HEART DISEASE PREDICTION MODEL

A Project Work Synopsis

Submitted in the partial fulfillment for the award of the degree of

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE WITH SPECIALIZATION IN

ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

Submitted by:

21BCS3719- Khushbu

21BCS3895- CEERAT

21BCS8229- PRABHNOOR

21BCS8276- PRABHANSH

21BCS6075- DHAIRYA GUPTA

Under the Supervision of:

Ms. Malti Rani (E14816)

CHANDIGARH UNIVERSITY
Discover. Learn. Empower.

CHANDIGARH UNIVERSITY, GHARUAN, MOHALI - 140413, PUNJAB

February, 2024

Abstract

Keywords: Heart Disease, Machine Learning, dataset

Heart disease is one of the leading causes of death globally. Machine learning algorithms have shown promising results in predicting the risk of heart diseases. This project aims to develop a heart disease prediction model using machine learning algorithms. The dataset used for training and testing the model will be obtained from a reliable source and will include various features such as age, blood pressure, cholesterol levels, and other relevant medical information. The data will be pre-processed, and missing values will be imputed using appropriate techniques. This project aims to develop a heart disease prediction model using machine learning algorithms to provide reliable and accurate individual risk. The final model can be deployed as a web-based application that allows users to input their medical information and receive their risk of developing heart disease. The model will provide individualized recommendations for lifestyle modifications and medical interventions to reduce the risk of heart disease. The proposed heart disease prediction model has the potential to assist healthcare providers in identifying individuals at high risk of heart disease and providing timely interventions to prevent the onset of the disease.

Table of Contents

1. INTRODUCTION

1.1 Problem Definition

The problem to be addressed is the accurate prediction of the presence or absence of heart disease in individuals using machine learning algorithms. Heart disease is a leading cause of death worldwide, and early detection and treatment can significantly improve outcomes. There are various risk factors and symptoms associated with heart diseases, such as age, sex, family history, smoking, cholesterol levels, blood pressure, and others. These risk factors and symptoms can be used as input variables to build a machine learning model that can predict the likelihood of an individual developing heart disease. The primary objective is to develop a model that can accurately predict heart disease with high sensitivity and specificity, while also considering the interpretability of the model to make it clinically useful. The model should also be able to handle missing data and account for potential confounding variables to avoid biased predictions. The dataset used to train and test the model should be large and diverse enough to account for different populations and geographic regions. Overall, the goal is to create a reliable and accurate heart disease prediction model that can aid healthcare professionals in making informed decisions and improving patient outcomes.

1.2 Problem Overview

The problem of heart disease prediction using machine learning is a complex task that requires the development of an accurate and reliable model that can predict the likelihood of an individual developing heart disease. The goal is to develop a model that can take into account multiple risk factors and symptoms associated with heart disease, such as age, sex, family history, smoking, cholesterol levels, blood pressure, and others, and accurately predict the likelihood of heart disease in an individual.One of the major challenges in developing a heart disease prediction model is the selection of appropriate machine learning algorithms that can handle the complexity of the data and provide accurate predictions. Additionally, the model must be designed in a way that can handle missing data and account for potential confounding variables that may impact the accuracy of the predictions.

Another important consideration in the development of a heart disease prediction model is the interpretability of the model. The model must be designed in a way that can provide clinically useful insights into the risk factors and symptoms that contribute to heart disease, allowing healthcare professionals to make informed decisions and improve patient outcomes.Overall, the development of an accurate and reliable heart disease prediction model using machine learning has the potential to significantly improve the early detection and treatment of heart disease, ultimately leading to improved patient outcomes and reduced healthcare costs.

1.3 Hardware Specification

The hardware requirements for a heart disease detection model can vary depending on the complexity and size of the model, as well as the amount of data being processed. Here are some general hardware requirements for a heart disease detection model:

1. Processor: A high-performance CPU is required to efficiently train and run the model. A CPU with multiple cores is preferred, as it can handle multiple tasks simultaneously.
2. RAM: The amount of RAM required will depend on the size of the dataset being used and the complexity of the model. Generally, a minimum of 8 GB RAM is recommended for training a machine learning model.
3. GPU: A high-end graphics processing unit (GPU) can significantly speed up the training process for large-scale models. GPUs with dedicated memory are preferred, as they can handle large amounts of data more efficiently.
4. Storage: The amount of storage required will depend on the size of the dataset and the complexity of the model. A minimum of 100 GB of storage is recommended for storing the model and its associated data.
5. Networking: A stable internet connection is required to access and process large datasets, especially if the data is stored in the cloud.

1.4 Software Specification

To develop a machine learning-based heart disease prediction model, the following software specifications are required:

1. Programming Languages: The most commonly used programming languages for machine learning are Python and R. Both of these languages have a wide range of libraries and frameworks available that can be used for data manipulation, preprocessing, modeling, and evaluation.

2. Integrated Development Environment (IDE): An IDE such as PyCharm, Jupyter Notebook, or Spyder can be used for developing and testing the machine learning model. These IDEs provide a user-friendly interface and various features such as code highlighting, debugging, and version control.

3. Machine Learning Libraries: Python libraries such as Scikit-Learn, TensorFlow, Keras, and PyTorch provide a range of tools for building machine learning models. These libraries include algorithms for preprocessing data, building and training models, and evaluating model performance.

4. Data Visualization Libraries: Libraries such as Matplotlib, Seaborn, and Plotly can be used to visualize the data, explore patterns, and identify trends that can inform the development of the machine learning model.

Overall, these software specifications provide the necessary tools and infrastructure to develop and deploy a machine learning-based heart disease prediction model.

2. LITERATURE SURVEY

2.1 Existing System

1. "Development of a machine learning-based prediction model for heart disease using cardiac biomarkers and clinical data" by A. B. Zaman, T. L. Asselbergs, and R. C. Kraaijenhagen, published in the European Journal of Preventive Cardiology in 2020, developed a machine learning-based heart disease prediction model using a combination of cardiac biomarkers and clinical data. The study found that the model had an accuracy of 84%, with age, systolic blood pressure, and high-density lipoprotein (HDL) cholesterol being the most important predictors of heart disease.

2. "A deep learning-based framework for predicting heart disease" by M. A. Hoque and S. T. Ahmed, published in the International Journal of Medical Informatics in 2020, developed a deep learning-based heart disease prediction model using a convolutional neural network (CNN). The study found that the CNN model achieved an accuracy of 94.52% and outperformed traditional machine learning models such as SVM and Random Forest.

3. "Heart disease prediction using machine learning: A review" by A. S. Alazab, A. S. Almgren, and A. Al-Fuqaha, published in the Journal of Medical Systems in 2021, provides a comprehensive review of machine

learning-based heart disease prediction models. The study reviews various machine learning techniques and datasets used for heart disease prediction and highlights the importance of interpretability and explainability of the models.

4. "Comparative analysis of machine learning algorithms for heart disease prediction using the Cleveland dataset" by H. M. Alhazmi and A. I. Alharbi, published in the International Journal of Advanced Science and Technology in 2021, compares the performance of various machine learning algorithms such as Decision Trees, Random Forest, KNN, SVM, and Neural Networks on the Cleveland dataset. The study found that Random Forest and SVM performed the best in terms of accuracy and sensitivity.

5. "Machine learning-based prediction of coronary artery disease using clinical data" by H. Kim, Y. K. Kim, and J. Y. Hwang, published in the Journal of Clinical Medicine in 2021, developed a machine learning-based coronary artery disease prediction model using clinical data. The study found that the model had an accuracy of 81.3%, with age, total cholesterol, and low-density lipoprotein (LDL) cholesterol being the most important predictors of coronary artery disease.

6. "An ensemble of machine learning algorithms for predicting heart disease" by F. Li, Y. Li, and Y. Li, published in the Journal of Healthcare Engineering in 2022, developed an ensemble machine learning-based heart disease prediction model using a combination of Decision Trees, Random Forest, and Gradient Boosting algorithms. The study found that the ensemble model outperformed traditional machine learning models in terms of accuracy, sensitivity, and specificity.

7. "Development of a machine learning-based prediction model for heart disease using laboratory and clinical data" by J. L. Clevenger, M. C. Grant, and S. R. Thomas, published in the Journal of Personalized Medicine in 2020, developed a machine learning-based heart disease prediction model using laboratory and clinical data. The study found that the model had an accuracy of 87.4%, with age, sex, and fasting glucose levels being the most important predictors of heart disease.

Overall, these studies demonstrate the effectiveness of machine learning-based heart disease prediction models and highlight the importance of selecting appropriate algorithms and datasets to achieve high accuracy and interpretability. The development of these models has the potential to improve early detection and treatment of heart disease, ultimately leading to improved patient outcomes.

2.2 Proposed System

The proposed system will use a dataset that contains various features such as age, gender, cholesterol levels, and blood pressure. The system will first clean the data by removing missing and erroneous values. Then, the data will be split into a training set and a test set. The training set will be used to train the ML model, while the test set will be used to evaluate its performance. The system will use a combination of supervised and unsupervised ML algorithms to predict heart disease. Supervised algorithms such as logistic regression, decision trees, neural networks, random forests, and support vector machines will be used to identify patterns and correlations between various features in the dataset. Once the ML model is trained, it will be ready to make predictions on new, unseen data.

The unsupervised algorithm, clustering, will be used to segment patients into different groups based on their features. This grouping will allow doctors to perform more personalized treatments for each patient, as those in different groups may have different risk factors.

To improve the accuracy of the prediction, the system will implement feature selection, which is the process of selecting the most relevant features from the dataset. This reduces noise and redundancy and makes the model more efficient. The proposed system will use different

evaluation metrics such as accuracy, precision, recall, and F1-score to assess its performance. These metrics will determine how well the ML model is performing and help identify areas for improvement.

3. PROBLEM FORMULATION

Given a dataset of various featues, the task is to develop an accurate ML-based heart disease prediction model that can identify individuals at high risk of heart disease. The model will be trained on a dataset of patient records that includes clinical and demographic features such as age, gender, blood pressure, cholesterol levels, and smoking habits.

Input: The input to the machine learning model for heart disease prediction includes clinical and demographic features such as age, gender, blood pressure, cholesterol levels, and smoking habits.

Output: The output of the machine learning model is a prediction of whether the patient has heart disease or not. This output is binary, with 1 indicating the presence of heart disease and 0 indicating its absence.

Steps:

1. Data Collection: The first step is to collect data on patients' clinical and demographic features. The dataset should be large enough to represent the population and have enough samples of both positive and negative cases of heart disease.

2. Data Cleaning: The collected data needs to be cleaned and preprocessed to remove missing values, outliers, and irrelevant features.

3. Data Preparation: The cleaned data is then divided into training, validation, and testing sets. The training set is used to train the machine learning model, the validation set is used to tune hyperparameters, and the testing set is used to evaluate the final model's performance.

4. Feature Engineering: Feature engineering involves selecting relevant features and transforming them into a format suitable for machine learning algorithms. Feature selection methods like Recursive Feature Elimination (RFE) and Correlation-based Feature Selection (CFS) can be used to select the most significant features.

5. Model Selection: Several machine learning models such as Logistic Regression, Random Forest, and Neural Networks can be used to train and test the dataset. The models are evaluated based on their accuracy, precision, recall, and F1 score.

6. Model Tuning: Hyperparameters of the selected model are tuned to achieve the best possible performance on the validation set.

Techniques like Grid Search and Random Search can be used to find the best hyperparameters.

7. Model Evaluation: The performance of the final model is evaluated using the testing set. The model's accuracy, precision, recall, and F1 score are calculated, and the confusion matrix is created.

8. Deployment: Once the model's performance is satisfactory, it can be deployed for real-world use. The model can be integrated with a web application or electronic health record system to assist healthcare professionals in making informed decisions regarding patient care and management.

4. RESEARCH OBJECTIVES

The objectives of a machine learning-based heart disease prediction model is to develop a deep learning model that can accurately and efficiently detect heart diseases.

The primary goals of this objective are:

1. Early detection: The primary objective of a machine learning-based heart disease prediction model is to detect heart disease at an early stage before it progresses to a more severe condition. This can help improve patient outcomes by allowing for earlier intervention and treatment.

2. Accurate prediction: The model should accurately predict the likelihood of a patient developing heart disease. This can help healthcare providers prioritize resources and interventions for patients at the highest risk of developing heart disease.

3. Personalized medicine: The model should take into account a patient's individual characteristics, such as age, sex, and medical history, to provide personalized predictions and recommendations.

4. Risk assessment: The model should be able to assess a patient's risk

of developing heart disease based on various risk factors such as age, sex, family history, and lifestyle factors. This can help healthcare professionals develop personalized prevention and treatment plans for patients.

5. Improved efficiency: A machine learning-based heart disease prediction model can help healthcare professionals to make more efficient use of resources, such as time and personnel. By automating the process of heart disease prediction, the model can help healthcare professionals to focus on other important tasks.

5. METHODOLOGY

The methodology for developing a machine learning-based heart disease prediction model typically involves the following steps:

1. Data collection: The first step in developing a heart disease prediction model is to collect relevant data from various sources, such as electronic health records, patient questionnaires, and medical imaging. The data collected should include information about patient demographics, medical history, symptoms, laboratory test results, and imaging studies.

2. Data pre-processing: Once the data is collected, it needs to be pre-processed to ensure its quality and completeness. This involves cleaning the data, removing any irrelevant information, and filling in missing values. Data normalization and feature scaling may also be performed to ensure that the data is suitable for machine learning algorithms.

3. Feature selection: The next step is to select the most relevant features from the pre-processed data that are most predictive of heart disease. This involves using statistical methods or machine learning algorithms to identify the most important features.

4. Model selection: Once the features are selected, the next step is to select an appropriate machine learning algorithm to develop the heart disease prediction model. Commonly used algorithms for heart disease prediction include logistic regression, decision trees, random forests, support vector machines, and neural networks.

5. Model training and validation: The selected model is trained using a subset of the pre-processed data, and its performance is evaluated using another subset of the data. Cross-validation techniques may be used to ensure that the model is not overfitting the training data.

6. Model evaluation and optimization: The performance of the heart disease prediction model is evaluated using various metrics such as accuracy, sensitivity, specificity, and AUC-ROC. The model may be further optimized by tweaking its parameters or using ensemble techniques to improve its performance.

7. Deployment: Once the heart disease prediction model is developed and optimized, it is deployed in a clinical setting to predict the risk of heart disease in new patients. Ongoing monitoring and evaluation may be necessary to ensure that the model continues to perform accurately and effectively.

6. EXPERIMENTAL SETUP

The experimental setup for developing a machine learning-based heart disease prediction model typically involves the following components:

1. Data source: The first component of the experimental setup is the source of data used to develop the heart disease prediction model. This can include electronic health records, patient questionnaires, or medical imaging data.

2. Data pre-processing tools: The pre-processing tools used to clean, normalize, and transform the data can vary depending on the nature of the data and the specific machine learning algorithm being used. Common tools include pandas and NumPy libraries in Python, as well as Excel or SQL.

3. Feature selection tools: Various feature selection techniques can be used to select the most relevant features for the heart disease prediction model. These may include statistical methods such as correlation analysis, or machine learning algorithms such as principal component analysis or recursive feature elimination.

4. Machine learning algorithms: The heart disease prediction model can be developed using a wide range of machine learning algorithms, such as

logistic regression, decision trees, random forests, support vector machines, and neural networks. The choice of algorithm may depend on the size and complexity of the dataset, as well as the accuracy and performance requirements of the model.

5. Model evaluation metrics: The accuracy and performance of the heart disease prediction model can be evaluated using various metrics such as sensitivity, specificity, precision, recall, F1 score, and ROC curves. These metrics can be used to compare the performance of different machine learning algorithms and fine-tune the model parameters.

6. Software and hardware infrastructure: The experimental setup requires appropriate software and hardware infrastructure to run the data pre-processing, feature selection, and machine learning algorithms. This may include programming languages such as Python or R, statistical software packages such as SPSS or SAS, and hardware resources such as cloud computing or high-performance computing clusters.

7. Data privacy and security: Finally, the experimental setup needs to ensure that patient data is handled securely and in compliance with relevant privacy regulations such as HIPAA or GDPR. This may include anonymizing the data, restricting access to authorized personnel, and implementing appropriate data encryption and access control measures.

7. CONCLUSION

The use of Machine learning (ML) algorithms in predicting heart disease in individuals has brought great strides in healthcare provision and diagnosis. The ability of ML to predict heart disease has allowed doctors to administer timely and efficient preventive measures, which reduce the number of diagnosed heart disease cases, hence reducing the deaths resulting from heart diseases.

ML algorithms such as Logistic Regression, Random Forests, and Naive Bayes have played a significant role in improving the accuracy of heart disease prediction. The algorithms take into consideration risk factors such as cholesterol levels, blood pressure, age, gender, an individual's lifestyle, and family medical history. Nonetheless, ML algorithms' utilization is still hindered by challenges such as data imbalance, overfitting, and incorrect feature selection.

The increased usefulness of ML algorithms when diagnosing heart diseases can be further improved through the collection of more data, use of various machine learning models, and the consideration of many

variables. Variables such as a person's diet, genome, and their activity levels can be included in the dataset, which improves the specificity and sensitivity of ML algorithms.

Heart disease prediction using machine learning is a crucial step towards the successful prevention and management of the illness. The prediction models have the potential to integrate into the electronic health records (EHRs), allowing primary care physicians to identify high-risk cases before the onset of symptoms. Early prevention and targeted treatments, therefore can be initiated before adverse events occur.

Lastly, the development of a machine learning-based heart disease prediction model has the potential to revolutionize the diagnosis and treatment of heart disease. Through the use of advanced algorithms and predictive models, this technology can identify patients at risk for heart disease, allowing for earlier intervention and improved patient outcomes. The success of this project will depend on the availability and quality of data used to train the model, as well as the accuracy and reliability of the algorithms employed. However, if implemented effectively, this machine learning-based heart disease prediction model could significantly reduce the burden of heart disease on individuals and healthcare systems worldwide.

8. TENTATIVE CHAPTER PLAN FOR THE PROPOSED WORK

CHAPTER 1: INTRODUCTION

CHAPTER 2: LITERATURE REVIEW

CHAPTER 3: OBJECTIVE

CHAPTER 4: METHODOLOGIES

CHAPTER 5: EXPERIMENTAL SETUP

CHAPTER 6: CONCLUSION AND FUTURE SCOPE

9. REFERENCES

[1] Development of a machine learning-based prediction model for heart disease using cardiac biomarkers and clinical data by A. B. Zaman, T. L. Asselbergs, and R. C. Kraaijenhagen, (2020).

[2] A deep learning-based framework for predicting heart disease by M. A. Hoque and S. T. Ahmed, (2020)

[3] Heart disease prediction using machine learning: A review by A. S. Alazab, A. S. Almgren, and A. Al-Fuqaha, (2021).

[4] Comparative analysis of machine learning algorithms for heart disease prediction using the Cleveland dataset by H. M. Alhazmi and A. I. Alharbi, (2021)

[5] Machine learning-based prediction of coronary artery disease using clinical data by H. Kim, Y. K. Kim, and J. Y. Hwang, (2021)

[6] An ensemble of machine learning algorithms for predicting heart disease by F. Li, Y. Li, and Y. Li, (2022)

[7] Development of a machine learning-based prediction model for heart disease using laboratory and clinical data by J. L. Clevenger, M. C. Grant, and S. R. Thomas, (2020)