



Hybrid CNN–LSTM Framework for Automated Detection of Pediatric Heart Murmurs

Asmaa Abduh AL-taib

Dept. Artificial Intelligence and Intelligent Systems
Gümüşhane University, GU
Gümüşhane , Turkey
asmaabdu221@gmail.com

Hayati Türe

Dept. Software Engineering
Trabzon University
Trabzon , Turkey
hayatiture@trabzon.edu.tr

Muhammed Emin Aslan

Dept. Software Engineering
Gümüşhane University, GU
Gümüşhane, Türkiye
measlan.tr@gmail.com

Abstract—This study presents a comprehensive deep learning framework for automated heart murmur detection from phonocardiogram (PCG) signals to address the subjective limitations of cardiac auscultation. Using the CirCor DigiScope dataset of 5,272 pediatric heart sound recordings from 1,568 subjects, we systematically evaluated multiple approaches including traditional machine learning models (Logistic Regression, SVM, KNN, Decision Tree, AdaBoost), basic deep learning architectures (standalone CNN and LSTM), and advanced transfer learning models (DenseNet121, EfficientNet B0/B1/B2, ResNet 50/101, MobileNet V2/V3, GhostNet, AttentionCNN). Our hybrid CNN-LSTM architecture combines mel-spectrogram spatial feature extraction with temporal sequence modeling using 128 traditional audio features and comprehensive preprocessing with data augmentation. The CNN-LSTM model achieved superior performance with 84.6% accuracy, 0.84 F1-score, 0.748 precision, and 0.800 recall, significantly outperforming traditional methods (best SVM: 0.716 F1-score) and advanced architectures (best DenseNet121: 0.72 F1-score), demonstrating robust classification with 0.879 AUC. Results indicate that hybrid deep learning architectures effectively automate cardiac auscultation, providing objective diagnostic tools with potential for improved healthcare accessibility and clinical decision-making.

Keywords—Convolutional Neural Networks; Deep Learning; Mel Spectrogram; Artificial intelligence; Recurrent Neural Networks; Heart Murmur Detection; Cardiac Auscultation; Long Short-Term Memory

I. INTRODUCTION

Cardiovascular diseases (CVDs) are a leading global cause of morbidity and mortality, necessitating early, accurate diagnosis [1]. Cardiac auscultation is a cost-effective screening tool [2], but its reliance on clinician expertise introduces variability, especially where specialists are scarce [3]. Digital phonocardiography (PCG) and AI—particularly ML and DL—

enable objective analysis, automating heart sound interpretation[4]. Most AI systems focus on binary murmur detection, but reliable multi-class classification is still needed.

Early automated analysis segmented heart sounds (S1, S2) and extracted features (statistical moments, spectral centroids, energy ratios) for SVM, KNN, and Random Forest classifiers, achieving moderate accuracy but limited in capturing complex PCG patterns [5]. DL models, especially CNNs, learn hierarchical features from raw signals [6], while LSTMs capture temporal dependencies [7]. Hybrid CNN-RNN models combine spatial and temporal learning, enhanced by data augmentation (time stretching, pitch shifting, noise injection) to improve generalization [8].

Multi-class classification frameworks further characterize murmurs (normal, systolic, diastolic) and valvular pathologies, using multi-branch CNNs, RNN-based hidden semi-Markov models, and hierarchical CNNs, achieving high sensitivity and specificity (e.g., 76.3% and 91.4%) [9]. However, clinically graded, fine-grained murmur classification remains limited.

This study presents a hybrid CNN-LSTM framework for automated murmur detection. PCG signals are converted to Mel spectrograms, with the model learning spatial and temporal patterns. Data augmentation improves robustness, demonstrating potential for accurate and interpretable automated cardiac auscultation.

II. THE PCG DATASET

This study utilized the publicly available CirCor DigiScope Phonocardiogram Dataset, extensively annotated and published on PhysioNet by Oliveira et al. (2022) for the PhysioNet/CinC Challenge 2022, serving as a benchmark for automated cardiac auscultation research. The CirCor DigiScope Dataset contains 5,272 PCG recordings from 1,568 pediatric subjects (0–21 years), lasting 4.8–80.4 s (mean 22.9 s), totaling 33.5 hours. Four auscultation sites capture the heart profile. Murmurs were

annotated for timing, shape, pitch, grade (Levine), quality, and location; S1 and S2 sounds were labeled via a semi-supervised approach with three algorithms and expert verification [10].

III. METHODS

This study aims to detect the presence or absence of cardiac murmurs, a task normally requiring expert cardiologists in controlled settings. Using advanced AI, 2D CNNs classify PCG recordings as murmur present or absent, learning hierarchical representations to capture subtle acoustic variations. The methodology covers dataset, preprocessing, model, and evaluation metrics. Fig. 1 shows Mel-scaled spectrograms of normal and murmur signals.

A. Preprocessing

The study analyzed PCG recordings from 942 patients, each with up to four .wav files (aortic, mitral, pulmonary, tricuspid; 5–30 s). Only patients with all four recordings were retained, yielding 574 patients (117 with murmurs, 457 without). All waveforms were standardized to 5 s, with data augmentation applied to balance murmur-positive samples. Long murmur recordings (>15 s) were split into three 5-s segments. Samples from the same patient were kept in the same subset to prevent data leakage. The final dataset included 510 training samples (47% positive), 124 validation (27% positive), and 124 testing samples (24% positive).

Feature Extraction

This study extracted 128 features from PCG signals across time, frequency, cepstrum, and statistical domains for all four auscultation sites to distinguish normal from murmur signals. Key features included the Amplitude Envelope (maximum frame amplitude, mean, median, standard deviation, 75th percentile) [11], Total Energy (overall loudness, computed by summing squares)(1),

$$E_x = \sum_n |x[n]|^2 \quad (1)$$

and Root Mean Squared Energy (less sensitive to outliers) (2)[12]. Additional features comprised Zero-Crossing Rate (signal sign changes, selected via RFE) [13], Skewness and Kurtosis (signal asymmetry and peakedness, selected via RFE), Peak Frequency (highest-amplitude frequency) [14], Onset Detection (S1, S2, and murmur peaks), Band Energy Ratio (low/high-band energy comparison) [15], Spectral Centroid (energy concentration), and the first 13 Mel-Frequency Cepstral Coefficients (MFCCs) capturing the spectral envelope on the Mel scale [16], highlighting their significance across all cardiac sites.

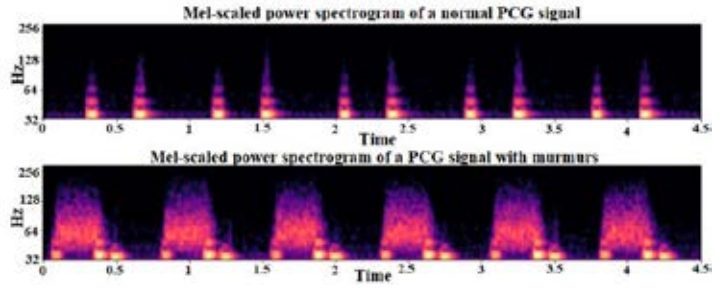


Fig. 1. Mel-scaled power spectrogram of a normal PCG signal. b) Mel-scaled power spectrogram of a PCG signal with murmurs.

$$RMSE = \sqrt{\frac{1}{N} \sum_{n=0}^{N-1} [x(n) - \tilde{x}(n)]^2} \quad (2)$$

B. Heart Murmur Classification

Supervised learning was applied to labeled audio from AV, PV, MV, and TV regions. Feature selection reduced redundancy via Lasso Logistic Regression, ANOVA F-value, and RFE (retaining 20 features). Data were split into training (510 samples, 47% positive), validation (124, 27%), and testing (124, 24%), with augmentation keeping all samples per patient in one subset.

C. Feature Selection

To improve model performance, three feature selection methods were applied: Lasso Logistic Regression (regularization shrinks coefficients to zero, optimized via cross-validated F1 score), ANOVA F-value (removes features independent of class), and RFE (Recursive Feature Elimination, iteratively removes least important features until 20 remain). Deep learning models included CNNs for spatial features, LSTMs for temporal dependencies, and CNN-LSTM hybrids. Transfer learning used GhostNet, EfficientNet (B0–B2), ResNet (50,101), MobileNet (V2,V3), DenseNet121, and AttentionCNN for efficiency, scalability, or specialized feature extraction. Training used stratified splits, hyperparameter tuning via EarlyStopping, ReduceLROnPlateau, and ModelCheckpoint, with Adam/RMSprop optimization and Categorical Crossentropy loss. Performance was evaluated with Accuracy, Precision, Recall, F1-Score, ROC-AUC, and Confusion Matrices.

IV. RESULTS

This study evaluates deep learning models for heart murmur detection using Mel spectrograms, demonstrating the effectiveness of various neural networks in classifying heart sounds as Absent, Present, or Unknown murmurs.

A. Dataset Characteristics

The dataset contained 942 patient recordings from four valve positions—Aortic (AV), Mitral (MV), Pulmonary (PV), and Tricuspid (TV)—with 73.8% absent, 19.0% present, and

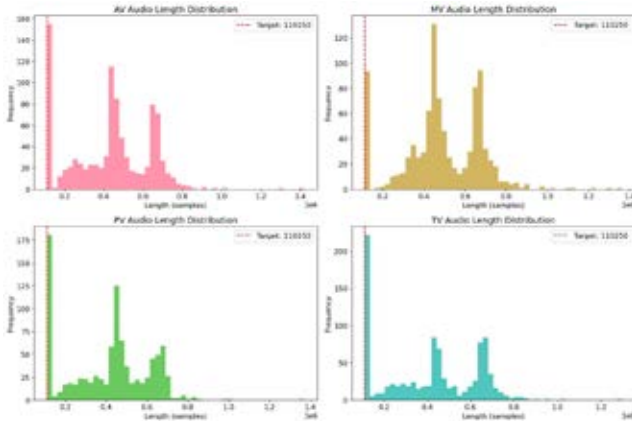


Fig. 2. Audio sample length distribution with target length for normalization

7.2% unknown murmurs. Preprocessing addressed class imbalance.

B. Mel Spectrogram Feature Extraction

Missing data were replaced with zero arrays, and all audio normalized to 5 seconds (Fig. 2). Data augmentation balanced classes. Mel spectrograms (22,050 Hz, 128 bins, 2,048 FFT, 512 hop, 128×216×1) and 14 conventional audio features (e.g., MFCCs, spectral centroid) were extracted.

C. Model Performance Comparison

● Traditional Machine Learning Models:

A Dummy baseline underperformed due to imbalance; tuned Logistic Regression, SVM, KNN, Decision Tree, LDA, QDA, and AdaBoost with 84 RFE-selected features improved F1 by ~10%. AV alone performed best, but combining all valves was superior.

● Basic Deep Learning Models

CNN, LSTM, and CNN-LSTM were tested; CNN-LSTM achieved the top F1 score of 0.84 (Table I).

● Advanced Model Architectures

GhostNet, EfficientNet (B0–B2), ResNet (50, 101), MobileNet (V2, V3), DenseNet121, and AttentionCNN were compared (Table II); DenseNet121 performed best due to dense connectivity.

D. Evaluation

The final CNN-LSTM model, configured with a learning rate of 0.001 and sigmoid activation, attained precision, recall, F1-score, and accuracy of 0.748, 0.800, 0.840, and 0.846,

respectively, on the test set, outperforming prior studies as shown in Table III.

Its probabilistic behavior was examined using the ROC curve, where class predictions result from the sign function at the decision boundary, while scikit-learn applies a sigmoid

TABLE I. COMPARISON OF TRADITIONAL MACHINE LEARNING MODELS AND THE CNN-LSTM DEEP LEARNING MODEL PERFORMANCE

Model	Model Selected	Training f1 Score	Validation f1 Score	Testing f1 Score
Logistic Regression	C=0.01, penalty = elasticnet	0.745	0.66	0.704
SVM	C=0.1, kernel=sigmoid	0.69	0.685	0.716
KNN		0.84		0.51
Decision Tree	criterion=entropy, max depth=50, min samples leaf=5, min samples split=20, splitter=random	0.8	0.68	0.49
LDA		0.82		0.51
QDA		0.97		0.30
ADABOOST	algorithm=SAM ME, nestimators = 100	0.93	0.68	0.66
CNN-LSTM		0.88	0.82	0.84

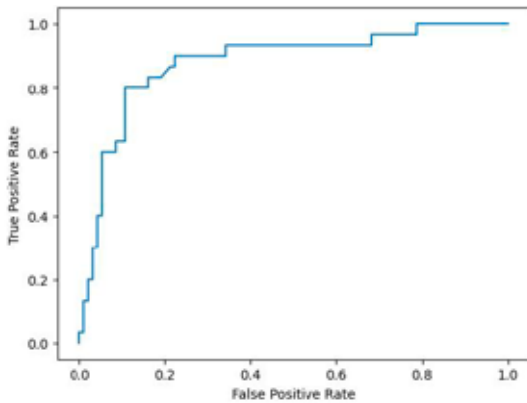
TABLE II. PERFORMANCE COMPARISON OF ADVANCED DEEP LEARNING ARCHITECTURES IN HEART MURMUR DETECTION

Model	Test Accuracy	Test Loss	Precision	Recall	F1 Score	Parameters	Model Size MB
DenseNet121	0.77	0.94	0.70	0.77	0.72	7565897	28.8
AttentionCNN	0.77	0.66	0.76	0.77	0.68	488718	1.86
GhostNet	0.75	0.69	0.56	0.75	0.64	2439341	9.30
EfficientNetB0	0.75	0.71	0.56	0.75	0.64	4839596	18.46
ResNet50	0.75	0.86	0.56	0.75	0.64	26216329	100
MobileNetV3Small	0.75	0.75	0.56	0.75	0.64	1238137	4.72
EfficientNetB1	0.65	0.98	0.66	0.65	0.65	7365264	28.09
MobileNetV2	0.33	1.66	0.62	0.33	0.40	2917449	11.12

transformation to the model's linear output, with parameters A and B estimated via Maximum Likelihood Estimation. This analysis yielded an AUC of 0.879, as depicted in Fig. 3.

TABLE III.COMPARISON OF OUR CNN–LSTM MODEL WITH PREVIOUS
 HEART MURMUR DETECTION STUDIES

Study	Method	Dataset	Accuracy	F1-score
Our Work	(2025)CNN-LSTM	CirCor DigiScope	84.6%	0.84
Huang et al. (2024)	2D CNN (Multi-task)	CirCor DigiScope	95.19%	-
Reyna et al. (2023)	Various CNN architectures	PhysioNet 2022	80.2%	0.78
Wang et al. (2024)	Branchformer with Uncertainty	PhysioNet 2022	79.8%	0.65
Li et al. (2023)	Neural Architecture Search	PhysioNet 2022	83.4%	0.83
Kumar et al. (2024)	CNN with Mel-spectrogram	PhysioNet 2022	81.8%	0.80


 TABLE III.COMPARISON OF OUR CNN–LSTM MODEL WITH PREVIOUS
 HEART MURMUR DETECTION STUDIES

V. DISCUSSION

This study presents a comprehensive framework for automated heart murmur detection, leveraging hybrid CNN–LSTM and advanced transfer learning models, including GhostNet, EfficientNet, ResNet, MobileNet, DenseNet121, and AttentionCNN. Cardiac sounds were classified as Absent, Present, or Unknown using 5-second Mel spectrograms combined with 14 traditional audio features, addressing class imbalance (73.8% Absent) via data augmentation and normalization. The CNN–LSTM effectively captured both spectral and temporal patterns, achieving an F1-score of 0.84 and outperforming baseline classifiers, while broader evaluation of multiple architectures extends prior work. Strengths include robust preprocessing, multi-dimensional feature representation, stratified splitting, and evaluation using F1-score, ROC-AUC, and visualizations, enhancing clinical relevance. Unlike prior studies limited in scope, this work offers a balanced comparison across models and explicitly addresses critical metrics. Limitations involve dataset imbalance and generalizability, motivating future work on advanced augmentation (SMOTE, GANs), external validation,

multimodal clinical integration, Explainable AI, and model compression for real-time deployment. Overall, the framework advances AI-driven cardiovascular diagnostics, enabling earlier murmur detection and broader healthcare accessibility.

REFERENCES

- [1] World Health Organization. Cardiovascular diseases (CVDs), 2021. <https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-cvds>
- [2] ScienceDirect. Heart auscultation—an overview. Accessed from <https://www.sciencedirect.com/topics/medicine-and-dentistry/heart-auscultation>
- [3] ScienceDirect. Extraction and assessment of diagnosis-relevant features for heart sound classification. Accessed from <https://www.sciencedirect.com/science/article/abs/pii/S104620232100181X>
- [4] LeCun, Y., Bengio, Y., & Hinton, G. Deep learning. *Nature*, 521(7553), 436–444, 2015.
- [5] Springer, D. B., Tarassenko, L., & Clifford, G. D. Logistic regression-HSMM based heart sound segmentation. *IEEE Transactions on Biomedical Engineering*, 63(4), 822–832, 2016.
- [6] Maknickas, V., & Maknickas, A. Recognition of normal and abnormal phonocardiographic signals using deep convolutional neural networks and Mel frequency spectral coefficients. *Physiological Measurement*, 38(8), 1671, 2017.
- [7] Bozkurt, B., Germanakis, I., & Stylianou, Y. A study on CNN-based automatic heart sound classification for pathological detection. *Computers in Biology and Medicine*, 100, 132–143, 2018.
- [8] Potes, C., Parvaneh, S., Rahman, A., & Conroy, B. Ensemble of feature-based and deep learning-based classifiers for detection of abnormal heart sounds. In 2016 Computing in Cardiology Conference (CinC), 621–624. IEEE, 2016.
- [9] A. Elola et al., IEEE Xplore. Beyond heart murmur detection: Automatic murmur grading from phonocardiogram. *IEEE Journal of Biomedical and Health Informatics*, 27(8), 3856–3866, 2023. DOI: <https://doi.org/10.1109/JBHI.2023.3275039>. Accessed from <https://ieeexplore.ieee.org/document/10122565>
- [10] PhysioNet. The CirCor DigiScope Dataset: An open-access pediatric heart sound database for murmur detection, 2022. <https://physionet.org/content/circor-heart-sound/1.0.3/>
- [11] He, L., & Dellwo, V. Amplitude envelope kinematics of speech signal: Parameter extraction and applications. University of Zurich, 2017. https://www.researchgate.net/publication/318275305_Amplitude_envelope_kinematics_of_speech_signal_parameter_extraction_and_application_s
- [12] Hodson, T. O. Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not. *Geoscientific Model Development*, 15, 5481–5487, 2022. <https://doi.org/10.5194/gmd-15-5481-2022>
- [13] McFee, B., Raffel, C., Liang, D., Ellis, D. P. W., McVicar, M., Battenberg, E., & Nieto, O. librosa: Audio and music signal analysis in Python. *Proceedings of the 14th Python in Science Conference*, 2015. <https://librosa.org/>
- [14] Erbe, C., Duncan, A., Hawkins, L., Terhune, J. M., & Thomas, J. A. Introduction to acoustic terminology and signal processing. In *Exploring Animal Behavior Through Sound: Volume 1*, 111–152. Springer, 2022. https://link.springer.com/chapter/10.1007/978-3-030-97540-1_4
- [15] ScienceDirect. Audio frame—an overview. Accessed August 10, 2025, from <https://www.sciencedirect.com/topics/engineering/audio-frame>
- [16] M. W. Lakdari, A. H. Ahmad, S. Sethi, G. A. Bohn, and D. J. Clink, ScienceDirect. Mel-frequency cepstral coefficients outperform embeddings from pre-trained convolutional neural networks under noisy conditions for discrimination tasks of individual gibbons. Accessed August 10, 2025, from <https://www.sciencedirect.com/science/article/pii/S1574954123004867>