Heart Disease Prediction using Deep Learning Methods

Xiao Fang, Ziao You, Weijie Chen, Yixuan Song, Kan Kim

June 5, 2024

Abstract

In today's world, heart diseases are becoming more prevalent and common, threatening the health of many people. Diagnosing heart disease at an early stage is an important topic for the medical field. In this paper, we explore the use of machine learning and deep learning techniques to classify the presence or absence of disease from heartbeat audio. We explore various machine learning algorithms and deep learning architectures, including Multilayer Perceptrons (MLPs), Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and ensemble models, trained on the CirCor DigiScope Phonocardiogram dataset. Our experiments demonstrate promising results, with the best-performing model achieving a test accuracy of around 69%.

1 Background

Heart disease is a prevalent health problem and early detection can significantly improve patient outcomes. Our project aims to develop a machine learning and deep learning based system to predict heart disease from heart sound recordings. Traditionally, doctors rely on auscultation (listening to heart sounds) and advanced imaging techniques such as MRI and CT scans for diagnosis. However, these methods can be time-consuming and resource-intensive. Our approach leverages the power of machine learning and deep learning to analyze heart sound recordings, providing a faster and more convenient alternative for initial screening.

2 Methodology

Our approach consists of three main parts: data collection and preprocessing, conventional machine learning, and deep learning experiments. We use the CirCor DigiScope phonocardiogram dataset and preprocess the data using various techniques, including filtering, thresholding, and data augmentation. For traditional machine learning, we explored algorithms such as Logistic Regression and SVM, which are trained on audio features and patient metadata.

2.1 Data Collection, Processing

We chose the CirCor DigiScope Phonocardiogram dataset, a collection of recordings by stethoscope and labeled by expert cardiologists indicating the patient's cardiac condition. It contains a total of 33.5 hours of audio data.

Data preprocessing is a crucial step in preparing the heart sound recordings for machine learning and deep learning models. We applied several techniques to enhance the quality and usability of the data. Pre-filtering involved removing portions of the audio that did not contain heartbeats, ensuring that only relevant segments were used for analysis. Frequency thresholding involved cutting off frequencies above 450Hz to reduce noise and focus on the relevant frequency range for heart sounds. Data augmentation techniques, such as extracting subsets of the audio with different durations, were employed to increase the diversity and size of the training data.

- MFCC - Spectral characteristics of sound.
- Chromagram - projected audio onto 12 bins representing the 12 distinct semitones (or chroma) of the musical octave.
- NMF - Decomposes a non-negative audio representation matrix into two lower-rank non-negative matrices, W (capturing the underlying patterns) and H (activations) in the data. We use the norm of the matrix of W as one of our feature representations.
- Band Power - Extract feature in each frequency range.

2.2 Conventional Machine Learning

Conventional machine learning approaches were explored as a baseline for heart disease prediction. We trained models like Kth Nearest Neighbors, Logistic Regression, Support Vector Machine, and Random Forest on a combination of audio features and patient metadata. These features included Mel-Frequency Cepstral Coefficients (MFCCs), chromagrams, mel-spectrograms, norm of NMF matrix, band power, and patient demographics. The models were evaluated using metrics such as accuracy, precision, recall, and F1-score. We also used two innovative training methods to help us get the optimal combination of models and features more efficiently.

- Train models on a mixture of features. We defined three different feature sets and tried models on them. The same model can be trained on different features so that we can know the most suitable feature for one type of model.
- Train different models on the same features. We defined the same features in the first step, then we trained all different models on them to get the best performance model.

On top of these training methods, we used the voting method. By aggregating the results of all model predictions and selecting the prediction with the highest number of votes as our final prediction, the robustness of the model can be significantly improved.

2.3 Deep Learning Experiments

Deep learning techniques were extensively explored for our project. We experimented with various architectures and input representations, such as 1D-CNNs on MFCCs, MLPs and CNNs on concatenated features (chromagrams, mel-spectrograms, MFCCs, and patient data), and 2D-CNNs on mel-spectrograms. Additionally, we investigated advanced models like multiscale CNNs and Conv-RNNs, which combine the strengths of convolutional and recurrent layers for analyzing sequential data.

- 1D-CNN with MFCCs. The architecture consists of 15 layers. We employ a series of Conv1D and pooling layers to extract features and reduce spatial dimensions, followed by fully connected layers that consolidate these features for a final two-class classification output. The input data is MFCCs features, and we eventually achieved 62.38% accuracy in the test set.

- MLP with Concatenated Feature - This experiment uses a two-hidden-layer CNN model, with a logistic activation function. We concatenate 4 categories of features and feed them to the model: Chromagram, Mel-Sspectrogram, mfcc, and patience metadata. Through hyper-parameter tuning, we hit a best performance of 66.46% in accuracy and 67.28% in F1 score.

- CNN with Concatenated Feature - Run multiple CNNs on different features and create another CNN with concatenated results from those CNNs as input.

- CNN on 2D-Mel-Spectrogram Feature - This experiment uses a 2D-CNN model, with three convolution blocks. We purely feed the 2D-Mel-Spectrogram feature to our model. All the samples are padded to the same length. We hit a best performance of 56.01% in accuracy and 49.64% in F1 score.

- Conv-RNN - Combining the spatial feature extraction capability of convolutional layers with the temporal processing strength of LSTM layers.

- Concatenated Multiple MLP - Combining multiple MLP with an output shape of 1 together with a linear layer to get the final result. Trained with feature extraction methods, with N feature sets as input, N MLPs will be created and combined, each MLP will be feeding from one feature set.

- RNN / LSTM - Classical DL models for time series data. Trained with raw audio data after preprocessing.

3 Advanced Models

To further enhance the performance of our heart disease prediction system, we have incorporated advanced deep learning architectures, ResNet and Multiscale CNN. ResNet is particularly noted for its ability to address the vanishing gradient problem, thereby facilitating more efficient training of deeper models. Conversely, Multiscale CNN employs an array of convolutional filters of varying sizes to capture features at multiple scales.

- RESNET - The implementation of ResNet within our framework is specifically tailored for one-dimensional input data, such as ECG signals. Our model architecture consists of multiple ResNet blocks, each taking ECG data points as input—representing various features—and producing outputs correlated to targeted class categories. Each block within the ResNet architecture processes the input through two convolutional layers, followed by batch normalization and ReLU activation functions, enhancing both the non-linearity and stability of the model during training. The ResNet model was trained using a diverse set of input features, including chromaticity maps, 1D and 2D Mel-spectrograms, Mel Frequency Cepstral Coefficients (MFCCs), band powers, and Non-Negative Matrix Factorization (NMF) representations. The best performance score for the RESNET model yielded 55.22% accuracy score.

- Multi-Scale CNN - In this experiment, we optimize our base 2D-CNN model. Two new components are introduced to the head of the network: The multi-Scale sampler, and the channel recalibration module. On top of the pipeline, the multi-scale sampler can enhance the model's ability to detect intricate details present at different levels of granularity. After that, an average-pooling layer will fusion the inputs from different samplers. Between the samplers and following convolution blocks, we add a channel recalibration module, which is designed to rectify the lack of cross-channel interaction in the extracted features. We still feed the 2D-Mel-Spectrogram feature to our model. Through hyper-parameter tuning, we hit a best performance of 58.22% in accuracy and 66.50% in F1 score.

4 Results and Discussion

In terms of performance, we are using accuracy as our main evaluation. Table 1 contains the best results from each of our main models. We applied a hard-voting on multiple machine-learning models including logisticRegression, SVC, KNeighborsClassifier, DecisionTreeClassifier, RandomForestClassifier, and GaussianNB and got a model with 63.07% accuracies as our baseline model. For the deep learning models we have. The best-performing model was a 1D CNN model with band power structure the CSV file contains more detailed information about patients, i.e. gender, and age as input, achieving a test accuracy of around 69%. We also tried some models without that CSV file as input but the best result from those models is only with 55.22% accuracy. This means additional information from the audio is significantly helpful for this topic. Additionally, we also tried to train similar models with time series data only and abstracted features from it. Which results in abstracted features that can get better results in most cases. In 1D CNN, we got models with 67% accuracy on average while training with different combinations of abstracted features, however, the similar CNN structure with time series can only get 54% accuracy on average. With those features abstracted from the audio data, we got several models beat our baseline machine-learning model. Single MLP gets 66.46% accuracy, 1D CNN gets 68.95% accuracy, Concat-CNN gets 64.22% accuracy, Conv-RNN gets 64.38%. While these results are encouraging, there is still room for improvement, and we plan to explore alternative feature representations and advanced architectures in future work.

Table 1: Results

Model	**Features**	**Accuracy**
Hard Voting ML	MFCC + CSV	63.07
Single MLP	Chromagram + 1D Mel-Spec + MFCC + Opensmile + CSV + Time Series	66.46
Multiscale 2D CNN	2D Mel-Spec	58.23
1D CNN	Bandpower Struct + CSV	68.95
Concat-CNN	Chromagram + 1D Mel-Spec + MFCC + Bandpower Struct + CSV	64.22
Conv-RNN	1D Mel-Spec + MFCC + Opensmile + CSV	64.38
ResNet	Chromagram + 1D Mel-Spec + MFCC + Time Series	55.22
Concat-MLP	2D Mel-Spec + Bandpower Struct + Opensmile + CSV	59.11
LSTM	Chromagram + 2D Mel-Spec + CSV + Time Series	53.85
RNN	Chromagram + Bandpower Struct + CSV + Time Series	50.00

5 Future Work

To enhance the performance of our current models, several strategies for improvement and expansion are proposed, building upon the initial promising results. Despite these initial successes, our models currently do not achieve the benchmarks set by other studies in the field. A critical area identified for further development is the advancement of feature engineering techniques, as we have come to realize that the current set of features may not be optimal. Preliminary investigations suggest that the Wavelet Scattering Transform (WST) could offer a more nuanced representation of time-series data, such as heart sound recordings. This method capitalizes on the temporal structure of the data, potentially uncovering subtle patterns associated with pathological states that are not captured by traditional feature extraction methods.

In addition to refining feature engineering, there is a need for rigorous hyperparameter optimization of our deep learning models. Systematic tuning of parameters can significantly influence model performance, suggesting that our current configurations may not be optimal. Future work will include the application of grid search and randomized search techniques to identify the most effective configurations.

Moreover, we plan to integrate advanced neural network architectures, such as attention mechanisms and transformers, which have demonstrated considerable success in various sequence modeling tasks. Their ability to model complex dependencies and highlight relevant features in sequence data is particularly promising for enhancing the interpretability and accuracy of heart disease prediction models.

Finally, to address potential overfitting and improve the robustness of our predictions, we will expand our dataset to include a larger and more diverse set of heart sound recordings, including those from adult populations. This broader dataset will enable the training of models that are more adaptable to varying patient demographics and clinical conditions, thus enhancing their generalizability.

6 Conclusion

This study uses deep learning techniques to test different methods to predict heart disease from phonocardiogram. We built multiple deep neural network models such as MLP, CNN, RNN.

Our experiments showed promising results. A 1D CNN trained with band power and CSV files shows the best performance of this experiment with an accuracy of about 68.95%. Our other notable models include multiple MLPs trained with different features (test accuracy 66.46%), CNN trained with concatenate features (test accuracy 64.22%), Conv-RNN (test accuracy 64.38%) and multi-scale CNN (test accuracy 58.23%).

Despite our current results, there is still room for improvement. Future work includes studying alternative feature representations. Such as wavelet scattering transform (WST), performing large-scale hyperparameter tuning, and studying the integration of attention mechanisms into the Transformer architecture. On the other hand. We only have a small dataset this time. If we could find or label a larger and more diverse dataset. We believe that we can improve it with even better accuracy.

Overall, this study demonstrates the feasibility and potential of using deep learning to predict heart disease from heart sounds. This provides a faster and more convenient alternative to traditional diagnosis, which will ultimately help improve the medical experience of patients with cardiovascular diseases.

References

[1] Khan, Younas, Qamar, Usman, Yousaf, Nazish, and Khan, Aimal. *Machine learning techniques for heart disease datasets: a survey*. Proceedings of the 2019 11th International Conference on Machine Learning and Computing, pp. 27-35, 2019.

[2] Oliveira, Jorge, Renna, Francesco, Costa, Paulo, Nogueira, Marcelo, Oliveira, Ana Cristina, Elola, Andoni, Ferreira, Carlos, Jorge, Alipio, Rad, Ali Bahrami, Reyna, Matthew, *et al.* *The CirCor DigiScope Phonocardiogram Dataset*. Version 1.0.0, 2022.

[3] Yildirim, Muhammed. *Automatic classification and diagnosis of heart valve diseases using heart sounds with MFCC and proposed deep model*. Concurrency and Computation: Practice and Experience, vol. 34, no. 24, p. e7232, 2022. Wiley Online Library.

[4] Balaha, Hossam Magdy, Shaban, Ahmed Osama, El-Gendy, Eman M, and Saafan, Mahmoud M. *A multi-variate heart disease optimization and recognition framework*. Neural Computing and Applications, vol. 34, no. 18, pp. 15907-15944, 2022. Springer.

[5] Nguyen, Minh Tuan, Lin, Wei Wen, and Huang, Jin H. *Heart sound classification using deep learning techniques based on log-mel spectrogram*. Circuits, Systems, and Signal Processing, vol. 42, no. 1, pp. 344-360, 2023. Springer.

[6] Al Bataineh, Ali and Manacek, Sarah. *MLP-PSO hybrid algorithm for heart disease prediction*. Journal of Personalized Medicine, vol. 12, no. 8, p. 1208, 2022. MDPI.

[7] Fradi, Marwa, Khriji, Lazhar, and Machhout, Mohsen. *Real-time arrhythmia heart disease detection system using CNN architecture based various optimizers-networks*. Multimedia Tools and Applications, vol. 81, no. 29, pp. 41711-41732, 2022. Springer.

[8] Rath, Adyasha, Mishra, Debahuti, Panda, Ganapati, and Satapathy, Suresh Chandra. *Heart disease detection using deep learning methods from imbalanced ECG samples*. Biomedical Signal Processing and Control, vol. 68, p. 102820, 2021. Elsevier.

[9] Su, Yu-Sheng, Ding, Ting-Jou, and Chen, Mu-Yen. *Deep learning methods in internet of medical things for valvular heart disease screening system*. IEEE Internet of Things Journal, vol. 8, no. 23, pp. 16921-16932, 2021. IEEE.

[10] Brunese, Luca, Martinelli, Fabio, Mercaldo, Francesco, and Santone, Antonella. *Deep learning for heart disease detection through cardiac sounds*. Procedia Computer Science, vol. 176, pp. 2202-2211, 2020. Elsevier.

[11] Kamble, Dipali Dilip, Kale, Pratiksha Hanumant, Nitture, Suraj Pandit, Waghmare, Komal Vijay, and Aher, Chetan Nimba. *Heart disease detection through deep learning model RNN*. In: *Smart Intelligent Computing and Applications, Volume 2: Proceedings of Fifth International Conference on Smart Computing and Informatics (SCI 2021)*, pp. 469-480, 2022. Springer.

[12] Charles, V Benhar, Surendran, D, and SureshKumar, A. *Heart disease data based privacy preservation using enhanced ElGamal and ResNet classifier*. Biomedical Signal Processing and Control, vol. 71, p. 103185, 2022. Elsevier.

[13] El-Ghaish, Hany and Eldele, Emadeldeen. *ECGTransForm: Empowering adaptive ECG arrhythmia classification framework with bidirectional transformer*. Biomedical Signal Processing and Control, vol. 89, p. 105714, 2024. Elsevier.