International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

A Comparative Analysis of Machine Learning Models for Heart Disease Prediction

Jayant Baid [a]*, *Atharva Rana* [b], *Sapna Gupta* [c]

[a,b,c] *Department of ITE, Maharaja Agrasen Institute of Technology*

ABSTRACT

Cardiovascular diseases continue to be the cause of death highlighting the importance of creating effective models, for early detection and intervention. The objective of this study was to design a machine learning algorithm of forecasting risk based on various patient conditions. The dataset included 14 attributes like age, gender, blood pressure, cholesterol levels among others. By analyzing the data at levels researchers aimed to refine existing models for more accurate predictions and greater practical benefits in healthcare settings. Rigorous testing and validation confirmed that the proposed predictive system excelled at identifying individuals at risk of heart diseases. This data centric approach promises disease identification, personalized treatment strategies, preventive care measures and a significant reduction, in the burden of cardiovascular illnesses.

Keywords: Heart disease, the use of machine learning, artificial intelligence, machine learning models

1. INTRODUCTION

Cardiovascular issues, which encompass conditions affecting the hearts functionality remain a concern, for health globally. The World Health Organization (WHO) reports that heart disease claims around 12 million lives and contributes to 31% of all deaths [1]. Despite advancements in healthcare, heart related problems. Impose a burden on both the healthcare sector and society. The early detection and prevention of heart ailments play a role in reducing morbidity and mortality rates. With an increasing awareness among scientists and medical professionals about the importance of interventions there is a growing reliance, on models to pinpoint high risk populations and institute preventive strategies. Utilizing machine learning algorithms these predictive models offer a means to forecast an individual's likelihood of developing heart disease based on their demographic information [2].

Numerous research studies have delved into exploring how machine learning algorithms can predict heart disease by considering a range of clinical and lifestyle factors. However, the effectiveness of these models could be influenced by variables, choices and the type of model utilized. Additionally, it is crucial to understand and assess how these predictive models impact decision making support and clinical applications. This study aims to enhance the development of models, for diagnosing heart disease and enhancing their precision and applicability, in settings. By leveraging machine learning techniques and diverse datasets this investigation sought to construct a model of identifying individuals at risk of heart disease. The research assesses the efficacy of machine learning algorithms in predicting risk through comprehensive experiments, including logistic regression support vector machine (SVM) decision trees and convolution techniques [3].

The findings of this research hold implications, for both practice and public health. Through the creation of predictive models, it becomes possible to enhance early detection, intervention and self-care methods thereby lessening the worldwide impact of heart disease. Moreover, this study underscores the value of integrating clinical information to address health issues and enhance individual well-being.

2. LITERATURE REVIEW

Cardiovascular disease prediction is a studied subject, due to its impact on public health. Previous research has explored algorithms and machine learning techniques to create models that can enhance early detection and intervention strategies. An overview of the current literature uncovers a diverse range of approaches and resources utilized in disease studies. One method involves utilizing demographic data for crafting models. For instance, Dey et al. (2016) utilized a dataset containing attributes, like age, gender, cholesterol levels, blood pressure and family history of heart disease to build models. Their study yielded outcomes in identifying heart disease emphasizing the significance of thorough data collection and precise variable selection.

The process of selecting features is crucial, in enhancing prediction models and minimizing errors. Various studies have explored the effectiveness of selection algorithms like feature elimination (RFE) and principal component analysis (PCA) in identifying features for assessing heart disease risk [4]. Jindal et al. Suggested using machine learning methods such as regression and K-NN to predict and categorize patients [5]. Alizadeh Sani (2013) compared support vector machine selection techniques for predicting heart disease emphasizing the significance of feature selection in enhancing model accuracy and interpretability [6].

Although significant progress has been made, predicting heart disease still presents challenges due to data distribution between cardiac patients) and bad patients, making model training and evaluation complex. Additionally, issues related to interpreting models and ensuring generalizability require exploration to facilitate the integration of models into clinical settings. In summary, the existing literature highlights the importance of machine learning techniques in cardiovascular disease prediction and provides insights into the development and evaluation of predictive models. Using diverse data sources, specific selection methods, and collaborative efforts, researchers continue to improve the accuracy and clinical utility of predictive models for cardiac diagnosis.

3. MACHINE LEARNING TECHNIQUES

Many machine learning algorithms have been used to predict heart disease, each with unique advantages and limitations. This study investigated the application of various machine learning algorithms to predict cardiovascular disease based on various clinical and demographic characteristics. The following methods are used:

3.1 Logistic Regression

Logistic regression is a statistical method that is widely used in binary classification problems and is therefore suitable for predicting the presence or absence of heart disease. Unlike some machine learning algorithms, it offers a high level of interpretation and allows understanding the impact of every aspect of the prediction. In this study, the model is based on age, gender, blood pressure, cholesterol level, etc. It takes many patient characteristics as input, such as: The variable (Y) represents the presence of heart disease (1) or absence (0) of heart disease. Logistic regression predicts the probability (P) that a patient has heart disease based on the combination of these characteristics.

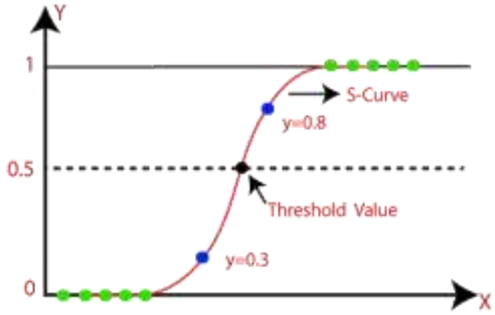


Fig. 1. Example of Logistic Regression

Its mathematical equation can be expressed as:

P(Y = 1 | X) = 1 / (1 + e^(-β0 + ΣβiXi))

Where:

β0 is the intercept term.

βi represents the coefficients associated with each feature (Xi).

e is the base of the natural logarithm (approximately 2.718). [14]

3.2 K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a non-parametric machine learning algorithm commonly used for task classification. In the context of cardiovascular disease prediction, ANN classifies new patient data based on their similarity to their nearest neighbors in the training data. The ANN model is trained on a dataset where each data point represents a patient with relevant characteristics (age, blood pressure, etc.) and indicates the presence or absence of heart disease. When information arrives indicating a new patient (with an unknown disease), the KNN algorithm calculates the distance between this new information and all points in the information. The parameter “K” is defined as the number of nearest neighbors to consider. This algorithm identifies K points that are similar to new patient data points (neighbors) in the training set based on distance measurements (i.e., Euclidean distance). Finally, a majority vote of the K-list of nearest neighbors determine the class list (with or without heart disease) for the new patient.

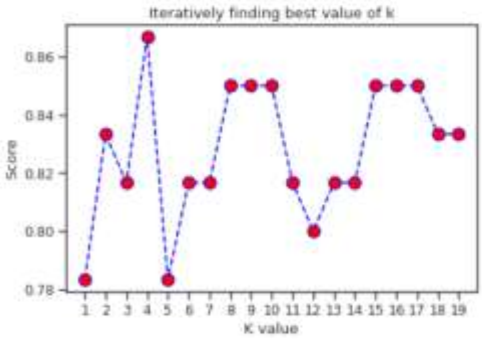
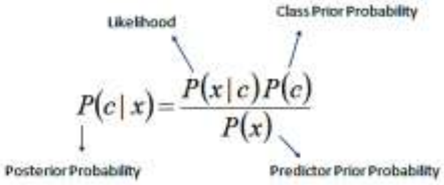


Fig. 2. Find best value of K on KNN

3.3 Naive Bayes

Naive Bayes is a machine learning algorithm commonly used for task classification. It is based on Bayes' theorem and assumes feature independence. Although this theory is not applicable to real data, it is still valid for some applications, including predicting heart disease. The Naive Bayes model is trained on the dataset, where each data point represents a patient with relevant characteristics (age, blood pressure, etc.) and the label indicates the presence or absence of heart disease. When new patient's data is received, Naive Bayes employs Bayes theorem to estimate the likelihood of the patient having heart disease. It calculates the probability of each feature value occurring given the presence or absence of heart disease. Assuming conditional independence simplifies this process. The model then assigns probabilities to these scenarios. Contrasts them with the likelihood of the patient not having heart disease to predict the probable outcome.



$$P(c\,|\,x)=\frac{P(x\,|\,c)P(c)}{P(x)}$$

$$P(c\,|\,X) = P(x_1\,|\,c)\times P(x_2\,|\,c)\times\cdots\times P(x_n\,|\,c)\times P(c)$$

Fig. 3. Formula for Naive Bayes

3.4 Decision Tree

Decision trees are widely used in the field of machine learning to categorize activities making them particularly effective, in predicting issues. They take on a tree structure where each node represents a decision based on patient data. As the model navigates through the tree using information it eventually reaches a leaf node that predicts whether heart disease is present or not. The decision tree algorithm learns from a dataset where each data point corresponds to a patient. Includes features and connections indicating the presence or absence of heart disease. It identifies points that differentiate between data points (related to health or heart disease). Based on these selected features at each node decision rules are established. Data points are then directed left or right within the tree based on how their values align with the decision rule. This process continues until certain stopping criteria are met, such as reaching a depth in the tree or grouping all data points of the class in one leaf.

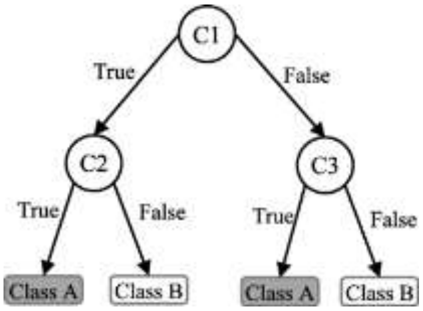


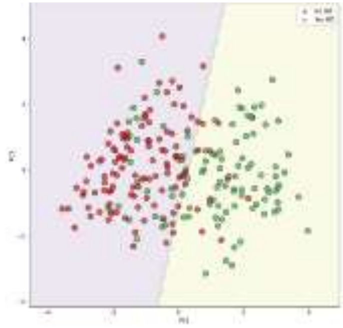
Fig. 4. Example of Decision Tree

When new patient data is received the system begins its journey, from the starting point. It moves through the structure based on the patient's attributes. At each stage it navigates along the path that aligns with the patient's details. Eventually it reaches a section displaying a list of categories for a patient (whether they have heart disease or not).

3.5 Support Vector Machines (SVM)

Support Vector Machine (SVM) is a technique, in machine learning that works well for classifying tasks and is particularly effective in predicting diseases. Unlike a decision tree that follows a step-by-step decision-making process SVMs goal is to identify the solution for separating data points belonging to classes (those with and without heart disease). The SVM algorithm is trained using a dataset where each data point represents a patient with features along with labels indicating the presence or absence of heart disease. The algorithm determines a plane that acts as a boundary between two groups (individuals and those with heart disease). The margin denotes the distance between this plane (hyperplane) and the closest data points, from each group, known as support vectors. SVM strives to establish a cut decision boundary that can effectively classify ambiguous data by maximizing the margin.

Fig. 5. Example of SVM



When a patient's data is received the model measures how far it is, from the established hyperplane. Using this distance and the hyperplane equation the model can forecast whether the new patient has heart disease.

3.6 Random Forests

Random forests stand out as a technique, in machine learning frequently applied in classification tasks, particularly useful for predicting heart disease [13]. Unlike a decision tree that tends to overfit the Random Forest method merges the forecasts from decision trees to craft superior and more precise models. To kickstart the forest algorithm, a basic decision tree is constructed manually. Each decision tree undergoes training using a variation (bootstrapping) of data points from the dataset with modifications. This implies that certain data points might be included times across trees while others may be entirely left out. Moreover, a subset of features (p) is chosen from all features (p), at each stage of the tree [7]. This injects variability into the feature selection process thereby diminishing the likelihood of overfitting on data [8]. Each decision tree is grown to its depth without pruning (minimum stopping criteria) to encompass an array of decision rules.

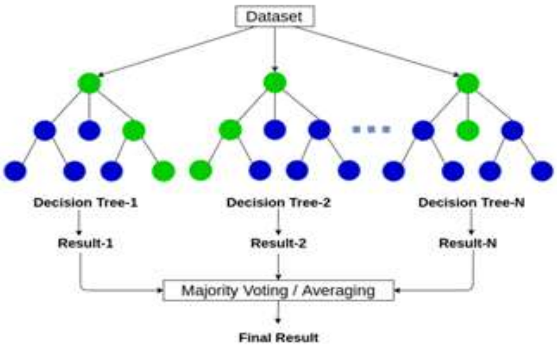


Fig. 6. Random Forest Concept

When information received about a new patient, it surpasses all the trees in the forest. Each tree makes a prediction (whether there will be a heart attack or not) based on the rules it has learned. Finally, the random forest model pools the prediction of all trees using majority voting. The category with the most votes becomes the final category for new patients.

4. METHODOLOGY

Data Collection: Data used in this study were collected from the UCI Heart Disease Cleveland dataset. The repository has 76 functions, but each published test involves the use of 14 of these functions [4,9,10,15,16]. Importantly, the Cleveland Repository is the only resource used by academic researchers to date. The main goal is to create a comprehensive database of clinical and demographic information about people and diseases related to heart disease.

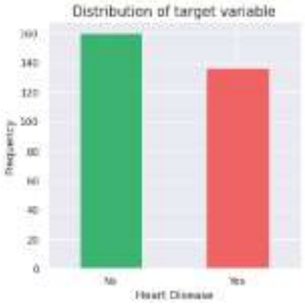


Fig. 7. No. of patients having heart disease

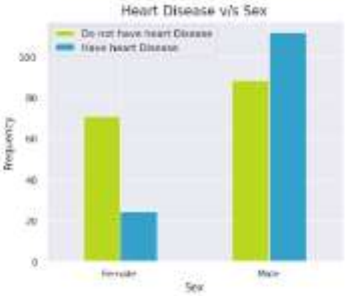


Fig. 8. Heart Disease vs Sex

Data Preprocessing: Before beginning the machine learning process, it is crucial to preprocess the data to ensure that it is well prepared for modeling. In this research various initial techniques were employed on the Cleveland Heart Disease dataset prior, to training and evaluating models [9]. This involved addressing missing data scaling features and other necessary steps.

Exploratory Data Analysis (EDA): Exploratory data analysis (EDA) is an important step in understanding the properties of data sets and identifying potential patterns or relationships that may exist between different data [11]. In this study, researchers used exploratory data analysis (EDA) to gain insights to undersrand product distribution, assess product quality and determine the model to pursue.

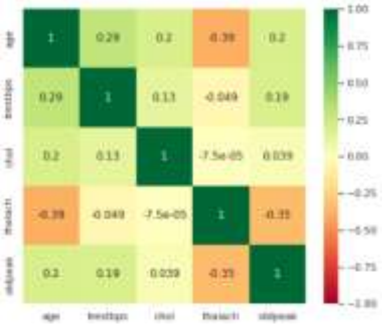


Fig. 9. Correlation Heatmap

Feature Selection: The objective of feature selection is to pinpoint the collection of attributes that accurately forecasts the outcome variable (, for example the presence or absence of heart disease) by reducing the scope. Utilize insights from exploratory data analysis to pick out features for constructing the model and subsequently employ methods, like correlation analysis, redundancy removal or importance assessment to gauge the significance of the forecast[13].

Model Selection: Choosing the right machine learning algorithm involves selecting the one for predicting outcomes considering factors, like effectiveness computing speed and interpretability. In this research different classification methods were evaluated for forecasting heart disease, such, as regression, random forests, support vector machines and other approaches.

Model Evaluation: Assessing models, for predicting disease is crucial in determining their effectiveness. This involves checking the model’s accuracy, reliability and generalizability through testing methods like precision, recall, F1 scores and confusion matrix [4,12]. The main aim of model evaluation is to measure how accurately the model can forecast risk and aid, in making decisions.

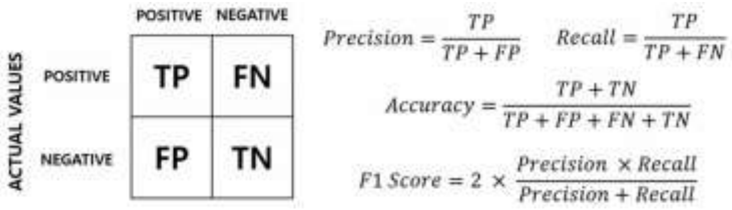

POSITIVE
NEGATIVE
ACTUAL VALUES
POSITIVE
NEGATIVE
TP
FN
FP
TN
$Precision = \frac{TP}{TP + FP}$
$Recall = \frac{TP}{TP + FN}$
$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$
$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$

Fig. 10. Formulas of Accuracy, Precision, Recall and F1 Score

4. Results

The results of this study, on disease cover the effectiveness of prediction models and important insights gained from evaluating the models.

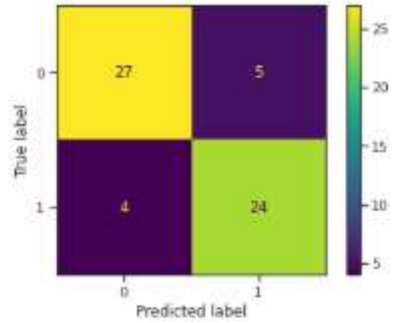


Fig. 11. Confusion Matrix of Logistic Regression

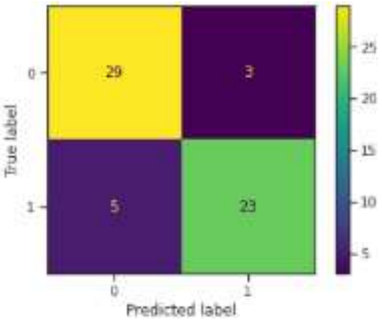


Fig. 12. Confusion Matrix of KNN

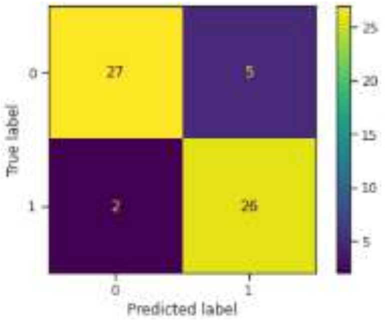


Fig. 13. Confusion Matrix of Naïve Bayes

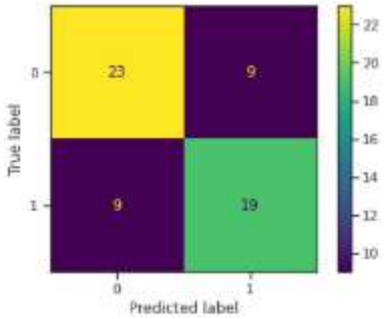


Fig. 14. Confusion Matrix of Decision Tree

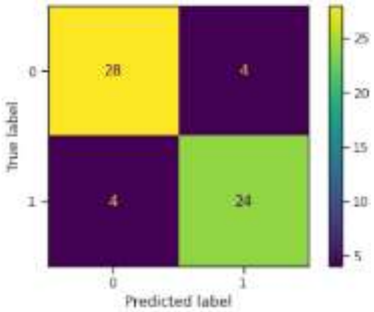


Fig. 15. Confusion Matrix of SVM

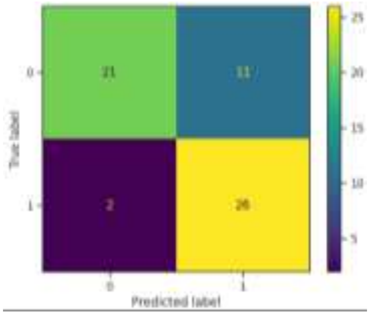


Fig. 16. Confusion Matrix of Random Forest

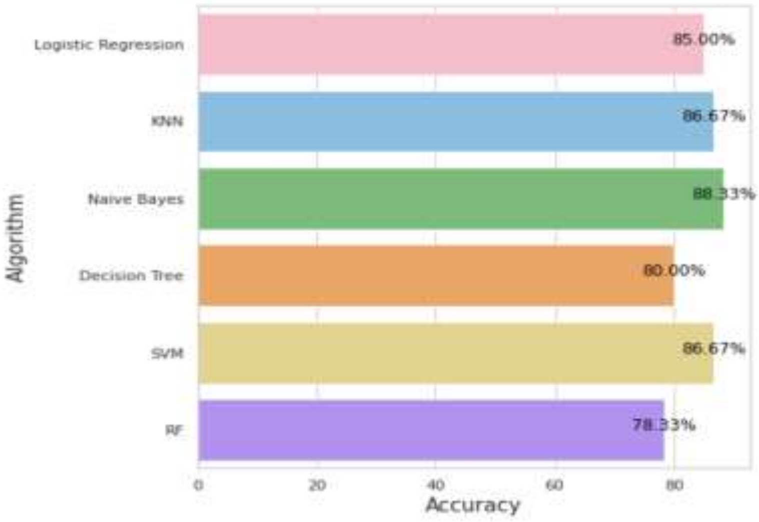


Fig. 17. Model Comparison

4. Conclusions

The main goal of this research is to create and assess a model that can anticipate issues using information and demographic data. By conducting analysis and evaluating models the study was able to gain valuable insights into the effectiveness of various machine learning techniques in predicting heart disease. The findings reveal that Naive Bayes achieved the accuracy among the comparison models at 88.33% followed by the K Nearest Neighbor Model (KNN) and Support Vector Machine (SVM), with accuracies of 86.67%. Logistic regression also demonstrated capabilities, with an accuracy rate of 85%. However, the decision tree and random forest model had accuracies at 80% and 78.33%, respectively.

These findings carry implications, for policymaking and healthcare decisions. The impressive accuracy of Naive Bayes KNN and SVM models suggests their potential in identifying individuals at risk of heart disease and guiding patient care [5,12]. Moreover, logistic regression, being a model showcases its effectiveness and could serve well in clinical risk evaluation. While decision trees and forest models may not match the accuracy levels of algorithms, they still offer insights into the underlying patterns and relationships within the data. Enhancing and fine tuning these models could enhance their performance and practicality, in medical research endeavors [8].

In general, this research contributes to the expanding field of predicting disease and underlines the significance of utilizing machine learning algorithms to enhance the assessment and monitoring of risks, in patients. By leveraging the capabilities of these models, physicians can effectively pinpoint individuals at risk, for heart conditions and put in place preventive strategies that lessen the impact of deaths caused by cardiovascular diseases and strokes. Nevertheless, it is crucial to recognize the constraints of our study, which include data inaccuracies, simplification of model assumptions and extrapolation to patient populations. Subsequent investigations should concentrate on overcoming these limitations and incorporating models to enhance the precision, applicability and clinical utility in predicting heart diseases. In essence this study showcases how machine learning holds promise in forecasting ailments. Through exploring algorithms and addressing identified shortcomings future studies can contribute towards creating robust and dependable diagnostic tools for enhancing patient care related to heart health.

Acknowledgements

The manuscript is not be submitted to any other journal for simultaneous consideration. The submitted work is original and is not published elsewhere in any form or language. No slicing is practiced. No data, text, or theories by others are presented as if they were ours. Proper acknowledgments to other works are given and no copyrighted material is used in the study. No animals or humans were harmed. The authors would like to thank the University Research Fund for its contribution. The authors declared that they have no conflict of interest.

References

Soni J, Ansari U, Sharma D, Soni S. (2011) *Predictive data mining for medical diagnosis: an overview of heart disease prediction*, Int. J. Comput. Appl, 17(8), pp. 43–48.

Haq, A.U., Li, J.P., Memon, M.H., Nazir, S., Sun, R. (2018) *A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms*, Mobile Information. Systems, pp. 1–21.

Hossain MI, Maruf MH, Khan MAR, Prity FS, Fatema S, Ejaz MS, Khan MAS (2023) *Heart disease prediction using distinct artificial intelligence techniques: performance analysis and comparison,* Iran J Comput Sci.

Spencer R., Thabtah F., Abdelhamid N., Thompson M. (2020) *Exploring feature selection and classification methods for predicting heart disease,* Digital Health.

Jindal, H., Agrawal, S., Khera, R., Jain, R., Nagrath, P. (2021) *Heart disease prediction using machine learning algorithms*. IOP Conf. Ser. Mater. Sci. Eng, 1022(1).

Roohallah Alizadehsani, Jafar Habibi, Mohammad Javad Hosseini, Hoda Mashayekhi, Reihane Boghrati, Asma Ghandeharioun, Behdad Bahadorian, Zahra Alizadeh Sani (2013) *A data mining approach for diagnosis of coronary artery disease*, Computer Methods and Programs in Biomedicine, 111(1), pp. 52-61.

Javeed A., Zhou S., Yongjian L., Qasim I., Noor A., Nour R. (2019) *An intelligent learning system based on random search algorithm and optimized random forest model for improved heart disease detection*, IEEE Access.

Zhenya Q., Zhang Z. (2021) *A hybrid cost-sensitive ensemble for heart disease prediction*, BMC Medical Informatics and Decision Making, 21(1), pp. 73.

Fitriyani N. L., Syafrudin M., Alfian G., Rhee J. HDPM (2020) *An effective heart disease prediction model for a clinical decision support system*, IEEE Access.

Ali L., Rahman A., Khan A., Zhou M., Javeed A., Khan J. A. (2019) *An automated diagnostic system for heart disease prediction based on statistical model and optimally configured deep neural network*, IEEE Access.

R. Indrakumari, T. Poongodi, Soumya Ranjan Jena (2020) *Heart Disease Prediction using Exploratory Data Analysis*, Procedia Computer Science, 173, pp. 130-139.

Mohan S., Thirumalai C., Srivastava G. (2019) *Effective heart disease prediction using hybrid machine learning techniques*, IEEE Access.

Barragán-Montero, A., Javaid, U., Valdés, G., Nguyen, D., Desbordes, P., Macq, B., Willems, S., Vandewinckele, L., Holmström, M., Lofman, F., Michiels, S. (2021) *Artificial intelligence and machine learning for medical imaging: a technology review*, Phys Med. 83, pp. 242–256.

Srivastava, N. (2005) *A logistic regression model for predicting the occurrence of intense geomagnetic storms*, Ann. Geophys. 23, pp. 2969–2974.

O. W. Samuel, G. M. Asogbon, A. K. Sangaiah, P. Fang, G. Li (2017) *An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction,* Expert Systems with Applications, 68, pp. 163–172.

R. Detrano, A. Janosi, and W. Steinbrunn (1989) *International application of a new probability algorithm for the diagnosis of coronary artery disease*, American Journal of Cardiology, 64(5), pp. 304–310.