

## Retraction

# Retracted: Analyzing the Performance of Machine Learning Techniques in Disease Prediction

### Journal of Food Quality

Received 19 December 2023; Accepted 19 December 2023; Published 20 December 2023

Copyright © 2023 Journal of Food Quality. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

- (1) Discrepancies in scope
- (2) Discrepancies in the description of the research reported
- (3) Discrepancies between the availability of data and the research described
- (4) Inappropriate citations
- (5) Incoherent, meaningless and/or irrelevant content included in the article
- (6) Manipulated or compromised peer review

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

### References

- [1] K. Phasinam, T. Mondal, D. Novaliendry, C. -H. Yang, C. Dutta, and M. Shabaz, "Analyzing the Performance of Machine Learning Techniques in Disease Prediction," *Journal of Food Quality*, vol. 2022, Article ID 7529472, 9 pages, 2022.

## Research Article

# Analyzing the Performance of Machine Learning Techniques in Disease Prediction

Khongdet Phasinam <sup>1</sup>, Tamal Mondal <sup>2</sup>, Dony Novaliendry <sup>3</sup>, Cheng-Hong Yang <sup>3</sup>,  
Chiranjit Dutta <sup>4</sup> and Mohammad Shabaz <sup>5</sup>

<sup>1</sup>Faculty of Food and Agricultural Technology, Pibulsongkram Rajabhat University, Phitsanulok, Thailand

<sup>2</sup>Department of Botany, Hiralal Mazumdar Memorial College for Women, Dakshineswar, Kolkata, India

<sup>3</sup>Department of Electronic Engineering, National Kaohsiung University of Science and Technology, Kaohsiung, Taiwan

<sup>4</sup>Faculty of Engineering and Technology, SRM Institute of Science and Technology, NCR Campus, Sadabad Jakhaiya, Uttar Pradesh, India

<sup>5</sup>Arba Minch University, Arba Minch, Ethiopia

Correspondence should be addressed to Mohammad Shabaz; [mohammad.shabaz@amu.edu.et](mailto:mohammad.shabaz@amu.edu.et)

Received 4 January 2022; Revised 29 January 2022; Accepted 8 February 2022; Published 3 March 2022

Academic Editor: Rijwan Khan

Copyright © 2022 Khongdet Phasinam et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The history of data stored can be used to forecast potential patterns and help companies make competitive decisions to increase their success and benefits. Many analysts look at healthcare sector data to identify and forecast illnesses in order to benefit patients and physicians in a variety of ways. This study is concerned with the diagnosis and estimation of heart disease. Heart disease is one of the most dangerous illnesses for humans, leading to death all over the world. Many different groups of researchers have used knowledge exploration methods in diverse fields to forecast heart disease and have shown acceptable degrees of precision. There were no real-time methods for analyzing and forecasting heart disease in its early stages. For the prediction of heart disease, decision trees are used to analyze various training and evaluation datasets. Classification algorithms such as Naive Bayes, ID3, C4.5, and SVM are being investigated. The UCI machinery heart disease data set is used in experimental studies.

## 1. Introduction

Every year, heart-related disorders [1] claim the lives of around a million individuals, making this the major cause. One out of every three deaths is caused by heart disease, and nearly half of them occur abruptly and without warning. The only indicator of heart disease is sudden death. A heart attack occurs when the internal organs of the heart or muscles begin to fail even though the person survives for a short length of time. This will cause serious abnormalities in the heart's components, resulting in health difficulties such as an increased risk of sudden cardiac arrest.

According to current predictions [2], India will have more heart disease encumbrances than any other country. In India, one out of every five deaths is caused by cardiac

disorders. Over the next three years, one out of every three deaths will be fatal.

There are several forms of cardiac disease, each of which damages various internal organs of the heart. As a result, any sort of heart illness may come within the category of cardiovascular illnesses [3], and some of the heart-related disorders are addressed here. Coronary heart disease is also known as coronary artery disease (CAD), and it is the most common kind of heart disease worldwide. It is a condition caused by fat accumulation in the blood veins and capillaries. It also inhibits blood from flowing into the veins and capillaries of the heart, resulting in an insufficient supply of oxygen and blood to the heart's internal organs.

Angina pectoris, often known as angina, is a medical term that refers to chest discomfort caused by a lack of blood

flow to the heart. It is a preventative sign of having cardiac issues. This sort of discomfort will last a few seconds or minutes. Congestive heart failure is defined as a condition in which the heart fails to circulate enough blood to the rest of the body's organs. Cardiomyopathy refers to the wearing out of the heart muscles or, in certain cases, a change in the appearance of the form of the muscle as a result of an inadequate cardiac pump. Cardiomyopathy can be caused by a variety of factors, including viral infections, alcohol abuse, and hypertension. CAD [4] is also known as a congenital defect within the heart, which refers to the formation of an irregular cardiac structure as a result of a fault in the internal elements of the heart. Furthermore, it is the type of congenital issue that occurs in newborns at birth. Arrhythmia is a form of cardiac disease characterized by the recurrent recurrence of heartbeats. Depending on the severity of the problem, the frequency of heartbeats might be fast, sluggish, or even imbalanced. These irregular heartbeats are often caused by a sort of faulty link in the heart's electrical circuit system.

Myocarditis is defined as an infection in the heart muscles caused by virus-related, bacterial, and fungal contamination that causes heart muscle pain. It is an uncommon condition with certain symptoms such as joint discomfort, limb swelling, or fever that may or may not be related to the heart. When compared to previous decades, heart disease has increased significantly and has become the leading cause of mortality. It is extremely difficult for healthcare practitioners to recognize rapidly and precisely [5]. As a result, it is critical to use computer knowledge in this analysis to assist healthcare professionals in detecting cancer in the early stages with increased accuracy. The activity of computationally extracting hidden and unknown knowledge patterns from huge data repositories of known information in order to develop prediction models is referred to as data mining [6]. It is a step toward discovering knowledge by utilizing database history. This found knowledge may be used by the organization's senior management to make strategic decisions to improve performance and profit. Data mining is a method of evaluating data from many perspectives and summarizing it into useful knowledge in a short period of time. Data mining approaches have been applied in a variety of applications. It appears to be the most efficient method of constructing healthcare support systems.

The major focus of the study is involved in analysis of the performance evaluation of machine learning techniques in disease prediction. The researchers aim to measure the overall performance of the different ML approaches so as to enhance the predictability of the model.

The novelty of the study lies in making an extensive and comparative analysis between the different ML models, and the researchers have used major determinants in order to make a better comparison between the models, such as accuracy results of classification algorithms and error rate results of classification algorithms; these aspects are intended to measure the effectiveness of disease prediction.

## 2. Research Gap

It has been noted that the previous studies were confined in analyzing the impact of ML approaches in the disease detection; however, this work is more confined in making a comparative analysis between the various ML approaches in the support in detection of diseases in the healthcare industry. It has been noted that the data mining employs a set of algorithms to identify hidden or unknown patterns in data, assisting in the transformation of massive amounts of data into valuable information for decision making. In the medical profession, there is a vast amount of data available. Despite the lack of effective healthcare analysis tools for identifying hidden information in patient's data, numerous data mining methods and methodologies are available to evaluate the massive volume of healthcare data in order to forecast numerous life-threatening diseases such as cancer, diabetes, liver illnesses, and heart disorders.

Hence, this research is more focused on understanding the critical aspects of the key ML approaches in enhancing the disease prediction for better healthcare services. Many researchers have developed data mining algorithms to identify and forecast these life-threatening illnesses in order to save people's lives. Existing studies and data mining tools forecast illnesses only broadly and not exactly.

## 3. Literature Survey

It has been stated by Khourdifi and Bahaj [7] that the proposed optimization tends to enhance the overall accuracy of the medical information. Based on the method has stated that the classification of accuracy has been increased and resulted in better performance. Furthermore, the models of KNN at 99.65% and RF at 99.6% are considered to have precise score, which are being achieved by the stated model. Also, Benjamin Fredrick David [8] has mentioned in the algorithm that is supposed to provide higher precision with the overall classification as being mentioned in the traditional system. The proposed machine learning approaches has enabled the understanding and predicting of heart diseases quickly. Furthermore, the exactness which has been traced using the random model has resulted in computing the coronary disease at the earliest. Shahi and Gurm [9] have mentioned that the data mining tools have been of much assistance in the prediction capabilities of heart diseases, and the researchers have mentioned that the application of critical algorithms has programmed in the determination of illness through critical analysis. The model has used different alternatives like Naive Bayesian principal Support vector machine and other tools. It has been further noted that the SVM is more reasonable and enables in offering better analysis of the sickness and support the practitioners to take profound action.

Moreover, few investigations have been carried out to analyze the clinical dataset which involves applying different classifiers and selections, as proposed by [10]. It has been noted that the model has applied different classification related to the heart disease where the accuracy is highly needed. Furthermore, as stated by Malav [10], the model is highly useful in predicting the coronary heart diseases. The implementation of the K means and neural networks has

supported in gathering and analyzing the pattern, thereby enabling in detecting the heart disease with an accuracy of nearly 97%.

Khateeb and Usman [11] used different classification methods for the Cleveland Heart Database dataset, such as KNN, Naive Bayesian, decision tree method, and bagging method. In this work, the selected resources are based on domain knowledge and nonspecific selection algorithms. This methodology increased the accuracy of the methods used by KNN and Naive Bayes but reduced the use of decision trees and exchange models. Sujatha [12] suggested evaluating the performance of ML algorithms according to warning accuracy, which includes switching between true positive and negative percentages and recall and accuracy. The study assesses the F1 score, which is the harmonic mean of recall and accuracy. The ROC curve is shown as a graph of the coverage between false negatives and false positives. Gomathi proposed a method for predicting many diseases using data recovery algorithms [13]. The proposed method suggests that the number of clinical trials can be reduced through data recovery methods for diseases such as breast cancer, diabetes, heart disease, and so on [14]. Shetty [15] developed a system that takes into account 13 risk factors for input data to predict and diagnose the occurrence of heart disease from the patient's medical data. After reviewing the information in the data set, preprocessing steps are performed such as deleting the information and consolidating the information. Kavitha and Kannan [16] proposed a system that uses principal component analysis in the classification method to detect heart disease, which includes functional extraction.

It was proposed to increase the accuracy of the expectations used by the classifiers and to reduce the computational cost of performing the prediction, minimizing the data dimension. The Naive Bayes algorithm provides an accuracy of 86.41% according to the method proposed by Vembandasamy [17] for the diagnosis of heart disease. The proposed technology uses a data set of 500 patients, with a division rate of 70%, using the Weka software to perform the classification. Dai [18] organized an analytical test and examined healthy habits that contribute to cardiovascular disease using manually classified information. Manual sorting is good or bad and suggests that, with ML rankings, the amount of data trained is 70% and the test data set is 30%. Intelligent optimization algorithms are used by various ML algorithms. The intelligent cardiac prediction system (IHDPS) is a model proposed by Palaniappan [19] using data extraction techniques such as neural networks, naive sinuses, and decision trees.

In recent years, the field of data mining has seen a surge in interest from all angles and disciplines. It is a means of gathering and determining a surprising quantity of intelligence from massive amounts of data. The key objectives for which data mining methods must be used for analysis and prediction are classification challenges for various kinds of applications [20].

The decision tree classification methodology is a widely utilized methodology with processes that are best suited for medicinal analysis. The C4.5 decision tree technique is a popular and successful classifier. The creation of

technologies can examine and infer "big data" sets in a computerized and adaptive manner, while giving accurate and actionable medical evidence, which is a crucial characteristic of precise treatment decisions. The goal of this study was to establish algorithms for the detection of cardiac illness and the prediction of mortality risk, and to see if such models outperformed conventional assessments.

ML systems can surpass conventional models in illness categorization and prediction. Such an approach has the potential to be more accurate. Methods for identifying patients at risk for diseases are easily automated. There are no good risk prediction algorithms, and some existing ones perform quite poorly. Future research should seek to test, automate, and validate the models in order to reduce the prevalence of undetected diseases and the burden of hostile results caused by delays in preventative actions [21].

Association rule mining is a popular mining approach for generating frequent itemsets, and it is especially useful in market basket research. The mining of frequent itemset technique can be used to anticipate the amount of hazard owing to heart-related disorders. A large number of cardiac patient datasets were used to forecast the risk level. In early phases, frequent itemsets were developed and utilized by healthcare professionals to diagnose and determine the amount of risk. This approach may be used to compute risk levels for any medical dataset, and the experimental findings offer an accurate forecast.

The created technology properly analyzes and forecasts the patients' risk level and saves the patient from harm. This approach is being tested on a heart illness database that has many thousands of records of various types of heart disorders. Prediction results are more accurate, motivating, and efficient than existing approaches. They are all the records of people who have various cardiac conditions. The prediction results are inspiring, and the method's efficiency in frequent itemset creation outperforms prior approaches [22].

Decision tree classification is a widely used methodology that is best suited for medical diagnosis. The C4.5 decision tree is a popular and successful classifier for pregnancy data categorization. The C4.5 classification algorithm is used to forecast a woman's risk of pregnancy. Complications during pregnancy have proven to be a big issue for women in today's world, resulting in the loss of both mother and fetus.

The accuracy of the C4.5 classifier's performance is tested. Other classification approaches can be used to analyze pregnancy data, but the C4.5 classifier is utilized because of its power, popularity, and efficiency, as well as the delicate nature of the pregnancy problem. The C4.5 classifier performs better and gives pregnant women an exact amount of risk, ensuring a safe and healthy pregnancy [23].

Classification is used to categorize each item in a batch of data into one of a preset set of classes that are used in larger applications to categorize various types of data. Different categorization approaches were used to compare Irvine datasets from the University of California. All techniques have been estimated based on their accuracy, execution time, and performance. The J48, regression testing, Bayes Net, and Naive Bayes Updatable algorithms were used for evaluation.



The performance of numerous classification approaches on datasets with assessment criteria like accuracy and execution time has been compared. It is believed that the performance of classification techniques varies depending on the dataset. The classifier's performance is affected by the following factors: (i) the dataset, (ii) the number of instances, (iii) attributes, and (iv) attribute types. J48 and Naive Bayes Updatable produced superior results when compared to other data sets. Future research will concentrate on the mix of classification approaches that may be utilized to improve performance [24].

Cardiovascular heart disease has become one of the leading causes of mortality, and early detection is critical. Angiography can provide a correct diagnosis, but it is quite expensive and has several negative effects. Several existing procedures have collected data from patients and utilized several mining algorithms to achieve high accuracy at a low cost and with few downsides. A database including 303 patient records and 54 characteristics was employed. According to the existing medical literature, the traits evaluated in this database are likely indications of CAD.

The datasets are cleaned using a method known as feature creation. To quantify prediction efficiency, the parameters gain and confidence are assessed. When compared to previous techniques, our technique has a higher accuracy rate of 94.8%. The goal for the future is to anticipate the state of particular arteries. It is more vital to diagnose disease-affected patients than it is to diagnose healthy people. To achieve more enhanced and fascinating results, larger datasets, new structures, and broader mining techniques may be used [25].

The difficulties with heart disease were explored for both sexes, utilizing mining rules to determine the cause. Using confidence as a measure, the University of California, Irvine (UCI) Cleveland dataset was compared to the ill and healthy datasets. Males are more likely than females to get cardiac disease. The appropriate property reflecting healthy and unwell circumstances was identified. It should be highlighted that pain and exercise-induced angina predict the incidence of heart disease in both men and women. The resting ECG is a significant aspect, and the flat slope only causes complications in females. According to the findings, men are more prone to CAD than women. Before the onset of menopause, women had no increased risk of heart attack compared to males of the same age [26].

CAD is the most common cause of death in the universe. Finding the interconnected risk concerns and following their progress through time are critical in the prompt avoidance and management of CAD. Using patient information, a method was developed to detect complications and causes of cardiac disorders. Various technologies, such as ML and rule-based systems, were employed. With intriguing datasets, the suggested model achieves a compelling performance. It provides an excellent opportunity to create a system for dealing with the challenging problem of identifying cardiac disease and associated risk factors. The features were examined in a variety of ways. The lessons learned from identifying risk factors have provided experience in creating comparable systems for future eras. The rule-based

approach needs few, if any, training cases. For delimited records, the cost is relatively low. However, this approach requires additional intelligence from professionals and specialists [27].

To create such a system for optimum classification, KNN is integrated with evolutionary algorithms. In complicated contexts, genetic algorithms function effectively and produce the best results. The findings demonstrated that this method improves the accuracy of detecting cardiac problems. An experimental result on seven distinct datasets demonstrates that this strategy is suitable for classification [28]. This prediction model assists medical practitioners in diagnosing cardiac illness with a limited number of variables.

The goal is to predict the existence of cardiac disease with the fewest possible characteristics. Many researches employ 13 characteristics, but our method uses 11 characteristics using Naive Bayes, the bagging algorithm, and the J48 decision tree classifier with no loss of accuracy. With 10-fold cross-validation, an unbiased estimate was achieved. The bagging algorithm rules are simple to understand since they generate human-readable classification rules, and it is one of the most successful mining strategies for diagnosing heart disease. With 0.5 seconds to generate the model, the bagging method achieves 85.03% accuracy. This technique may be used to discover patients as well [29].

This approach to detecting breast cancer involves reducing the number of features to the best amount using information gain and then applying the revised dataset to the adaptive neuro-fuzzy inference system (ANFIS). When compared to other procedures, the acquired accuracy of this methodology is around 98.24%. It is difficult to use information technology to diagnose patients with cardiac disorders. The primary emphasis is on classification accuracy in order to assess the effectiveness of suggested techniques. In the future, classification speed and computing costs will be taken into account for further optimization [30]. The pros and cons of the previous studies have been listed in Table 1.

For each application, mining comprises classification methods, clustering techniques, and frequent pattern mining. As a result of this, the classification approach receives a greater reaction due to its true properties. The main goal is to build an effective classifier for absolute prediction. The advancement of related approaches has demonstrated the importance of ML approaches for massive datasets. There are a variety of ML methods accessible, and the most accurate approaches are picked by taking into account the benefits and limitations of each algorithm. It was presented how to evaluate several classification algorithms utilizing Weka to analyze the performance of several classifiers. The SVM performs admirably. The goal is to use Weka to estimate and test the five carefully selected categorization methods. It was discovered that the best method for using the dataset is an SVM, which has the capacity to significantly improve the existing techniques for use in the medical sector [31].

Many classification algorithms' efficacy in predicting the existence or likelihood of developing heart disease has been evaluated. A review investigation was carried out into 303

TABLE 1: Pros and cons from previous studies

Sl. No.	Authors and methodologies	Pros	Cons
1	Adaptive neuro-fuzzy inference system. Kavitha and Kannan [16]	Enhancing the accuracy and performance	The cost is high when compared with other methods
2	Naive Bayes. Dai et al. [20]	The model applies a specific likelihood of occurrence of the specified task	It is a probability estimation and hence may not be applied for chronic diseases
3	K-nearest neighbour. Ross et al. [26]	One of the most commonly used ML algorithms	There are more advanced models that are currently in place
4	ANN. Kharche et al. [29]	The interconnection can be effectively rewired, and hence it can adapt to the changing environment	The more hidden layers will impact the overall output and interpretation of the data

domains. Logistic regression (LR), artificial neural networks (ANNs), and decision trees (DTs) were evaluated in terms of efficiency. Thirteen patient characteristics were chosen for testing. The performance of categorization algorithms was compared using graphical representations. It was discovered that the ANN produced a greater area in the baseline and curve. The error rate of logistic regression is 0.22, the error rate of an artificial neural network is 0.198, and the error rate of a decision tree is 0.21. It has been discovered that the ANN is the best approach for classifying records since it has the lowest mistake rate and maximum accuracy. The fundamental advantage of NN is that it describes the relationship between the possible signs and warnings. Because NN is a black box, it is difficult to explain and validate in practice. The advantages of NN are its consistent choices and its ability to handle large datasets [32].

#### 4. Methodology

This section presents a framework for prediction of heart disease. As shown in Figure 1, this framework consists of heart disease data set, data preprocessing component, feature extraction component, and classification algorithms.

The study is more confined in measuring the overall performance of the machine language systems and in enabling the prediction of diseases based on the major factors, such as SVM, C4.5, ID3, and Naive Bayes. The researchers intend to apply the descriptive category of analysis so as to measure the overall performance of these algorithms. These algorithms are compared using determinants, such as accuracy results of classification algorithms and error rate results of classification algorithms.

The main classification algorithms used in the study are as follows:

**4.1. The Iterative Dichotomiser 3.** ID3 is a decision tree technique used to categorize instances by examining their values and attributes. The tree is built from the top down, starting with a collection of instances and necessary characteristics [33]. The instances were partitioned at each node of the tree based on the results of the property testing. This is done recursively until the set in a particular subtree is similar to the instances in the similar category, at which point it is marked as a leaf node.

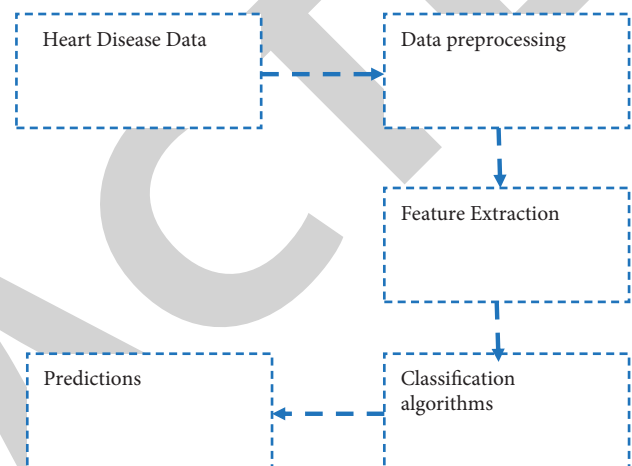


FIGURE 1: A framework for heart disease prediction.

The property to test is chosen based on information-theoretic criteria that will seek to maximize information gain by decreasing entropy at each node [34]. The term “entropy” refers to the amount of entropy as the measure of the dataset’s uncertainty, which ranges between zero and one. The value of 1 of entropy indicates that the data examples belong to the same class labeled homogenous [35]. The group of records is sent into the decision tree as input. These datasets have a higher entropy score, indicating that the data is random.

**4.2. C4.5 Classification Algorithm.** It generates a DT from the provided dataset. C4.5 decision trees are a popular approach to defining information from an ML methodology because they provide a rapid and strong classification method. This algorithm gives several chances for tree trimming. Pruning produces few and easily comprehended results. Tree trimming might be employed to prevent overfitting. The main strategy described in the C4.5 algorithm is an iterative classification that continues until the data is categorized as close or perfectly as possible by creating pure leaf nodes. This strategy achieves the greatest accuracy from the training data but also introduces needless rules that solely assign a certain behavior to the data. If the rules are tested on different data, they will be less effective.

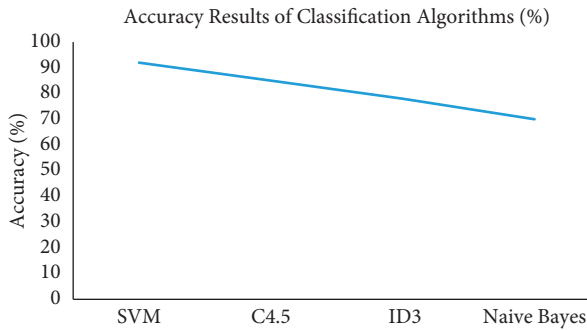


FIGURE 2: Graphical representation of accuracy results of classification algorithms (%).

**4.3. Naive Bayes Classification Algorithm [18].** This approach is mostly used when the dimensionality of the provided input is high. This classification model is concerned with a simple probabilistic model based on the Bayes proposal with robust independence assumptions. It uses the Bayes statement to determine the likelihood of a result occurring by considering the likelihood of an alternative event that has already happened.

The most important aspects of ML are classification and prediction where the world brimming with AI and ML consciousness encompassing nearly everything around. Naive Bayes is a basic yet surprisingly incredible algorithm for predictive examination. It is a classification strategy dependent on Bayes, the hypothesis with suspicion of freedom among predictors. It involves two sections such as Naive and Bayes, in straightforward terms. In the Naive Bayesian method, the classifier will accept when the nearness of a precise feature of a class is random with the presence of another attribute. Regardless of whether these features rely upon one another or upon the presence of different features, these properties autonomously add to the probability that is the reason named after seeing that Naive.

Especially for huge datasets, we can use Naive Bayes model that is easy to build. In probability theory and measurements, Bayes hypothesis, which is then again known as the Bayes law or the Bayes rule, depicts the likelihood of an event dependent on earlier information on the conditions that may be identified with the event. Bayes theorem is a technique to build out the conditional probability. By considering the prior knowledge of a condition that relates to the event, the conditional probability of the event occurrence can be calculated by using Bayes theorem. The conditional likelihood is the likelihood of an event happening given that it has some relationship to at least one different event. Bayes hypothesis is marginally more nuanced more or less, and it gives the real likelihood of an event. From the given information, the Naive Bayes method defines the speculation  $H$  and proof  $E$ , and Bayesian theorem expresses the connection involving the likelihood of the theory before getting the proof  $P(H)$  and the likelihood of the speculation in the wake of getting the proof is given by. This relates the probability of theory before cutting the proof

TABLE 2: Accuracy results of classification algorithms.

Machine learning algorithms	Accuracy (%)
SVM	92
C4.5	85
ID3	78
Naive Bayes	70

that is  $P(H)$  to the likelihood of the speculation in the wake of getting the proof that is  $P(H|E)$ , thus  $P(H)$  is known as the prior probability, while  $P(H|E)$  is known as the posterior probability, and the factor  $P(E|H)$  that narrates both is known as the likelihood ratio. Figure 2 shows the delineation of the Naive Bayes method. Now, using this term Bayes theorem specifies differently as the procedure probability equals the prior probability times the likelihood ratio. [36–38].

**4.4. Support Vector Machine.** The support vector method is now widely used for efficient multidimensional function approximation. The basic idea behind support vector machines (SVM) is to determine a classifier or regression machine that minimizes the training set error. The general method is to fix the training set errors associated with architecture and then to use a method to minimize the generalization errors [39]. The primary advantage of SVM as adaptive models for binary classification and regression is that they provide a classifier with a minimal VC dimension, which implies a low expected probability of test set errors. The SVM constructs a binary classifier from a set of labeled patterns called training examples [40].

The various applications of SVM that are generally used with it are detecting face, normal text in hypertext classification, image classification, and bio-information. The support vector machine is specific to supervised learning ML model learns from the past input data and makes future predictions as output. SVM is a method that looks at data and sorts it into one of two categories. In the larger picture of the ML model and under supervised learning we can see that the support vector fits in under classification deciding what yes-and-no is and there is also a regression version but it is primarily used for classification. SVM exists in both linear and non-linear forms. There are two data sets like train dataset and test dataset, which are involved in SVM.

## 5. Result Analysis

In the experimental study, the UCI machinery heart disease data set [41] is employed. Many previous studies have made use of databases in two ways. Some studies used the entire set of 75 traits, while others focused on the 13 or 14 most important traits for analysis and prediction. To identify the necessary qualities, a thorough investigation is necessary. The dataset under consideration for implementation in this study is described.

The Cleveland database containing 303 records is used as input in Naive Bayes, ID3, C4.5, and SVM algorithms. The

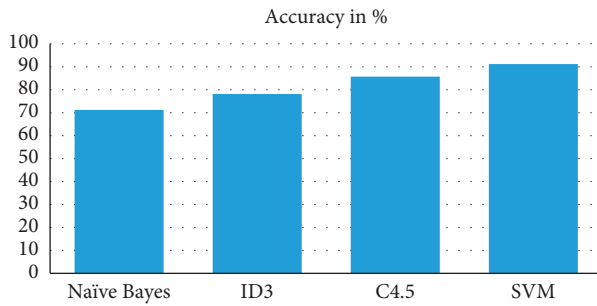


FIGURE 3: Accuracy results of classification algorithms.

TABLE 3: Results of classification algorithms' error rate.

Machine learning algorithms	Accuracy (%)
SVM	8
C4.5	14
ID3	22
Naive Bayes	28

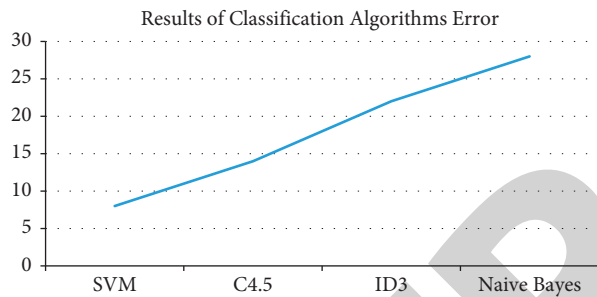


FIGURE 4: Error rate results of classification algorithms.

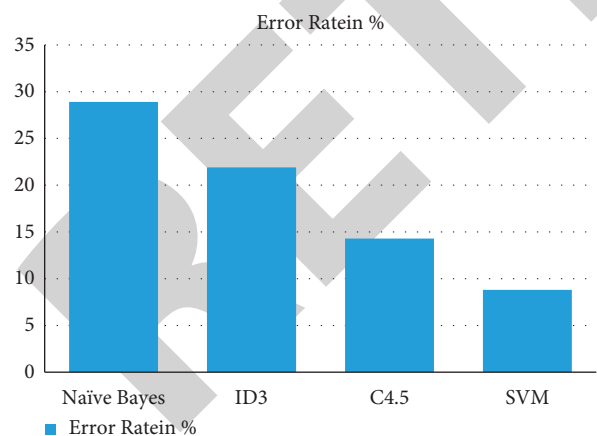


FIGURE 5: Error rate results of classification algorithms.

accuracy and error rate achieved are listed in Table 2 and are graphically shown in Figures 2 and 3. Also, the error rate results of classification algorithms are listed in Table 3 and are graphically represented in Figures 4 and 5.

Overall, we can say that SVM classification algorithms have better results and fewer errors than ID3, C4.5, and Naive Bayes classification algorithms. Therefore, SVM classification algorithms are used by most algorithm developers.

## 6. Conclusion and Future Scope

Heart disease is one of the most serious human disorders, causing mortality all over the world. Many different groups of researchers have utilized knowledge exploration methods in a variety of domains to anticipate cardiac disease with acceptable levels of precision. In the early days, there were no real-time tools for assessing and forecasting cardiac disease. Decision trees are used to assess multiple training and assessment datasets for the prediction of heart disease. Naive Bayes, ID3, C4.5, and SVM are among the classification methods examined. In the experimental investigation, the UCI machinery heart disease data set was employed. It is found that the SVM outperforms the other three algorithms in terms of accuracy and error rate.

This study is focused on analyzing only the machine-learning approaches; however, other AI approaches like deep learning, newer technologies like robotics, and process automation can also be considered in the study for making a comparative analysis between the models; this will enable in identifying the diseases more proactively and enable in enhancing the effectiveness in healthcare industry.

### Data Availability

The data shall be made available on request.

### Conflicts of Interest

The authors declare that they have no conflict of interest.

### References

- [1] A. Singh and R. Kumar, "Heart disease prediction using machine learning algorithms," in *Proceedings of the 2020 International Conference on Electrical and Electronics Engineering (ICEE3)*, pp. 452–457, Gorakhpur, India, January 2020.
- [2] V. Sharma, S. Yadav, and M. Gupta, "Heart disease prediction using machine learning techniques," in *Proceedings of the 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, pp. 177–181, Greater Noida, India, December 2020.
- [3] P. Motarwar, A. Duraphe, G. Suganya, and M. Premalatha, "Cognitive approach for heart disease prediction using machine learning," in *Proceedings of the 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, pp. 1–5, Vellore, India, February 2020.
- [4] R. Manne and S. C. Kantheti, "Application of artificial intelligence in healthcare: chances and challenges," *Current Journal of Applied Science and Technology*, vol. 40, no. 6, pp. 78–89, 2021.
- [5] R. J. P. Princy, S. Parthasarathy, P. S. Hency Jose, A. Raj Lakshminarayanan, and S. Jeganathan, "Prediction of cardiac disease using supervised machine learning algorithms," in *Proceedings of the 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pp. 570–575, Madurai, India, May 2020.
- [6] H. Li, J. Sun, and J. Wu, "Predicting business failure using classification and regression tree: an empirical comparison with popular classical statistical methods and top classification mining methods," *Expert Systems with Applications*, vol. 37, no. 8, pp. 5895–5904, 2010.



- [7] Y. Khourdifi and M. Bahaj, "Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization," *International Journal of Intelligent Engineering and Systems*, vol. 12, no. 1, pp. 242–252, 2019.
- [8] H. Benjamin Fredrick David, "Heart disease prediction using data mining techniques," *ICTACT Journal on Soft Computing*, vol. 09, no. 01, pp. 1817–1823, 2018.
- [9] M. Shahi and K. Gurm, "Heart disease prediction system using data mining techniques," *Oriental Journal of Computer Science and Technology*, vol. 6, pp. 457–466, 2017.
- [10] M. Fatima and M. Pasha, "Survey of machine learning algorithms for disease diagnostic," *Journal of Intelligent Learning Systems and Applications*, vol. 09, no. 01, pp. 1–16, 2017.
- [11] A. Malav, K. Kadam, and P. Kamat, "Prediction of heart disease using k-means and artificial neural network as a hybrid approach to improve accuracy," *International Journal of Engineering and Technology*, vol. 9, no. 4, 2017.
- [12] Khateeb and Usma, "Efficient heart disease prediction system using K-nearest neighbor classification technique," in *Proceedings of the International Conference on Big Data Internet Things*, Tetouan Morocco, March 2017.
- [13] J. Sujatha, "Performance evaluation of machine learning algorithms in the classification of Parkinson disease using voice attributes," *International Journal of Applied Engineering Research*, vol. 12, pp. 10669–10675, 2017.
- [14] D. Gomathi and S. Priyaa, "Multi disease prediction using data mining techniques," *International Journal of System and Software Engineering*, vol. 4, no. 2, pp. 12–14, 2016.
- [15] A. Shetty and C. Naik, "Different data mining approaches for predicting heart disease," *International Journal of Innovative in Science Engineering and Technology*, vol. 5, pp. 277–281, 2016.
- [16] R. Kavitha and E. Kannan, "An efficient framework for heart disease classification using feature extraction and feature selection technique in data mining," in *Proceedings of the 2016 International Conference on Emerging Trends in Engineering, Technology and Science (ICETETS)*, Pudukkottai, India, February 2016.
- [17] G. Nagrajan, "Effective prediction model for heart disease using machine learning algorithm," *International Journal of Engineering Research and Technology*, vol. 7, no. 01, pp. 01–08, 2016.
- [18] J. Vijayashree and N. C. S. NarayanaLyengar, "Heart disease prediction system using data mining and hybrid intelligent techniques: a review," *International Journal of Bio-Science and Bio-Technology*, vol. 8, no. 4, pp. 139–148, 2016.
- [19] K. Vembandasamy, R. Sasipriya, and E. Deepa, "Heart diseases detection using naïve Bayes algorithm," *IJISSET-International Journal of Innovative Science, Engineering & Technology*, vol. 2, pp. 441–444, 2015.
- [20] W. Dai, T. S. Brisimi, W. G. Adams, T. Mela, V. Saligrama, and I. C. Paschalidis, "Prediction of hospitalization due to heart diseases by supervised learning methods," *International Journal of Medical Informatics*, vol. 84, no. 3, pp. 189–197, 2015.
- [21] P. V. Ankur Makwana, "Identify the patients at high risk of re-admission in hospital in the next year," *International Journal of Science and Research*, vol. 4, pp. 2431–2434, 2015.
- [22] Prerana, Shivaprakash, and Swetha, "Prediction of heart disease using machine learning algorithms- Naïve Bayes, introduction to PAC algorithm, comparison of algorithms and HDPS," *International Journal of Science and Engineering*, vol. 3, no. 2, pp. 90–99, 2015.
- [23] G. Kose, H. Sever, M. Bal, and A. Ustundag, "Comparison of different inference algorithms for medical decision making," *International Journal of Computational Intelligence Systems*, vol. 7, no. Supplement 1, pp. 29–44, 2014.
- [24] M. Kumari and S. Godara, "Comparative study of data mining classification methods in cardiovascular disease prediction," *International Journal of Computer Science & Technology*, vol. 2, 2011.
- [25] S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," *International Journal of Computer Science and Network Security*, vol. 8, no. 8, pp. 1–6, 2008.
- [26] E. G. Ross, N. H. Shah, R. L. Dalman, K. T. Nead, J. P. Cooke, and N. J. Leeper, "The use of machine learning for the identification of peripheral artery disease and future mortality risk," *Journal of Vascular Surgery*, vol. 64, no. 5, pp. 1515–1522, 2016.
- [27] M. Ilayaraja and T. Meyyappan, "Efficient data mining method to predict the risk of heart diseases through frequent itemsets," *Procedia Computer Science*, vol. 70, pp. 586–592, 2015.
- [28] B. N. Lakshmi, T. S. Indumathi, and N. Ravi, "A study on C.5 decision tree classification algorithm for risk predictions during pregnancy," *Procedia Technology*, vol. 24, pp. 1542–1549, 2016.
- [29] D. Kharche, K. Rajeswari, and D. Abin, "Comparison of different datasets using various classification techniques with weka," *International Journal of Computer Science and Mobile Computing*, vol. 3, no. 4, pp. 389–393, 2014.
- [30] R. Alizadehsani, J. Habibi, M. J. Hosseini et al., "A data mining approach for diagnosis of coronary artery disease," *Computer Methods and Programs in Biomedicine*, vol. 111, no. 1, pp. 52–61, 2013.
- [31] V. Lahoura, H. Singh, A. Aggarwal et al., "Cloud computing-based framework for breast cancer diagnosis using extreme learning machine," *Diagnostics*, vol. 11, no. 2, p. 241, 2021.
- [32] H. Yang and J. M. Garibaldi, "A hybrid model for automatic identification of risk factors for heart disease," *Journal of Biomedical Informatics*, vol. 58, pp. S171–S182, 2015.
- [33] D. Koundal, B. Sharma, and Y. Guo, "Intuitionistic based segmentation of thyroid nodules in ultrasound images," *Computers in Biology and Medicine*, vol. 121, Article ID 103776, 2020.
- [34] V. Chaurasia and S. Pal, "Data mining approach to detect heart diseases," *International Journal of Advanced Computer Science and Information Technology (IJACSIT)*, vol. 2, no. 4, pp. 56–66, 2013.
- [35] D. Chandna, "Diagnosis of heart disease using data mining algorithm," *IJCSIT International Journal of Computer Science and Information Technologies*, vol. 5, no. 2, pp. 1678–1680, 2014.
- [36] H. Bouali and J. Akaichi, "Comparative study of different classification techniques: heart disease use case," in *Proceedings of the 2014 13th International Conference on Machine Learning and Applications (ICMLA)*, pp. 482–486, IEEE, Detroit, MI, USA, December 2014.
- [37] A. Khemphila and V. Boonjing, "Comparing performances of logistic regression, decision trees, and neural networks for classifying heart disease patients," in *Proceedings of the 2010 International Conference on Computer Information Systems and Industrial Management Applications (CISIM)*, pp. 193–198, IEEE, Krakow, Poland, October 2010.
- [38] A. K. Borah, R. K. Deka, and A. Gohain Barua, "Implications of cognitive studies of fluid mechanics of coronary artery

bypass surgery: a more realistic view," *International Journal of Advanced Science and Technology (IJAST)*, vol. 6, 2009.

- [39] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81-106, 1986.
- [40] V. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, New York, NY, USA, 1998.
- [41] <https://archive.ics.uci.edu/ml/datasets/heart+disease>.

RETRACTED