HEART DISEASE PREDICTION USING MACHINE LEARNING

## A PROJECT REPORT

***Submitted by***

**Vinay Partap Singh(22BDO10040)** **Yash Kumar (22BDO10009)**

**Vibhanshu(22BDO10017)**

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

# BACHELOR OF ENGINEERING

**IN**

COMPUTER ENGINEERING



## Chandigarh University



# BONAFIDE CERTIFICATE

Certified that this project report **“HEART DISEASE PREDICTION USING MACHINE LEARNING”** is the bonafide work of “**Vinay Partap Singh(22BDO10040), Yash(22BDO10009), and Vibhanshu (22BDO10017)”** who carried out the project work under my/our supervision.

**SIGNATURE**

**HEAD OF THE DEPARTMENT**

AIT-CSE

**SIGNATURE**

Ms. Shilpa Sharma

**SUPERVISOR**

Submitted for the project viva-voce examination held on

### INTERNAL EXAMINER EXTERNAL EXAMINER

**Table of contents**

List of tables (1)

List of figures (2)

Abstract (3)

### Chapter 1 Introduction……………………………………………………(4)

### Chapter 2 Literature review………………………………………………(7)

### Chapter 3 Design and process

* 1. Concept generation (14)
  2. Evaluation of features/specifications (19)
  3. Machine earning (24)
     1. Supervised Learning (24)
     2. Unsupervised Learning (28)
  4. Survey Method (29)
  5. Literature section and scope definition (30)
  6. Feature selection (30)
  7. Model evaluation (31)
  8. Ethical considerations (32)
  9. Integrating Machine Learning With HR Processes (32)

### Chapter 4 Result analysis and validation (34)

**Chapter 5 Conclusion and future scope** (36)

References (38)

# LIST OF FIGURES

Fig. 1 Proposed System

Fig2 Machine Learning

Fig 3. DT Structure

Fig 4 Unsupervised learning workflow

# ABSTRACT

Heart disease is a significant global cause of mortality, and predicting it through clinical data analysis poses challenges. Machine learning (ML) has emerged as a valuable tool for diagnosing and predicting heart disease by analyzing healthcare data. Previous studies have extensively employed ML techniques in medical research for heart disease prediction. In this study, eight ML classifiers were utilized to identify crucial features that enhance the accuracy of heart disease prediction. Various combinations of features and well-known classification algorithms were employed to develop the prediction model. Neural network models, such as Naïve Bayes and Radial Basis Functions, were implemented, achieving accuracies of 94.78% and 90.78% respectively in heart disease prediction. Among the state-of-the-art methods for cardiovascular problem prediction, Learning Vector Quantization exhibited the highest accuracy rate of 98.7%. The motivation behind predicting Cardiovascular Heart Disease lies in its potential to save lives, improves health outcomes, and allocates healthcare resources efficiently. The key contributions encompass early intervention, personalized medicine, technological advancements, the impact on public health, and ongoing research, all of which collectively work toward reducing the burden of CHD on both individual patients and society as a whole analysis.

# CHAPTER-1 INTRODUCTION

The healthcare industry generates a lot of data about patients, illnesses, and diagnoses, but it isn't being used correctly to produce the desired results. Heart disease and stroke are two of the main causes of death. According to a WHO report, cardiovascular diseases directly kill more than 17.8 million people every year. Because there isn't enough analysis, the healthcare industry's huge amounts of patient, illness, and diagnosis data don't have the effect on patient health that was hoped for[1](https://www.nature.com/articles/s41598-023-40717-1#ref-CR1). Heart and blood vessel diseases, or CVDs, include coronary artery disease, myocarditis, vascular disease, and other conditions. Stroke and heart disease kill 80% of all people who die from CVD. Three-quarters of all people who die are under the age of 70. The main things that put you at risk for cardiovascular disease are your gender, smoking, age, family history, poor diet, lipids, lack of physical activity, high blood pressure, weight gain, and drinking alcohol[2](https://www.nature.com/articles/s41598-023-40717-1#ref-CR2). High blood pressure and diabetes are two examples of things that can be passed down and make you more likely to get cardiovascular disease. Some of the other things that raise the risk are being inactive, being overweight, not eating well, having back, neck, and shoulder pain, being very tired, and having a fast heartbeat. Most people have chest pain, shoulder pain, arm pain, shortness of breath, and a general sense of weakness. As it has been for a long time, chest pain is the most common sign that the heart isn't getting enough blood[3](https://www.nature.com/articles/s41598-023-40717-1#ref-CR3). This kind of chest pain is called angina in medicine. Some tests, like X-rays, Magnetic Resonance Imaging (MRI), and angiography, may help figure out what is wrong. On the other hand, sometimes important medical equipment is not easily accessible, which limits what can be done in an emergency. When it comes to figuring out what's wrong with your heart and treating it, every second counts[4](https://www.nature.com/articles/s41598-023-40717-1#ref-CR4). Heart disease diagnostics aren't as good as they could be, and there is a huge need for better big-data analysis in cardiovascular system redesign and patient outcomes. But noise, incompleteness, and irregularities in the data make it hard to draw clear, accurate, and well-grounded conclusions from them. Because of recent improvements in technologies like big data, information storage, and retrieval, computerised intelligence plays an important role in cardiology. In order to draw conclusions from the data mined with different ML models, researchers used pre-processing techniques[5](https://www.nature.com/articles/s41598-023-40717-1#ref-CR5). Using a common set of algorithms and their variations, which are used to keep track of hereditary cardiac disorders and healthy controls, it is possible to predict when the first stage of heart failure will start. Classification technique, DT, SVC, LR, and RF machines are all types of algorithms that can be used to predict cardiac arrest. When it comes to machine learning, there are three main ways to think: The three main types of machine learning are task-driven supervised ML (classification/regression), data-driven unsupervised ML (clustering), and error-driven reinforcement learning (RL). Coronary artery disease is a very common disease of the main blood vessels that bring blood to the heart muscle. Plaques, which are made up of lipoproteins, can build up in the arteries of the heart, which can lead to coronary artery disease. Atherosclerosis is the name for the buildup of these plaques[6](https://www.nature.com/articles/s41598-023-40717-1#ref-CR6). Atherosclerosis slows the flow of blood through the veins to the chest and other organs. It goes up if you have heart disease, angina, or a stroke. Men and women may have different warning signs and symptoms of coronary artery disease. For example, men are more likely than women to have chest pain. In addition to chest pain, women are more likely to experience shortness of breath, nausea, and sudden exhaustion. Heart failure, chest tightness, chest pressure, and chest pain can all be signs of coronary artery disease[7](https://www.nature.com/articles/s41598-023-40717-1#ref-CR7). The Heart Disease Prediction System incorporates the Naive Bayesian Classification technique to assist in making decisions. By analyzing a vast database of past heart disease cases, the system uncovers valuable insights. This model is highly efficient in identifying patients at risk of heart disease. It possesses the ability to respond to intricate queries, showcasing its strengths in terms of interpretability, access to comprehensive information, and accuracy[8](https://www.nature.com/articles/s41598-023-40717-1#ref-CR8). Making accurate and timely decisions is crucial in the medical field, especially when treating patients. Machine learning (ML) techniques play a significant role in predicting diseases by leveraging the extensive data generated by the healthcare industry. In India, heart disease is a leading cause of mortality, and the World Health Organization (WHO) emphasizes the importance of timely intervention to predict and prevent strokes. This paper focuses on predicting cardiovascular disease with enhanced accuracy by employing ML techniques such as Decision Tree and Naïve Bayes, in conjunction with risk factors. The dataset utilized in this study is the Heart Failure Dataset, which comprises 13 attributes[9](https://www.nature.com/articles/s41598-023-40717-1#ref-CR9). The author investigated how well two algorithms, Support Vector Machine (SVM) and Naive Bayes, performed in predicting the occurrence of heart disease and the survival status of patients. The algorithms were applied to a dataset that included sixteen attributes from the University of California, Irvine's Centre for Machine Learning and Intelligent Systems. To assess the models' performance, a confusion matrix was used to visualize metrics like accuracy, recall, precision, and error. Additionally, statistical analysis was carried out by utilizing the receiver operating characteristic (ROC) curve and calculating the area under the curve to demonstrate the accuracy of the models[10](https://www.nature.com/articles/s41598-023-40717-1#ref-CR10). In this research paper, a system is introduced that employs a radial basis function neural network to accurately predict eight different types of cardiac arrhythmias. The primary focus of the study is the analysis of heart rate time series data, and the proposed algorithm is specifically designed to predict specific arrhythmias, namely Left bundle branch block, Atrial fibrillation, Normal Sinus Rhythm, Right bundle branch block, Sinus bradycardia, Atrial flutter, Premature Ventricular Contraction, and Second-degree block. The heart rate time series data utilized in the study is sourced from the MIT-BIH arrhythmia database. Both linear and nonlinear features are extracted from the heart rate time series of each individual arrhythmia. Training of the radial basis function neural network (RBFN) is conducted using 70% of the feature datasets, while the remaining 30% is dedicated to predicting the occurrence of the eight cardiac diseases. The proposed approach demonstrates an impressive overall prediction accuracy of 96.33%, surpassing the performance of existing methods documented in the literature[11](https://www.nature.com/articles/s41598-023-40717-1#ref-CR11). A novel method known as Radial Basis Classification is introduced for the classification of heart disease using clinical databases. Conventional classifiers that involve multiple attributes tend to have a large number of parameters, making it difficult to determine the ideal attributes. To address this, the concept of Multivariate Function Classifier Ideas is proposed, aiming to encourage a more cohesive stochastic trend and minimize the likelihood of errors or unforeseen results. This formula proves beneficial for arranging multidimensional data and enhancing the accuracy of grouping in the analysis phase. The results of the study indicate that the suggested calculation method offers higher precision compared to previous approaches[12](https://www.nature.com/articles/s41598-023-40717-1#ref-CR12). The backpropagation neural network has demonstrated satisfactory performance in predicting accuracy. However, to further enhance accuracy and determine the specific type of heart disease, the paper integrates the CBR technique with the ANN. By leveraging historical patient records, a level of accuracy reaching 97% is attained. This research not only utilizes CBR to enhance accuracy but also to predict the type of heart disease. The CBR output encompasses both the identified type of heart disease and the recommended medication. This enables a comparison between the original medication and the medication suggested by the RBF (Radial Basis Function). The medication prescribed using this approach exhibits a comparative accuracy of 98%[13](https://www.nature.com/articles/s41598-023-40717-1#ref-CR13). Symptoms include trouble breathing, pain in the upper back, neck, jaw, or throat, and pain, numbness, weakness, or a chill in the limbs. Due to the narrowing of blood vessels in certain parts of the body, it is possible to have coronary artery disease and not know it until you have a heart attack, angina, stroke, or heart failure. Keep an eye out for signs of heart problems, and if you're worried, talk to your doctor. If you get checked out often, heart (cardiovascular) disease may be found earlier[14](https://www.nature.com/articles/s41598-023-40717-1#ref-CR14). This proposed method uses supervised ML classifiers to show how different models can predict the presence of cardiovascular disease and evaluate the performance of these classifiers, such as the random forest, decision tree, support vector machine, XGBoost, radial basis function, k-nearest neighbour, naïve bayes and learning vector quantization.

The goal of predicting Cardiovascular Heart Disease is to develop accurate and reliable models that can assess an individual's risk of developing various cardiovascular conditions, enabling early intervention, personalized treatment, and ultimately reducing the burden of heart disease on public health.

# CHAPTER 2 LITERATURE REVIEW

Heart rate variability (HRV) has emerged as a reliable predictor for congestive heart failure (CHF). However, challenges remain in effectively extracting temporal features and efficiently classifying high-dimensional HRV representations. To address these challenges, this study proposes an ensemble method that utilizes short-term HRV data and deep neural networks for CHF detection. The research incorporates five publicly available databases: BIDMC CHF database (BIDMC-CHF), CHF RR interval database (CHF-RR), MIT-BIH normal sinus rhythm (NSR) database, fantasia database (FD), and NSR RR interval database (NSR-RR). Three different lengths of RR segments (N = 500, 1000, and 2000) are employed to evaluate the proposed method. Initially, expert features are extracted from the RR intervals (RRIs). Subsequently, a network based on long short-term memory-convolutional neural networks is constructed to automatically extract deep-learning (DL) features. Finally, an ensemble classifier is used to detect CHF using the aforementioned features. Blindfold validation is conducted on three CHF subjects and three normal subjects, resulting in accuracies of 99.85%, 99.41%, and 99.17% for N = 500, 1000, and 2000 length RRIs, respectively, utilizing the BIDMC-CHF, NSR, and FD databases15. In this publication, there is a summary of past studies and an analysis of how well the algorithm works. Before training and testing different algorithms, the suggested architecture processes the data that comes in first. The author suggests using Adaboost because it makes every ML method look better. Also, the author agreed that settings could be fine-tuned to improve accuracy. Researchers came up with a deep learning strategy for analysing and spotting cardiac conditions by using the UCI dataset. They went on to say that deep neural networks could help improve the analysis and diagnosis of cardiovascular disease as a whole. Compared to other ways to improve model performance, they found that the Talos Hyper process worked the best16. The KNN, RF, SVM, and DT algorithms were studied as ML models for predicting heart disease with high accuracy, high recall, and high precision. As shown in their estimation method for cardiac disorders, which is hosted on the UCI ML library, SVM-based categorization was the most accurate. We looked at the results of four machine learning techniques and one neural network (NN) for spotting heart disease. This study compared algorithms for predicting cardiac dose based on things like reliability, recall, accuracy, and F1. The Deep NN algorithm was able to spot heart problems 98% of the time. In order to show that the algorithm is useful for predicting illness, they focused on how it could be used with a medical dataset. The researchers came to the conclusion that boosting and bagging are good ways to improve the performance of classifiers that aren't very good at predicting the risk of heart disease. The results showed that the accuracy of predictions went up a lot after feature selection was used, which improved the procedure17. Ensemble approaches were used to improve the accuracy of bad classifiers by no more than 7%. In recent years, ML algorithms have gotten a lot of praise for how accurate and useful they have become at making predictions. It is critical to be able to create and recommend models with the greatest accuracy and efficiency possible18. Since hybrid models use many ML techniques and data systems, they may be able to accurately predict health problems. Weedy classifiers worked better when they used bagging and boosting, and their ability to predict cardiovascular disease risk was rated well when they worked together. They made the hybrid model by using majority voting with the Bayes Net, NB, C4.5, MLP, and RF classifiers19. With 85.48 percent of the time, the model that was made is right. In addition to learning models, the UCI cardiovascular disease dataset has recently been used with ML methods like RF and SVM. Accuracy went up when a lot of classifiers were added to the voting-based model20. Based on the data, using the weak classifiers led to an increase of 2.1% in accuracy. We used ML classification methods to figure out how people with long-term conditions would do. They found that the Hoeffding classifier can predict coronary disease with an accuracy of 88.56 percent. Overall, they found that when the hybrid model was used with the desired features, it was 87.41% accurate. We used an SVM model and the Fisher score method to choose features based on the mean.

We used a lot of different classification methods and feature sets to make this one-of-a-kind prediction model. The proposed HRFLM used an ANN with a deep network and 13 clinical features as inputs. Data mining techniques like DT, SVM, NN, and KNN were also looked into. Researchers have found that it's helpful to use SVM to predict who will get sick. There was a new method called "vote," and a hybrid method that combines LR and NB was talked about. The HRFLM strategy worked out to be 88.7% effective22. We were able to make a model to predict death from cardiac failure that takes into account a wider range of risk factors by improving the random survival forest23. The IRSF used a split criterion and a stop criterion that were new to the field to tell the difference between survivors and people who didn't make it. Data mining has also been used to find out if someone has a cardiovascular disease24. Heart diseases are still diagnosed using Bayesian, DT classifiers, NN, association law, KNN, SVM, and ML algorithms. SVM was right 99.3% of the time. Several classifiers based on machine learning have been made to predict how long a patient will live25. Characteristics that were linked to the most important risk factors were rated, and the results were compared to traditional bio statistical testing. Researchers came to the conclusion that serum creatinine levels and ejection fraction are the two most important things to look at when trying to make accurate predictions26. The ML algorithm was used to make a model for finding CVD. In this study, we cleaned and looked at the data in four different ways. The DT and RF methods got an accuracy rate of 99.83%, while the SVM and KNN methods only got accuracy rates of 85.32% and 84.49%, respectively. Using the ensemble method, another study predicted CHF by looking at HRV and using deep neural networks to fill in knowledge gaps in unrelated areas. Overall, the method suggested was 99.85% right. In a recent publication27, different types of data were used to make an intelligence framework. These were principal component analyses and RF-based MLA. The FAMD was applied to RF in order to value the relevant properties and predict illness. The suggested method is correct 93.44% of the time, sensitive 89.28% of the time, and specific 96.74% of the time. In order to test their theory, the authors used a set of 303 cases that were made by adding to the Cleveland dataset. In tests, the suggested DT algorithm did 75.5% better than the baseline algorithm. Heart disease is often referred to as "cardiovascular disease"28. Several researchers are trying to make it easier to tell if someone has heart disease. Their research on heart disease covers a lot of ground. The author used data from the Hungarian and Statlog sets to classify CVD using the reduced error pruning tree (REP tree), R tree, M5P tree, logistic regression (LR), J48, naive bayes (NB), and JRIP. People use random forest (RF), decision tree (DT), and linear regression (LR). Support vector machine (SVM), CART, linear discriminant analysis (LDA), gradient boosting (XGB), and random forest (RF) are all used29. The goal of this study is to find a way to figure out how likely someone is to get heart disease. The results show that SVM does better than LR because it gets 96% accuracy while LR only gets 92% accuracy. The author says that the DT model always does better than the NB model and the SVM model. SVM has been shown to be 87% accurate, DT to be 90% accurate, and LR to be the most accurate at predicting when heart disease will happen, compared to DT, SVM, NB, and k-nearest neighbour (KNN). Table 1, represents the overall performance metric comparison of state-of-the-art methods.

| **Year** | **Author Name** | **Online Database** | **Classification Type** | **Performance Metric** | **Accuracy** |  |
| --- | --- | --- | --- | --- | --- | --- |
| 2022 | [20](https://www.nature.com/articles/s41598-023-40717-1#ref-CR20) | IoT based data | K-NN, DT, RF, MLP, NB, L-SVM | Accuracy, sensitivity, F1 score | 96.12 |  |
| 2022 | [21](https://www.nature.com/articles/s41598-023-40717-1#ref-CR21) | Di-ScRi database | Evimp functions, Multivariate adaptive regression | Accuracy, Specificity, Sensitivity, F1 score | 91.2 |  |
| 2022 | [22](https://www.nature.com/articles/s41598-023-40717-1#ref-CR22) | Hungarian-Statlog database | LR, NB, RF REP, M5P Tree, J48, JRIP | RMSE, MAE | 89.7 |  |
| 2022 | [23](https://www.nature.com/articles/s41598-023-40717-1#ref-CR23) | UCI repository | KNN, DT, LR, NB, SVM | Accuracy, Sensitivity, F1-Score, Specificity | 93.23 |  |
| 2022 | [24](https://www.nature.com/articles/s41598-023-40717-1#ref-CR24) | Congenital heart disease database of 3910 Singleton | RF-fetal echocardiography | RMSE, MAE | 95.02 |  |
| 2022 | [25](https://www.nature.com/articles/s41598-023-40717-1#ref-CR25) | Pathogen, Host feature | LR, KNN, SVM, RF | Accuracy, sensitivity, F1 score | 94.08 |  |
| 2022 | [26](https://www.nature.com/articles/s41598-023-40717-1#ref-CR26) | Heart Disease (Kaggle Repository) | KNN, RF, ANN, Ada, GBA | RMSE, MAE | 90.91 |  |
| 2021 | [27](https://www.nature.com/articles/s41598-023-40717-1#ref-CR27) | Heart Cleveland (UCI repository) | LR, DT, RF, SVM, HRFLM | Accuracy, Sensitivity, F1-Score, Specificity | 96.22 |  |
| 2021 | [28](https://www.nature.com/articles/s41598-023-40717-1#ref-CR28) | UCI Cleveland database | RF, DT, LR | Accuracy, sensitivity, F1 score | 94.21 |  |
| 2021 | [29](https://www.nature.com/articles/s41598-023-40717-1#ref-CR29) | UCI repository | SVM, NB, DT | Sensitivity, accuracy | 94.11 |  |

Table 1

The RF-based method is 97% accurate at predicting congenital heart disease, with a specificity of 88% and a sensitivity of 85%. They were able to find CVD with 94% accuracy, 95% specificity, and 93% sensitivity by using LR, MARS, EVF, and CART-ML. RF was used to predict drug targets in host-host and host–pathogen interactions related to CVD caused by microorganisms. Several ensembles and hybrid representations have been put forward to solve the problem of predicting heart disease. Based on the suggested method[30](https://www.nature.com/articles/s41598-023-40717-1#ref-CR30), CVD from the Mendeley Institute, the Cleveland datasets, and the IEEE Port are all processed with a high level of accuracy (96%, 88.24%, and 93%, respectively). The author put together the LR and RF algorithms to predict heart disease and got an accuracy of 88.7%. In this study, researchers want to find out more about how calcium in the coronary arteries and plaque in the carotid arteries are related. Both are linked to a higher risk of heart disease, but they may not be causing any symptoms yet. Machine learning and the internet of things are often used to predict and diagnose illnesses right now. The author was able to predict heart problems 94% of the time with the help of mobile devices and the deep learning method. The author employs machine learning classifiers and the Internet of Things to predict heart infections before they occur[31](https://www.nature.com/articles/s41598-023-40717-1#ref-CR31). At the end of the day, we want to show that ML could be a good way to solve the problem at hand. We can use ML to look at cases related to illnesses and health problems by looking at hundreds of healthcare datasets. Researchers have worked on sophisticated computer perception for reliable healthcare to find out how machine vision practises help human needs, such as psychosocial health, specific movement, exposure-induced fatigue, frequently having to watch live actions, image analysis, deep learning, pattern classification, and how language understanding and computer animation work with robotics[32](https://www.nature.com/articles/s41598-023-40717-1#ref-CR32). The authors noticed and wrote about how users learn about sharp interfaces and virtual reality tools, which leads to the development of complex restorative systems that can do human activities and recognise them. The work backs up the direct method of machine vision in the healthcare sector. This includes the technology behind intelligent wheelchairs, possible help for the visually impaired, and other object tracking solutions that have recently been used to monitor health and safety[33](https://www.nature.com/articles/s41598-023-40717-1#ref-CR33). Scientists used support vector machines, generalised boosting machines, logistic regression, light boosting machines, and random forests to see how likely someone was to get cardiovascular disease. RF was the best way to predict who would get heart disease. It was right 88% of the time. Our method is put up against the current study. This is the first and only study to compare the accuracy of seven different ML classifiers for predicting cardiovascular illness. These methods include the most cutting-edge ones like learning vector quantization, RBF neural networks, and logistic regression. So, it is now possible to use a system that is both accurate and useful for predicting heart problems. Also, we suggest using the best machine learning classifier when making smart systems for predicting CHD[34](https://www.nature.com/articles/s41598-023-40717-1#ref-CR34). The key features of cardiovascular illnesses include high morbidity, disability, and death, and the etiology of heart disease remains an unresolved worldwide issue. Therefore, accurate early prediction of anticipated outcomes in individuals affected by cardiac illness is necessary. In this work, we employed ML modelling to predict cardiac disease. This study focuses on predicting heart disease using ML classifiers. The authors first address the dataset problem, and subsequently enhance and standardize it for tokenization and lowercase conversion. The datasets were then utilized to train and test the classifiers, assessing their performance to achieve the highest level of accuracy. These algorithms must meet strict admission criteria, including modernity, representativeness, and high maturity. Previously, we employed Naive Bayes and Radial Basis Functions by examining the works of prior researchers. We investigated whether these approaches had been utilized on the UCI heart dataset by earlier researchers.

The proposed work contributions:

1. i.

The authors commence by discussing datasets, which are subsequently standardized and enhanced. These datasets are then employed to train and test several classifiers to determine the one with the highest accuracy.

1. ii.

Subsequently, the authors utilize the correlation matrix to classify the optimal values or features.

1. iii.

The third step involves applying the ML classifiers to the pre-processed dataset, aiming for the highest achievable accuracy through parameter modifications.

1. iv.

In the fourth and final step, the suggested classifiers are assessed for accuracy, precision (specificity), recall (sensitivity), and F-Measure.

Ultimately, the suggested classifiers outperform the state-of-the-art classifiers presented in Table [1](https://www.nature.com/articles/s41598-023-40717-1#Tab1) in terms of accuracy.

# CHAPTER 3 DESIGN AND PROCESS

## Concept Generation

With the utilization of the heart dataset, we employed ML classifiers to predict the presence of coronary heart disease. The dataset was obtained from the UCI repository, and feature engineering was applied for data pre-processing before selecting the features. Subsequently, we divided it into training and test datasets, using around 70% of the total data for training and the remaining portion for testing. The training dataset is used to create a model that predicts heart disease, while the test dataset is utilized to evaluate the classifiers. Prior to transforming categorical variables into numerical values for classification, a thorough dataset analysis was conducted. The dataset was labelled as "normal" and "diseased" in Step 1. The "diseased" label indicates the presence of heart disease, while the "normal" label indicates the absence of heart disease. In Step 2, data cleaning was performed during the training phase. Data pre-processing involved handling missing values by calculating the mean due to the presence of partial and missing values. Step 3 involved data visualization using Exploratory Data Analysis (EDA) to examine relationships between various attributes. Notably, we identified that the correlation for FBS is relatively low. Moving to Step 4, ML classifiers were applied to the pre-processed dataset, and the classifiers' performance was evaluated using a variety of parameters. As previously mentioned, the dataset was split into test and training sets to respectively assess the classifiers and develop the model. The employed classifiers demonstrated varying levels of accuracy in detecting the presence of heart disease. Figure [1](https://www.nature.com/articles/s41598-023-40717-1#Fig1) illustrates the stages of our proposed working approach.

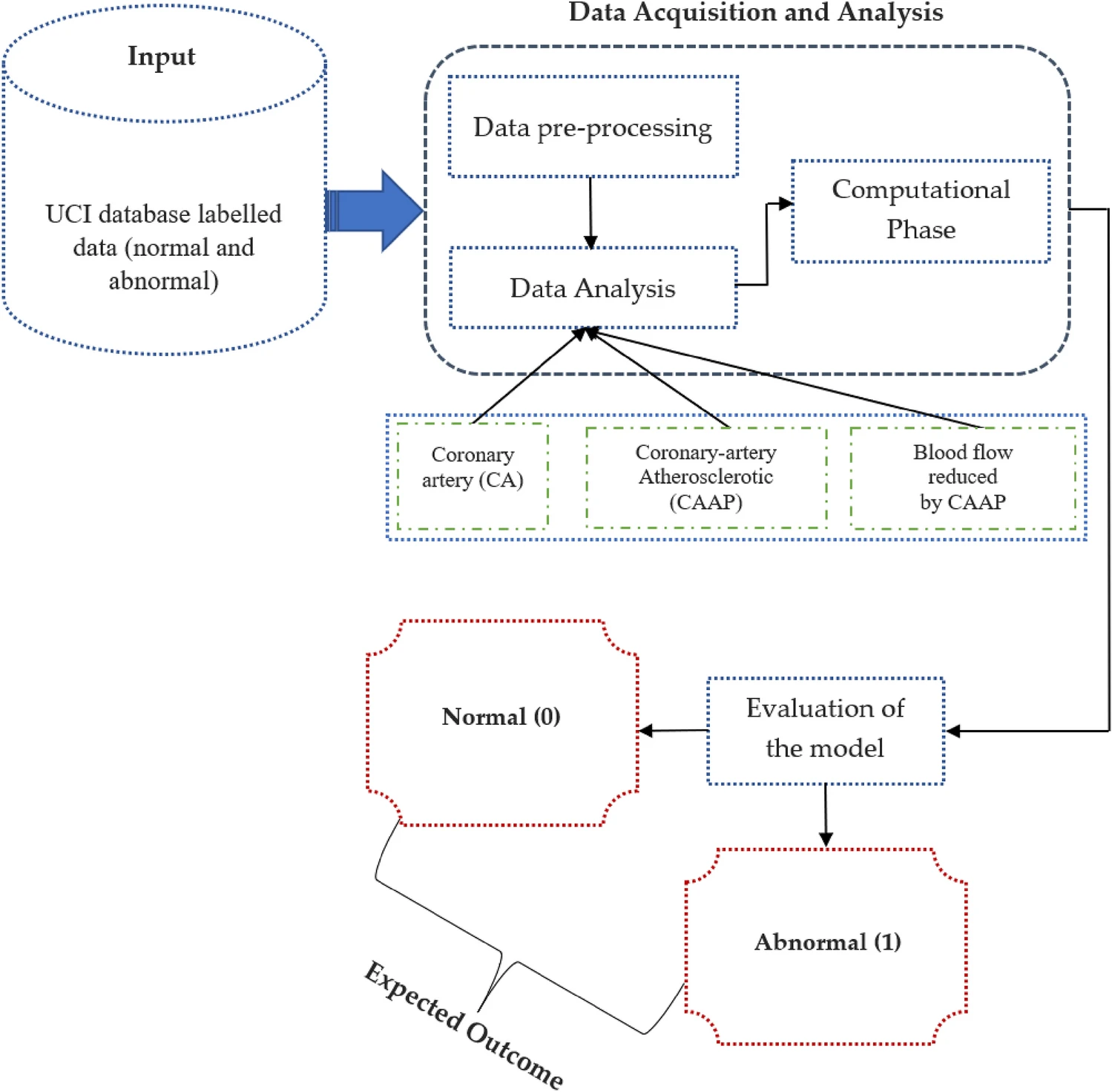


Fig. 1 Proposed System

This study investigates ML techniques such as Naive Bayes, SVM, voting, XGBoost, AdaBoost, bagging, DT, KNN, RF, and LR classifiers. These algorithms can aid doctors and data analysts in making correct diagnoses of cardiac disease. This article incorporates recent data on cardiovascular illness, as well as relevant journals, research, and publications. The methodology, as in[1](https://www.nature.com/articles/s41598-024-74656-2#ref-CR1), provides a framework for the suggested model. The methodology is a set of steps that transform raw data into consumable and identifiable data patterns. The proposed approach consists of three stages: the first stage is data collection; the second stage extracts specific feature values; and the third stage is data exploration, as shown in Fig. [1](https://www.nature.com/articles/s41598-024-74656-2#Fig1). Depending on the procedures employed, data preprocessing deals with the missing values, cleansing of the data, and normalization[2](https://www.nature.com/articles/s41598-024-74656-2#ref-CR2). We then classified the pre-processed data using the ten classifiers (A1, A2,., A10). Finally, after putting the suggested model into practice, we evaluated its performance and accuracy using a range of performance measures. This model developed a Reliable Prediction System for Heart Disease (RPSHD) using a variety of classifiers. This model uses 13 medical factors for prediction, among which are age, sex, cholesterol, blood pressure, and electrocardiography[3](https://www.nature.com/articles/s41598-024-74656-2#ref-CR3)This research employs both the CHDD and a private dataset for heart disease prediction. The CHDD dataset has 303 samples, while the private dataset has 200, and they have the same features. The combined dataset contains 503 records, and 13 features are associated with each one (including demographic, clinical, and laboratory parameters). The datasets have many features that can be used for heart disease prediction including *age*,*gender*,*blood pressure*,*cholesterol levels*,*electrocardiogram readings-ECG*,*chest pain*,*exercise-induced angina*,*blood sugar with fasting condition*,*max heart rate achieved*,*oldpeak*,*coronary artery*,*thalassemia*,*and other clinical and laboratory measurements*, as shown in Table [2](https://www.nature.com/articles/s41598-024-74656-2#Tab2). The outcome variable known as *“Target”* takes a binary value and refers to the *heart disease predicting* feature (i.e., it indicates whether or not cardiac disease is present).

| **Feature**  **no.** | **Feature name** | **Feature**  **code** | **Description** | **Values type** |
| --- | --- | --- | --- | --- |
| 1 | Age | AGE | Age of patient | Number of years |
| 2 | Gender | GEN | Patient sex | Female = 0, male = 1 |
| 3 | Chol | CHOL | Evaluation of a patient’s cholesterol levels | mg/dl |
| 4 | Trestbps | BRP | Blood resting pressure | Mm |
| 5 | CP | CPT | Chest pain types | Typical angina = 1, atypical angina = 2, nonanginal pain = 3, asymptomatic = 4 |
| 6 | Fbs | FBS | Blood sugar in fasting case | < or > 120 mg/dl  (true = 1, false = 0) |
| 7 | Thalach | MHR | Maximum rate achieved on heart | Continuous |
| 8 | RestEcg | REC | Electrocardiograph by resting | 0 = no abnormalities, 1 = normal, 2 = left ventricular hypertrophy (possible or certain) |
| 9 | Oldpeak | OP | ST depression when compared to rest taken quantity | Continuous |
| 10 | Exang | EIA | Angina caused by exercise | 1 = there is pain, 0 = there is no pain |
| 11 | Ca | CMV | Count of main vessels colored by fluoroscopy | 0–3 |
| 12 | Slope | PES | Peak exercise ST segment slope | Up sloping = 0, flat = 1, down = 2 |
| 13 | Thal | TS | Thallium stress | Negative = 0, positive = 1, inconclusive = 2 |
| 14 | Target | | target variable representing diagnosis of heart disease using the angiographic disease status. | 0 = no heart disease (< 50% diameter narrowing)  1 = heart disease (> 50% diameter narrowing) |

Table 2

In this research, preprocessing was performed on collected data. The CHDD has four inaccurate CMV records and two erroneous TS entries. Incorrect data is updated to reflect the best possible values for all fields. Then, StandardScaler is employed to normalize all the features to the relevant coefficient, ensuring each feature has a zero mean and one variance. By considering the patient’s history of cardiac problems and following other medical concerns, an organized and composed augmented dataset was chosen.

The dataset studied in this research is a combination of accessible public WBCD and chosen private datasets. Partitioning the two datasets in this way allows us to use the holdout validation method. In this study, 25% of the data is in the test dataset, compared to 75% in the training dataset. The mutual information method is used in this research to measure the interdependence of variables. Larger numbers indicate greater dependency and information gathering.

## Evaluation and selection of Features/Specifications

The proposed system mentioned by us is an algorithm which aims to combine the various prediction models to create a hybrid system which continuously tracks progress in real time and has the ability to predict to provide an optimized evaluation and statistics in most efficient manner.

Creating a hybrid prediction system involves integrating multiple models that capture different aspects of employee grades. Here's a high-level outline of an algorithm for such a system:

* Model Selection:

Identify and select various traffic flow prediction models, including: Statistical models (e.g., ARIMA, SARIMA, LSTM, Prophet)

Machine learning models (e.g., Random Forest, Gradient Boosting, Support Vector Machines)

Deep learning models (e.g., Convolutional Neural Networks, Recurrent Neural Networks).

* Model Training and Validation:

Train each selected model using historical traffic data and validate their performance using appropriate metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), etc.

Tune hyperparameters of the models to optimize their performance.

* Ensemble Building:

Develop an ensemble method to combine predictions from multiple models. Ensemble methods can include:

Use this feedback to update the models and fine-tune the system parameters.

* Deployment and Scalability:

Deploy the hybrid prediction system in a scalable infrastructure capable of handling large volumes of real-time data.

Monitor system performance and scalability to ensure smooth operation under varying loads.

The problem of learning from unstable data has increased with the continued expansion of data availability. This is related to the performance of learning algorithms. The nature of the problem is about efficient transformation. Standard algorithms have deviated to methods such as wrong cost analysis due to even distribution policies. This is a recurring problem. If learning from imbalanced datasets cannot be achieved, prediction accuracy will not increase. In this respect, both internal and external imbalances create problems [18]. Hybrid approaches to this issue are gaining increasing popularity. The minority class may be more important in data mining than the non-minority class because it contains very important and useful information and can affect the general information [1].

1. SMOTE SMOTE is an approach where the Minority class is instantiated by creating synthetic instances. Oversampling is done by getting a sample of each minority class and presenting synthetic samples along line segments joining any of the nearest neighbors [2]. In order to reduce the problems caused by oversampling, it was deemed appropriate in SMOTE to create new minority samples instead of weighting the data points. Interpolation is made between neighboring minority class instances. In this way, it focuses on the concept of feature space [3]
2. Random Forest classifier : This method is preferred to use in the classification and regression analysis that based on creating more than one decision trees. This method stands up to “ensemble learning”. In this learning method, a vast number of decision tree is created with the selection more than one subset in the dataset is called as “bagging method”. Especially random forest is constituted with created a vast number of decision tree and these are combined in order to obtain decision tree which is giving the best results as a result of the classification process of the data is carried out [4].
3. Hyperparameters are settings that control the learning process and model configuration before training begins. Unlike model parameters, which are learned from the data during training (e.g., weights in neural networks or split thresholds in decision trees), hyperparameters are specified by the researcher and govern aspects of the algorithm, like its complexity and capacity to learn. Proper tuning of hyperparameters can significantly impact model performance and generalizability.

After training the data using various machine learning methods, several evaluation metrics are generated to assess the performance, success, and overall effectiveness of the models. These metrics are typically derived from the confusion matrix, which is a summary table used to evaluate the results of classification tasks [5]. The key evaluation metrics that help measure the power and success of the models include accuracy, precision, recall, specificity, and F1 score [6-7]. Each metric provides a unique insight into how well the model is performing, especially in handling different classes in the dataset.

Accuracy

Accuracy measures the proportion of correctly classified data points out of the

total number of data points. It is a general indicator of how well the model is performing. Essentially, it tells you how many instances were correctly identified by the model across both classes. However, accuracy can be misleading if the dataset is imbalanced, as it does not account for the model’s ability to identify both classes.[8] The formula for accuracy is:

Accuracy = TP + TN / TP + FN + FP + TN Where:

* + TP = True Positive
  + TN = True Negative
  + FP = False Positive
  + FN = False Negative

Precision

Precision is the ratio of data points classified as positive that are actually positive. This metric is particularly useful when the cost of false positives is high, as it indicates how many of the predicted positives were truly positive. It helps to determine the model's reliability in identifying positive instances. Precision is calculated using the formula:

Precision = TP/ TP + FP Where:

* + TP = True Positive
  + FP = False Positive

Recall

Recall (also known as sensitivity or true positive rate) measures the proportion of actual positives that are correctly predicted as positive. It is a critical metric

when the cost of false negatives is high, as it reflects the model’s ability to identify all relevant positive instances. Recall is especially useful when identifying rare events, such as fraud detection or disease diagnosis.[8] It is calculated as:

Recall = TP/ TP +FN Where:

* + TP = True Positive
  + FN = False Negative

Specificity

Specificity (also known as the true negative rate) evaluates how well the model identifies negative instances. It is the ratio of actual negatives that are correctly predicted as negative. Specificity is important when you want to ensure that the model does not incorrectly classify negative instances as positive. It is calculated with the following formula:

Specificity = TN/ TN =FP Where:

* + TN = True Negative
  + FP = False Positive

F1 Score

The F1 Score is a harmonic mean of precision and recall, and it provides a balance between the two metrics. It is particularly useful when you need to balance the trade-off between precision and recall, especially in cases where the classes are imbalanced. A high F1 score indicates both high precision and recall, making it an important metric for evaluating model performance in situations where both false positives and false negatives have significant consequences.[8]

The formula for the F1 score is:

F1 Score = 2 \* Precision \* Recall/ Precision + Recall

## Machine Learning

Machine Learning (ML) techniques are considered statistical models that are utilized to make classifications and predictions based on the data provided [24]. ML is an area of AI that focuses on the development of prediction algorithms depending on the fair discovery of patterns within huge datasets and without being designed specifically for a particular job [25]. ML models are classified into three categories according to the learning techniques they employ: supervised learning, unsupervised learning, and reinforced learning (RL). In addition, ML algorithms might be further subdivided into several subgroups depending on distinct learning approaches

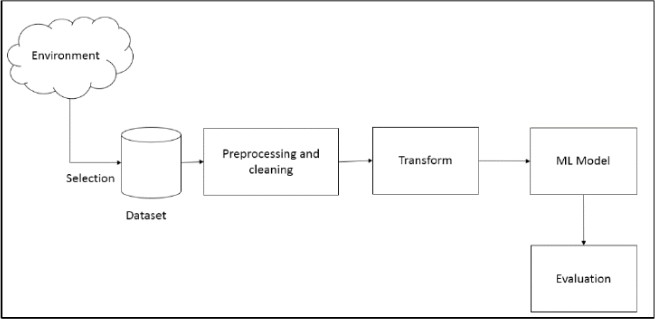


Fig.2 Machine learning

### Supervised learning

In the tasks that depend on supervised learning, a labeled dataset known as feature vectors and their corresponding predicted output labels are supplied to the model. The objective of these models is to create an inference function that maps feature vectors into output labels. When the ML model training is complete, it can make predictions based on new data. Continuous or discrete predictions can be generated using supervised learning algorithms [24]. Support Vector Machine (SMV), KNN,

Logistic Regression, Linear Regression, Decision Trees (DT), Random Forests (RF), and Naive Bayes are examples of supervised learning approaches [25].

### Logistic regression

Logistic regression is a supervised learning approach used to differentiate between two or more groups [27]. It provides, in terms of 0 and 1, the likelihood that an event will occur based on the values of the input variables (i.e., it gives the binomial outcome). For instance, predicting whether or not an e-mail is categorized as spam is a binomial result of Logistic Regression. In addition, Logistic Regression can produce multinomial outcomes, such as predicting the preferred cuisine (Chinese, Italian, Mexican, etc.). In addition, Logistic Regression can produce ordinal results, such as rating a product from 1 to 5.

Therefore, Logistic Regression is concerned with categorical target variable prediction [33]. Logistic Regression provides several benefits, including ease of implementation, computational efficiency, training efficiency, and regularization simplicity. In Logistic Regression, input features do not require scaling. In addition, Logistic Regression is immune to data noise and multi-collinearity. Logistic Regression, on the other hand, is unsuitable for nonlinear problems since its decision surface is linear, and sensitive to overfitting, and all independent variables must be recognized for it to work successfully [33]

### Linear Regression.

Regression is an example of a supervised learning technique in which the value of the output variable is decided by the values of the input variable and the utilized labeled datasets. Regression can be used to model and predict continuous variables. In linear regression, an attempt is made to fit a straight hyperplane to the data set if the relationship between the variables of a dataset is linear [33]. Linear Regression is calculated according to (2) [32]:

F(x) = mx + b + e

where x is the independent variable, F(x) is the dependent variable, m is the slope of the line, b is the y-intercept, and e is the error term.

The best prediction accuracy may be achieved using the Linear Regression algorithm if the following steps are followed to prepare the training data [32]:

* + Assume that the dependent and independent variables are linear, i.e., apply any of the available data transformation techniques to make the data linear.
  + Remove noisy data and outliers using a technique for cleaning data.
  + To minimize overfitting, do pair-wise correlation and exclude the most linked variables.
  + Apply Gaussian distribution to the training data to generate more accurate predictions.
  + Rescale inputs to improve the reliability of the prediction.

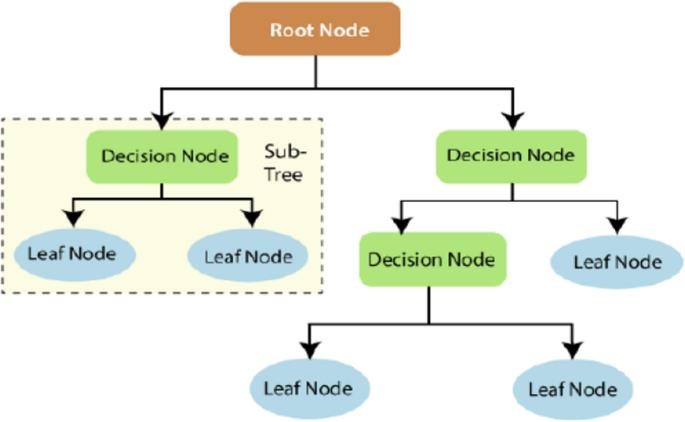
From the above discussion, it is clear that the Linear Regression algorithm is straightforward to comprehend. In addition, the ideal linear relationship between dependent and independent variables is demonstrated. In contrast, Linear Regression can only predict the numeric output. It is inappropriate for nonlinear data and highly sensitive to outliers. Also, data must be independent [32].

### Decision trees

Classifier-generating systems are one of the most popular strategies in data mining [34]. In data mining, classification algorithms are capable of processing vast quantities of data. It can be used to create assumptions about categorical class names, categorize information based on training sets and class labels, and classify newly accessible data [35].

DTs are one of the powerful approaches utilized in numerous domains, including ML, image processing, and pattern recognition [36]. DT is a model

that sequentially as well as cohesively combines a set of basic tests in which a numerical characteristic is compared with a threshold value [37]. In addition, DT is a common classification model in Data Mining [38]. Every tree is composed of nodes and branches. Each node represents an attribute inside a group to be categorized, and each branch provides a possible value for the node [39]. Figure 7 illustrates the structure of DT.



### Fig 3 DT structure

1. **Random Forest**

RF is an ensemble classifier since it employs many DTs to compensate for the shortcomings of a single DT [45,46,47,48,49]. The 'vote' of all trees is utilized to determine the final class for each unknown. This eliminates the possibility that a single tree may not be ideal. Therefore, adding numerous trees should result in a global optimum [50]. For the formation of each tree in the "forest", the bootstrap approach is used for resampling. In addition, on each node split, a subset of features is randomly selected, and the split variable selection occurs over this subset.

The projected value for classification is the majority vote, and the average, for regressions [51,52,53,54]. On RF models, there are two parameters for tuning:

mtry, which is the number of features that are randomly picked to consider in each split; and ntree, which is the trees count in the model. The mtry parameter has a tradeoff: large values increase the correlation among trees but improve the accuracy of each tree [51]. The unused elements are called the Out of Bag (OOB) samples, which can be employed for validation in this case, each tree predicts over its OOB samples, and the final result is an average over the outcomes of the trees [55].

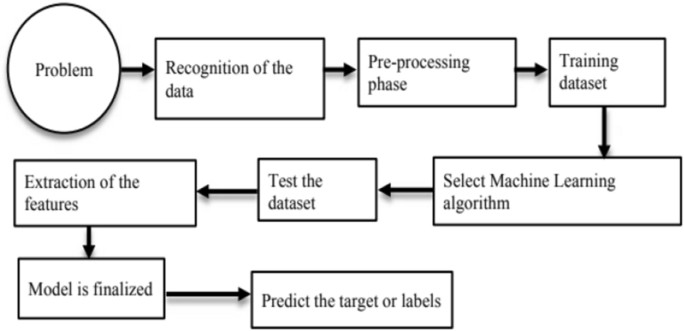
There are two options for estimating the relevance of each variable and ranking them accordingly. The initial choice is to utilize the OOB samples. In this option, the accuracy is calculated over the set of each tree and its corresponding OOB samples, a variable is randomly permuted among samples, and the accuracy is recalculated on the new set. Applying this to the set of all trees and average for each variable yields a metric for comparing relevance. This metric for comparison is known as the Permutation Importance Index (PIM) or Variable Importance Measure (VIM). The alternative is to calculate the split improvement for each tree and node using a measure (e.g., the Gini Index) and use these values to compare the significance of the variables [55].

RFs offer high flexibility and prediction rates. It also does not overfit the data when the number of trees is considered. Alternatively, a graphical representation is not feasible as in DTs [55].

### Unsupervised learning

In unsupervised learning, there is no output label information contained in the dataset. The purpose of these models is to infer the link between data and/or to uncover hidden variables [25]. These strategies are mostly used to reduce the size of a dataset by extracting key features. Reducing the number of features helps prevent problems such as high computational cost and multi-collinearity [57]. Figure 8 depicts unsupervised learning, in which the machine guesses the result according to

past experiences and learns from information previously provided to anticipate the real-valued outcome.



**Fig 4 Unsupervised learning workflow**

## Survey Methodology

The articles reviewed in this paper have been published in high-quality conferences and journals of IEEE, Elsevier, Springer, and IOP publishing. Machine learning, deep learning, Employee promotion prediction, Employee statistics, Enterprise resource planning and Employee evaluation using imbalanced dataset are some of the search terms used to find these articles. The articles examined in this survey are directly relevant to the application of ML approaches in prediction. Both empirical and literature reviews on the abovementioned subjects were considered for this work.

This survey compares various forecasting and evaluation techniques. This study provides a detailed discussion of the approaches and algorithms which are utilized for predictions, performance measurements, and tools used for these procedures.

Statistical methods, AI, and data mining techniques have been widely employed

recently to evaluate data and anticipate. Previous findings demonstrated that no single technology could evaluate enormous datasets only by itself. Therefore, according to the data structure and its volume, the proper technology must be applied to extract the best insight from the collected data.

## Literature Selection and Scope Definition

To ensure a comprehensive and balanced review, a systematic literature selection process was implemented. This process involved searching multiple databases such as IEEE Xplore, ACM Digital Library, Google Scholar, and SpringerLink, using a range of keywords related to employee evaluation and promotion through machine learning. Keywords included terms like "employee performance prediction," "machine learning in human resources," "promotion prediction models," and "bias in HR analytics."

Inclusion and Exclusion Criteria

To filter the vast literature, the following criteria were applied:

* Inclusion: Studies published in the last ten years, peer-reviewed articles, works involving machine learning applications specific to employee performance, and studies that offered both technical and interpretive insights into the models used.
* Exclusion: Studies focused solely on traditional statistical methods without machine learning, articles not offering empirical results, and papers with limited data on model evaluation metrics.

Each study was reviewed for relevance, ensuring that the selected works covered a range of machine learning applications from different sectors, including IT, finance, and healthcare.

## Feature Selection and Engineering

Feature selection is crucial in developing machine learning models that are both

accurate and fair. This review examines how different studies approach feature selection and engineering, focusing on commonly used attributes like:

* Demographic Data: Age, gender, and educational background, while useful, are sensitive features that can introduce bias. The review examines strategies for responsibly including or excluding these features.
* Performance Metrics: Studies commonly use annual performance ratings, attendance records, and project evaluations. The review discusses how metrics vary by industry and position, affecting the adaptability of models across different organizational contexts.
* Skill and Competency Data: Skills, certifications, and professional development activities are highly predictive of promotion potential. Feature engineering techniques, such as one-hot encoding for categorical data and normalization for continuous variables, are explored to standardize input data.

## Model Evaluation and Validation Approaches

A critical part of this review involves examining how studies evaluate the effectiveness and generalizability of machine learning models in promotion prediction.

Evaluation Metrics

Common evaluation metrics include accuracy, precision, recall, and F1-score. Studies are analyzed to understand how different metrics reflect model performance in a promotion context:

* Accuracy: This basic metric, while indicative of overall model performance, may not adequately capture the nuances in promotion decisions, especially when datasets are imbalanced.
* Precision and Recall: In HR applications, recall is often crucial, as a high recall ensures fewer eligible employees are overlooked. Studies using recall as a primary metric are reviewed for insights into promotion fairness.
* AUC-ROC: The area under the receiver operating characteristic curve is discussed for its utility in balancing false positive and false negative rates, which is crucial in employee evaluations.

Validation Techniques

Cross-validation methods like k-fold cross-validation and train-test splits are reviewed. Studies implementing robust validation practices are highlighted for their reliability in real-world HR applications. Techniques for handling imbalanced data, such as oversampling and SMOTE (Synthetic Minority Over-sampling Technique), are also reviewed to understand how they help ensure model fairness.

## Ethical Considerations and Bias Mitigation

Addressing bias is essential in using machine learning for employee evaluation. This review examines bias detection and mitigation techniques as discussed in selected studies.

* Bias Detection Techniques

The review covers studies that implement statistical bias detection methods, such as disparate impact analysis, to evaluate if certain demographic groups face unfair disadvantages. Techniques for de-biasing, including re-weighting and data augmentation, are explored.

* Fairness-aware Algorithms

Fairness-aware machine learning, where algorithms are explicitly adjusted to minimize bias, is a growing area. The review examines algorithms designed to treat sensitive attributes neutrally, reducing potential biases that could affect fairness in promotion predictions.

## 3.9. Integrating Machine Learning with HR Processes

The practical implementation of machine learning in HR departments is a focal point of this review. Studies addressing the technical and organizational challenges

involved in incorporating machine learning systems into HR workflows are reviewed, including

* Data Privacy and Security: The review discusses studies focusing on secure data handling practices essential for sensitive employee information.
* Real-time Data Processing: The benefits and challenges of real-time evaluation models in employee assessment are examined, particularly for organizations with dynamic performance evaluation needs.
* Interpretability and Transparency: Studies that highlight the importance of interpretable models in HR are reviewed, as interpretability helps ensure HR teams and employees trust the promotion recommendations provided by machine learning systems.

# CHAPTER 4

**RESULT ANALYSIS AND VALIDATION**

In this chapter, the results of ML classifiers on various evaluation requirements, such as accuracy, recall, and F-measure, are addressed. Examples of these evaluation constraints include: In addition to this, the performance of machine learning classification models is assessed using the dataset, which includes information on heart disease. k-NN did not do very well, although RBF, NB, and LVQ fared better than the other classifiers when compared to their overall performance. As can be seen in Table [7](https://www.nature.com/articles/s41598-023-40717-1#Tab7), the most important assessment criteria that were taken into consideration in this study to evaluate the performance of the ML classifier are the sensitivity, accuracy, specificity, recall, precision, and F-measure ratings. As a consequence of this, the specificity and sensitivity of the targeted class are calculated in order to evaluate the accuracy with which the given method is projected to perform. The "TP" (true positive), "TN" (true negative), "FN" (false negative), and "FP" (false positive) rates are used to compute the accuracy, precision, recall, and F measure in ML. These measures are determined by the quality of the data. Each correct positive and correct negative prediction is further subdivided into correct positive and correct negative forecasts. Every model correctly predicted the TP, TN, FP, and FN outcomes. The letters TP stand for diseased, which means infected. FN is an illness that is not believed to be related to cardiovascular disease. The FP illness is one that has been predicted but has never been seen in humans. In the actual world, TN does not exist as a disease, and this is not anticipated to change in the foreseeable future. The performance of ML approaches in terms of accuracy is listed in Table [7](https://www.nature.com/articles/s41598-023-40717-1#Tab7). By associating the performances of these classifiers, we observed that radial basis functions, naive bayes, and learning vector quantization, as well as their relatedness to other ML classifiers, led these models to achieve almost 90.06%, 94.16%, and 98.07% accuracy, respectively.

| **Classification techniques** | **Performance metric parameters** | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **Precision** | **Accuracy** | **Sensitivity** | **Specificity** | **Recall** | **F-measure** |
| Random forest | 88.07 | 88.78 | 87.91 | 87.1 | 85.31 | 87.89 |
| **Proposed learning vector quantization** | **98.07** | **98.78** | **97.91** | **97.1** | **95.31** | **97.89** |

Table 3

| **Classification techniques** | **Performance metric parameters** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Precision** | **Accuracy** | **Sensitivity** | **Specificity** | **Recall** | **F-measure** |
| Decision tree | 89.07 | 89.78 | 88.91 | 88.1 | 86.31 | 88.89 |
| **Proposed learning vector quantization** | **98.07** | **98.78** | **97.91** | **97.1** | **95.31** | **97.89** |

Table 4

# CHAPTER 5 CONCLUSION AND FUTURE SCOPE

In this study, machine learning classifiers are utilised to determine whether or not a patient has heart problems. The dataset was taken from the repository at UCI. Following data collection, they will go through cleaning and pre-processing steps. Following this step, machine learning models are used for predictive analysis. We investigated the potential of these eight applied machine learning methods for making accurate predictions about cardiac disease. The inclusion criteria for these algorithms are that they be mature, representative, and at the state of the art in their respective fields. We have previously used the Naive Bayes and RBF neural networks, but other scholars have not used them on the UCI cardiovascular disease dataset. As a result, we have achieved a higher level of accuracy than they have, as shown in the table titled "state of the art," which compares our results to those of other researchers. The final findings demonstrate that when the learning machine classifiers were put to use, the Naive Bayes and RBF neural networks achieved an accuracy of 94.78% when attempting to forecast the presence of coronary cardiovascular disease. However, the Learning Vector Quantization method achieved the highest categorization accuracy of 98.78%, with a specificity of 97.1% and sensitivity of 97.91%, a precision of 98.07% and 95.31%, and 97.89% F1score and F-measure values, respectively.

In the future, our research aims to further enhance the reliability of our conclusions by incorporating additional datasets. We will explore the use of metaheuristic techniques and nature-inspired algorithms to optimize the parameters of machine learning classifiers and deep learning methods. This optimization process will enable us to more effectively evaluate the presence of heart disease across various heart disease-related datasets. Additionally, we will focus on improving the accuracy of existing algorithms to enhance their performance in detecting heart disease. By leveraging these advancements, we aim to provide more robust and accurate methods for the diagnosis and evaluation of heart disease.

# REFERENCES

1. Goundar, S., Nayyar, A., Maharaj, M., Ratnam, K. and Prasad, S. (2021) How Artificial Intelligence is Transforming the ERP Systems. Enterprise Systems and Technological Convergence: Research and Practice, South Pacific, Information Age Publishers, 1-15.
2. Chopra, R., Sawant, L., Kodi, D. and Terkar, R. (2022) Utilization of ERP Systems in Manufacturing Industry for Productivity Improvement. Sciencedirect, 62, 1238-1245.

https://doi.org/10.1016/j.matpr.2022.04.529

1. Chaoliang, H. and Qi, Z. (2021) Optimization of Supply Chain Efficiency Management Based on Machine Learning and Neural Network. Neural Computing and Applications, 33, 1-10.
2. Yan, A.Y., Suraksha, G.B., Tana, C.L.C. and Klaus, S.B. (2022) Integrating Machine Learning, Modularity and Supply Chain Integration for Branding 4.0. Sciencedirect, 104, 136-149.

https://doi.org/10.1016/j.indmarman.2022.04.013

1. Iota, K.N. and Robert, P. (2021) Artificial Intelligence, Machine Learning and Big Data in Finance. OECD.

https://[www.oecd.org/finance/financial-markets/Artificial-intelligence-](http://www.oecd.org/finance/financial-markets/Artificial-intelligence-) machine-learning-big-data-in-finance.pdf

1. Dhivya, R.S. and Sujatha, P. (2023) Employee Performance Prediction for Workforce Planning using Ensemble Hybrid Model. In 2023 10th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi.
2. Alqahtani, F.A. and Almaleh, A. (2022) Analysis and Prediction of Employee Promotions Using Machine Learning. 2022 5th International Conference on Data Science and Information Technology (DSIT), Shanghai, 22-24 July 2022.

https://doi.org/10.1109/DSIT55514.2022.9943959

1. Alsobaey, T.M. and Al-Alawi, A.I. (2023) The Use of Data Mining Techniques to Predict Employee Performance: A Literature Review. In 2023 International Conference On Cyber Management and Engineering (CyMaEn), Bangkok, 26-27 January 2023.

https://doi.org/10.1109/CyMaEn57228.2023.10050937

1. Kimberling, E. (2024) The Benefits of ERP for Human Resources Management.

https://[www.thirdstage-consulting.com/erp-for-human-resources-](http://www.thirdstage-consulting.com/erp-for-human-resources-) management/

1. Zambuk, F.U. and Abdullahi, A. (2023) Salary Prediction Model Using Principal Component Analysis and Deep Neural Network Algorithm. International Journal of Innovative Science and Research Technology, 8, 10.
2. Shaprapawad, S. and Borugadda, P. (2024) Education and Society Salary Prediction Using Machine Learning Regression Algorithms. Education and Society, 48, 88-94.
3. Ashwin, J. (2021) What Is a Pythonista and How to Become One. https://pythonistaplanet.com/pythonista/
4. Nasr, M., Shaaban, E. and Samir, A. (2019) A Proposed Model for Predicting Employees’ Performance Using Data Mining Techniques: Egyptian Case Study. International Journal of Computer Science and Information Security (IJCSIS), 17, 1-10.
5. M. Dağdeviren, Bulanik analitik hiyerarşi prosesi ile personel seçimi ve bir uygulama. Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi, 22(4), 2007.
6. A. Kibria, A. S. Kazi, A. A. Banbhan, A. K, Shahani & I. Junejo, Investigating linkages of performance appraisal, employee promotion

and job satisfaction with employee performance in banking sector of Pakistan. Journal of Contemporary Issues in Business and Government Vol, 27(2), 2021.

1. R. B. R. Mohammad and M., H., K. Daisy, Perceived Fairness of Performance Appraisal, Promotion Opportunity and Nurses Turnover Intention: The Role of Organizational Commitment, Asian Social Science; Vol. 11, No. 9, 2015.
2. https://[www.kaggle.com/code/flaviocavalcante/employees-](http://www.kaggle.com/code/flaviocavalcante/employees-) evaluation-for-promotion-eda- ml/data?select=employee\_promotion.csv
3. H. He and E. A. Garcia, “Learning from Imbalanced Data,” IEEE Trans. Knowledge and Data Engineering, vol. 21, issue 9, pp. 1263-

1284, 2009

1. N. V. Chawla, SMOTE: Synthetic Minority Over-sampling Technique, 2002.
2. A. Fer andez, SMOTE for Learning from Imbalanced Data: Progress andChallenges, Marking the 15-year Anniversary, 2018.
3. B. Krawczyk, Learning from imbalanced data: open challenges and future directions, 2016.
4. F. Rayhan, S. Ahmed, A. Mahbub, R. Jani, S. Shatabda & D.M. Farid,.

Cusboost: cluster-based under-sampling with boosting for imbalanced classification. In 2017 2nd International Conference on Computational Systems and Information Technology for Sustainable Solution

(CSITSS) (pp. 1-5). IEEE, 2017.

1. M. Somvanshi, A review of machine learning techniques using decision tree and support vector machine. IEEE, 2017.
2. S. Agatonovic-Kustrin, Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. Elsevier, 2000.
3. T. Chen, XGBoost: A Scalable Tree Boosting System. ACM, 2016.
4. T. M. Oshiro, How Many Trees in a Random Forest?, 2012.
5. S.Umadevi, & D. Marseline, A Survey on Data Mining Classification Algorithms. International Conference on Signal Processing and Communication, 2017, 64-268.
6. I. A. Zriqat, , A. M. Altamimi & M. Azzeh, A Comparative Study for Predicting Heart Diseases Using Data Mining Classification Methods, 2017, 868-879