Heart Disease Prediction from heart beat audio signals using Machine Learning and Network Analysis

Ligari D. • Alberti A. [1]

[1] *Department of Computer Engineering, Data Science, University of Pavia, Italy*
Course of Advanced Biomedical Machine Learning

Github page: https://github.com/DavideLigari01/advanced-biomedical-project
Date: June 28, 2024

Abstract — Heart disease remains one of the leading causes of mortality worldwide, making early diagnosis crucial. This study aims to predict heart diseases by analyzing heartbeat audio signals using machine learning and network analysis. We utilized a dataset from the PASCAL Classifying Heart Sounds Challenge 2011, which includes normal heart sounds, murmurs, extra heart sounds, extra systoles, and artifacts. Various preprocessing techniques such as noise reduction, resampling, and segmentation were applied to ensure data quality. Features were extracted using methods like Mel-Frequency Cepstral Coefficients (MFCC), Chroma, RMS, ZCR, and spectral features. Multiple machine learning models including LightGBM, XGBoost, CatBoost, Random Forest, and Multilayer Perceptron were trained and evaluated. The best performing model achieved high accuracy in distinguishing between different heart sound categories. This research highlights the potential of machine learning in cardiac diagnostics and provides a foundation for future advancements in the field.

Keywords ——TO BE DEFINED—

CONTENTS

1. Introduction

1.1. Problem Domain

1.2. Research Question

1.3. Previous Research

2. Methods

2.1. Source of Data

2.1.1. Type of sources

2.1.2. Classes

2.1.3. Data Distribution

2.2. Data Preprocessing

2.3. Feature Extraction

2.3.1. Features Type

2.3.2. Sampling Rate Selection

2.3.3. Extraction Interval Selection

2.3.4. Number of Features per Type

2.4. Feature Selection

2.5. Models

2.5.1. Metrics used

2.5.2. Prevention Model

2.5.3. Support Model

2.6. Tools and Software

- Scikit-learn - Numpy - Pandas - Matplotlib - Seaborn - Scipy - XGBoost - CatBoost - PyTorch - Torchaudio - Librosa - TensorFlow - Keras - Shap - Imblearn - Other Utility Libraries (e.g. joblib, os, sys, etc.)

3. Results

3.1. Prevention Model

3.2. Support Model

3.2.1. Explainability

3.3. Other Experiments

3.3.1. CNNs

3.3.2. Tiered Ensemble Model

3.3.3. Data Augmentation

4. Discussion

4.1. Positioning in Existing Research

4.2. Limitations

5. Conclusion

5.1. Overall Impression

5.2. Future Work

6. APPENDIX