Heart Disease Diagnosis Prediction Using Machine Learning and Data Mining Techniques

DR. S. K. SINGH[1], POONAM JAIN[2], GANESH PATIL[3], VIKAS MANOJ KUMAR PANDEY[4]
[1]*H.O.D, Department of IT, Thakur College of Science and Commerce, Thakur Village, Kandivali (East), Mumbai, Maharashtra, India*
[2]*Assistant Professor, Department of IT, Thakur College of Science and Commerce, Thakur Village, Kandivali (East), Mumbai, Maharashtra, India*
[3,4]*PG Student, Department of IT, Thakur College of Science and Commerce, Thakur Village, Kandivali (East), Mumbai, Maharashtra, India*

Abstract- ***Cardiovascular diseases, including heart disease, are a leading cause of mortality worldwide. Timely and accurate diagnosis of heart disease is of paramount importance in providing effective healthcare. This project explores the application of machine learning and data mining techniques for the diagnosis and prediction of heart disease. The primary objective is to develop predictive models that can assist healthcare professionals in making informed decisions, potentially leading to early intervention and improved patient outcomes. The project involves the collection of relevant medical data, such as patient demographics, clinical test results, medical history, and lifestyle factors. Data pre-processing techniques are employed to clean and transform the data for analysis. Feature engineering and selection methods are applied to identify the most informative variables for heart disease prediction. A variety of machine learning algorithms are considered, including logistic regression, decision trees, random forests, support vector machines, and neural networks. The models are trained on a portion of the data and evaluated using separate testing datasets. Performance metrics, including accuracy, precision, recall, and F1 score, are used to assess the models' predictive capabilities. Hyperparameter tuning is performed to optimize model performance, and the best-performing model is deployed in a clinical setting, such as a hospital or healthcare system. Continuous monitoring and updates ensure the model's relevance in real-world applications. The challenges of data privacy, model interpretability, and the need for domain expertise are addressed, ensuring that the developed models comply with healthcare regulations and maintain ethical standards in handling patient data. This project demonstrates the potential of machine learning and data mining to aid in heart disease diagnosis and prediction, contributing to more effective healthcare practices and potentially saving lives.***

Indexed Terms- ***Prediction, Machine Learning, Data Mining, Logistic Regression, K-Nearest Neighbors' Classifier, Random Forest, Support Vector Machine (SVM).***

I. INTRODUCTION

Cardiovascular diseases, particularly heart disease, stand as one of the leading causes of morbidity and mortality worldwide. In many regions, including developed countries, heart disease accounts for a significant portion of healthcare expenditure and places a substantial burden on healthcare systems. Prompt and accurate diagnosis of heart disease is essential for ensuring timely and effective treatment, which can significantly improve patient outcomes and reduce healthcare costs. The traditional approach to diagnosing heart disease typically involves an amalgamation of clinical assessments, medical history, and various diagnostic tests such as electrocardiography (ECG), blood pressure measurements, and blood lipid profiles. While these methods have proven to be valuable, they are often subjective and may not fully harness the wealth of data available in modern healthcare settings. The advent of data science and machine learning techniques has revolutionized the healthcare industry by offering the potential to derive insights and predictions from large volumes of patient data. Machine learning and data

mining techniques can effectively capture complex patterns and relationships within medical datasets, potentially enabling more accurate and early diagnosis of heart disease. This paradigm shift motivates the development of this project, which aims to leverage the power of these technologies to improve heart disease diagnosis and prediction. The motivation behind this project is multifaceted: Early Diagnosis: Early detection of heart disease can significantly improve patient outcomes by enabling timely interventions and lifestyle modifications. Machine learning can assist in identifying subtle risk factors that might be missed through traditional diagnostic methods. Data Utilization: Modern healthcare systems generate vast amounts of patient data, including electronic health records, medical imaging, and wearable device data. This project seeks to harness this wealth of information to improve heart disease diagnosis and prediction. Efficiency and Precision: Machine learning models can augment healthcare professionals' decision-making processes, providing accurate and objective assessments, reducing the likelihood of errors, and optimizing resource allocation. Healthcare Costs: By facilitating early intervention and targeted treatments, the project has the potential to reduce the economic burden of heart disease on healthcare systems and patients alike.

II. LITERATURE REVIEW

Prediction of heart disease using data mining techniques has been an ongoing effort for the past two decades. Most of the papers have implemented several data mining techniques for the diagnosis of heart disease such as Decision Tree, Naive Bayes, neural network, kernel density, automatically defined groups, bagging algorithm, and support vector machines showing different levels of accuracy (Yan, Zheng et al. 2003; Andreeva 2006; Das, Turkoglu et al. 2009; Sitar-Taut, Zdrenghea et al. 2009; Raj Kumar and Reena 2010; Srinivas Rani et al. 2010) on multiple databases of patients from around the world. One of the bases on which the papers differ are the selection of parameters on which the methods have been used. Many authors have specified different parameters and databases for testing the accuracies. In particular, researchers have been investigating the application of the Decision Tree technique in the diagnosis of heart disease with considerable success. Sitair-Taut et al. used the Weka tool to investigate applying Naive Bayes and J48 Decision Trees for the detection of coronary heart disease. Tu et al. used the bagging algorithm in the Weka tool and compared it with the J4.8 Decision Tree in the diagnosis of heart disease. In [9], the decision-making process of heart disease is effectively diagnosed by a Random forest algorithm. In [10] based on the probability of decision support, the heart disease is predicted. As a result, the author concluded that decision tree performs well and sometimes the accuracy is similar in Bayesian classification. In year 2013, S. Vijiyaraniet. al. [2] performed a work, An Efficient Classification Tree Technique for Heart Disease Prediction. This paper analyzes the classification tree techniques in data mining. The classification tree algorithms used and tested in this paper are Decision Stump, Random Forest and LMT Tree algorithm. The objective of this research was to compare the outcomes of the performance of different classification techniques for a heart disease dataset.

III. K-Nearest Neighbors (KNN)

K-Nearest Neighbors, or KNN, is an intuitive machine learning algorithm employed for both classification and regression tasks. It finds its relevance in healthcare, particularly when the relationships between data points have spatial significance. KNN is distinctive for its non-parametric and instance-based approach. Here's how it operates. The 'K' in KNN signifies the number of nearest neighbors considered when making predictions. In classification, KNN identifies the K closest data points from the training set to the new data point and assigns a class label based on the majority class among these neighbors. KNN depends on a distance metric (e.g., Euclidean distance, Manhattan distance, or cosine similarity) to measure the similarity or dissimilarity between data points. The choice of distance metric is a pivotal factor influencing the algorithm's performance. In classification, KNN adopts a majority voting mechanism. If most of the K nearest neighbors belong to a particular class, the algorithm assigns that class label to the new data point. For regression tasks, KNN calculates the average of the target values of its nearest neighbors.

Selecting the right value for 'K' is a critical decision in KNN. A smaller 'K' (e.g., 1) yields a more sensitive

model susceptible to noise, while a larger 'K' provides a smoother decision boundary but may oversimplify the problem. Logistic regression is a statistical modeling technique designed for binary classification tasks. In the realm of healthcare, especially in predicting heart disease, it serves as a vital tool for estimating the likelihood of heart disease based on patient features. Logistic regression is known for its simplicity and interpretability. Logistic regression is tailored for binary classification problems. In the case of heart disease prediction, it's used to predict the probability of a patient having heart disease (1) or not having heart disease (0). Logistic regression employs the logit transformation, which maps the linear combination of input features to a value in the range (0, 1), representing probabilities. This transformation is essential for modeling the likelihood of belonging to one of the two classes. Logistic regression operates by constructing a model that estimates the probability of a binary outcome. It assumes a linear relationship between input features and the log odds of the outcome. The logistic function (sigmoid) transforms the linear combination of features into a probability score. The formula for logistic regression is as follows:
$P(Y=1) = 1 / (1 + e^{(\beta0 + \beta1X1 + \beta2X2 + ... + \betanXn)})$
P(Y=1): Probability of the target variable being 1 (e.g., the presence of heart disease). e: Base of the natural logarithm.
β0, β1, …, βn:
Coefficients representing the strength and direction of the relationship between features (X1, X2, ..., Xn) and the log-odds of the target variable. Logistic Regression (modelL) is instantiated, trained on training data, and then used to make predictions on the testing data. The accuracy of the Logistic Regression model is calculated, and a classification report is generated to provide precision, recall, F1-score, and support for each class. In essence, KNN finds its predictions based on the proximity of data points, while Logistic Regression estimates probabilities using a logit transformation, making it particularly suited for binary classification tasks in healthcare, such as predicting heart disease

IV. METHODOLOGY

Import Libraries: The script starts by importing necessary libraries such as pandas (for data manipulation), seaborn and matplotlib (for data visualization), and various modules from the sklearn library (for machine learning tasks).

Data Sources: The "heartdisease.csv" dataset contains patient data that is used to build a machine learning model for heart disease prediction. In this context, the sources for data collection include: heartdisease.csv

Dataset: This dataset may contain various patient attributes, clinical measurements, and a binary label indicating the presence or absence of heart disease. The data in this CSV file serves as the primary source for this project.

Read Data: The code reads a CSV file named "Heartdisease.csv" into a pandas Data Frame

Data Preprocessing:
Once data is collected, it may require preprocessing to make it suitable for analysis:

A. Handling Missing Data: Address any missing values in the dataset by imputing values, removing affected records, or using appropriate techniques depending on the nature of the missing data.
B. Outlier Detection: Identify and handle outliers that might skew the analysis. Outliers can impact model performance and should be carefully managed.

Data Transformation: Normalize or standardize numerical features to ensure consistent scales. This step is essential when combining data from different sources to ensure uniformity.

EDA (Exploratory Data Analysis): Exploratory data Analysis is a critical step in understanding the dataset, uncovering patterns, and gaining insights into the data's distribution. It involves visualizing the data to identify relationships between features and potential predictors of heart disease.

Data Visualization: Use various data visualization techniques to gain insights into the dataset. This includes Histograms: Plot histograms for numerical features to visualize their distributions. For instance, you can create a histogram of age to see the age distribution in the dataset. Box Plots: Create box plots to identify outliers and visualize the distribution of numerical data, especially those related to heart

disease risk factors. Count Plots: Count plots can be used to visualize the distribution of categorical features. For example, you can create a count plot to show the distribution of gender in the dataset. Correlation Heatmap: Create a correlation heatmap to identify relationships between numerical features. This can help discover which attributes are strongly correlated with heart disease. Pair Plots: Pair plots are useful for visualizing relationships between pairs of numerical features. They can help identify patterns and potential predictors of heart disease. Violin Plots: Violin plots combine box plots and kernel density estimates to visualize the distribution of data. They are useful for showing the distribution of numerical data across different categories, such as heart disease status.

V. RESULTS

The objective of a project focused on heart disease diagnosis and prediction using machine learning and data mining techniques typically includes a combination of clinical, research, and technological goals. These objectives aim to address specific healthcare challenges and improve the accuracy and efficiency of diagnosing and predicting heart diseases. Detect heart diseases at an early stage, allowing for timely interventions and treatments to improve patient outcomes and reduce mortality rates. Develop machine learning models that outperform traditional diagnostic methods by providing more accurate predictions, reducing the likelihood of false positives and false negatives. collectively aim to improve the accuracy of heart disease diagnosis and prediction, enhance patient care, and contribute to the broader goal of reducing the global burden of heart disease. It is noteworthy that the highest accuracy is achieved by KNN with an accuracy score 81% logistic regression provides an accuracy of 83%

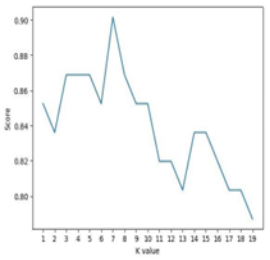

Fig.1 KNN Output

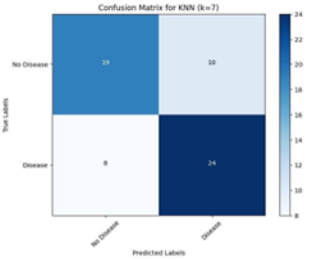

Fig.2 Confusion Matrix

Algorithm	Accuracy	F1-Score
K-nearest neighbour	83%	90%
Logistic Regression	81%	79%

Fig.3 Accuracy Table

CONCLUSION

A project focused on heart disease diagnosis and prediction using machine learning and data mining techniques represents a significant endeavor with the potential to have a profound impact on healthcare. In conclusion, a project focused on heart disease diagnosis and prediction represents a significant step toward improving healthcare outcomes and addressing a global health concern. The project's successes, ethical considerations, and potential for future contributions make it a valuable endeavor in the field of healthcare and medical research.

REFERENCES

[1] C. S. Dangare and S. S. Apte, “Improved study of heart disease prediction system using data mining classification techniques,” International Journal of Computer Applications, vol. 47, no. 10, pp. 44–48, 2012.

[2] S. Palaniappan and R. Awang, “Intelligent heart disease prediction system using data mining techniques,” pp. 108–115, 2008.

[3] Y. E. Shao, C.-D. Hou, and C.-C. Chiu, “Hybrid intelligent modelling schemes for heart disease classification,” Applied Soft Computing, vol. 14, pp. 47–52, 2014.

[4] M. Shouman, T. Turner, and R. Stocker, “Using data mining techniques in heart disease diagnosis and treatment,” pp. 173–177, 2012.;3

[5] P. V. Ankur Makwana, “Identify the patients at high risk of re-admission in hospital in the next year,” International Journal of Science and Research, vol. 4, pp. 2431– 2434, 2015.

[6] J. Nahar, T. Imam, K. S. Tickle, and Y.-P. P. Chen, “Computational intelligence for heart disease diagnosis: A medical Knowledge driven

approach," Expert Systems with Applications, vol. 40, no. 1, pp. 96–104,2013.

[7] Y. Xing, J. Wang, Z. Zhao, and Y. Gao, "Combination data mining methods with new medical data to predicting outcome of coronary heart disease," pp. 868–872, 2007.

[8] Combination data mining methods with new medical data to predicting outcome of coronary heart disease," in Convergence Information Technology, 2007. International Conference on. IEEE, 2007, pp. 868–872.

[9] Y. E. Shao, C.-D. Hou, and C.-C. Chiu, "Hybrid intelligent modelling, schemes for heart disease classification," Applied Soft Computing, vol. 14, pp. 47–52, 2014. Diabetic foot ulcer wound tissue detection and classification. S Patel, R Patel, D Desai - … conference on innovations in …, 2017 - ieeexplore.ieee.org

[10] A machine learning model for early detection of diabetic foot using thermogram images. A Khandakar, MEH Chowdhury, MBI Reaz… - … in biology and medicine, 2021 – Elsevier