SciNexuses

Journal Homepage: sciencesforce.com/scin

SciNexuses Vol. 1 (2024) 9–15

Paper Type: Original Article

# Harnessing Machine Learning for Accurate Cardiovascular Disease Prediction

**Ahmed Abdelhafeez [1,*], Abdullah Ashraf [1] and Hussam Elbehiery [1]**

[1] Faculty of Information Systems and Computer Science, October 6th University, Giza, 12585, Egypt;
Emails: aahafeez.scis@o6u.edu.eg; Abdullahasakr@gmail.com; H.elbehiry.csis@o6u.edu.eg.

**Received:** 12 Oct 2023 **Revised**: 30 Dec 2023 **Accepted**: 26 Jan 2024 **Published**: 29 Jan 2024

## Abstract

Cardiovascular disease (CVD) is a life-threatening disease rising considerably in the world. Early detection and prediction of CVD as well as other heart diseases might protect many lives. This requires tact clinical data analysis. The potential of predictive machine learning algorithms to develop the doctor's perception is essential to all stakeholders in the health sector since it can augment the efforts of doctors to have a healthier climate for patient diagnosis and treatment. We used the machine learning (ML) algorithm to carry out a significant explanation for accurate prediction and decision-making for CVD patients. Simple random sampling was used to select heart disease patients from the Khyber Teaching Hospital and Lady Reading Hospital, Pakistan. ML methods such as decision tree (DT), random forest (RF), logistic regression (LR), Naïve Bayes (NB), and support vector machine (SVM) were implemented for classification and prediction purposes for CVD patients in Pakistan. We performed exploratory analysis and experimental output analysis for all algorithms. We also estimated the confusion matrix and recursive operating characteristic curve for all algorithms. performance of the proposed ML algorithm was estimated using numerous conditions to recognize the most suitable machine learning algorithm in the class of models. RF algorithm had the highest accuracy of prediction, sensitivity, and recursive operative characteristic curve of 85.01%, 92.11%, and 87.73%, respectively, for CVD. It also had the least specificity and misclassification errors of 43.48% and 8.70%, respectively, for CVD. These results indicated that the RF algorithm is the most appropriate algorithm for CVD classification and prediction. Our proposed model can be implemented in all settings worldwide in the health sector for disease classification and prediction.

**Keywords:** Cardiovascular Disease, Machine Learning, Risk Assessment, Data Mining, Predictive Modeling, Supervised Learning, Clinical Data.

## 1 | Introduction

Heart disease is the major cause of death globally. More people die annually from CVDs than from any other cause, an estimated 12 million people die from heart disease every year [1]. Heart disease kills one person every 34 seconds in the United States. Heart attacks are often a tragic event and are the result of blocking blood flow to the heart or brain. People at risk of heart disease may show elevated blood pressure, glucose, and lipid levels as well as stress. All these parameters can be easily measured at home by basic health facilities [2]. Coronary heart disease, Cardiomyopathy, and cardiovascular disease are the categories of heart disease.

Corresponding Author: aahafeez.scis@o6u.edu.eg

https://doi.org/10.61356/j.scin.2024.1267

The word "heart disease" includes a variety of conditions that affect the heart and blood vessels and how the fluid gets into the bloodstream and circulates there in the body. CVD) causes many diseases, disability, and death [3]. Diagnosis of the disease is important and complex work in medicine. Medical diagnosis is considered a crucial but difficult task to be done efficiently and effectively [4]. The automation of this task is very helpful. Unfortunately, all physicians are not experts in any subject specialists, and beyond the scarcity of resources, there are some places. Data mining can be used to find hidden patterns and knowledge that may contribute to successful decision-making [5]. This plays a key role for healthcare professionals in making accurate decisions and providing quality services to the public. The approach provided by the health care organization to professionals who do not have more knowledge and skills is also very important [6]. One of the main limitations of existing methods is the ability to draw accurate conclusions as needed. In our approach, we are using different data mining techniques and machine learning algorithms, Naïve Bayes, k Nearest Neighbor (KNN), Decision tree, Artificial Neural Network (ANN), and Random Forest to predict heart disease based on some health parameters [7].

# 2 | Material and Methods

## 2.1 | Material and Methods

The data were collected from the two largest teaching hospitals, the Lady Reading Hospital (LRM) and the Khyber Teaching Hospital (KTH), in Khyber Pakhtunkhwa (KPK), one of the four provinces of Pakistan. Ethical approval for the inclusion of heart disease patients was sought from the Human Ethical Committees of the two teaching hospitals [8]. The ethics approval certificate number for the Lady Reading Hospital is B371/12/07/2022, while that of the Khyber Teaching Hospital is A418/12/07/2022. A simple random sampling technique was employed in the collection of sample units included in the survey. The sample data consisted of a total of 518 randomly selected heart disease patients. Data is analyzed using Anaconda Navigator's Jupiter Notebook. It is an open-source software where we can implement multiple machine-learning algorithms by importing libraries. We can also download the needed libraries by anaconda prompt. It allows us to create live code, perform visualizations, process data, and plot graphs.

### 2.1.1 | Variables in the Study

The CVD data included the individual output with corresponding factors. The all-inclusive dataset contained the following attributes: age, gender, height, weight, systolic, diastolic, cholesterol, glucose, smoke, alcohol intake, physical activity, cardiovascular disease, and body mass index (BMI). The response variable, CVD, was classified into two categories "presence" and "absence." Furthermore, the data was cleaned of noise, inconsistencies, or any missing observations [9]. We found a few missing observations in the data because some of the patients were discharged from the ward without any proper residential address or mobile/telephone numbers to trace them. As a result, it was very difficult to contact them. Since our analysis is based on complete data, we replaced the missing data by implementing the usual statistical method such as using median/mode for the categorical data to replace the missing values with the corresponding value [10]. Thus, the data cleaning was completed using the corresponding statistical tools for the preprocessing stage. Different data mining techniques were utilized in association, classification, clustering, pattern evaluation, and prediction. In the methods section below, we have discussed the techniques extensively.

## 2.2 | Material and Methods

### 2.2.1 | Classification

Classification is the process of categorizing a given set of data into classes. Classification can be performed for both structured and unstructured data. Predicting the class of the provided data points is the first step in the procedure [11]. Common names for the classes include target, label, and categories. Different statistical and mathematical procedures such as linear programming, decision trees, and neural networks involve classification. notwithstanding, CVD detection can be recognized through classification procedures because it has two categories, that is, one has CVD or not [12].

### 2.2.2 |Decision Tree Algorithm

The decision tree (DT) is one of the most important predictive modeling and classification methods in learning algorithms that are widely used in practical approaches in supervised learning techniques [13].

### 2.2.3 |Random Forest Algorithm

A random forest (RF) is a classifier consisting of a collection of tree-structured classifiers {h (x; €k); k 1, 2, ...} where €k are independent and identically distributed random vectors where each tree casts a unit vote for the most popular class at the input of the predictor, x [14].

### 2.2.4 |Logistic Regression Algorithm

The logistic regression (LR) model is the most accurate in the case of the dichotomous categorical response variable in the machine learning (ML) algorithm, the LR model can be used for classification purposes [15].

### 2.2.5 |Support Vector Machine Algorithm

Support Vector Machine (SVM) Algorithm. Among the different classification techniques, the support vector machine (SVM) is well known for its discriminative power for classification. The SVM is widely considered in recent times due to its efficiency in most different pattern classification techniques [16].

## 3 |Results and Discussion

The descriptive analysis of the attributes at the aggregate and age levels of the responses of all randomly selected patients with heart disease in the study has illustrated the numerical output of the cardiovascular disease-associated risk factors. which indicates the variability in the age proportion of the CVD-affected patients. The exploratory analysis revealed that almost 52.1% of the respondents had CVD at an aggregate level. Furthermore, there was a noticeable variation in the proportion of heart disease concerning different factors such as gender, physical activity, smoking, and so on that correlated with CVD. For instance, a maximum of 4.25% of 60-year-old patients were estimated to have CVD, whereas a maximum of 0.19% of 45-year-old patients had it.

Figure 1 shows the gender, cholesterol level, and glucose levels for all randomly selected CVD patients in the study. The figure shows that a greater proportion of the patients had CVD. Figure 2 presents a line graph for the proportion of gender concerning the age of patients. The figure shows that CVD is predominant in males compared to females since a greater proportion of the males had the disease. Moreover, the proportion of CVD patients increased from forty years to sixty-one years, which confirms the result of [17].

To achieve our goal, we employed the binary classifier based on a supervised machine learning algorithm for classification to predict the association for the appropriate class of patients [18–20] as proposed by [21] and [22]. Table 1 indicates the output of the predictive models that were used for the prediction of CVD.

All five ML algorithms (i.e., DT, SVM, NB, LR, and RF) were used to build the CVD prediction model in two different stages. In the initial stage, the data were split into two separate 70% and 30% groups for training and validation, respectively. In the second stage, however, the data were split into 75% and 25% for training and validation, respectively. The RF model had the highest accuracy of 85.01% with a 95% confidence interval of (0.6608, 0.8043), followed by DT with 83.72% accuracy with a 95% confidence interval of (0.654, 0.7986). The SVM and LR algorithms had the same accuracy of 83.08%, respectively, with a 95% confidence interval of (0.654 and 0.7986). The NB had the lowest accuracy of 74.74% with a 95% confidence interval of (0.567, 0.7221). This shows that the RF algorithm is the best predictor of CVD patients. Our outcome confirms the results obtained by the authors in [23- 25].

Sensitivity, mathematically defined as the ratio of the total number of true-positive patients to the sum of the number of true-positive and false-negative patients, was used to find the proportion of true patients suffering from CVD [26, 28]. Similarly, the specificity is described according to respondents who are not affected by cardiovascular disease. Specificity, mathematically defined as the ratio of the total number of true negatives to the sum of the number of true negatives and false-positive patients [29], was also used to determine the true proportion of true patients who are not suffering from CVD [30]. The RF algorithm estimated sensitivity and specificity as 86.11% and 65.48%, respectively. That is, our algorithm correctly classified 86.11% of the patients who have CVD but failed to identify 13.89% as having CVD. Similarly, the test correctly classified 65.48% of patients as not having CVD while 34.52% of them were misclassified.

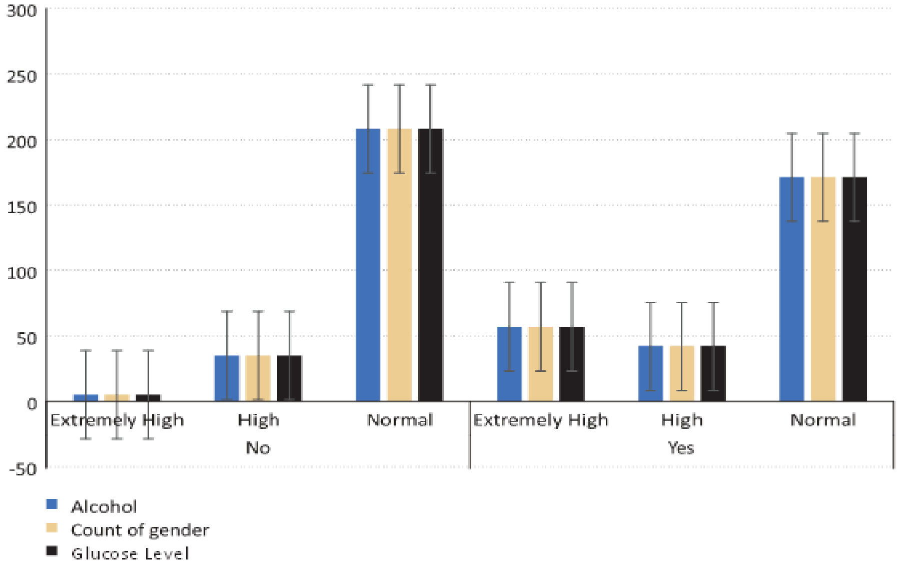

**Figure 1.** Bar graph with error bars for patient CVD status with gender, cholesterol level, and glucose level.

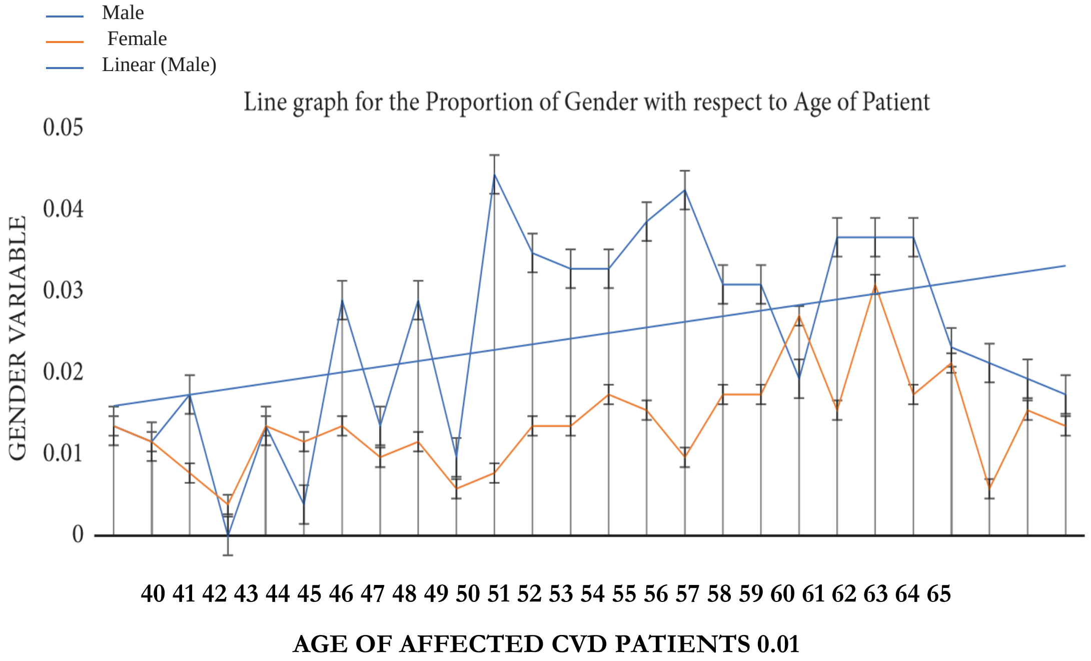

**Figure 2.** Line chart with error bars for the proportion of gender concerning the age of patients.

**Table 1.** An experimental output of the predictive models for CVD patients.

| Output | DT | SVM | NB | LR | RF |
|---|---|---|---|---|---|
| **Accuracy** | 0.85 | 0.84 | 0.74 | 0.9 | 0.88 |
| **95% confidence interval** | (0.6608, 0.8043) | (0.654, 0.7986) | (0.567, 0.7221) | (0.654, 0.7986) | (0.6745, 0.8158) |
| **Sensitivity** | 0.9028 | 0.8472 | 0.8889 | 0.8333 | 0.8611 |
| **Specificity** | 0.5952 | 0.631 | 0.4405 | 0.6429 | 0.6548 |
| **+Predicted value** | 0.6566 | 0.663 | 0.5766 | 0.6667 | 0.6813 |
| **−Predicted value** | 0.8772 | 0.8281 | 0.8222 | 0.8182 | 0.8862 |
| **Prevalence** | 0.4615 | 0.4615 | 0.4615 | 0.4615 | 0.4615 |
| **Detection rate** | 0.4167 | 0.391 | 0.4103 | 0.3846 | 0.3974 |
| **Detection prevalence** | 0.6346 | 0.5897 | 0.7115 | 0.5769 | 0.5833 |

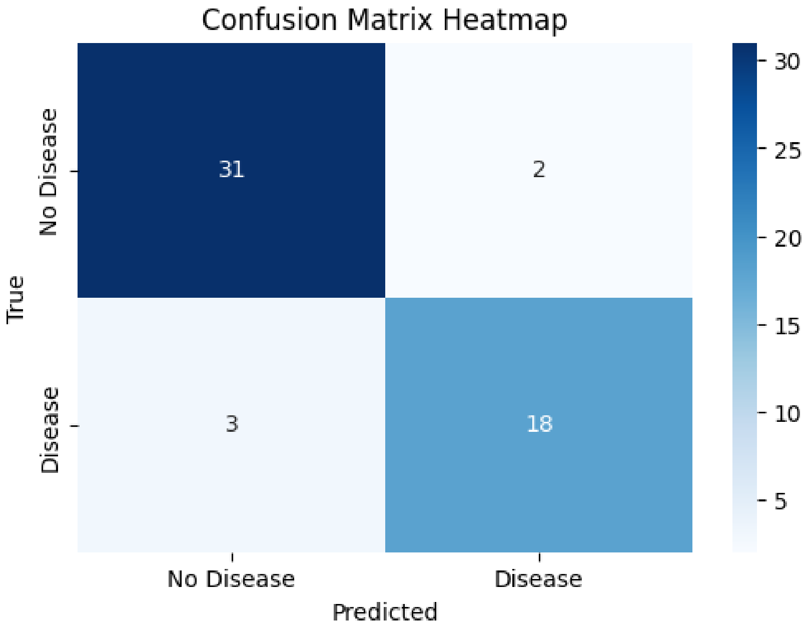

**Figure 3.** Confusion matrix.

## 4 | Conclusions

Heart diseases are considered a significant apprehension in medical data analysis. The potential of predictive machine learning algorithms to develop the doctor's perception is essential to all stakeholders in the health sector since it can augment the efforts of doctors to have a healthier climate for patient diagnosis and treatment. This study investigated the performance of predictive ML algorithms for CVD patients. CVD is one of the leading causes of mortality worldwide. We used data from the Lady Reading Hospital and the Khyber Teaching Hospital in Khyber Pakhtunkhwa Province, Pakistan. Ethical approval for the inclusion of heart disease patients was sought from the Human Ethical Committees of the two teaching hospitals. Five machine learning algorithms (i.e., DT, RF, LR, NB, and SVM) were implemented for the classification and prediction of CVD. We performed exploratory analysis and experimental output analysis for all algorithms. We also estimated the confusion matrix as shown in Figure 3 and recursive operating characteristic curve for all algorithms. The performance of the proposed ML algorithm was estimated using numerous conditions to recognize the most suitable machine learning algorithm in the class of models. The LR algorithm had the highest accuracy of prediction 90 % for CVD.

## Acknowledgments

The author is grateful to the editorial and reviewers, as well as the correspondent author, who offered assistance in the form of advice, assessment, and checking during the study period.

## Author Contributions

"Conceptualization, A.A. and H.E.; Methodology, A.A.; Software, H.E.; Validation, A.A., and H.E.; formal analysis, A.A.; investigation, H.E.; resources, A.A.; data maintenance, H.E.; writing-creating the initial design, A.A.; writing-reviewing and editing, H.E.; visualization, A.A.; monitoring, H.E.; project management, A.A.; funding procurement, A.A. All authors have read and agreed to the published version of the manuscript.

## Funding

This research has no funding source.

## Data Availability

The datasets generated during and/or analyzed during the current study are not publicly available due to the privacy-preserving nature of the data but are available from the corresponding author upon reasonable request.

## Conflicts of Interest

The authors declare that there is no conflict of interest in the research.

## Ethical Approval

This article does not contain any studies with human participants or animals performed by any of the authors.

## References

[1] M. G. Tektonidou, "Cardiovascular disease risk in antiphospholipid syndrome: thrombo-inflammation and atherothrombosis," Journal of Autoimmunity, vol. 128, Article ID 102813, 2022.

[2] World Health Organization, The World Health Report: 2000: Health Systems: Improving Performance, World Health Organization, Geneva, Switzerland, 2000.

[3] M. A. Said, Y. J. van de Vegte, M. M. Zafar, et al., "Contributions of interactions between lifestyle and genetics on coronary artery disease risk," Current Cardiology Reports, vol. 21, no. 9, pp. 1–8, 2019.

[4] M. De Paoli, D. W. Wood, M. K. Bohn, et al., "Investigating the protective effects of estrogen on $\beta$-cell health and the progression of hyperglycemia-induced atherosclerosis," American Journal of Physiology-Endocrinology and Metabolism, vol. 323, no. 3, pp. E254–E266, 2022.

[5] S. Jo´zwik, A. Wrzeciono, B. Cie´ ´slik, P. Kiper, J. Szczepanska-´ Gieracha, and R. Gajda, "The use of virtual therapy in cardiac rehabilitation of male patients with coronary heart disease: a randomized pilot study," Healthcare, vol. 10, no. 4, p. 745, 2022.

[6] H. Gulfam Ahmad and M. Jasim Shah, "Prediction of cardiovascular diseases (cvds) using machine learning techniques in health care centers," Azerbaijan Journal of High-Performance Computing, vol. 4, no. 2, pp. 267–279, 2021.

[7] C.D. Patnode, N. Redmond, M. O. Iacocca, and M. Henninger, "Behavioral counseling interventions to promote a healthy diet and physical activity for cardiovascular disease prevention in adults without known cardiovascular disease risk factors: updated evidence reports and systematic review for the US Preventive Services Task Force," JAMA, vol. 328, no. 4, pp. 375–388, 2022.

[8] M. Kavitha, G. Gnaneswar, R. Dinesh, Y. R. Sai, and R. S. Suraj, "Heart disease prediction using hybrid machine learning model," in Proceedings of the 2021 6th International Conference on Inventive Computation Technologies (ICICT), pp. 1329–1333, Coimbatore, India, January 2021.

[9] F. M. Zahid, S. Ramzan, S. Faisal, and I. Hussain, "Gender-based survival prediction models for heart failure patients: a case study in Pakistan," PLoS One, vol. 14, no. 2, Article ID e0210602, 2019.

[10] K. Hill, "Review of the World Health Report 2000: health systems: improving performance, by World Health Organization," Population and Development Review, vol. 27, no. 2, pp. 373–376, 2001.

[11] N. Al-Milli, “Backpropagation neural network for prediction of heart disease,” Journal of Theoretical and Applied Information Technology, vol. 56, no. 1, pp. 131–135, 2013.

[12] S. Bashir, Z. S. Khan, F. H. Khan, A. Anjum, and K. Bashir, “Improving heart disease prediction using feature selection approaches,” in Proceedings of the 2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST), pp. 619–623, Islamabad, Pakistan, January 2019.

[13] A. Aleem, G. Prateek, and N. Kumar, “Improving heart disease prediction using feature selection through genetic algorithm,” in Advanced Network Technologies and Intelligent Computing ANTIC, I. Woungang, S. K. Dhurandher, K. K. Pattanaik, A. Verma, and P. Verma, Eds., Springer, Berlin, Germany, pp. 765–776, 2021.

[14] J. Soni, U. Ansari, D. Sharma, and S. Soni, “Predictive data mining for medical diagnosis: an overview of heart disease prediction,” International Journal of Computer Applications, vol. 17, no. 8, pp. 43–48, 2011.

[15] W. M. Jinjri, P. Keikhosrokiani, and N. L. Abdullah, “Machine learning algorithms for the classification of cardiovascular disease-a comparative study,” in Proceedings of the 2021 International Conference on Information Technology (ICIT), pp. 132–138, Amman, Jordan, July 2021.

[16] M. N. Uddin and R. K. Halder, “An ensemble method based multilayer dynamic system to predict cardiovascular disease using machine learning approach,” Informatics in Medicine Unlocked, vol. 24, Article ID 100584, 2021.

[17] M. Kumar, S. Shambhu, and A. Sharma, “Classification of heart disease patients using data mining techniques,” International Journal of Research in Electronics and Computer Engineering, vol. 6, no. 3, pp. 1495–1499, 2018.

[18] K. Sudhakar and D. M. Manimekalai, “Study of heart disease prediction using data mining,” International Journal of Advanced Research in Computer Science and Software Engineering, vol. 4, no. 1, pp. 1157–1160, 2014.

[19] U. F. Nyaga, J. J. Bigna, V. N. Agbor, M. Essouma, N. A. Ntusi, and J. J. Noubiap, “Data on the epidemiology of heart failure in Sub-Saharan Africa,” Data in Brief, vol. 17, pp. 1218–1239, 2018.

[20] R. Prasad, P. Anjali, S. Adil, and N. Deepa, “Heart disease prediction using logistic regression algorithm using machine learning,” International Journal of Engineering and Advanced Technology, vol. 8, no. 3S, pp. 659–662, 2019.

[21] C. M. Wu, M. Badshah, and V. Bhagwat, “Heart disease prediction using data mining techniques,” in Proceedings of the 2019 2nd International Conference on Data Science and Information Technology (DSIT 2019), pp. 7–11, New York, NY, USA, July 2019.

[22] A. D. Gordon, Classification, Chapman and Hall/CRC, London, UK, 2nd edition, 1999.

[23] E. M. De Villiers, C. Fauquet, T. R. Broker, H. U. Bernard, and H. Zur Hausen, “Classification of papillomaviruses,” Virology, vol. 324, no. 1, pp. 17–27, 2004.

[24] X. Liu, X. Wang, Q. Su et al., “A hybrid classification system for heart disease diagnosis based on the RFRS method,” Computational and Mathematical Methods in Medicine, vol. 2017, Article ID 8272091, 11 pages, 2017.

[25] T. G. Dietterich, “Machine learning,” Annual Review of Computer Science, vol. 4, no. 1, pp. 255–306, 1990.

[26] P. H. Swain and H. Hauska, “The decision tree classifier: design and potential,” IEEE Transactions on Geoscience Electronics, vol. 15, no. 3, pp. 142–147, 1977.

[27] M. T. Huyut and H. Ustu¨nda¨ g, “Prediction of diagnosis and prognosis of COVID-19 disease by blood gas parameters using decision trees machine learning model: a retrospective observational study,” Medical Gas Research, vol. 12, no. 2, pp. 60–66, 2022.

[28] S. Christa, V. Suma, and U. Mohan, “Regression and decision tree approaches in predicting the effort in resolving incidents,” International Journal of Business Information Systems, vol. 39, no. 3, pp. 379–399, 2022.

[29] Y. Y. Song and Y. Lu, “Decision tree methods: applications for classification and prediction,” Shanghai Arch Psychiatry, vol. 27, no. 2, pp. 130–135, 2015.

[30] A. Akavia, M. Leibovich, Y. S. Resheff, R. Ron, M. Shahar, and M. Vald, “Privacy-preserving decision trees training and prediction,” ACM Transactions on Privacy and Security, vol. 25, no. 3, pp. 1–30, 2022.

**Disclaimer/Publisher’s Note:** The perspectives, opinions, and data shared in all publications are the sole responsibility of the individual authors and contributors, and do not necessarily reflect the views of Sciences Force or the editorial team. Sciences Force and the editorial team disclaim any liability for potential harm to individuals or property resulting from the ideas, methods, instructions, or products referenced in the content.