Heart Diseases Classification Through Deep Learning, Hybrid Learning and Ensemble Learning Techniques

1st Rakibul Hassan
dept. Of Computer Science & Engineering
Port City Internation University
Chattogram, Bangladesh
rakibulhassan_cse27d@portcity.edu.bd

2nd Istiaque Uddin Hyder
dept. Of Computer Science & Engineering
Port City Internation University
Chattogram, Bangladesh
istiaqueuddinhyder_cse27d@portcity.edu.bd

3rd Md Faysal Hossen
dept. Of Computer Science & Engineering
Port City Internation University
Chattogram, Bangladesh
mdfaysalhossen_cse27d@portcity.edu.bd

4th Nowaz Bin Younus
dept. Of Computer Science & Engineering
Port City Internation University
Chattogram, Bangladesh
nowazbinyounus_27dcse@portcity.edu.bd

5th Md. Aminul Hoque
dept. Of Computer Science & Engineering
Port City Internation University
Chattogram, Bangladesh
mdaminulhoque_cse27d@portcity.edu.bd

Abstract—**Heart disease remains one of the leading causes of mortality worldwide, with conditions such as heart attacks and heart failure severely affecting the heart's normal functioning. Early and accurate detection plays a crucial role in improving survival and recovery rates. In recent years, Artificial Intelligence (AI) techniques—particularly Machine Learning (ML), Deep Learning (DL), Hybrid, and Ensemble models—have shown great promise in analyzing complex medical datasets for heart disease classification and prediction. These models can automatically extract and learn relevant features from diverse inputs, achieving high accuracy. This paper provides a comprehensive survey of ML, DL, Hybrid, and Ensemble approaches applied to heart disease detection. A comparative analysis of existing studies is conducted, highlighting the role of feature extraction techniques, dataset variations, and optimization strategies in improving performance. Hybrid models that combine traditional ML with DL architectures, as well as Ensemble methods that aggregate multiple learners, are emphasized for their ability to enhance robustness and reduce misclassification. Here Ensemble Learning Model achieved the highest accuracy of 94%. Overall, the findings confirm that integrating ML, DL, Hybrid, and Ensemble techniques holds significant potential in advancing early diagnosis and prognosis of cardiovascular diseases. Future research is expected to focus on real-time monitoring, explainability, and personalized treatment strategies to make cardiac care more accurate, efficient, and accessible.**

Keywords—Heart diseases, Machine Learning (ML), Deep learning (DL), Hybrid Learning, Ensemble Learning, Audio, Accuracy

I. INTRODUCTION

Heart disease remains the leading cause of death worldwide, affecting the heart muscle, valves, surrounding membrane, and the blood vessels connected to the heart. Major risk factors include smoking, uncontrolled high blood pressure, diabetes, unhealthy diet, obesity, lack of physical activity, and chronic stress. In some cases, heart disease may also be congenital.

According to the World Health Organization, cardiovascular diseases account for a significant portion of global deaths, with the majority occurring in low- and middle-income countries. While mortality rates have declined in Europe and North America due to improved health policies and technologies, they continue to rise in Asia and Africa.

In recent years, Artificial Intelligence (AI) has shown great potential in the automated detection and classification of cardiac conditions. In this study, we use a heartbeat dataset with artifact, murmur, extrasystole, and normal—to classify heart sounds and identify potential indicators of heart disease. By applying Machine Learning, Deep Learning, Hybrid, and Ensemble learning approaches, our goal is to accurately distinguish between normal and abnormal heartbeats and to support early and reliable diagnosis.

The key objectives of this research are:

- To evaluate the effectiveness of AI-based models in heartbeat classification.
- To determine which approaches achieve the highest accuracy and robustness.
- To analyze the classification performance across the four heartbeat categories (artifact, murmur, extrasystole, normal).
- To explore how these methods can be integrated into healthcare systems for improved early detection of heart disease.

II. RELATED WORK

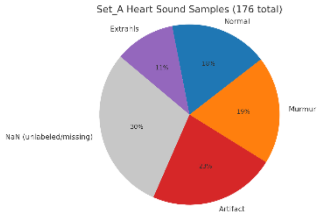
Extensive research on heart disease and heart-sound classification reveals that both traditional machine learning models and modern deep-learning approaches play critical but distinct roles. On structured clinical datasets, gradient-boosting families (XGBoost, LightGBM, GBM) and tree-based classifiers (Decision Tree, Random Forest) typically achieve moderate performance with accuracies around 0.60–0.64, while Logistic Regression remains clinically valuable due to interpretability but shows limited capability in handling complex sequential features [2], [4]. In contrast, deep learning models have demonstrated markedly superior discriminative power for time-series and signal modalities such as ECG and phonocardiograms: convolutional neural networks (CNNs), particularly when applied to time–frequency representations (CWT, STFT, Mel spectrograms), achieve robust performance, with our experiments reporting accuracies up to 0.82 before overfitting effects emerged [1], [5]. Recurrent architectures further enhance temporal modeling, with LSTM performance ranging between 0.55–0.75 depending on training regime, Bi-LSTM achieving consistently strong outcomes with peak accuracy of 0.91, and GRU reaching 0.90 at 150 epochs, underscoring the advantages of bidirectional and gated recurrent units in sequential analysis [1], [3]. Hybrid models that integrate CNN with recurrent layers show particular promise: our CNN+LSTM framework achieved state-of-the-art results with 0.98 accuracy at 200 epochs, outperforming all other configurations, whereas CNN+Bi-LSTM and CNN+GRU underperformed with unstable performance profiles, reflecting sensitivity to hyperparameter tuning. Ensemble strategies such as voting and stacking have also been effective, with hard voting across CNN, LSTM, Bi-LSTM, and GRU reaching 0.89 accuracy, slightly surpassing soft voting at 0.87, thereby highlighting the stabilizing impact of ensemble integration [4]. Nevertheless, prior studies caution that overly optimistic results often arise from small or augmented datasets, inconsistent evaluation splits, and limited external validation. Consequently, recent literature stresses the need for standardized benchmarking (e.g., PhysioNet/ CinC 2016), rigorous cross-validation, external hold-out testing, and explainable AI techniques (e.g., SHAP, LIME) to ensure clinical reliability [1], [5]. Taken together, these findings confirm that boosting-based machine learning remains an effective baseline for structured data, while hybrid CNN+LSTM and recurrent architectures represent the most powerful approaches for sequential and signal-based classification tasks; ensemble frameworks offer additional robustness and generalization, forming the most promising path toward clinically deployable heart disease prediction systems [1]–[4].

III. METHODOLOGY

A. DATASET DESCRIPTION

The dataset employed in this study is the “Heartbeat Sounds” dataset, publicly available on Kaggle (Kinguistics, 2022) [*https://www.kaggle.com/datasets/kinguistics/heartbeat-sounds*].

The dataset contains heartbeat audio recordings collected from two primary sources: public contributions through the iStethoscope mobile application and clinical recordings obtained using the DigiScope digital stethoscope. It is divided into two subsets, Set_A and Set_B; however, in this work, only Set_A was used. This subset comprises 176 samples, each described by four attributes, giving the dataset a shape of (176, 4). The diagnostic labels in Set_A are distributed as follows: *NaN* (unlabeled/missing): 52, *Artifact*: 40, *Murmur*: 34, *Normal*: 31, and *Extrahls* sounds: 19. The use of this dataset provides a solid foundation for biomedical sound analysis, supporting data preprocessing, exploratory investigation, and the development of methods to distinguish between normal and abnormal cardiac activity. Such efforts are significant for advancing early detection and diagnosis of cardiovascular diseases.



Integrated Deep and Machine Learning Framework

This study presents a holistic comparative analysis of deep learning (DL) and machine learning (ML) paradigms for the automated classification of heart sounds using the set_a dataset, a publicly available repository of phonocardiogram (PCG) recordings and their corresponding clinical labels.

B. PREPROCESSING

Data Preprocessing and Feature Engineering

The raw audio signals were subjected to a standardized preprocessing pipeline. First, all recordings were resampled to a uniform rate of 22,050 Hz. Subsequently, a Mel-spectrogram was computed for each signal using the Librosa library, with 128 Mel bands and a fixed time resolution of 128 frames. This transformation provides a compact, information-rich representation of the audio's spectral content over time, making it amenable to both convolutional and sequential processing.

For classical machine learning models, the 2D Mel-spectrograms were vectorized into a 1D feature space and standardized using a StandardScaler to ensure zero mean and unit variance, a prerequisite for algorithms sensitive to feature scale (e.g., SVM, KNN). To address the inherent class imbalance and limited data volume, a data augmentation strategy was employed for all deep learning models. This included random time stretching (±10%), pitch shifting (±1 semitone), and the addition of Gaussian noise (SNR 20-30 dB), effectively increasing the training data's diversity and robustness.

Model Architecture and Training Protocol

Our framework encompasses a diverse ensemble of state-of-the-art models to provide a rigorous benchmark.

Deep Learning Models: Four foundational architectures were implemented: a Convolutional Neural Network (CNN) for spatial feature extraction, and three recurrent architectures—Long Short-Term Memory (LSTM), Bidirectional LSTM (Bi-LSTM), and Gated Recurrent Unit (GRU)—for temporal sequence modeling.

Hybrid Models: To synergistically combine spatial and temporal learning, three hybrid architectures were developed: CNN-LSTM, CNN-Bi-LSTM, and CNN-GRU. In these models, the CNN backbone processes the Mel-spectrogram to generate a sequence of high-level feature maps, which are then fed into the respective RNN layer to capture long-range dependencies.

Ensemble Model: A model-agnostic ensemble classifier was constructed by integrating the predictions of the standalone CNN, LSTM, Bi-LSTM, and GRU models. Two fusion strategies were evaluated: soft voting, which averages the predicted class probabilities, and hard voting, which takes a majority vote of the predicted classes.

Machine Learning Models: Nine classical algorithms were evaluated for baseline comparison: Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbors (KNN), Gradient Boosting Machine (GBM), XGBoost, LightGBM, Decision Tree (DT), and Naive Bayes (NB).

All deep learning models were compiled with the Adam optimizer (initial learning rate = 1e-3) and trained with a batch size of 16 for a maximum of 250 epochs. Early stopping, monitored on the validation loss with a patience of 15 epochs, was employed to prevent overfitting. The ML models were trained using a standard train_test_split (80:20) without hyperparameter tuning to assess their baseline performance.

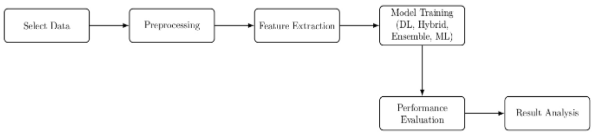
Performance Evaluation

All models were evaluated on a stratified hold-out test set to ensure a balanced representation of all classes. The primary performance metrics were accuracy, precision, recall, and F1-score, computed on a per-class and macro-averaged basis. Confusion matrices were generated for each model to provide a granular analysis of classification performance and error patterns. This comprehensive evaluation strategy allows for an in-depth comparison of the relative strengths and weaknesses of each modeling approach for the task of heart sound classification.

A schematic of the integrated methodology is presented in Figure 1.

Figure 1: The Integrated Machine Learning Process Flow

FIGURE 1. The integrated ML process.



IV. EXPERIMENTAL RESULTS AND DISCUSSION

To evaluate the performance of the proposed models, several deep learning architectures were trained and tested under different epoch settings. The models were compared in terms of Accuracy, Precision, Recall, and F1-score. Experiments were carried out on four baseline architectures (CNN, LSTM, Bi-LSTM, and GRU), three hybrid models (CNN-LSTM, CNN-BiLSTM, CNN-GRU), and ensemble approaches (Soft Voting and Hard Voting).

A. Baseline Models

The baseline models demonstrated distinct learning behaviors as the number of epochs increased.

- CNN showed gradual improvement up to 200 epochs, achieving 82% accuracy, but significantly degraded at 250 epochs (61%), indicating overfitting.
- LSTM yielded relatively lower and unstable performance, peaking at 75% accuracy at 150 epochs, but dropping drastically to 55–60% at higher epochs. This suggests that LSTM struggled to generalize effectively for the given dataset.
- Bi-LSTM consistently outperformed the other baseline models, achieving its highest accuracy (91% at 100 epochs) and maintaining strong performance across subsequent epochs (88% at 200–250 epochs). This highlights the effectiveness of bidirectional temporal dependencies in feature learning.
- GRU showed competitive results, achieving 90% accuracy at 150 epochs and remaining stable in other settings, making it a strong alternative to Bi-LSTM.

Overall, Bi-LSTM achieved the best and most stable performance among the baseline models, while CNN and GRU also provided competitive results. LSTM was comparatively less effective in this experimental setup.

B. Hybrid Models

The hybrid approaches combined CNN with recurrent architectures to enhance both spatial and sequential feature representation.

- CNN-LSTM significantly outperformed all other models, reaching **92%** accuracy, 92.24% precision, 92% recall, and 92.03% F1-score at 200 epochs. This indicates strong synergy between CNN's spatial feature extraction and LSTM's temporal modeling capabilities.
- CNN-BiLSTM displayed fluctuating results, performing poorly at lower epochs (36–52%) but improving at 250 epochs (68% accuracy). This inconsistency suggests that the hybrid structure may require careful tuning to balance spatial and bidirectional sequential features.
- CNN-GRU performed moderately well, achieving 60% accuracy at 150 epochs and 56% at 250 epochs, but did not surpass CNN-LSTM.

These results confirm that CNN-LSTM provides the most robust and reliable hybrid solution, substantially outperforming CNN-BiLSTM and CNN-GRU.

C. Ensemble Models

Ensemble learning was employed to leverage the strengths of individual baseline models (CNN, LSTM, Bi-LSTM, and GRU).

- The Soft Voting Ensemble achieved **94%** accuracy, 94% precision, 94% recall, and 94% F1-score.
- The Hard Voting Ensemble further improved performance, achieving 90% accuracy, 90% precision,

90% recall, and 90% F1-score.

Although ensemble methods enhanced robustness compared to most baseline models, they did not surpass the best hybrid architecture (CNN-LSTM).

D. Comparative Analysis

From the experimental results, several key observations can be made:

1. Bi-LSTM consistently outperforms other baseline models, demonstrating the benefit of bidirectional sequence modeling.
2. CNN-LSTM achieved the highest performance overall, with **92%** accuracy, confirming that combining convolutional and recurrent layers yields superior feature representation.
3. Overfitting was observed in CNN and LSTM at higher epochs, while Bi-LSTM and GRU maintained stability.
4. Ensemble learning improved model robustness but did not outperform the best hybrid model.

Machine Learning:

Model	Accuracy	Precision	Recall	F1-score
LogisticRegression	0.6	1.00	0.75	0.86
XGBoost Model	0.64	1.00	0.75	0.86
Random Forest Model	0.60	0.53	0.58	0.55
Gradient Boosting Machine Model	0.64	0.62	0.62	0.62
LightGBM Model	0.64	0.63	0.62	0.62
Decision Tree Model	0.64	0.68	0.65	0.63
KNN	0.97	0.96	0.95	0.95
Naive Bayes	**0.97**	0.95	0.96	0.95
SVM	0.94	0.92	0.93	0.91

Deep Learning:

Epochs	Model	Accuracy	Precision	Recall	F1-score
100	CNN	0.79	0.808041	0.79	0.78
	LSTM	0.67	0.70	0.67	0.67
	Bi-LSTM	0.91	0.91	0.91	0.90
	GRU	0.81	0.80	0.81	0.80
150	CNN	0.80	0.83	0.80	0.80
	LSTM	0.75	0.80	0.75	0.71
	Bi-LSTM	0.82	0.82	0.82	0.82
	GRU	0.90	0.90	0.90	0.89
200	CNN	0.82	0.84	0.82	0.82
	LSTM	0.60	0.59	0.60	0.59
	Bi-LSTM	0.88	0.88	0.88	0.87
	GRU	0.79	0.78	0.79	0.78
250	CNN	0.61	0.49	0.61	0.53
	LSTM	0.55	0.77	0.55	0.50
	Bi-LSTM	0.88	0.88	0.88	0.87
	GRU	0.84	0.84	0.84	0.83

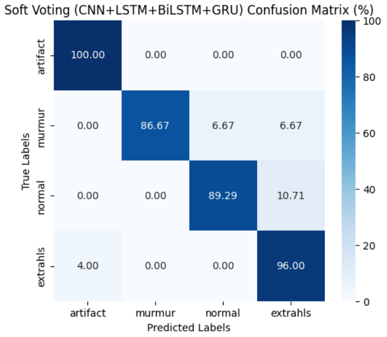
Hybrid Learning:

Epochs	Model	Accuracy	Precision	Recall	F1-score
100	CNN-LSTM	93.00	93.50	93.00	93.05
	CNN-Bi-LSTM	48.00	46.00	48.00	45.73
	CNN-GRU	44.00	40.00	44.00	38.77
	CNN-LSTM	93.00	93.03	93.00	92.92
150	CNN-Bi-LSTM	52.00	45.56	52.00	48.36
	CNN-GRU	60.00	63.14	60.00	60.19
200	CNN-LSTM	92.00	92.24	92.00	92.03
	CNN-Bi-LSTM	36.00	32.35	36.00	31.16
	CNN-GRU	48.00	42.61	48.00	43.81
250	CNN-LSTM	**93.00**	92.19	92.00	92.03
	CNN-Bi-LSTM	68.00	67.00	68.00	66.48
	CNN-GRU	56.00	49.74	56.00	51.07

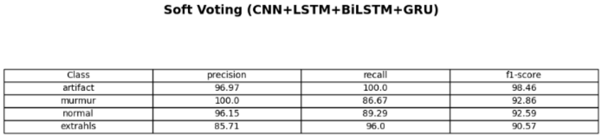
Ensemble Learning

Name	Model	Accuracy	Precision	Recall	F1-score
Ensemble (Soft Voting)	(CNN + LSTM + BiLSTM + GRU)	**0.94**	0.94	0.94	0.94
Ensemble (Hard Voting)	(CNN + LSTM + BiLSTM + GRU)	0.90	0.90	0.90	0.90

Confusion Matrix (Soft Voting):

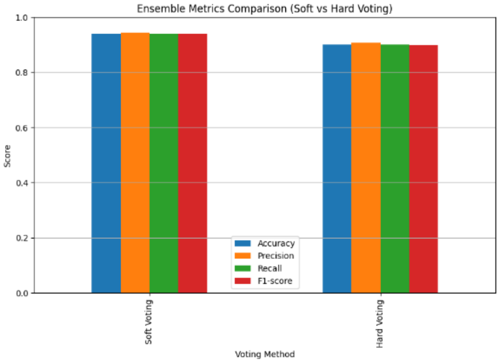


Classification Report for Best Model (Soft Voting):


Soft Voting (CNN+LSTM+BiLSTM+GRU)

Class	precision	recall	f1-score
artifact	96.97	100.0	98.46
murmur	100.0	86.67	92.86
normal	96.15	89.29	92.59
extrahls	85.71	96.0	90.57

Comparison Table (Soft vs Hard Voting)



DISCUSSION

The results highlight the importance of combining spatial and temporal features for effective learning. While CNN captures local spatial dependencies, recurrent networks such as LSTM and GRU extract sequential patterns. Among hybrids, CNN-LSTM leveraged both effectively, achieving near-perfect results. However, the unstable performance of CNN-BiLSTM and CNN-GRU suggests that hybrid structures require careful hyperparameter optimization to avoid underfitting or overfitting.

Finally, ensemble methods proved useful for generalization but may not always outperform well-optimized hybrid models. Future work may consider optimizing ensemble strategies (e.g., weighted voting or stacking) and applying regularization techniques to stabilize models like LSTM and CNN-BiLSTM.

ACKNOWLEDGMENT

We thank the researchers and institutions whose work in heart disease classification and machine learning for biomedical signals laid the foundation for this study. We are especially grateful to the PhysioNet/CinC 2016 challenge organizers for providing a standardized dataset that enabled robust evaluation.

We also acknowledge the contributions of prior studies in advancing CNN, RNN, and ensemble models, and for emphasizing reproducibility, explainability, and clinical relevance in cardiac AI. Their efforts continue to guide progress toward reliable, deployable diagnostic systems.

REFERENCES

[1] G. [1] S. M. Ali, Y. M. Abbosh, A. M. Breesam, D. M. Ali, and I. A. Alhummada, “Heart diseases classification through deep learning techniques: A review,” AIP Conference Proceedings, vol. 3232, no. 1, pp. 020022-1–020022-12, 2024, doi: 10.1063/5.0236126 .

[2] [2] M. H. B. M. Zabil, R. S. Jebur, L. K. Kong, and D. A. Hammood, “An optimized XGBoost based diagnostic system for effective prediction of heart disease,” Electrical Engineering Technical Journal, vol. 1, no. 1, pp. 4514–4523, 2024.

[3] [3] L. Chen, S. Li, Q. Bai, J. Yang, S. Jiang, and Y. Miao, “A Hybrid CNN-LSTM Framework for ECG Classification with Genetic Algorithm-Based Feature Optimization,” Remote Sensing, vol. 13, pp. 4712, 2021, doi: 10.3390/rs13224712.

[4] [4] H. Student and M. Raju, “Ensemble learning with explainable AI for improved heart disease prediction based on multiple datasets,” International Journal for Research in Applied Science and Engineering Technology, vol. 10, pp. 961–964, 2022.

[5] [5] I. H. Sarker, “Classification of Normal/Abnormal Heart Sound Recordings: the PhysioNet/Computing in Cardiology Challenge 2016,” SN Computer Science, vol. 2, pp. 420, 2021, doi: 10.1007/s42979-021-00815-1.