DESIGN AN ML MODEL FOR PREDICTING HEART DISEASE AND INTEGRATE THE MODEL

Abstract

The prognosis of heart disease is one of the most challenging problems in contemporary medicine. Nearly one person dies from heart disease every minute in the modern world. To process vast amounts of data, the healthcare sector needs data science. The study's findings show how well the model predicts heart disease and outperforms other methods while providing information on the most important risk variables. In general, this study advances the rapidly expanding field of machine learning (ML) applications in healthcare and highlights the significance of responsible model creation and use for improved patient care and outcomes. Due to the difficulty of forecasting cardiac illness, automation of the technique is essential to reduce risks and provide the patient with early warning. A properly split dataset is used to train the chosen ML model, and hyper parameter adjustment is done to enhance performance. The model's prediction accuracy is evaluated using metrics including accuracy, precision, recall, F1 score, and area under the curve. The model is used in the real world once it performs satisfactorily.

Keywords: Heart Disease, medical dataset, medicine, prognosis, ML model.

Authors

Dr. Mariyan Richard A
Assistant Professor
Department of Masters of Computer Applications
Nitte Meenakshi Institute of Technology
Bangalore, Karnataka, India.

Ms. Joy Lavinya
Associate Professor
Department of Masters of Computer Applications
Nitte Meenakshi Institute of Technology
Bangalore, Karnataka, India.

Dr. Prasad Naik Hamsavath
Professor
Department of AI & DS
BGS College of Engineering & Technology
Bangalore, Karnataka, India.

I. INTRODUCTION

The study's recommendations for future research center mostly on various data mining methods for predicting heart illness. The heart is the primary muscle in the human body. In essence, it regulates how blood moves through our body. Any heart condition can make other parts of the body more painful. Any disorder that interferes with the heart's normal function is referred to as heart disease. Heart disease is one of the leading causes of death in the modern world. Smoking, consuming alcohol, eating a lot of fat, and leading a sedentary lifestyle are all risk factors for heart disease. According to the World Health Organization, heart disease claims the lives of more than 10 million people annually. The fundamental issue in modern healthcare is the provision of high-quality services and efficient, accurate diagnoses. The developed ML model for heart disease prediction offers healthcare practitioners a valuable tool for early detection and risk assessment. State the objective of your research paper: Developing an ML model for heart disease prediction and deploying it effectively. By providing accurate predictions, this model can aid in making informed decisions and improving patient outcomes Despite being the largest cause of death globally in recent years, cardiac illnesses are also the ones that can be effectively managed and controlled. The proper time of discovery impacts how effectively a disease will be treated overall. The recommended strategy seeks to recognize these heart abnormalities early in order to avoid detrimental effects.

For analysis and knowledge extraction, records of a substantial collection of medical data gathered by medical practitioners are available. Heart disease, encompassing a range of conditions such as coronary artery disease, heart failure, and arrhythmias, has a significant impact on global health. Despite numerous efforts to combat this condition, heart disease remains a leading cause of disability and premature death in both developed and developing countries. Early detection and timely intervention are vital in reducing the burden of heart disease on individuals and healthcare systems. ML-based predictive models can play a pivotal role in achieving this goal by identifying high-risk patients, enabling timely interventions, and potentially preventing adverse cardiac events. Machine learning techniques help in the early diagnosis and prediction of heart illness after data analysis. This study evaluates the performance of different ML techniques, such as Naive Bayes, Decision Tree, Logistic Regression, and Random Forest, in order to predict cardiac sickness at an early stage.

Developing accurate and reliable predictive models for heart disease poses several challenges. One of the foremost challenges is the scarcity of high-quality and diverse datasets. Although medical data is valuable, it is often sensitive and protected by privacy regulations, limiting its accessibility for research purposes. Furthermore, the complex nature of heart disease demands the incorporation of various clinical, demographic, and lifestyle factors into the model, necessitating feature selection techniques to identify the most influential predictors. Additionally, the risk of algorithmic biases and interpretability issues requires careful consideration to ensure fairness and transparency in the model's predictions.

To develop a machine learning model capable of accurately predicting the risk of heart disease in individuals.

To explore and apply feature selection techniques to identify the most relevant and informative predictors for heart disease prediction.

To evaluate the performance of the developed ML model using appropriate metrics and compare it with existing approaches.

To address ethical considerations, such as data privacy and model interpretability, in the deployment of the predictive model in clinical settings.

Regardless of age, heart disease is increasing in both men and women. However, other factors including gender, diabetes, and BMI also play a part in this sickness. In this study, we tried to predict and analyze heart disease by considering factors like age, gender, blood pressure, heart rate, diabetes, and so on. Due to the numerous factors at play, heart disease is difficult to predict.

Coronary heart disease or coronary artery disease are terms used to describe the constriction of the coronary arteries. The heart receives blood and oxygen from the coronary arteries. Many people are made ill or risk death as a result of it. It is a prevalent form of cardiac disease. High blood sugar levels brought on by diabetes can damage blood vessels and the nerves that control the heart and blood vessels. Patients with long-term diabetes are more likely to experience problems.

II. LITERATURE REVIEW

The medical dataset was evaluated by Monika Gandhi et al. using naive bayes, decision trees, and neural networks. There are several different aspects at play. Therefore, the number of features needs to be decreased. Feature selection can be used to achieve this. They claim that doing this saves time. They used neural networks and decision trees.

Thomas, R. J. For the purpose of predicting cardiac illness, Theresa Princy used the K closest neighbour method, neural networks, naive Bayes, and decision trees. To determine the risk of developing heart disease, they used data mining tools.

This comprehensive review explores the application of machine learning algorithms in predicting various cardiovascular diseases, including heart disease. It delves into the performance, strengths, and limitations of algorithms like decision trees, naive Bayes, and artificial neural networks, providing valuable guidance for researchers in choosing suitable models.

Sellappan Palaniyappan and Rafiah Awang constructed Intelligent Heart Disease Prediction Systems (IHDPS) using decision trees, Naive Bayes, and artificial neural networks. Both tabular and graphical representations of the results are presented in order to improve visualisation and facilitate interpretation. It also aids in lowering treatment costs by offering efficient treatments. Finding underlying linkages and patterns has frequently gone unexploited. Advanced data mining methods were used to solve this problem.

III. PROPOSED MODEL

The proposed model for heart disease prediction is designed to leverage the power of machine learning algorithms to accurately and efficiently identify individuals at risk of heart disease. To achieve this, a combination of tree-based ensemble models, logistic regression, and L1 regularization is used to create a robust and interpretable predictive tool. The proposed study evaluates the effectiveness of the four classification algorithms and uses it to forecast cardiac disease. Accurately determining whether a patient has cardiac disease is the aim of this investigation. Data from the patient's health report are entered by the healthcare practitioner. A model that predicts the chance of acquiring heart disease incorporates the data. shows the entire process.

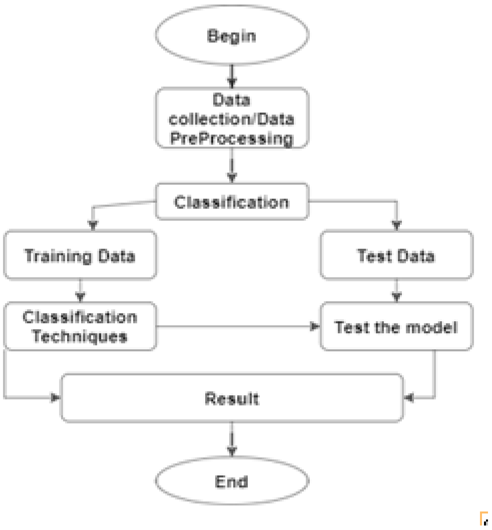

Figure 1

1. **Data Collection and Preprocessing:** Data collection is the first crucial step in the development of an ML model for heart disease prediction. Data preprocessing and preparation are crucial steps in building a successful machine learning model for heart disease prediction. These steps involve cleaning, transforming, and organizing the data in a way that ensures the model's effectiveness and generalization. In this project, data collection entails acquiring clinical and demographic information from patients with known heart disease status, including those with and without heart disease.

    To ensure the dataset's representativeness and diversity, data may be collected from multiple healthcare institutions, research studies, or publicly available datasets. The data collection process should adhere to ethical guidelines, obtain appropriate informed consent, and protect patient privacy by de-identifying sensitive information.

    Data preprocessing is a fundamental stage that prepares the raw data for analysis and model training. The main goals of data preprocessing in this heart disease prediction project are

    - **Outlier Detection and Treatment:** Identify and assess outliers in the data that deviate significantly from the majority of data points. Depending on the nature of the data and the domain knowledge, outliers may be removed or transformed to reduce their impact on the model.
    - **Feature Selection:** Identify the most informative features that contribute significantly to the prediction task. Removing irrelevant or redundant features can improve model efficiency and generalization.

2. **Classification:** In the context of heart disease prediction, classification methods refer to machine learning algorithms used to categorize patients into distinct classes based on their heart disease status (e.g., presence or absence of heart disease). classification is a fundamental task in machine learning where the goal is to assign input data points to one of several predefined classes or categories classification algorithms can be employed to determine whether an individual is at risk of developing heart disease based on various clinical and demographic features. Accuracy, precision, recall, and just a few of the various metrics that are used to calculate and examine the performance of each algorithm.

The various algorithms that were investigated in this study are given below:

- **Decision Tree:** Decision Trees are widely used in various domains, including healthcare, due to their simplicity, transparency, and ability to handle both numerical and categorical data. Because of their speed, dependability, clarity, and minimal data preparation requirements, decision trees are often employed. Decision Trees split the data into subsets based on the values of features. The algorithm searches for the best feature and value to split the data in a way that maximizes the purity of the resulting subsets. Following the matching branch to the value shown by the comparison result, a hop is performed to the next node.

- **Logistic Regression:** For binary classification problems, the classification method known as logistic regression is widely used. Despite its name, it is primarily used for classification, not regression. It is particularly well-suited for problems where the dependent variable (target) is binary, with two possible outcomes, such as "Yes" or "No," "True" or "False," or "1" or "0."regression is a useful tool for classifying data since it uses 13 independent variables. The trained logistic regression model can be used for risk assessment, early detection, and decision support in clinical settings, aiding healthcare professionals in providing personalized care and interventions to patients at higher risk of heart disease.

- **Naive Bayes:** Naive Bayes is a probabilistic classification algorithm based on Bayes' theorem with a strong assumption of feature independence. Despite its simplicity, Naive Bayes has proven to be effective in many real-world classification tasks, including heart disease prediction Naive Bayes can be surprisingly effective for many real-world classification tasks, particularly in situations with limited data and high-dimensional feature spaces. It is commonly used in text classification, spam filtering, sentiment analysis, and other natural language processing tasks, but it is also applicable to various other domains, including healthcare for tasks like disease prediction. Simplicity and Speed: Naive Bayes is computationally efficient and requires relatively less training data compared to other algorithms. Scalability: It can handle large and high-dimensional datasets efficiently. Interpretability: Naive Bayes provides clear and interpretable results, allowing users to understand the reasoning behind the model's predictions. Good for Text Data: Naive Bayes works well for text classification tasks, where the independence assumption is often reasonable.

IV. DATASET

The dataset used in the heart disease prediction project is a crucial component as it forms the foundation for training and evaluating the machine learning models. The model will be constructed using the training data, and its performance will be assessed using the testing data. The dataset contains a collection of records, with each record representing an individual patient. Each patient's information is described by various clinical and demographic features, and the target variable indicates the presence or absence of heart disease for that individual. s an AI language model, I don't have direct access to current data sets. However, I can mention some publicly available data sets that are commonly used in heart disease prediction and classification tasks. Please note that the availability and terms of use of these data sets may vary, so ensure to check the data set's documentation and licensing agreements before using them for research or any other purposes.

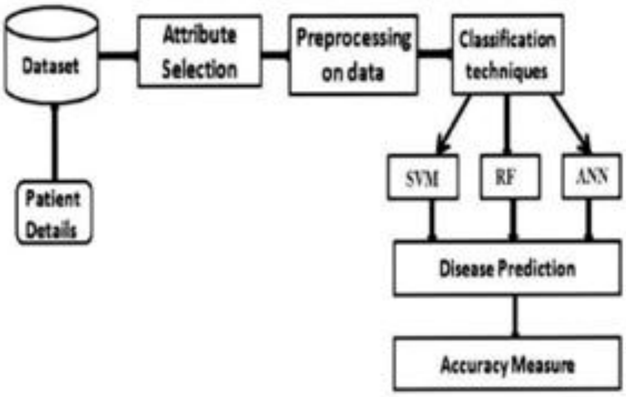

Figure 2

1. **Random Forest:** Random Forest is a popular and powerful ensemble learning method used for both classification and regression tasks. It is an extension of decision trees that builds multiple trees and combines their predictions to improve accuracy and reduce overfitting. Random Forest is widely used in various domains, including healthcare, due to its robustness, flexibility, and ability to handle high-dimensional data. Bagging in random Forest employs the technique of bagging, where multiple decision trees are trained on different subsets of the training data. Each tree is constructed using a random sample (with replacement) from the original dataset. This process is known as bootstrap sampling. The goal of bagging is to create diverse and independent trees, reducing the variance in the model and improving generalization. n addition to using random subsets of the data, Random Forest introduces randomness in feature selection during tree construction. At each split of a tree, the algorithm considers only a random subset of features instead of using all available features. This feature randomness further enhances the diversity of the trees and helps the model to capture different patterns in the data.

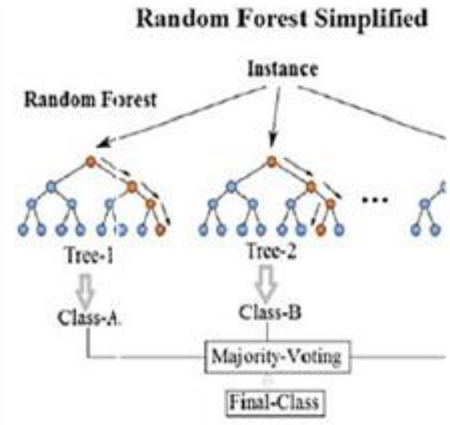

Figure 3

2. **Support Vector Machines:** SVM is widely used in various domains, including image recognition, text classification, and medical diagnosis, including heart disease prediction. A support vector machine is a supervised classifier. n binary classification, SVM aims to find the optimal hyperplane that best separates the data points of different classes. The optimal hyperplane is the one that maximizes the margin, i.e., the distance between the hyperplane and the nearest data points of each class. These data points are called support vectors.

In the context of heart disease prediction, SVM can be applied to a dataset containing patient characteristics, medical history, and relevant clinical features. The algorithm learns the optimal hyperplane that best separates patients with heart disease from those without based on the feature values. Recent investigations have found that they are quite popular. The great overall empirical performance has led to the development of its popularity.

3. **Python:** Python has a vast ecosystem of libraries and frameworks specifically tailored for machine learning and data analysis. Some of the most popular libraries include: Scikit-learn a comprehensive library for various machine learning algorithms, including support for classification, regression, and model evaluation. TensorFlow and Keras Popular deep learning libraries for building and training neural networks, including multi-layer perceptron's (MLPs) and more complex architectures. Pandas a powerful library for data manipulation and analysis, useful for preprocessing and cleaning the heart disease data. NumPy Provides support for numerical operations and array manipulation, often used in conjunction with Pandas for data handling. Matplotlib and Seaborn for data visualization and generating plots to present research findings. Given these advantages, Python is an excellent choice for the research paper's implementation of the ML model for heart disease prediction and for conducting experiments to evaluate the model's performance.

4. **MS Excel:** Microsoft Excel is a spreadsheet application available for both Windows and Mac OS X. Microsoft Excel can be a useful tool in the context of the research paper on developing an ML model for heart disease prediction. While Python is the primary programming language for building and training the machine learning model, Excel can complement the research Data Visualization: Excel's charting features allow researchers

to create simple visualizations to get a quick overview of the data distribution and relationships between variables. While more advanced data visualization can be achieved in Python using libraries like Matplotlib and Seaborn, Excel can serve as a quick visualization tool for initial insights. Data Transformation and Feature Engineering: Excel can be used for basic feature engineering tasks, such as calculating new variables or combining existing ones. For example, researchers can create new columns representing BMI (Body Mass Index) from height and weight data or calculate age groups from birth dates. Summary Statistics Excel's functions can calculate summary statistics like mean, median, standard deviation, and percentiles. These statistics can help researchers better understand the distribution and characteristics of the data. However, it's important to note that Excel has limitations compared to Python when it comes to more complex data analysis and machine learning tasks. Python and its specialized libraries offer more advanced and scalable solutions for model development, tuning, and validation.

V. ETHICAL CONSIDERATIONS

When conducting a research project involving machine learning for heart disease prediction, several ethical considerations need to be addressed to ensure the project's integrity, fairness, and responsible use of data and technology. Here are some key ethical considerations for this type of project:

1. **Data Privacy and Informed Consent:** Ensure that the data used for the research is obtained with proper informed consent from the patients. Patient privacy and data confidentiality must be maintained throughout the project. Personal identifying information should be anonymized or encrypted to protect patient identities.

2. **Bias and Fairness:** Be cautious of potential biases in the data that could lead to unfair or discriminatory outcomes. Biases may arise from historical data, data collection methods, or model training. Efforts should be made to detect and mitigate biases to ensure fair predictions for all individuals, regardless of gender, race, or other sensitive attributes.

3. **Transparency and Interpretability:** Machine learning models used for medical purposes must be transparent and interpretable. Healthcare professionals and patients should be able to understand how the model arrives at its predictions. Black-box models, like deep neural networks, may require additional measures to explain their decisions.

4. **Generalization and Validation:** It is crucial to evaluate the model's performance on diverse and representative datasets. The model should not only perform well on the training data but also generalize to new and unseen patient data. Validation on separate datasets can help ensure the model's reliability.

5. **Clinical Validation and Expert Involvement:** Involving medical experts and clinicians throughout the research project is essential. Medical professionals can provide domain-specific insights, validate the model's predictions, and ensure the project aligns with medical best practices.

6. **Data Bias and Health Disparities:** Analyze the potential impact of data bias on health disparities. Biased data could exacerbate existing health disparities if the model is biased against certain patient populations. Efforts should be made to ensure equal and fair

representation of all groups in the data.Responsible Deployment and Use: If the model is intended for real-world deployment, consider the potential consequences of its use. Ensure that healthcare professionals understand the model's limitations and use it as a decision-support tool rather than a replacement for medical expertise.

VI. FEATURE SELECTION

Feature selection is a critical step in building an accurate and efficient predictive model. It involves identifying and choosing the most relevant and informative features (input variables) from the dataset, while eliminating irrelevant or redundant ones. Effective feature selection not only improves the model's performance but also reduces training time and complexity.

1. **Univariate Feature Selection:** This method evaluates each feature independently based on statistical tests or scoring functions.

   Common techniques include Chi-square test for categorical features and ANOVA F-test for continuous features. The features with the highest scores or the lowest p-values are selected.

2. **Recursive Feature Elimination (RFE):**RFE is an iterative method that recursively removes the least important features from the dataset. The process involves training the model, ranking the features by their importance, and eliminating the least important ones. This continues until the desired number of features is reached or model performance stops improving.

3. **L1 Regularization (Lasso Regression):**L1 regularization adds a penalty term to the cost function based on the absolute magnitude of the model's coefficients.

   As a result, some feature weights may be driven to exactly zero, effectively eliminating those features from the model.Features with non-zero coefficients are selected as the most important ones.

4. **Tree-Based Feature Importance:** Tree-based algorithms like Decision Trees and Random Forests can provide a feature importance score. Features that are more frequently used for splits in the trees are considered more important. You can use this information to rank and select the top features.

5. **Correlation Analysis:** Features that are highly correlated with the target variable tend to be good predictors. Additionally, features with high inter correlations among themselves may be redundant, and one of them can be removed. Pearson correlation coefficient or Spearman rank correlation can be used for continuous variables, while point-biserial correlation can be used for categorical variables.

6. **Feature Importance from Ensemble Models:** If you are using ensemble models like Gradient Boosting Machines (GBM) or XGBoost, they can provide a feature importance ranking. These models can be trained to rank features based on their impact on reducing the model's loss function.

7. **Domain Expertise:** In some cases, domain knowledge can guide the feature selection

process. Consulting with healthcare professionals and experts in cardiology can help identify the most relevant features.

VII. CONCLUSION

Throughout the research process, we have explored various machine learning algorithms, including logistic regression, decision trees, random forest, naive Bayes, support vector machines, and artificial neural networks. Each of these algorithms brings unique strengths to the task of heart disease prediction, enabling accurate risk assessment and personalized interventions. Data preprocessing and preparation played a crucial role in ensuring the quality and integrity of the dataset. Ethical considerations were diligently addressed to protect patient privacy, mitigate biases, and promote fairness in the predictions. The involvement of medical experts and clinicians throughout the research ensured that the model aligns with clinical best practices and maintains interpretability and transparency. The development and deployment of an ML model for heart disease prediction hold the potential to transform healthcare practices and positively impact the lives of countless individuals. By combining the power of data-driven algorithms with ethical considerations and medical expertise future work will benefit from using a larger dataset than that used in this analysis and building a web application based on the Random Forest technique. to enhance outcomes and help doctors successfully and accurately forecast heart disease.

REFERENCES

[1] Jafar Alzubi, Anand Nayyar, Akshi Kumar. "Machine Learning from Theory to Algorithms: An Overview", Journal of Physics: Conference Series, 2018

[2] Fajr Ibrahem Alarsan., and Mamoon Younes 'Analysis and classification of heart diseases using heartbeat features and machine learning algorithms',Journal Of Big Data,2019;6:81.

[3] Internet source [Online].Available (Accessed on May 1 2020): http://acadpubl.eu/ap

[4] Mr. ChalaBeyene, Prof. Pooja Kamat, "Survey on Prediction and Analysis the Occurrence of Heart Disease Using Data Mining Technique", International Journal of Pure and Applied Mathematics, 2018.

[5] Mohan, Senthilkumar, Chandrasegar Thirumalai, and Gautam Srivastava, "Effective heart disease prediction using hybrid machine learning techniques" IEEE Access 7 (2019): 81542-81554.

[6] Samuel AL 1959 Some studies in machine learning using the game of checkers IBM Journal of research and development 3 210-29 Jul CrossrefGoogle Scholar

[7] Mitchell T. 1997 Machine Learning (McGraw Hill) 2Google Scholar

[8] Hayashi Chikio 1998 Data Science, Classification, and Related Methods. Studies in Classification, Data Analysis, and Knowledge Organization (Japan: Springer) What is Data Science? Fundamental Concepts and a Heuristic Example 40-51 01-01 Crossref Google Scholar.

[9] dhage Sandhya N. and Raina Charanjeet Kaur A review on Machine Learning Techniques International Journal on Recent and Innovation Trends in Computing and Communication (IJRITCC) 4 395-399 March 16 Google Scholar

[10] Dey Ayon Machine Learning Algorithms: A Review International Journal of Computer Science IJCSIT Google Scholar

[11] Samuel AL. Some studies in machine learning using the game of checkers. IBM Journal of research and development. 1959 Jul;3(3):210-29.

[12] Mitchell, T. (1997). Machine Learning. McGraw Hill. p. 2. ISBN 978-0-07-042807-2.

[13] Hayashi, Chikio (1998-01-01). "What is Data Science? Fundamental Concepts and a Heuristic

[14] Example". Data Science, Classification, and Related Methods. Studies in Classification, Data Analysis, and Knowledge Organization. Springer Japan. pp. 40–51. doi:10.1007/978-4-431-65950-1_3. ISBN9784431702085

[15] Sandhya N. dhage, Charanjeet Kaur Raina, "A review on Machine Learning Techniques", March 16Volume 4 Issue 3 , International Journal on Recent and Innovation Trends in Computing and

Communication (IJRITCC), ISSN: 2321-8169, PP: 395 – 399

[16] AyonDey , “Machine Learning Algorithms: A Review”, (IJCSIT) International Journal of Computer Science

[17] A report by Royal Society, April 2017, “Machine learning: the power and promise of computers that learn by example “, ISBN: 978-1-78252-259-1.and Information Technologies, Vol. 7 (3) , 2016, 1174-1179

[18] Minton S, Zweben M. Learning, Planning, and Scheduling: An Overview. In Machine Learning Methods for Planning 1993 (pp. 1-29).

[19] Sejnowski T. Net talk: A parallel network that learns to read aloud. Complex Systems. 1987;1:145-68.