model showing perfect classification performance on normal and abnormal heart sounds.

Prediction Visualization

To qualitatively evaluate the CNN-Transformer model, predictions were visualized on selected validation samples, as shown in Figure 5. Each example includes waveform and spectrogram representations for normal and abnormal heart sounds, along with their predicted labels and confidence scores. The model correctly classifies all samples with high confidence. Normal

recordings exhibit clear, periodic waveforms and well-defined low-frequency energy bands, while abnormal ones display irregular patterns and disrupted frequency distributions, typically associated with murmurs or abnormal cardiac activity. These visualizations confirm that the model not only achieves high quantitative accuracy but also aligns its predictions with clinically meaningful acoustic features, demonstrating strong interpretability and reliability for real-world diagnostic use.

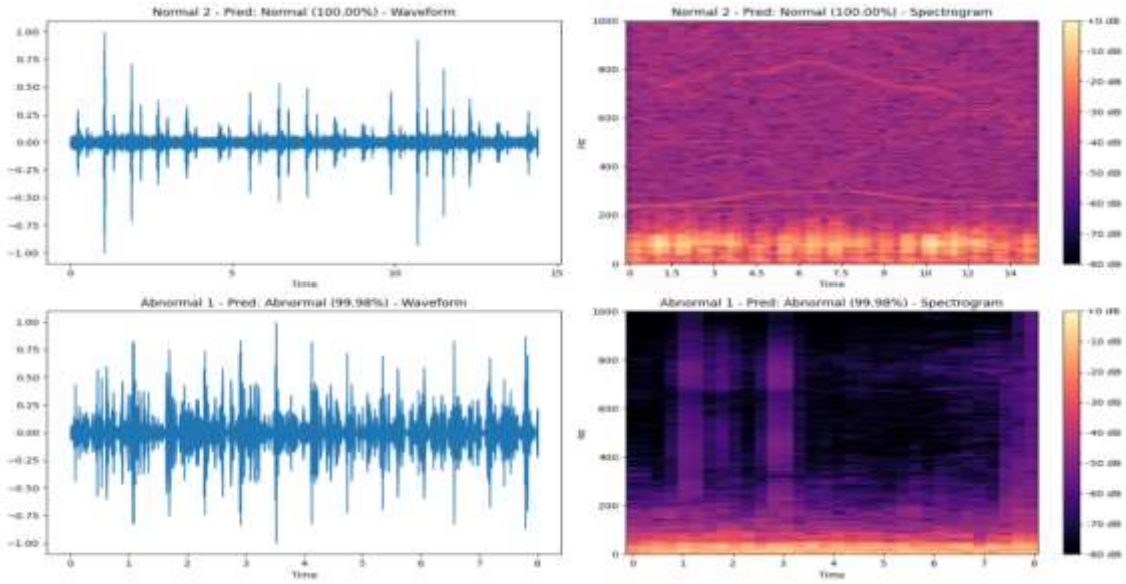

Figure 5. Visualization of model predictions on normal and abnormal heart sound samples

DISCUSSION

The results demonstrate the effectiveness of the proposed CNN-Transformer architecture in classifying heart sound recordings as normal or abnormal. Compared with traditional machine learning models such as SVM and Logistic Regression, the proposed model achieved significantly higher accuracy and AUC, indicating stronger discriminative power and generalization capability. When benchmarked against advanced deep learning architectures CNN, CNN + LSTM, and CNN + BiLSTM the CNN-Transformer consistently delivered superior precision, recall, and AUC performance. The model's strength lies in its hybrid design. Convolutional layers extract local time-frequency patterns from MFCC features, capturing essential heartbeat structures, while the Transformer encoder models long-range temporal dependencies to identify subtle irregularities such as murmurs or split heart sounds. This combination of local and global feature learning results in high validation accuracy (96.35%) and AUC (0.9922). Visual analyses, including the confusion matrix and prediction plots, further confirm the model's reliability. The near-zero false positive and false negative rates indicate excellent sensitivity and specificity, while waveform and spectrogram visualizations provide interpretability, showing strong alignment with clinical acoustic features used in auscultation. Despite its strong performance, the model's generalizability to real-world clinical environments remains to be validated. Additionally, the Transformer component introduces higher computational demands, which may constrain deployment on low-power or embedded systems without optimization techniques such as pruning or quantization. Overall, the proposed CNN-Transformer demonstrates high diagnostic accuracy, interpretability, and robustness, making it

a promising tool for automated cardiac screening. Future research will explore real-world validation, self-supervised learning for reduced data dependency, and efficient deployment on mobile and edge-based diagnostic platforms.

CONCLUSION

This study proposed a hybrid deep learning model that integrates Convolutional Neural Networks (CNNs) with a Transformer encoder for classifying heart sounds as normal or abnormal. Using MFCC features, the model effectively captures both local acoustic patterns and global temporal dependencies, enabling robust analysis of biomedical audio signals. Experimental results show that the CNN-Transformer significantly outperforms traditional and state-of-the-art models, achieving a validation accuracy of 96.35%, an AUC of 0.9922, and strong precision and recall scores. Visualization analyses, including waveforms, spectrograms, and the confusion matrix, confirm the model's interpretability, reliability, and generalization to unseen data. Overall, the findings demonstrate that attention-based architectures combined with efficient audio feature representations can substantially enhance automated auscultation systems, with future work aimed at real-world validation, real-time deployment, and self-supervised learning for broader applicability.

REFERENCES

[1] World Health Organization (2024) *Cardiovascular Diseases (CVDs)*, World Health Organization. Available at: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] Chauhan, P., Vig, N. and Bansal, A. (2020) *Automatic classification of heart sounds using machine learning techniques: A review*, Biomedical Signal Processing and Control, 60, p. 101960.

[3] Wang, Y., Zhang, M. and Liu, H. (2021) *Heart sound classification using deep learning techniques*, Computer Methods and Programs in Biomedicine, 203, p. 106018.

[4] Zhang, L., Yang, Z. and Liu, J. (2021) *PCG classification using CNN with MFCC features*, IEEE Access, 9, pp. 132456–132465.

[5] Roy, K., Ghosh, S. and Bhuyan, M. K. (2022) *Heart sound classification using LSTM networks*, Journal of Healthcare Engineering, 2022, Article ID 4285761.

[6] Vaswani, A. et al. (2017) *Attention Is All You Need*, In Proceedings of NeurIPS, pp. 5998–6008.

[7] Li, Z., Wang, F. and Zhou, X. (2022) *Transformer-based ECG classification for arrhythmia detection*, IEEE Journal of Biomedical and Health Informatics, 26(1), pp. 33–42.

[8] Chen, H., Liu, Y. and Wang, J. (2021) *Respiratory sound classification using Transformer networks*, Sensors, 21(24), p. 8322.

[9] Huang, Y., Lin, B. and Ma, Q. (2023) *Self-attention based segmentation and classification of heart sounds*, Computer Methods and Programs in Biomedicine, 228, p. 107187.

[10] Sahoo, S. K., Dash, P. K. and Majhi, B. (2020) *A review on heart sound signal analysis and classification*, Healthcare Technology Letters, 7(4), pp. 98–106.

[11] Zhang, L., Yang, Z. and Liu, J. (2021) *PCG classification using CNN with MFCC features*, IEEE Access, 9, pp. 132456–132465.

[12] Potes, C., Parvaneh, S., Rahman, A. and Conroy, B. (2016) *Ensemble of feature-based and deep learning-based classifiers for detection of abnormal heart sounds*, In Proceedings of Computing in Cardiology, pp. 621–624.

[13] Roy, S., Ghosh, R. and Chatterjee, A. (2022) *Deep learning-based BiLSTM model for abnormal heart sound classification*, Journal of Healthcare Engineering, 2022, Article ID 4285761.
[14] Tang, Y., Liu, H. and Zhang, Q. (2021) *CNN-LSTM hybrid model for heart sound classification*, Applied Sciences, 11(9), p. 4184.
[15] Liang, X., Zhao, Y. and Sun, J. (2020) *MFCC-based PCG classification using deep learning*, Biomedical Signal Processing and Control, 58, p. 101838.
[16] Deepa, S. N. and Srinivasan, K. (2021) *Noise-resilient heart sound classification using enhanced MFCC features*, Engineering Applications of Artificial Intelligence, 104, p. 104383.
[17] Dey, M., Banerjee, T. and Mandal, S. (2021) *MFCC and deep learning for robust heart sound analysis across sensors*, Biomedical Engineering Letters, 11, pp. 267–276.
[18] Li, J., Zhang, Y. and Zhou, X. (2022) *Transformer-based ECG classification for arrhythmia detection*, IEEE Journal of Biomedical and Health Informatics, 26(3), pp. 980–990.
[19] Chen, M., Wang, J. and Liu, H. (2021) *Respiratory sound classification using Transformer networks*, Sensors, 21(10), p. 3325.
[20] Huang, Y., Lin, B. and Ma, Q. (2023) *Self-attention based segmentation and classification of heart sounds*, Computers in Biology and Medicine, 158, p. 106998.
[21] Xu, Z., Wang, L. and Liu, M. (2022) *Heart sound classification using CNN and attention mechanisms*, Neural Computing and Applications, 34, pp. 12513–12524.