Comprehensive Machine Learning and Deep Learning Framework for Enhanced Home-Based Heart Disease Prediction

Chenxi Liu
Applied Computer Science
Fairleigh Dickinson University
Vancouver, Canada
c.liu2@student.fdu.edu

Xiaoyi Wang
Applied Computer Science
Fairleigh Dickinson University
Vancouver, Canada
x.wang6@student.fdu.edu

Junrong Huang
Applied Computer Science
Fairleigh Dickinson University
Vancouver, Canada
j.huang2@student.fdu.edu

Xiaoyu Zhang
Applied Computer Science
Fairleigh Dickinson University
Vancouver, Canada
x.zhang3@student.fdu.edu

Qianying Zhao
Applied Computer Science
Fairleigh Dickinson University
Vancouver, Canada
z.qianying@student.fdu.edu

***Abstract*—This project aims to enhance a heart disease prediction framework using machine learning and deep learning models. By conducting advanced Exploratory Data Analysis (EDA), tuning model hyperparameters, and incorporating deep learning techniques like Deep Neural Networks (DNNs) or Convolutional Neural Networks (CNNs), we will improve the accuracy and generalization of predictions. A larger dataset will also be integrated to further enhance the system's robustness for real-world, home-based applications.**

I. Introduction

A. Problem Statement

The original study [1] focused on heart disease prediction using various machine learning models such as Logistic Regression, K-Nearest Neighbor, Support Vector Machine, Naive Bayes, Decision Tree, Random Forest, and XGBoost. However, these models were used with default parameters, and the dataset was relatively small, limiting the model's generalization capabilities. Furthermore, the Exploratory Data Analysis (EDA) was basic, which could lead to suboptimal feature selection and model performance.

B. Importance

Heart disease is one of the leading causes of death worldwide. Developing a reliable, home-based prediction system can provide early detection, allowing for timely medical intervention, which can significantly reduce the burden on healthcare systems and improve patient outcomes.

C. Challenges

Developing a system capable of delivering accurate predictions using limited data and default model parameters, as seen in the original study, presents challenges. Enhancing prediction accuracy requires comprehensive data exploration, algorithmic tuning, and deep learning model development.

D. Solution

Our solution involves conducting a comprehensive EDA, applying hyperparameter tuning to existing machine learning models such as Random Forest and XGBoost, and introducing deep learning models like DNNs or CNNs. We will also integrate a larger dataset, such as the Framingham Heart Study, to improve model generalization. The use of advanced feature engineering and data augmentation techniques will further enhance the model's predictive power.

E. Contribution

Our contributions include improving upon the original heart disease prediction framework by integrating deep learning models, conducting extensive data analysis, and fine-tuning existing models. These enhancements aim to increase the accuracy, robustness, and generalization of the system, making it more applicable to real-world, home-based care scenarios.

II. Related Work

A. Existing Solutions

Several machine learning models have been applied to heart disease prediction, including Logistic Regression, K-Nearest Neighbor, Support Vector Machine, Naive Bayes, Decision Tree, Random Forest, and XGBoost. Random Forest models achieved the best accuracy in the original paper (90.16%). Additionally, deep learning models like DNN and CNN have been explored in recent research but were not applied in the original study.

B. Comparison

Our approach will differ by including a deeper exploration of data patterns through advanced EDA techniques and experimenting with hyperparameter tuning. Additionally, we will

incorporate deep learning models and test more machine learning algorithms, introducing larger and more diverse datasets, ensuring that the best-performing model is chosen.

III. Design and Implementation

A. Solution Strategy

We will begin by performing comprehensive EDA to uncover hidden correlations, patterns, and relationships in the data. This will involve statistical analyses, visualizations, and feature engineering. Following this, hyperparameter tuning will be applied to improve the performance of existing machine learning models such as Random Forest and XGBoost. A deep learning model will also be developed and optimized using techniques like dropout regularization, batch normalization, and adaptive learning rates.

B. Product Description

The final product will be a hybrid machine learning and deep learning framework that can be deployed for at-home heart disease prediction. The system will consist of models trained on both the Cleveland dataset and larger public datasets, offering more accurate and reliable predictions.

C. Algorithms

We will use the following algorithms:

- Machine Learning: Random Forest, XGBoost, Gradient Boosting Machines, Support Vector Machines.
- Deep Learning: Deep Neural Network (DNN), potentially expanding to CNN for advanced feature extraction.

D. Implementation

The models will be implemented using Python, with libraries such as pandas, NumPy, Matplotlib, Seaborn for data analysis, and Scikit-Learn, XGBoost, and Pytorch for machine learning and deep learning. We will also use Jupyter Notebook as our development environment.

E. Methodology

- Hardware: We will use local development machines and potentially cloud-based environments for larger datasets.
- Operating system: macOS.
- Programming language: Python.
- Libraries: Scikit-Learn, XGBoost, Pytorch, NumPy, Matplotlib, Seaborn, pandas.

References

[1] G. Kumar Sahoo, K. Kanike, S. K. Das, and P. Singh, "Machine learning-based heart disease prediction: A study for home personalized care," in *2022 IEEE 32nd International Workshop on Machine Learning for Signal Processing (MLSP)*, 2022, pp. 01–06.