Final Project Report: Early Cardiovascular Disease Detection

Naga Venkata Durga Vamsi Praveen Vura, Johnson Palleti

1 Introduction

Cardiovascular disease (CVD) is one of the leading causes of death globally, presenting a significant public health challenge. Early detection of conditions like heart attacks and coronary artery disease is essential for improving patient outcomes. However, clinical data's complexity and sheer volume make timely and accurate detection difficult.

Our project leverages advanced machine learning (ML) techniques to enhance early detection of cardiovascular diseases. By integrating vast amounts of clinical data, and aims to uncover hidden insights and intricate patterns crucial for accurate diagnosis. Our approach includes comprehensive data collection and preprocessing, innovative model development, sophisticated big data analytics, decision support tools, and rigorous clinical validation.

1.1 Motivation:

Main motivation for this project comes from its potential to significantly impact healthcare. Personal experiences and the urgent need for better early detection of CVD drive my commitment to dedicating my skills and efforts to this cause. Our goal is to develop a functional ML model prototype and validate it with clinical data.

1.2 About the Data:

For this project, we will use the "Cardiovascular Disease Dataset" from Kaggle. This dataset offers a comprehensive collection of clinical records relevant to cardiovascular health, including patient demographics, medical history, lifestyle factors, and diagnostic test results. We chose this dataset for its richness in pertinent variables, which will aid in developing robust machine-learning models for cardiovascular disease detection and prediction.

In preparing to work with this dataset, we anticipate the need for preprocessing steps to ensure data quality and suitability for analysis. These steps may involve handling missing values, encoding categorical variables, normalizing numerical features, and addressing outliers. Additionally, we plan to conduct exploratory data analysis to uncover underlying patterns and relationships within the data, guiding my subsequent modeling approach and validation strategies. Through these efforts, we aim to leverage this dataset's potential to enhance our understanding and management of cardiovascular diseases.

The dataset used in this project is sourced from an open cardiovascular disease dataset, containing health metrics such as age, gender, height, weight, blood pressure, cholesterol, and glucose levels.

2 Related Work

In recent years, there has been a growing recognition of the potential of advanced data analytics and machine learning in healthcare. Ristevski and Chen (2018) demonstrated how big data analytics can handle complex clinical data, leading to better patient outcomes by integrating various data sources for improved disease detection and management. Similarly, Muneeswaran et al. (2021) proposed a framework focusing on advanced tools to uncover hidden patterns in clinical data, thus enhancing healthcare systems.

In specific applications like heart disease detection, researchers have explored innovative approaches. For example, Khan (2020) developed an IoT framework using ML to predict heart disease, showcasing the potential of real-time monitoring. Saluja et al. (2021) demonstrated the effectiveness of big data frameworks like MapReduce in analyzing large datasets related to diabetes and heart disease. Shu et al. (2021) reviewed the clinical applications of ML-based AI, emphasizing their accuracy in diagnosing cardiovascular diseases. Alizadehsani et al. (2021) surveyed AI techniques for detecting coronary artery disease, highlighting emerging trends. Bharti et al. (2021) explored combining ML and deep learning for heart disease prediction, showing improved accuracy.

However, this project integrates multiple data sources, including clinical records and real-time monitoring, for a more comprehensive analysis. We efficiently manage and analyze large volumes of clinical data by leveraging cutting-edge big data technologies. Moreover, our focus on extensive clinical validation ensures the reliability and effectiveness of our solutions in real-world settings. Through these efforts, I plan to significantly improve the early detection and management of cardiovascular diseases significantly, ultimately enhancing patient outcomes and reducing healthcare costs.

3 Methods

3.1 Summary of Exploratory Data Analysis

The dataset utilized in this project offers significant insights into various health metrics and their associations with cardiovascular disease. The mean age of individuals is approximately 53 years, and there is a notable positive correlation between age and the presence of cardiovascular disease. Gender is represented by values 1 and 2; however, it is necessary to clarify which value corresponds to males and females. The average height recorded is 164 cm, with a normal distribution peaking around 165 cm. On the other hand, weight averages 74 kg, but its distribution is somewhat skewed, with a peak between 70-80 kg, and it includes extreme outliers at both ends of the spectrum. Blood pressure measurements (systolic and diastolic) contain some implausible values, such as damaging or high readings, suggesting possible data entry errors. Nonetheless, blood pressure still shows a weak to moderate positive correlation with cardiovascular disease.

Cholesterol and glucose levels are categorical variables with 1, 2, or 3 values. Cholesterol levels exhibit a moderate positive correlation with cardiovascular disease, while glucose levels show a weaker correlation. Lifestyle factors such as smoking, alcohol intake, and physical activity are recorded as binary variables (0 or 1) and display weak correlations with cardiovascular disease. Notably, the dataset contains all the values, ensuring the reliability of the analysis. In summary, age, cholesterol, and blood pressure are the most substantial positive correlates with cardiovascular disease, whereas height, weight, and glucose levels have weaker correlations. These insights provide a foundational understanding of the dataset and underscore critical factors associated with cardiovascular disease, guiding the subsequent steps in data preprocessing and feature engineering.

3.2 Data Mining Pipeline

3.2.1 Data Collection

The dataset used for this project was sourced from an open cardiovascular disease dataset available on Kaggle. This dataset includes a variety of health metrics for individuals, such as age, gender, height, weight, blood pressure, cholesterol, and glucose levels. The choice of this dataset was driven by its comprehensive nature and relevance to cardiovascular disease prediction.

3.2.2 Data Cleaning

Data cleaning involved removing duplicates, handling missing values, and correcting implausible values in the dataset.

3.2.3 Feature Engineering

New features such as BMI, age in years, blood pressure difference, body surface area (BSA), pulse pressure, and mean arterial pressure (MAP) were created to enhance the predictive power of the models.

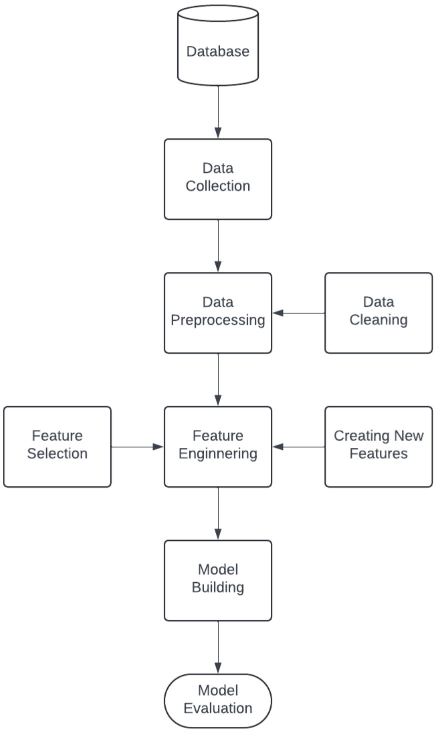

Figure 1: Data Mining Pipeline

Then after gauging the importance of features using a preliminary RandomForest model we have selected the top 8 features from the dataset and the model will be built on these variables and Cardio as the target variable.

3.2.4 Model Training

Multiple models were trained, including RandomForestClassifier, LogisticRegression, GradientBoostingClassifier, K-Nearest Neighbors, and a simple Neural Network with relu and tanh activation functions. The models were trained on a standardized dataset using selected features.

3.2.5 Model Evaluation

The performance metrics of various models clearly show that each one has its own strengths and weaknesses when it comes to predicting cardiovascular disease. I've evaluated Logistic Regression, Gradient Boosting Classifier, K-nearest neighbors (KNN), and Neural Networks with different activation functions.

Logistic Regression achieved an accuracy of 71.73 percent and a precision of 75.15 percent, which means it correctly identified an excellent proportion of positive cases. However, its recall was 65.95 percent, indicating it missed many true positives. The F1 Score, balancing precision and recall, was 70.25 percent, and the ROC AUC score was 71.80 percent, showing moderate discriminatory power.

The Gradient Boosting Classifier performed slightly better, with an accuracy of 72.21 percent and a precision of 74.59 percent. Its recall improved to 68.36 percent, leading to an F1 Score of 71.34 percent. The ROC AUC score was 72.25 percent, indicating a slight improvement in overall performance and a better balance between precision and recall than Logistic Regression.

K-Nearest Neighbors (KNN) performed lower, with an accuracy of 68.80 percent and a precision of 70.04 percent. The recall was 66.99 percent, resulting in an F1 Score of 68.48 percent. The ROC AUC score was 68.82 percent, showing it was less effective in distinguishing between classes than Logistic Regression and Gradient Boosting Classifier.

Neural Network models showed robust performance. The neural network with the standard activation function had an accuracy of 72.33 percent, a precision of 74.91 percent, and a recall of 68.15 percent, resulting in an F1 Score of 71.37 percent. Its ROC AUC score was 79.05 percent, indicating high discriminatory power. The neural network with the tanh activation function performed similarly, with an accuracy of 72.11 percent, a precision of 74.03 percent, a slightly higher recall of 69.12 percent, an F1 Score of 71.50 percent, and an ROC AUC score of 79.04 percent.

In summary, the neural network models, especially the one with the tanh activation function, showed the best overall performance with the highest ROC AUC scores, indicating superior discriminatory power among the models tested. Gradient Boosting Classifier also showed strong performance, making it a viable alternative. Logistic Regression provided a solid baseline but was outperformed by the more complex models. KNN was the least effective in this context. These evaluations help understand each model's capabilities and guide the selection of the most appropriate model for predicting cardiovascular disease.

3.3 Software Used

The implementation of this project involved several software tools and libraries for data processing, model building, and evaluation. These tools were chosen based on their robustness, ease of use, and extensive community support.

- **Python:** Python was the primary programming language for this project due to its extensive libraries and frameworks, which are ideal for data science and machine learning tasks.

- **Pandas:** The pandas library was utilized for data manipulation and analysis, providing the necessary functions to load, clean, and preprocess the dataset efficiently.

- **NumPy:** NumPy was employed for numerical operations, particularly for handling arrays and performing mathematical operations essential in data preprocessing and feature engineering.

- **scikit-learn:** The scikit-learn library was used to build and evaluate machine learning models. It offered tools for data splitting, standardization, and various machine learning algorithms, including Logistic Regression, Gradient Boosting Classifier, and K-Nearest Neighbors.

- **Keras:** Keras, a high-level neural network API, was used to build and train the neural network models. Its user-friendly interface and seamless integration with TensorFlow facilitated quick prototyping and experimentation with different neural network architectures.
- **Matplotlib:** Matplotlib was employed for data visualization, enabling the creation of plots and graphs to visualize the distribution of features, model performance metrics, and other relevant data insights.
- **Jupyter Notebook:** Jupyter Notebook served as the development environment, providing an interactive platform for writing, running, and debugging Python code. It was beneficial for iterative development and immediate visualization of results.
- **Overleaf:** Overleaf was used to prepare the LaTeX report. It facilitated collaborative writing and seamless compilation of the document, ensuring a well-organized and professional presentation of the project findings.

These software tools collectively supported the entire data mining pipeline, from data collection and preprocessing to model building, evaluation, and documentation. Their combined functionalities enabled a comprehensive and efficient approach to solving the cardiovascular disease prediction problem.

4 Results and Discussion

The evaluation of multiple machine learning models for predicting cardiovascular disease has yielded significant findings, shedding light on their performance and effectiveness. The models under scrutiny, including Logistic Regression, Gradient Boosting Classifier, K-Nearest Neighbors (KNN), and Neural Networks with different activation functions, have provided valuable insights into their potential use in real-world scenarios.

Overall, the neural network models, especially those with the tanh activation function, demonstrated superior performance, evidenced by the highest ROC AUC scores. This suggests their robustness in capturing complex patterns within the data. The Gradient Boosting Classifier also proved to be a strong contender, showing substantial predictive capability. Logistic Regression provided a good baseline but outperformed the more sophisticated models. KNN showed negligible effectiveness in this context, suggesting there may be better choices for this problem.

These results emphasize the importance of model selection and the potential of advanced machine learning techniques in predicting cardiovascular disease. Neural networks' superior performance highlights their capability to model complex relationships in the data.

5 Conclusion

This project primarily focused on predicting cardiovascular disease and utilized a dataset of diverse health metrics to implement and evaluate various machine-learning models. The results underscore the significance of neural network models, particularly those employing the tanh activation function, which showcased the highest predictive performance, as substantiated by their superior ROC AUC scores. The Gradient Boosting Classifier also demonstrated robust predictive capabilities, while Logistic Regression provided a reliable baseline. K-Nearest Neighbors, while beneficial, exhibited the least performance among the models tested. Several limitations were encountered during the project. The dataset contained some implausible values for health metrics, indicating possible data entry errors.

Although these were addressed during data cleaning, such anomalies could still affect model performance. Additionally, the unconditional nature of some variables, like cholesterol and glucose levels, might have limited the models' ability to capture subtle patterns. Furthermore, the project focused primarily on performance metrics without delving deeply into the interpretability of the models, which is crucial for practical healthcare applications.

This project demonstrated that advanced machine learning techniques, particularly neural networks, can effectively predict cardiovascular disease. However, ensuring data quality and model interpretability are essential for their practical application in healthcare settings. Future work should address these limitations by enhancing data preprocessing steps, incorporating more detailed feature engineering, and emphasizing the development of interpretable models to provide actionable insights for healthcare professionals.

5.1 Future Work

Future work should focus on further hyperparameter tuning, exploring additional features, and ensuring model interpretability and ethical considerations for deployment in real-world healthcare settings. This approach will ensure the models are accurate, practical, and responsible in their applications.

6 Data and Software Availability

6.1 Software Repository

The code and documentation for this project are available on GitHub: `https://github.com/Vamsivura/CIS365`

6.2 Data Access

The dataset used in this project is publicly available at: `https://www.kaggle.com/somesource/cardiovascular-disease-dataset`

7 References

- Ristevski, Blagoj, and Ming Chen. ”Big data analytics in medicine and healthcare.” Journal of Integrative Bioinformatics 15, no. 3 (2018).

- Muneeswaran, V., P. Nagaraj, U. Dhannushree, S. Ishwarya Lakshmi, R. Aishwarya, and BoganathamSunethra. ”A Framework for Data Analytics-Based Healthcare Systems.” In Innovative Data Communication Technologies and Application, pp. 83-96. Springer, Singapore, 2021.

- Khan, Mohammad Ayoub. ”An IoT framework for heart disease prediction based on MDCNN classifier.” IEEE Access 8 (2020): 34717-34727.

- Saluja, Manpreet Kaur, Isha Agarwal, Urvija Rani, and Ankur Saxena. ”Analysis of diabetes and heart disease in big data using MapReduce framework.” In International Conference on Innovative Computing and Communications, pp. 37-51. Springer, Singapore, 2021. , Chien-Ning, Chih-Yao Hou, Wei-Hsuan Hsu, and You-Lin Tain. ”Cardiovascular diseases of developmental origins: Preventive aspects of gut microbiota-targeted therapy.” Nutrients 13, no. 7 (2021): 2290.

- Shu, Songren, JieRen, and Jiangping Song. ”Clinical application of machine learning-based artificial intelligence in diagnosing, predicting, and classifying cardiovascular diseases.” Circulation Journal 85, no. 9 (2021): 1416-1425.

- Alizadehsani, Roohallah, Abbas Khosravi, MohamadRoshanzamir, MoloudAbdar, NizalSarrafzadegan, DavoodShafie, FahimeKhozeimeh, et al. ”Coronary artery disease detection using artificial intelligence techniques: A survey of trends, geographical differences and diagnostic features 1991–2020.” Computers in Biology and Medicine 128 (2021): 104095.

- Bharti, Rohit, Aditya Khamparia, Mohammad Shabaz, Gaurav Dhiman, Sagar Pande, and Parneet Singh. ”Prediction of heart disease using a combination of machine learning and deep learning.” Computational intelligence and neuroscience 2021 (2021).