PERFORMANCE ANALYSIS OF DIFFERENT CLASSIFICATION ALGORITHMS FOR HEART DISEASE PREDICTION

Mrs. A. Vijaya Shanthi[1], G. Harika[2], A. Sowmya[3], P. Chandini[4], D. Sai Laxmi[5], UG Students[2,3,4,5],
Department of Electronics and Communication Engineering, Vignan's Institute of Engineering for Women
Mrs. A. Vijaya Shanthi[1], Assistant Professor, Department of Electronics and Communication Engineering, Vignan's Institute of Engineering for Women.

ABSTRACT
The use of machine learning technology has shown promising results in addressing healthcare challenges and improving biomedical research. In particular, it has shown potential in predicting diseases at an early stage, enabling effective management and treatment of symptoms. One area where this technology can be applied is in the early detection of heart disease, which has seen a rise in the number of fatalities. Researchers have used different supervised machine-learning techniques such as K-nearest neighbors, Naive Bayes, support vector machines, neural networks, and random forest classifiers to predict heart disease based on data from the University of California, Irvine (UCI) Machine Repository. The findings indicate that the random forest classifier outperformed the other supervised classifiers in terms of accuracy, precision, and sensitivity.
Keywords: Heart disease,Machine learning, University of California, Irvine (UCI) Machine Repository, K-nearest neighbors, Naive Bayes, support vector machines, neural networks, Random Forest Classifier.

INTRODUCTION

Human diseases are medical conditions that can impair bodily functions and manifest in various signs of illness. Among the leading causes of death worldwide, cardiovascular diseases (CVDs) such as heart attacks and strokes are prevalent. The World Health Organization (WHO) reports that CVDs claim the lives of approximately 17 million individuals annually. Heart disease is a type of CVD that can affect different areas of the heart, including the muscles, valves, and internal electrical pathways that regulate muscle contractions. It is a significant cause of death in many countries, such as India, the UK, the US, Canada, and Australia. Risk factors associated with heart disease include age, gender, family history, smoking, chemotherapy drugs, high blood pressure, high blood cholesterol levels, diabetes, obesity, physical inactivity, stress, and poor hygiene. While genetics can contribute to the development of heart disease, lifestyle factors play a critical role. Recognizing the risk factors is crucial to preventing the onset of heart disease and improving patient outcomes. One way to predict heart disease is by utilizing machine learning to identify unseen patterns and provide clinical insights to aid physicians in planning and delivering care. Feature selection is a critical step in this process, as irrelevant or redundant features can significantly affect classification accuracy. After data pre-processing and feature extraction, the selected features were supplied to the K-nearest neighbor (KNN), Naive Bayes classifier, neural networks, and support vector machines for classification purposes. This combination of methods led to improved classification accuracy and a better diagnosis of heart disease. By reducing the number of features and increasing classification accuracy, this approach offers a promising solution for predicting heart disease and improving patient outcomes. In this work, we analyze and estimate the use of various machine learning algorithms in predicting the occurrence of coronary heart disease. To achieve this objective, all the available attributes in the dataset were combined to develop a classification model. The study findings revealed that the machine learning models employed were successful in efficiently predicting the risk of heart disease. The results of this research can have significant implications for the early detection and prevention of coronary heart disease, ultimately improving the quality of patient care.

LITERATURE SURVEY

Various approaches are utilized by researchers to apply machine learning and data-mining techniques in addressing health-related problems, with a particular focus on predicting and categorizing chronic illnesses. These methods vary from researcher to researcher, and may involve different algorithms and models to achieve their desired outcomes.

[1] Sibo Prasad Patro, Gouri Sankar Nayak, Neelamadhab Padhy "Heart Disease Prediction by using novel optimization algorithm: A Supervised learning prospective". The purpose of this study is to develop a framework for predicting heart disease using major risk factors and various classification algorithms, including Naïve Bayes, Bayesian Optimized Support Vector Machine, K-Nearest Neighbors, and Salp Swarm Optimized Neural Network. The heart disease dataset from the UCI Machine Repository is used for this research, and the proposed optimized algorithm is tested for its effectiveness in early detection and monitoring of heart disease. The use of classification algorithms enables the prediction of optimal results based on training and test data. Optimization techniques are utilized to analyze the data and improve the accuracy of the prediction model.

[2] Khaled Mohamad Almustafa et al. proposed Prediction of heart disease and classifiers, Khaled Mohamad Almustafa conducted a study aimed at predicting heart disease using minimal attributes and various classification algorithms. The dataset utilized in this research was sourced from Cleveland, Switzerland and consisted of 76 attributes and a class attribute of 1025 patients. However, only 14 features were used in the analysis. To obtain the most accurate classification accuracy for predicting heart disease cases, the researchers employed several classification algorithms, including k-nearest neighbor, decision tree, Naïve Bayes, SVM, and stochastic gradient descent. The study sought to determine the best-performing algorithm for predicting heart disease using minimal attributes.

[3] In a study conducted by Juan-Jose Beunza et al. a supervised machine learning algorithm was employed to predict clinical events with high validity and accuracy. The study utilized data from the Framingham heart study, which contained 4240 observations, with a focus on identifying risk factors for heart disease through data mining techniques. Various machine learning algorithms were applied to the dataset, including RapidMiner and R-Studio. Notably, a neural network model was implemented to address the issue of missing data. The study demonstrated the effectiveness of machine learning algorithms in predicting clinical events, especially when combined with data mining techniques.

[4] Youness Khourdifi et al. conducted a study on using machine learning for predicting heart disease and highlighted the benefits of optimization algorithms in dealing with complex non-linear problems. To improve the accuracy of heart disease classification, the researchers employed a method called Fast Correlation-Based Feature Selection (FCBF) to filter out redundant features. Various classification algorithms, including support vector machine, k-nearest neighbor, naive Bayes, and random forest, were utilized in conjunction with particle swarm optimization and ant colony optimization processes. The study demonstrated the effectiveness of combining feature selection techniques with optimization algorithms in improving the accuracy of heart disease classification.

[5] Yekkala et al. employed Particle Swarm Optimization (PSO) in conjunction with particle methods such as Random Forest, Ada-Boost, and Bagged Tree to improve the accuracy of predicting heart disease. The dataset comprised 270 samples and 14 attributes and had already been processed. PSO was used as a feature selection method to eliminate irrelevant and missing data. The results indicated that using Bagging Trees on PSO can enhance the learning accuracy of heart disease prediction. Overall, Yekkala et al.'s study demonstrated the potential of PSO in combination with particle methods to improve the accuracy of predicting heart disease.

[6] Shao et al.utilized 13 risk factors to predict heart disease. Unlike existing approaches, this study proposed a novel hybrid framework that combines three methods: Multivariate Adaptive Regression (MAR), Logistic Regression (LR), and Artificial Neural Network (ANN) to achieve various risk factors. The hybrid framework initially reduces the encoded values of risk factors using LR and MAR. The remaining encoded factors are then used to train the ANN. The study's simulation results demonstrate that the hybrid approach outperforms the conventional single-stage neural network.

PROPOSED METHODOLOGY

In this section, we provide an overview of various techniques employed in our research. This includes a detailed description of the methods used, as well as an explanation of the dataset utilized.

Dataset.

The UCI Heart disease dataset is a widely used dataset in the field of cardiovascular research. It is a compilation of four distinct databases, but the most commonly used subset is the UCI Cleveland dataset. This dataset contains a total of 76 features, including demographic, clinical, and laboratory parameters. However, most published studies utilize only a subset of 14 features, which have been identified as the most relevant predictors of heart disease. These features include age, sex, chest pain type, resting blood pressure, serum cholesterol level, fasting blood sugar level, resting electrocardiographic results, maximum heart rate achieved, exercise-induced angina, ST depression induced by exercise, slope of the peak exercise ST segment, number of major vessels, thalassemia, and the presence of heart disease. The UCI Heart disease dataset has been extensively used for developing predictive models and evaluating the effectiveness of machine learning algorithms in detecting heart disease.

Table 1: Dataset Feature's information

Feature	Description
Age	Age of a particular individual in years.
Gender	Gender of the particular individual.
Cp	chest pain is classified into four types: (1) Typical angina, (2) Atypical angina, (3) Non-angina pain, (4) Asymptomatic.
Trestbps	"Trestbps" refers to the resting blood pressure, resting blood pressure is the measurement of blood pressure when the individual is at rest.
Chol	"Chol" refers to the level of the total cholesterol in the blood of an individual.
FBS	"FBS" refers to the level of fasting blood glucose of an individual.
Restecg	"Restecg" refers to a resting electrocardiogram, which is a test that records the electrical activity of the heart while an individual is at rest.
Thalach	"Thalach" refers to the maximum heart rate achieved during exercise. It refers to the highest heart rate that a person can reach during physical activity which is typically measured in beats per minute (bpm).
Exang	"Exang" refers to angina pectoris, which is a type of chest pain that occurs when the heart muscle is not getting enough oxygen-rich blood.
Oldpeak	"Oldpeak" is known as the ST depression induced by exercise or stress.
Slope	"Slope" refers to the slope of the ST segment on an ECG during exercise stress testing.
Ca	"Ca" refers to the number of major vessels colored by fluoroscopy.
Thal	"Thal" refers to thallium stress test results.
Target	"Target" refers to the outcome variable that is being predicted.

BLOCK DIAGRAM OF THE PROPOSED WORK

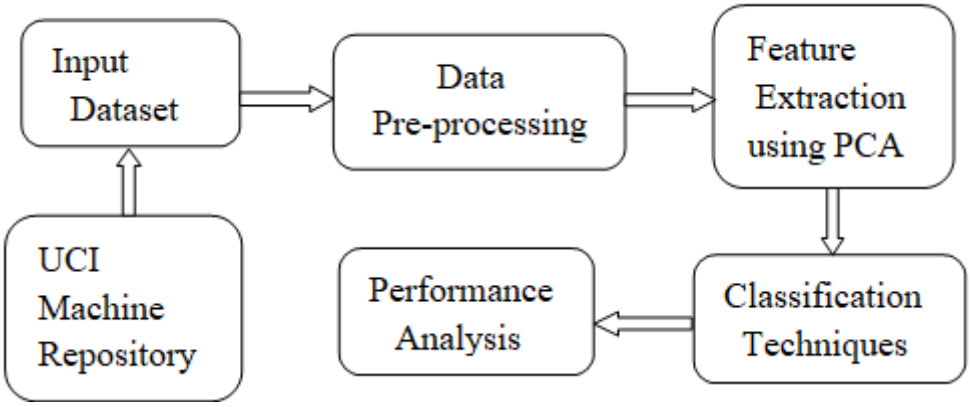

Figure 1: Block diagram of the proposed work

MACHINE LEARNING

Machine learning is a branch of artificial intelligence that allows computer systems to learn and improve performance without explicit programming. Through the development of algorithms and models, machine learning systems can analyze large datasets to identify patterns and relationships that humans may not be able to detect. This information can be used to make accurate predictions and informed decisions. The process of machine learning involves inputting vast amounts of data into an algorithm, which learns to recognize patterns and predict outcomes based on the data. Three primary types of machine learning are supervised learning, unsupervised learning, and reinforcement learning. Supervised learning involves using labeled data to teach an algorithm to make predictions by example. The labeled data serves as a supervisor, guiding the algorithm to accurately predict future output.

MACHINE LEARNING ALGORITHMS

Machine learning algorithms are computer programs designed to use statistical techniques to enable computers to learn from data and enhance their performance on a particular task. Specifically, in this work, supervised machine learning algorithms are utilized to accomplish the task at hand.

NAIVE BAYES

The Naive Bayes Classifier is a commonly used algorithm in machine learning that is well-suited for classification tasks. Its foundation lies in the Bayes theorem, and it assumes that an object's features are unrelated to one another. By considering the likelihood of a feature occurring in a specific class, the Naive Bayes classifier can calculate the probability of that object belonging to that class. This approach is beneficial for classification tasks as it can rapidly generate precise predictions based on input data. Additionally, the Naive Bayes Classifier is known for its simplicity and ease of implementation, which makes it a preferred choice for applications that require speed and efficiency.

$P(y / x1,x2,\ldots\ldots,xn) = (P(x1,x2,\ldots..,xn / y) * P(y)) / P(x1,x2,\ldots\ldots,xn)$

Where:

- $P(y / x1, x2,\ldots...,xn)$ is the posterior probability of class y given the features $x1,x2,\ldots\ldots,xn$.
- $P(x1,x2,\ldots..,xn / y)$ is the likelihood of observing the features $x1,x2,\ldots..,xn$ given class y.
- P(y) is the prior probability of class y.
- $P(x1,x2,\ldots.,xn)$ is the evidence, which is a normalizing constant that ensures that the posterior Probabilities sum up to 1 over all possible classes.

SUPPORT VECTOR MACHINE

Support Vector Machine (SVM) is a widely-used supervised learning algorithm that is primarily used for classification problems in machine learning. The goal of the SVM algorithm is to create a decision boundary or hyper plane that can effectively separate the n-dimensional space into different classes, allowing for

accurate classification of new data points in the future. The SVM algorithm uses the training data to find the hyper plane and then classifies the new data points by projecting them onto the hyper plane. SVM is a powerful algorithm that works well with both linear and nonlinear data and has been widely used in many applications, including image classification, text classification, and bioinformatics. SVM can be of two types, Linear SVM and Non-linear SVM. Linear SVM is used for datasets that can be classified into two classes by using a single straight line, while non-linear SVM is used for datasets that cannot be classified using a straight line. Non-linear SVM utilizes different types of kernel functions to transform the data into higher-dimensional space, where it can be classified using a hyper plane.

K-NEAREST NEIGHBORS

K-Nearest Neighbor (KNN) is a supervised learning algorithm used for classification tasks. It determines the target class of a data point by measuring its similarity with other training data points in the model. This similarity is assessed based on the features or characteristics of the data points. The algorithm then calculates the distance between the unclassified data points and the classified ones using metrics like Euclidean distance or Manhattan distance. KNN selects the K smallest distances and tallies the class that appears most often. The unclassified data point is then classified based on the most common class. The algorithm works by collecting unclassified data points and determining their proximity to classified data points. The K closest data points are then analyzed to find the most frequent class, which is assigned to the unclassified data point. This process is repeated for all unclassified data points in the dataset. KNN is a simple yet powerful algorithm suitable for small to medium-sized datasets and is capable of binary and multiclass classification tasks.

NEURAL NETWORKS

Neural networks are a type of machine learning method that mimics the structure of the human brain. They consist of interconnected nodes or neurons that are arranged in a layered structure. The goal of neural networks is to create a self-learning system that can continuously improve its performance by learning from its mistakes. These networks process data through layers of interconnected neurons, allowing them to detect patterns and make accurate predictions. Neural networks are particularly powerful for solving complex problems, such as image recognition and natural language processing, as they can analyze vast amounts of data and identify patterns that may be difficult for humans to detect. With their ability to learn and adapt to new information, neural networks have become a valuable tool for various applications, from speech recognition to self-driving cars.

RANDOM FOREST CLASSIFIER

Random Forest is a widely used supervised machine learning algorithm that is based on the combination of multiple decision trees. The approach, known as bagging, involves using random subsets of features and threshold values for each tree, improving the overall accuracy of the model. The algorithm assesses feature importance scores to determine which features contribute most to the model's accuracy. One of its key advantages is its ability to handle outlier values and noisy or missing data, making it ideal for complex datasets. With its ability to handle high-dimensional data and provide accurate predictions, Random Forest is an effective tool for various applications, including image classification, text mining, and predictive analytics.

RESULTS AND DISCUSSIONS

To ensure the accuracy and reliability of machine learning models, it is common practice to split the dataset into two parts, namely the training and testing datasets. The purpose of this split is to train the model using the training dataset and evaluate its performance on the unseen testing dataset, which represents new data. This approach allows for estimating the model's generalization capability, which is crucial for practical applications. The evaluation of the model's performance is usually done using a confusion matrix and various performance metrics such as accuracy, precision, and sensitivity. By comparing and contrasting the results of

different supervised machine learning classifiers on the same dataset, it is possible to determine the best and most accurate model for predicting cardiac disease.

CONFUSION MATRIX

The confusion matrix is a tool used to assess a model's performance. It consists of four terms that determine the performance metrics: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).TP refers to when the model correctly predicts the positive class, while TN refers to when the model correctly predicts the negative class. FP refers to when the model incorrectly predicts the positive class, and FN refers to when the model incorrectly predicts the negative class.These terms are used to calculate various metrics such as accuracy, precision, recall, and F1-score, which help to evaluate the effectiveness of a model in making predictions. Overall, the confusion matrix is an important tool for understanding the strengths and weaknesses of a model and identifying areas for improvement.

	0	1
0	True Negative (TN)	False Positive (FP)
1	False Negative (FN)	True Positive (TP)

Figure 2: Block Diagram of the Confusion Matrix

EVALUATION PARAMETERS

Accuracy: Accuracy is a performance metric that measures the proportion of correct predictions made by a model in relation to the total number of instances it has made predictions for. In simpler terms, it is the percentage of correctly predicted instances out of all instances evaluated by the model.

Accuracy = (TP+TN) / (TP+FP+FN+TN).

Precision: Precision is a performance metric that measures the proportion of individuals who were predicted by a model to be at risk of developing CHD and actually had a risk of developing CHD. In other words, precision measures the accuracy of the positive predictions made by the model.

Precision = (TP) / (TP+FP).

Sensitivity: Sensitivity, also known as recall or true positive rate, is a performance metric that measures the proportion of individuals who were actually at risk of developing CHD and were correctly identified as such by the algorithm. In other words, sensitivity measures the model's ability to correctly identify true positive cases among all positive cases.

Sensitivity = (TP) / (TP+FN).

Table 2: Performance Metrics Comparison of the Classifiers.

Classifier	Accuracy	Precision	Sensitivity
Naïve Bayes	85.25%	91.18%	83.78%
Support Vector Machine	81.97%	88.24%	81.08%
K-Nearest Neighbors	67.21%	67.65%	71.88%
Neural Networks	83.61%	82.35%	84.85%
Random Forest Classifier	93.50%	97.25%	91.38%

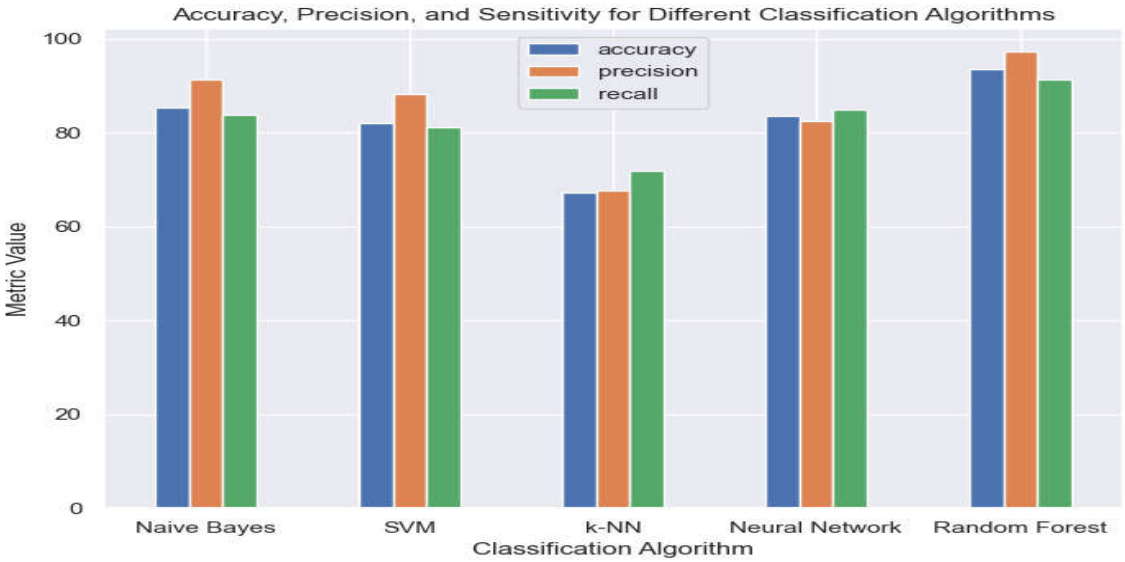

Figure 3: Comparative results graph representation for various classification methods

CONCLUSION

The aim of this study was to compare the effectiveness of five different machine learning classifiers in predicting heart disease based on a given dataset. The performance of the classifiers was assessed using various measures such as the confusion matrix, accuracy, precision, and sensitivity. The study involved building five machine learning classifiers, Random Forest, Naive Bayes, Support Vector Machine, K-Nearest Neighbor, and Neural Networks, and comparing their results. Among these classifiers, Random Forest had the highest accuracy rate of 93.50%, followed by Naive Bayes with an accuracy rate of 85.25%. Neural Networks had an accuracy rate of 83.61%, Support Vector Machine had an accuracy rate of 81.97%, and K-Nearest Neighbor had the lowest accuracy rate of 67.21%. The results of the study showed that Random Forest was the most effective classifier for predicting heart disease. The method could be beneficial in identifying patients at risk of heart disease, leading to early intervention and a reduced risk of mortality.

FUTURE SCOPE

Heart disease prediction models can be further improved by using some of the deep learning algorithms to achieve higher accuracy precision, and sensitivity. A bigger dataset ensures more precise and accurate findings. This is critical since medical diagnosis is a very delicate problem that requires high degrees of accuracy precision and sensitivity.A web application that integrates these methods and uses a larger dataset than the one used in this study might be developed in the future. As a result, healthcare providers will be better able to predict and treat cardiac abnormalities with more precision and efficiency.

REFERENCES

1. Sibo Prasad Patro, Gouri Sankar Nayak, Neelamadhab Padhy “Heart Disease Prediction by using novel optimization algorithm: A Supervised learning prospective”, 26 (2021) 100696.
2. Almustafa, K. M. Prediction of heart disease and classifiers' sensitivity analysis, 02 July 2020.BMC Bioinf, 21.
3. J.J. Beunza, E. Puertas, E. García-Ovejero, G. Villalba, E. Condes, G. Koleva, M.F. LandechoComparison of machine learning algorithms for clinical event prediction (risk of coronary heartdisease) J Biomed Inf, 97 (2019), p. 103257.
4. Y. Khourdifi, M. Bahaj Heart disease prediction and classification using machine learningalgorithmsoptimized by particle swarm optimization and ant colony optimizationInt. J. Intell. Eng. Syst., 12 (1) (2019), pp. 242-252.
5. I. Yekkala, S. Dixit, M.A. Jabbar August. Prediction of heart disease using ensemble learning andParticle Swarm Optimization 2017 international conference on smart technologies for smart nation(SmartTechCon), IEEE (2017), pp. 691-698.
6. Shao YE, Hou CD, Chiu CC. Hybrid intelligent modeling schemes for heart diseaseclassification. Appl Soft Comput 2014; 14:47–52.
7. Latha C B C and Jeeva S C 2019 Improving the accuracy of prediction of heart disease risk based onensemble classification techniques Inf. in Med. Unl. 16100203.
8. Singh A and Kumar R 2020 February heart disease prediction using machine learning algorithmsInternational Conf. on Electrical and Electronics Engineering 452-7.
9. Kohli P S and Arora S 2018 Application of machine learning in diseases prediction 4th InternationalConf. on Computing Communication and Automation.
10. Mohan S K, Thirumalai C and Srivastava G 2019 Effective heart disease prediction using hybridmachinelearning technique IEEE Access 7 81542-54.
11. Nagaraj M Lutimath, Chethan C, Basavaraj S Pol., Prediction of Heart Disease using MachineLearning, International Journal of Recent Technology and Engineering,8, (2S10), pp 474-477.