Open Access Library Journal
2024, Volume 11, e11490
ISSN Online: 2333-9721
ISSN Print: 2333-9705

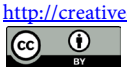
A Comprehensive Analysis of Machine Learning Techniques for Heart Disease Prediction

Elchin Asgarov

Department of Computer Science, The University of Manchester, Manchester, UK
Email: elchinasgrv@gmail.com

How to cite this paper: Asgarov, E. (2024) A Comprehensive Analysis of Machine Learning Techniques for Heart Disease Prediction. *Open Access Library Journal*, **11**: e11490.
https://doi.org/10.4236/oalib.1111490

Received: March 26, 2024
Accepted: April 27, 2024
Published: April 30, 2024



Abstract

Heart disease is one of the most important problems the world faces. It is an ongoing problem and it is leading to the cause of death globally. To solve this issue, predicting early heart disease is important. This research focuses on supervised machine learning techniques as a potential tool for heart disease prediction. This study has done a comprehensive review of 30 articles published between 1997 to 2023 about machine learning techniques to predict heart disease. The common problem is that authors use different data sets, and different numbers of parameters to train and test these models. These two factors could affect the model's accuracy. To compare different models, I only used articles that analyze more than one method using the same data to prevent bias. Some traditional machine learning methods such as Artificial Neural Network, and K-Nearest Neighbor demonstrated significant variation in accuracy, occasionally reaching as high as 100% but sometimes falling below 60% in specific situations which is inconsistent. Compared to these models, Hybrid Models show consistent accuracy, with a minimum accuracy rate of 88%, suggesting that they could be a better approach to predicting heart disease.

Subject Areas

Machine Learning

Keywords

Machine Learning, Supervised Learning, Hybrid Model, Heart Disease Prediction, Cardiovascular Disease

1. Introduction

The heart is vital to our existence because it is our main engine that circulates

blood through our entire body. Critical body organs, such as the brain, are at risk for disorders that can quickly have fatal consequences. Due in part to changes in our lifestyles, the stress we feel at work, and our eating habits, more people suffer from heart problems today. Over the last decade, heart disease, which is known as cardiovascular disease, has been the main cause of death globally. A report by the World Health Organization estimates, about 17.9 million deaths worldwide are caused by cardiovascular disorders each year. Some of them, coronary artery disease and strokes account for 80% [1]. A wide range of factors, such as genetics, work habits, and lifestyle choices, greatly impact the development of heart disease. Heart disease can be significantly predicted by lifestyle factors like smoking, drinking too much alcohol or caffeine, stress, and not doing physical activity, as well as physiological factors like being overweight, and having high blood pressure. Implementing preventative measures to avoid deaths requires prompt, accurate, and early detection of cardiac disease [2]. We can apply many methods of machine learning to predict heart disease problems.

The main objective of machine learning is to enable computers to learn from data on their own. This means increasing their skills without the need for human direction. This is achieved through the subset of machine learning. Unsupervised learning, supervised learning, and reinforcement learning are the three main categories of machine learning methods. These methods use an alternate method to solve problems and extract information from data.

Figure 1 illustrates several machine learning methodology types, each having a unique method of learning and prediction.

1) Supervised Learning: This machine learning method makes use of well-labeled datasets to create an obvious link between each training set of data and the corresponding result. The mechanism develops the ability to predict future events through input evaluation. Once expertise, it uses techniques like regression and classification to uncover the narrative contained within the data.

2) Unsupervised Learning: This approach examines unlabeled data to find inherent classifications and hidden patterns. Known as "unsupervised" since it doesn't require explicit guidance, it achieves success in identifying frequently missed patterns that naturally develop. Clustering is an example of unsupervised learning.

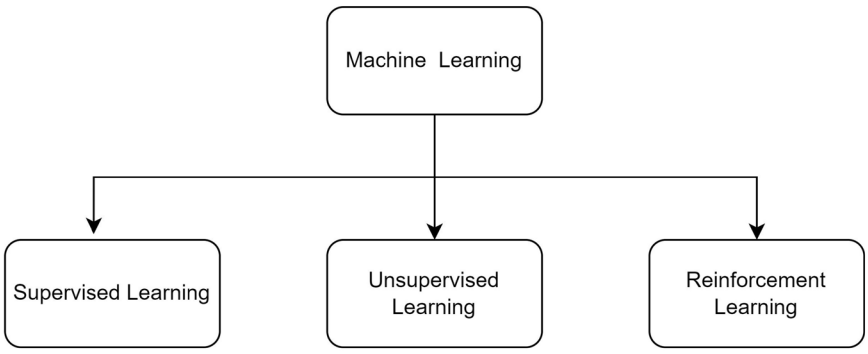


Figure 1. Machine learning paradigms.

3) Reinforcement learning: Like the one that came before two, this type of learning is different. This paradigm's main objective is to learn through committing mistakes and using feedback from its actions. Reinforcement learning has widespread application in areas like gaming and robotics.

The primary aim of this study is to meticulously analyze the accuracy of various supervised machine learning techniques in predicting heart disease, offering a comparison of their performance. This study seeks to identify which method provides the most reliable predictions for clinical use by evaluating models ranging from traditional algorithms to more complex deep learning networks. Through this research, the aim is to bridge the gap between technical machine learning advancements and their implementation in diagnostic practices, ultimately contributing to improved patient outcomes in heart disease management.

2. Methods

2.1. Naïve Bayes' Classifier

The Naïve Bayes classifier offers an easy way of classification through the use of supervised learning principles. It relies on the idea that every single typical in the dataset operates independently of all of them.

$$P(X/Y) = \frac{P(Y/X) \times P(X)}{P(Y)} \tag{1}$$

Equation (1) is a formula for Bayes theorem. The posterior probability, written as $P(X/Y)$ in the Naïve Bayes technique, estimates the likelihood of a happening after considering the specific proof. The event's prior probability, or $P(X)$, reflects its likelihood at first before taking into account new information. On the other hand, $P(Y/X)$, often known as the likelihood, expresses the likelihood of coming across a specific piece of evidence in the case that it transpires. $P(Y)$, sometimes referred to as the predictor prior probability, estimates the initial likelihood of finding the evidence independent of the occurrence [2].

Figure 2 illustrates the Naïve Bayes Classifier, a probabilistic machine learning model used for classification issues. The collection of geometric shapes in blue, red, and green on the left side of the diagram illustrates a dataset with a variety of characteristics. These overlapped shapes show that the raw data remains unclassified. The statistical analysis process is symbolized by the Naïve Bayes Classifier box at the middle of the diagram. The result of the classification process can be observed in the right in the organized arrangement of shapes, that are divided into three categories: triangles, squares, and circles. This shows how the classifier can foresee each shape's category based on its features.

2.2. Decision Tree

A decision tree is a highly famous method in the field of machine learning. It is excellent in evaluating and categorizing data points because it uses a framework for decision-making that is similar to a tree's architecture. This method shows

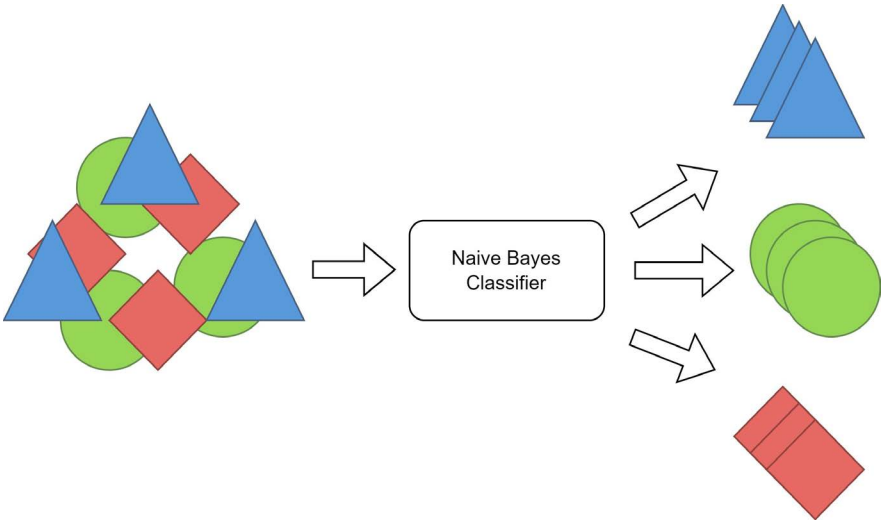


Figure 2. An illustration of Naïve Bayes classifier.

itself as an essential tool in precision-critical machine learning applications despite providing an easily understood visual reference to assist with judgments. This approach quickly creates a complex choice tree by repeatedly dividing the dataset into increasingly more detailed parts. This tree is distinguished for containing an array of branching methods and node kinds, which when combined make for a thorough decision-making process [3]. Three distinct types of nodes are employed by the decision tree model for analysis.

- The root node offers the basis for all other nodes to operate properly (decision node).
- The interior node controls multiple variables (chance node).
- The leaf node indicates the outcome of every evaluation (outcome node) [2].

The Decision Tree model is illustrated in **Figure 3**. There are three different types of nodes shown. Decision Node, marked by a blue oval, is located at the highest possible level. It acts as the starting point for the basic decision-making process. The decision node is the original guide for two separate routes, labeled "Option 1" and "Option 2", that eventually end in the Chance Nodes. They are shown graphically in green. Outcome nodes, which are illustrated as yellow circles, are generated from each Chance Node. As the final stages of the decision-making process, the previously mentioned nodes indicate potential outcomes that choices made at the Decision Node and the probabilistic effects at the Chance Nodes.

2.3. K-Nearest Neighbor

This method is particularly appropriate for scenarios requiring classification where the distribution of the data is unknown or insufficiently known. The fundamental operation of the algorithm is to identify that "k" data points are closest to the query point with no target value from the training set. Afterward, the algorithm gives the mean value of these nearby points to the query point [4].

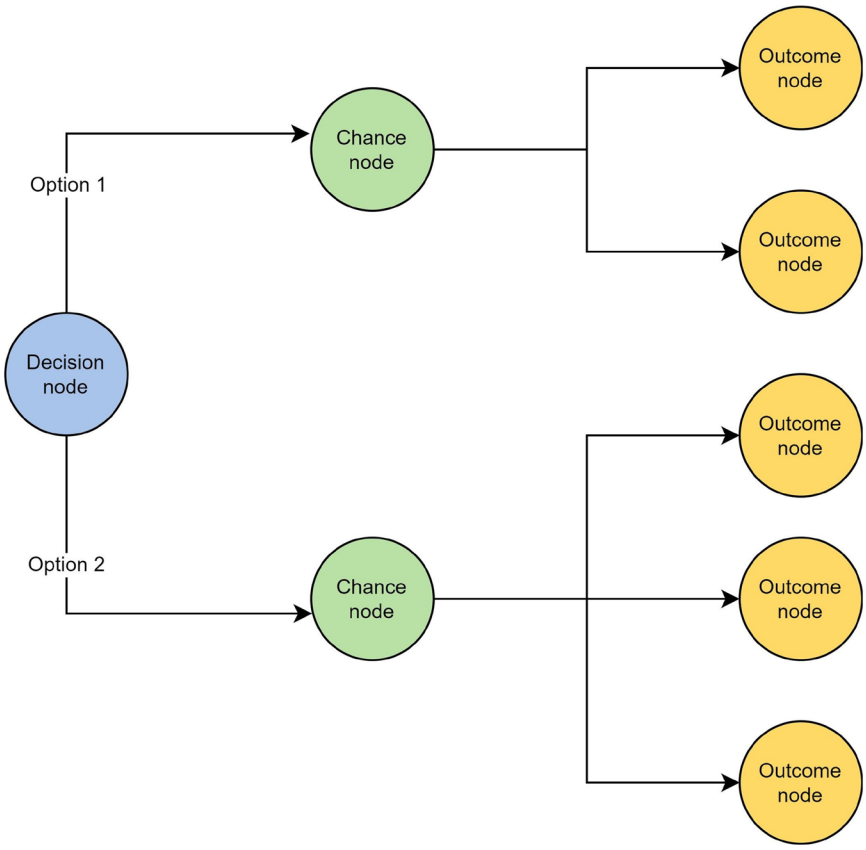


Figure 3. An illustration of a decision tree.

Figure 4 is the KNN classification concept and it shows the process as follows: The two-dimensional space represents a feature space in which every data point (represented by the colored dots) corresponds to a case that has been grouped into one of two distinct colors (yellow and green, respectively). The red star indicates the point of data that requires categorization. Circles marked by "K = 3" and "K = 6" with radii having a variety of nearest neighbors demonstrate how the classification is made by the value of "K". The category to which the new data point (as represented by the red star) is predicted is dependent on the value of "K". For "K = 3," the three closest points to the red star determine its category. For "K = 6," the six closest points would be considered.

2.4. Random Forest

The Random Forest algorithm is well recognized as an outstanding method of classification for supervised learning, with strong capabilities for carrying out regression tasks in addition. The system uses many decision trees to generate predictions, with each tree giving a vote toward the final expected classification. In the algorithm, the prediction that receives the most total votes is selected. An increased number of trees usually leads to improved accuracy.

While Random Forest demonstrates proficiency in dealing with difficulties with classification and successfully manages datasets comprising missing values,

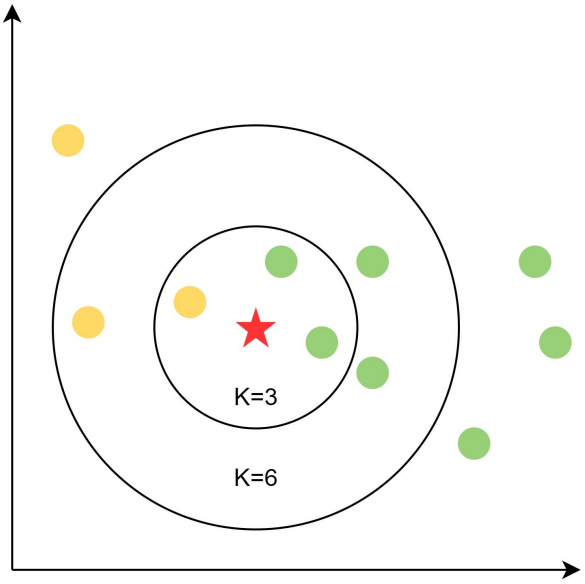


Figure 4. An illustration of k-nearest neighbor.

its predictive accuracy may suffer as a result of its reliance on extensive datasets and extensive trees, which may compromise the transparency of the generated outcomes [2].

Figure 5 illustrates a decision tree, that is an essential component in the field of machine learning used for classification and regression. At the highest point, the blue oval symbolizes the decision node. This acts as the first phase in the process of decision-making. It indicates an initial assessment predicated on a characteristic that ideally divides the data. Lines originate from the decision node and end at chance nodes, which are denoted by green circles. The chance nodes depict the conditions or tests that apply to extra characteristics of the data. They facilitate the dataset's further division into more homogeneous subsets that are more comparable. The chance nodes will be followed by extra chance nodes or yellow circles that are designated as outcome nodes. The outcome nodes, which are the ultimate nodes of the tree and reflect conclusions or predictions, are chosen by following the path from the decision node to the chance nodes. Then all the results are averaged and the final result is shown.

2.5. Hybrid Model

Hybrid machine learning models represent a more powerful instrument that combines the strengths of several algorithms. Their goal is to utilize the different advantages of each method. In a comparable style, one element of the hybrid system might have the capacity of recognizing patterns, comparable to a neural network, while another adds accuracy, similar to a support vector machine.

A composite model has been built via the combination of the random forest and decision tree systems' capabilities. This novel function model is constructed using probabilities acquired from the random forest method. Taking

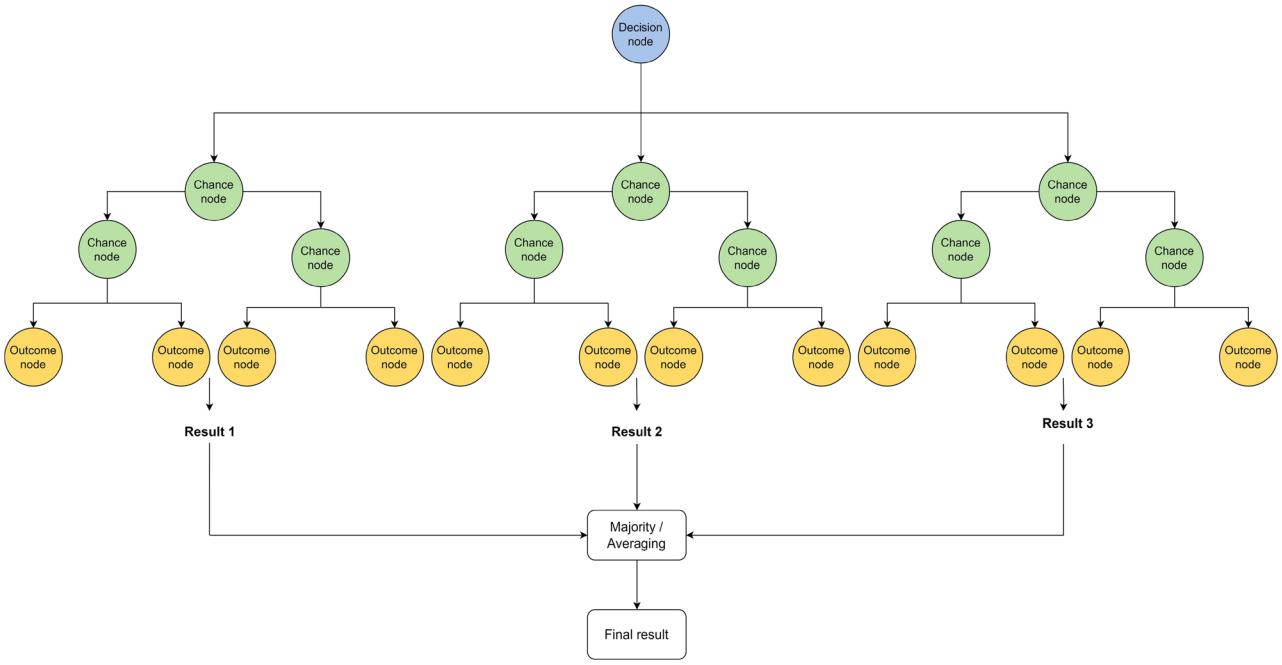


Figure 5. An illustration of a random forest algorithm.

the incorporation of the probabilistic findings generated by the random forest algorithm into the training dataset, the decision tree algorithm analyzes the resulting data. The decision tree determines probabilities which are then mutually applied to the test dataset [5].

Another study provides a novel hybrid approach called Hybrid Random Forest with Linear Model (HRFLM). HRFLM seeks to enhance the accuracy of heart attack forecasts. HRFLM approach utilizes every available feature without imposing any constraints on feature selection. Experiments were undertaken as part of a research effort to investigate the characteristics of machine learning algorithms using this hybrid method. The results obtained from these experiments provide proof that the HRFLM method exhibits an increased level of efficacy for forecasting cardiac disease compared to the existing methods [6].

2.6. Logistic Regression

Logistic Regression (LR) is a powerful classification tool that is especially popular among supervised learning algorithms (See **Figure 6** and **Figure 7**). Called an extension of traditional regression analysis, it has been specifically developed to deal with binary outcomes that show the existence presence, or absence of an event. LR computes the likelihood that a specific new input is classified into a given category. Since it operates on probabilities, the range of its values is restricted to one-to-one. Thus, when using LR in binary classification tasks, an upper limit must be set for distinguishing between two probable categories. To illustrate, the input could be classified as "class A" if the calculated probability passes 0.5; inversely, it might be classified as "class B" otherwise [3].

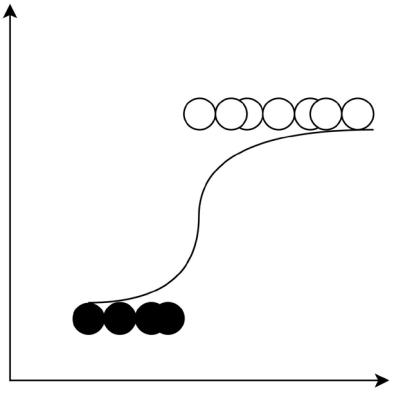


Figure 6. An illustration of logistic regression.

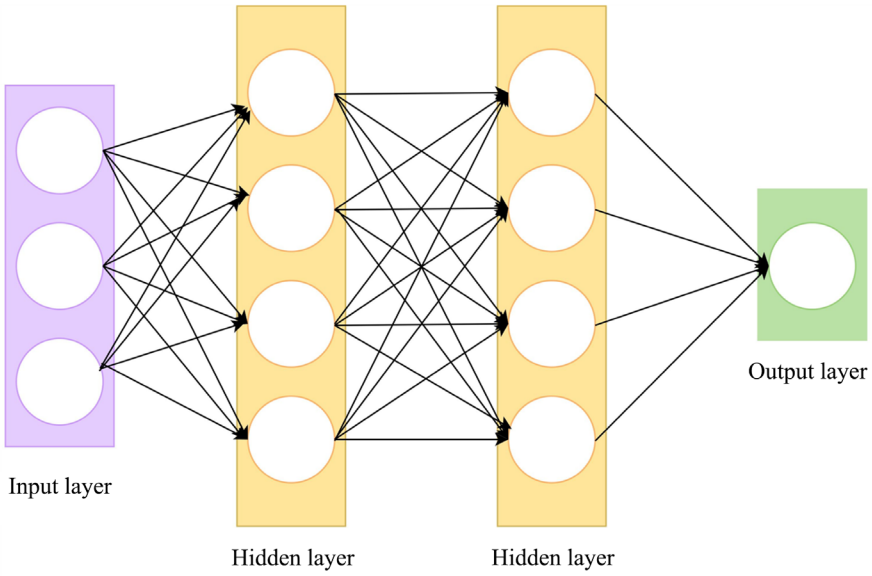


Figure 7. An illustration of an artificial neural network.

2.7. Artificial Neural Network

A collection of machine learning techniques is used by artificial neural networks (ANNs) to imitate brain-like activities. These structures are fresh and take inspiration from the operation of neural networks in the human brain. A complex network of neurons capable of processing, maintaining, and adapting to new information develops in the human brain via synapses; this mechanism is facilitated by neuroplasticity. Similar ANNs consist of multiple nodes that are interconnected to create a network. The operation of each node is replicated by the output from a different node, which aids the execution of complex computations. Several hidden layers, each performing a unique transformation, are interspersed with the input and output layers that are crucial to the architecture of ANNs. The transfer of signals throughout the network is affected by the weighting of the edges or connections between these parts. Weights are modified throughout the training phase to enable the ANN to acquire information from data and produce accurate forecasts of outcomes [3].

2.8. Support Vector Machine

Regarded as a classifier and an indicator with a predefined target variable, the Support Vector Machine is an extensively acknowledged approach in the discipline of supervised learning. SVM tries to find the ideal hyperplane inside the space of features to accomplish an independent separation between distinct classes as a component of its classification ability. In a support vector machines framework, data points obtained from the training set are converted into a multi-dimensional space with the widest disparity possible among distinct categories. After this arrangement, during the validation phase, extra data points are introduced and classified according to their location in this gap [4].

The operational principle of a Support Vector Machine (SVM), a potent classification technique in the field of machine learning, can be seen in the figure given. SVM classification works by identifying the hyperplane that divides a dataset into classes that perform best.

Figure 8 illustrates two distinct classes within a dataset symbolized by stars and triangles, accordingly. Each dimension of the feature space, represented by the axes of the graph, correlates to a distinct feature that the data. In this dataset, the data elements from two distinct categories are denoted by triangles and stars. The decision limit or hyperplane as determined by the SVM is represented by the straight line. The SVM seeks to maximize the margin, which represents the distance between the line and the adjacent data point to each of both categories when finding the best location for this line. The points in the closest distance to the line on both ends are referred to as support vectors. The previously mentioned information points exert an influence on the orientation and position of the hyperplane. The line represents the decision boundary, which is set by the SVM's optimization process. It shows the optimal method for separating classes while taking into consideration the dimensionality of the feature space.

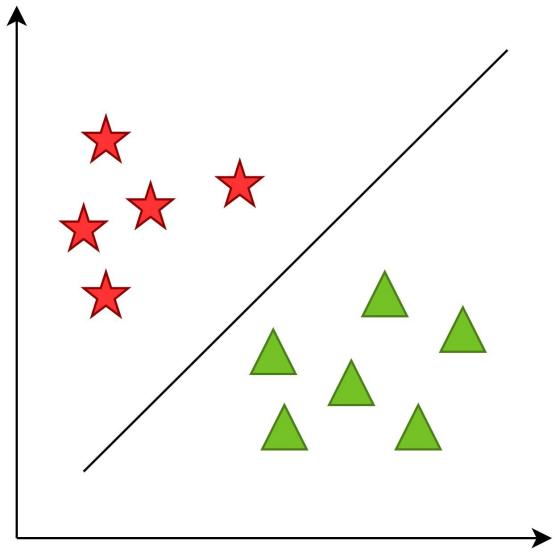


Figure 8. A simple illustration of a support vector machine.

3. Results

In this comprehensive review, I analyzed a total of 30 research articles spanning from 1997 to 2023. The annual distribution of these articles is presented in the bar chart (see Figure 9), showcasing the fluctuations in publication volume over the years.

Figure 10 offers a comprehensive graphical representation of the frequency distribution of machine learning methods. With an impressive total of 22, the Naïve Bayes (NB) algorithm emerges as the most frequently employed method, due to its uncomplicated probabilistic approach. Decision trees (DT) are additionally substantially represented, in a frequency of 20. This may be credited to its comprehension and its capacity to replicate how people make decisions via the division of a data set into more manageable subsets. At place 15, Support Vector Machines (SVM) play an important role in the chart. Random Forests

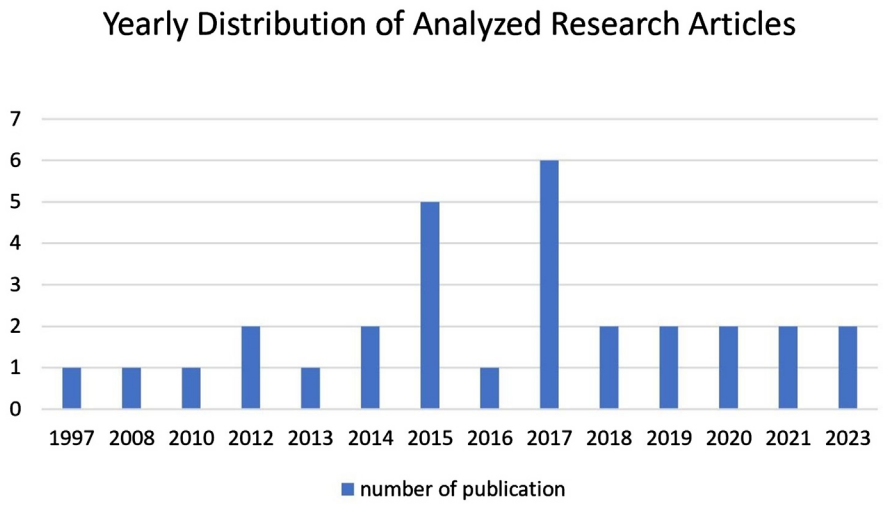


Figure 9. Yearly distribution of articles.

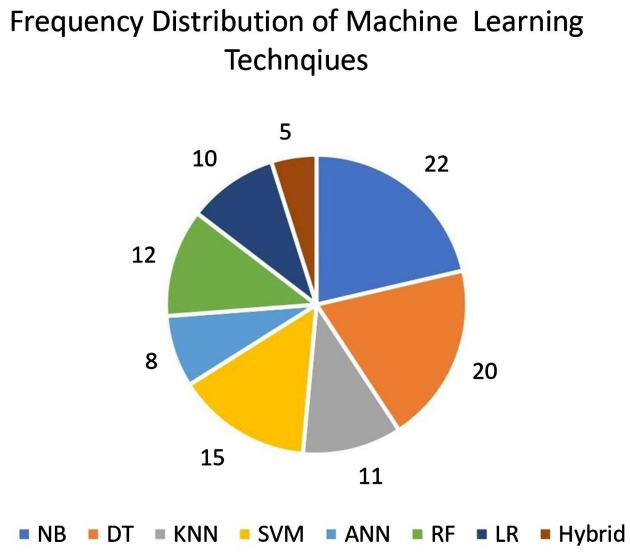


Figure 10. Distribution of machine learning techniques.

(RF) exhibit an average frequency of 12. RFs, in their role as an ensemble of Decision Trees, are highly regarded for their efficacy and ability to reduce overfitting. Their position inside the distribution implies that they have an edge when faced with complex datasets that derive gains from the ensemble methodology, thus enhancing performance without requiring considerable fine-tuning. The fact that the k-nearest Neighbors (KNN) and Logistic Regression (LR) appear in a frequency of 11 and 10 respectively suggests both of these techniques are seen as beneficial due to their simple implementation and interpretability. The frequency of Artificial Neural Networks (ANN) is unexpectedly low at 8. At the lowest frequency of 5, hybrid techniques indicate an exploratory application. Hybrid models, which integrate traits of multiple algorithms, strive to make use of the unique strengths of each one to obtain higher accuracy contrasted with what could be accomplished through independent models.

The highest accuracy rates for various machine learning techniques are shown in Figure 11, giving a straightforward glance at predictive modeling performance benchmarks. At the highest level of accuracy, Support Vector Machines (SVM), k-nearest Neighbors (KNN), Artificial Neural Networks (ANN), and Decision Trees (DT) are said to have attained a score of 100%. SVMs have become known for their outstanding efficiency in high-dimensional spaces, which could potentially explain their extraordinary efficacy. The fact that KNN and DT can generate precise forecasts regardless of their simplicity could indicate that they were used on datasets that had clear classification boundaries or were adequately resistant to noise and outliers. Achieving 98.49%, Random Forests (RF) falls just short of the optimal accuracy threshold. The high accuracy of RF serves as a case study of the effectiveness of ensemble learning for creating accurate and overfit-resistant predictive models. Hybrid techniques, which combine the advantageous aspects of various machine learning techniques, display a precision of 98.40%. Naïve Bayes (NB) and Logistic Regression (LR) lag with accuracy levels of 96.5% and 90%, respectively.

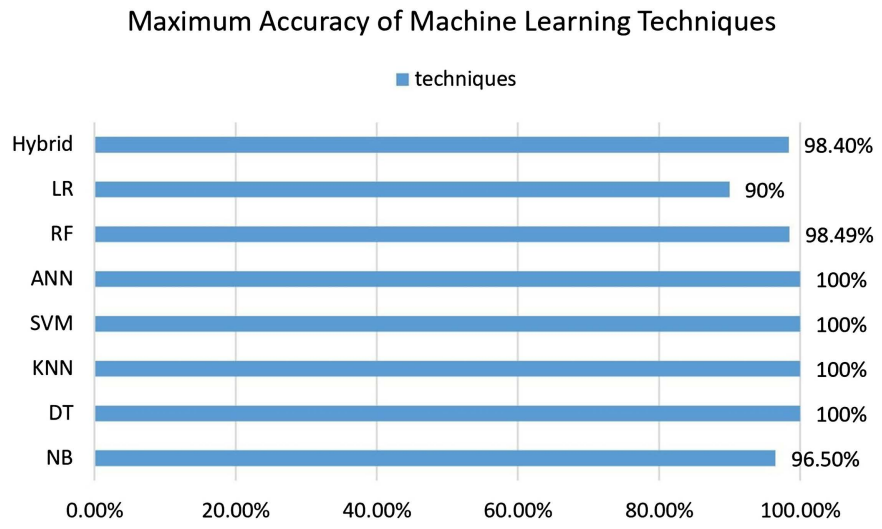


Figure 11. Maximum accuracy of each technique.

Figure 12 shows the minimum level of accuracy obtained via different machine-learning methodologies. In terms of efficiency, hybrid techniques hold the highest position with an accuracy rate of 88%. The observed high rate indicates that by combining different methods of learning, a model could be produced that is more exactly configured for understanding the intricate details of the data relating to specific complex problems. By deploying a variety of algorithms, this approach has a chance to improve the model's resistance against a wider range of data types and reduce the likelihood of overfitting. The Random Forest (RF) algorithm places second in terms of accuracy, attaining 73.20% accuracy. Thereafter, Support Vector Machines (SVM) attain a 67.71% rate of precision. On the contrary, Logistic Regression (LR), which is distinguished by its simplicity and directness, attains an accuracy rate of 63%. The function is frequently used in tasks requiring binary classification and gives the probability related to the output predictions. The accuracy of Artificial Neural Network (ANN) is recorded at 62.78%. Approximately 53% and 63%, are the accuracy at which k-nearest Neighbors (KNN) and Logistic Regression (LR) are assessed respectively. Although decision trees (DT) are renowned for their simplicity and high level of comprehension 59.77%, they are extremely sensitive to the training data and susceptible to overfitting. With a 45.85% accuracy rate, Naïve Bayes (NB) is positioned at the bottom of the hierarchy (See Table 1).

4. Discussion

The current investigation carried out an in-depth review of 30 scholarly articles in which the authors assessed the performance of various machine-learning techniques. However, a thorough investigation has revealed two basic constraints that are essential to understanding the findings. At first, it was discovered that while all authors utilized similar algorithms, they chose different datasets for model training. The existence of variability in the data can considerably alter the accuracy metrics, thereby diminishing the validity of a direct comparison between the results. Variations in the data, including inaccuracies in sample

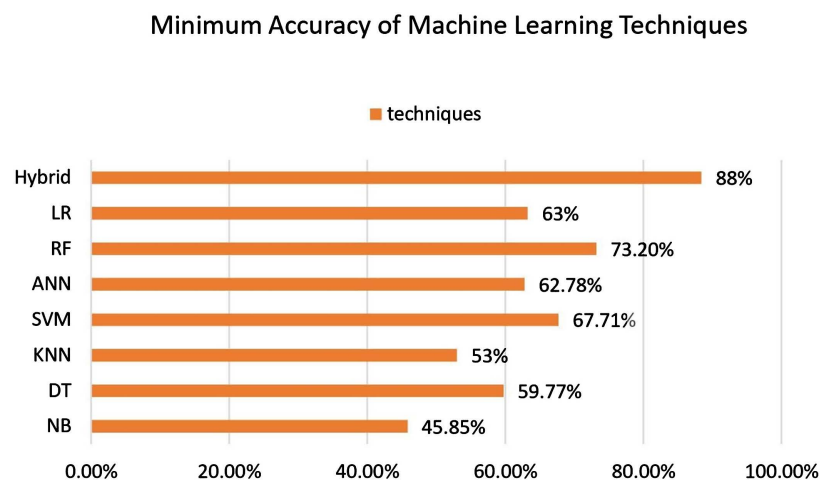


Figure 12. Minimum accuracy of each technique.

Table 1. Summary of findings.

Reference	Algorithm	Accuracy (%)	Best one	Year
Venkatalakshmi and Shivsankar [7]	NB, DT	NB = 85.03, DT = 84.01	NB	2014
Patel *et al.* [8]	NB, DT	NB = 96.5, DT = 99.2	DT	2013
Khateeb and Usman [9]	NB, DT, KNN	NB = 66.66, DT = 76.89, KNN = 79.2	KNN	2017
Lu *et al.* [10]	NB, DT, SVM, ANN	NB = 82.32, DT = 84.35, SVM = 86.62, ANN = 86.06	SVM	2018
Marikani and Shyamala [11]	NB, DT, KNN, SVM, RF	NB = 81.7, DT = 95.4, KNN = 75.7, SVM = 100, RF = 96.3	SVM	2017
Dangare and Apte [12]	NB, DT, ANN	NB = 94.44, DT = 96.6, ANN = 99.25	ANN	2012
Bhatla and Jyoti [13]	NB, DT, ANN	NB = 90.74, DT = 99.62, ANN = 100	ANN	2012
Anbarasi *et al.* [14]	NB, DT	NB = 96.5, DT = 99.2	DT	2010
Kim *et al.* [15]	SVM, ANN, LR	SVM = 67.71, ANN = 62.78, LR = 62.23	SVM	2015
Mansoor *et al.* [16]	RF, LR	RF = 89, LF = 89	RF, LR	2017
Mustaqeem *et al.* [17]	NB, KNN, SVM, RF	NB = 74.43, KNN = 76.7, SVM = 74.47, RF = 76.5	RF	2017
Toshniwal *et al.* [18]	NB, SVM, RF	NB = 88.44, SVM = 98.41, RF = 98.49	RF	2015
Forsen *et al.* [19]	RF, LR	RF = 73.2, LR = 76.7	LR	2017
Palaniappan *et al.* [20]	NB, DT, ANN	NB = 86.12, DT = 80.4, ANN = 85.68	NB	2008
Long *et al.* [21]	NB, SVM, ANN	NB = 83.3, SVM = 75.9, ANN = 77.8	NB	2015
Guvenir *et al.* [22]	NB, KNN	NB = 50, KNN = 53	KNN	1997
Samad *et al.* [23]	NB, DT, KNN	NB = 45.9, DT = 59.8, KNN = 67	KNN	2014
Khan *et al.* [24]	KNN, SVM	KNN = 73.8, SVM = 68.8	KNN	2015
Aravinthan and Vanitha [25]	NB, ANN	NB = 81.3, ANN = 82.6	ANN	2016
Dwivedi [26]	NB, DT, KNN, SVM, ANN, LR	NB = 83, DT = 77, KNN = 80, SVM = 82, ANN = 84, LR = 85	LR	2018
Shah *et al.* [2]	NB, DT, KNN, RF	NB = 88.2, DT = 73.6, KNN = 73.7, RF = 84.2	NB	2020
Otoom *et al.* [27]	NB, SVM	NB = 84.5, SVM = 85.1	SVM	2015
Mohan *et al.* [6]	NB, DT, SVM, RF, LR, HM	NB = 75.8, KNN = 85, SVM = 86.1, RF = 86.1, LR = 82.9, HM = 88.4	HM	2019
Ali *et al.* [28]	DT, KNN, RF, LR	DT = 100, KNN-100, RF = 100, LR = 89.6	DT, KNN, RF	2021
Kavitha *et al.* [5]	DT, RF, HM	DT = 79, RF = 81, HM = 88	HM	2021
Singh and Kumar [29]	DT, KNN, SVM, LR	DT = 79, KNN = 87, SVM = 83, LR = 78	KNN	2020
Pouriyeh *et al.* [30]	NB, DT, SVM	NB = 83.5, DT = 77.6 SVM = 84.15	SVM	2017
Fitriyani *et al.* [31]	NB, DT, SVM, RF, LR, HM	NB = 83.17, DT = 76.09, SVM = 71.06, RF = 82.14, LR = 84.85, HM = 98.40	HM	2019
Shrivastava *et al.* [32]	DT, SVM, RF, LR, HM	DT = 80, SVM = 88.33, RF = 85, LR = 90, HM = 96.7	HM	2023
Doppala *et al.* [33]	NB, DT, KNN, SVM, RF, LR, HM	NB = 89, DT = 90, KNN = 87, SVM = 91, RF = 89, LR = 89, HM = 96	HM	2023

size, feature space, and intrinsic data distribution, could significantly affect the computational performance. Therefore, an algorithm's outstanding results on one dataset may be coupled with poor performance on another, a circumstance that does not naturally indicate the algorithm's ability but rather indicates its suitability for the specific characteristics of the data. Also, an observed discrepancy has been found in the selection and modification of algorithm parameters between the different research projects. The parameterization of machine learning models is a vital step in their execution, as the amount and nature of these parameters have the potential to significantly impact the performance of the model. Significant variations in results can result from the use of different parameters or even a large number of parameters, which confuses the task of assigning performance differences to the algorithms themselves as compared to the parameter selection. Despite these challenges, important findings have been gathered from the literature review. A noteworthy tendency is the growing preference for hybrid models, which consistently show exceptional accuracy, with the lowest documented value reaching a remarkable 88%. This emphasizes an important chance for hybrid models in real-world executions. Some models reach 100% accuracy in some tests but they also sometimes cannot even reach 60%. It is the reason using the Hybris model, 88% is the lowest accuracy is better than choosing a method that is sometimes accurate and sometimes poorly accurate. Data indicates that hybrid models are resilient due to their ability to combine multiple learning strategies and reduce the drawbacks of individual models. They provide a robust solution for a variety of predictive tasks. The constant display of exceptional performance by hybrid models implies that they may prove to be particularly useful in areas where accuracy is critical and the consequences of error are significant. The results suggest that further investigation is justified regarding hybrid models, with a heightened focus on their applicability in heart disease prediction.

5. Conclusion

The study revealed significant differences in efficiency among multiple machine-learning techniques when implemented in a similar dataset. Hybrid machine-learning models have grown as an indicator for assessing efficiency, exhibiting extraordinary reliability with a minimum accuracy rate of 88%. The finding implies that hybrid models, which combine various algorithmic abilities, exhibit more adaptability to variations in data and factors. Thus, they provide a more reliable and consistent methodology for predicting heart disease.

Conflicts of Interest

The author declares no conflicts of interest.

References

[1] Seckeler, M.D. and Hoke, T. (2011) The Worldwide Epidemiology of Acute Rheu-

matic Fever and Rheumatic Heart Disease. *Clinical Epidemiology*, **3**, 67-84. https://doi.org/10.2147/CLEP.S12977

[2] Shah, D., Patel, S. and Bharti, S.K. (2020) Heart Disease Prediction Using Machine Learning Techniques. *SN Computer Science*, **1**, 1-6. https://doi.org/10.1007/s42979-020-00365-y

[3] Uddin, S., Khan, A., Hossain, M.E. and Moni, M.A. (2019) Comparing Different Supervised Machine Learning Algorithms for Disease Prediction. *BMC Medical Informatics and Decision Making*, **19**, Article No. 281. https://doi.org/10.1186/s12911-019-1004-8

[4] Ramalingam, V., Dandapath, A. and Karthik, R.M. (2018) Heart Disease Prediction Using Machine Learning Techniques: A Survey. *International Journal of Engineering & Technology*, **7**, 684-687. https://doi.org/10.14419/ijet.v7i2.8.10557

[5] Kavitha, M., Gnaneswar, G., Dinesh, R., Sai, Y.R. and Suraj, R.S. (2021) Heart Disease Prediction Using Hybrid Machine Learning Model. 2021 6*th International Conference on Inventive Computation Technologies* (*ICICT*), Coimbatore, 20-22 January 2021, 1329-1333. https://ieeexplore.ieee.org/document/9358597 https://doi.org/10.1109/ICICT50816.2021.9358597

[6] Mohan, S., Thirumalai, C. and Srivastava, G. (2019) Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. *IEEE Access*, **7**, 81542-81554. https://doi.org/10.1109/ACCESS.2019.2923707

[7] Venkatalakshmi, B. and Shivsankar, M.V. (2014) Heart Disease Diagnosis Using Predictive Data Mining. *International Journal of Innovative Research in Science, Engineering and Technology*, **3**, 1873-1877. https://api.semanticscholar.org/CorpusID:44216820

[8] Patel, S., Kumar Yadav, P. and Shukla, D. (2013) Predict the Diagnosis of Heart Disease Patients Using Classification Mining Techniques. *IOSR Journal of Agriculture and Veterinary Science*, **4**, 61-64. https://doi.org/10.9790/2380-0426164

[9] Khateeb, N. and Usman, M. (2017) Efficient Heart Disease Prediction System Using K-Nearest Neighbor Classification Technique. *Proceedings of the International Conference on Big Data and Internet of Thing*, London, 20-22 December 2017, 21-26. https://doi.org/10.1145/3175684.3175703

[10] Lu, P., *et al.* (2018) Research on Improved Depth Belief Network-Based Prediction of Cardiovascular Diseases. *Journal of Healthcare Engineering*, **2018**, e8954878. https://doi.org/10.1155/2018/8954878

[11] Marikani, T. and Shyamala, K. (2017) Prediction of Heart Disease Using Supervised Learning Algorithms. *International Journal of Computer Applications*, **165**, 41-44. https://doi.org/10.5120/ijca2017913868

[12] Dangare, C.S. and Apte, S.S. (2012) Improved Study of Heart Disease Prediction System Using Data Mining Classification Techniques. *International Journal of Computer Applications*, **47**, 44-48. https://doi.org/10.5120/7228-0076

[13] Bhatla, N. and Jyoti, K. (2012) An Analysis of Heart Disease Prediction Using Different Data Mining Techniques. *International Journal of Engineering*, **1**, 1-8.

[14] Anbarasi, M., Anupriya, E. and Iyengar, N. (2010) Enhanced Prediction of Heart Disease with Feature Subset Selection Using Genetic Algorithm. *International Journal of Engineering Science and Technology*, **2**, 5370-5376.

[15] Kim, J., Lee, J. and Lee, Y. (2015) Data-Mining-Based Coronary Heart Disease Risk Prediction Model Using Fuzzy Logic and Decision Tree. *Healthcare Informatics Research*, **21**, 167-174. https://doi.org/10.4258/hir.2015.21.3.167

[16] Mansoor, H., Elgendy, I.Y., Segal, R., Bavry, A.A. and Bian, J. (2017) Risk Prediction

Model for In-Hospital Mortality in Women with ST-Elevation Myocardial Infarction: A Machine Learning Approach. *Heart & Lung*, **46**, 405-411. https://doi.org/10.1016/j.hrtlng.2017.09.003

[17] Mustaqeem, A., Anwar, S.M., Majid, M. and Khan, A.R. (2017) Wrapper Method for Feature Selection to Classify Cardiac Arrhythmia. 2017 39*th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (*EMBC*), Jeju, 11-15 July 2017, 3656-3659. https://ieeexplore.ieee.org/document/8037650

[18] Toshniwal, D., Goel, B. and Sharma, H. (2015) Multistage Classification for Cardiovascular Disease Risk Prediction. In: Kumar, N. and Bhatnagar, V., Eds., *Big Data Analytics*, Springer, Berlin, 258-266. https://doi.org/10.1007/978-3-319-27057-9_18

[19] Forsen, H., *et al.* (2017) Evaluation of Machine Learning Methods to Predict Coronary Artery Disease Using Metabolic Data.

[20] Palaniappan, S. and Awang, R. (2008) Intelligent Heart Disease Prediction System Using Data Mining Techniques. 2008 *IEEE/ACS International Conference on Computer Systems and Applications*, Doha, 31 March-4 April 2008, 108-115. https://doi.org/10.1109/AICCSA.2008.4493524

[21] Long, N.C., Meesad, P. and Unger, H. (2015) A Highly Accurate Firefly Based Algorithm for Heart Disease Prediction. *Expert Systems with Applications*, **42**, 8221-8231. https://doi.org/10.1016/j.eswa.2015.06.024

[22] Guvenir, H.A., Acar, B., Demiroz, G. and Cekin, A. (1997) A Supervised Machine Learning Algorithm for Arrhythmia Analysis. *Computers in Cardiology*, Lund, 7-10 September 1997, 433-436.

[23] Samad, S., Khan, S.A., Haq, A. and Riaz, A. (2014) Classification of Arrhythmia. *International Journal of Electrical Energy*, **2**, 57-61. https://doi.org/10.12720/ijoee.2.1.57-61

[24] Khan, A., Khan, S.A., Shaukat, A. and Akhtar, M. (2015) Identifying Best Feature Subset for Cardiac Arrhythmia Classification. 2015 *Science and Information Conference* (*SAI*), London, 28-30 July 2015, 494-499.

[25] Aravinthan, K. and Vanitha, M. (2016) A Comparative Study on Prediction of Heart Disease Using Cluster and Rank Based Approach. *International Journal of Advanced Research in Computer and Communication Engineering*, **5**, 421-424.

[26] Dwivedi, A.K. (2016) Performance Evaluation of Different Machine Learning Techniques for Prediction of Heart Disease. *Neural Computing and Applications*, **29**, 685-693. https://doi.org/10.1007/s00521-016-2604-1

[27] Otoom, A., Abdallah, E., Kilani, Y., Kefaye, A. and Ashour, M. (2015) Effective Diagnosis and Monitoring of Heart Disease Effective Diagnosis and Monitoring of Heart Disease. *International Journal of Software Engineering and Its Applications*, **9**, 143-156.

[28] Ali, M.M., Paul, B.K., Ahmed, K., Bui, F.M., Quinn, J.M.W. and Moni, M.A. (2021) Heart Disease Prediction Using Supervised Machine Learning Algorithms: Performance Analysis and Comparison. *Computers in Biology and Medicine*, **136**, Article ID: 104672. https://doi.org/10.1016/j.compbiomed.2021.104672

[29] Singh, A. and Kumar, R. (2020) Heart Disease Prediction Using Machine Learning Algorithms. 2020 *International Conference on Electrical and Electronics Engineering* (*ICE*3), Gorakhpur, 14-15 February 2020, 452-457. https://ieeexplore.ieee.org/abstract/document/9122958 https://doi.org/10.1109/ICE348803.2020.9122958

[30] Pouriyeh, S., Vahid, S., Sannino, G., De Pietro, G., Arabnia, H. and Gutierrez, J. (2017) A Comprehensive Investigation and Comparison of Machine Learning

Techniques in the Domain of Heart Disease. 2017 *IEEE Symposium on Computers and Communications* (*ISCC*), Heraklion, 3-6 July 2017, 204-207. https://doi.org/10.1109/ISCC.2017.8024530

[31] Fitriyani, N.L., Syafrudin, M., Alfian, G. and Rhee, J. (2020) HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System. *IEEE Access*, **8**, 133034-133050. https://doi.org/10.1109/ACCESS.2020.3010511

[32] Shrivastava, P.K., Sharma, M., Sharma, P. and Kumar, A. (2022) HCBiLSTM: A Hybrid Model for Predicting Heart Disease Using CNN and BiLSTM Algorithms. *Measurement*: *Sensors*, **25**, Article ID: 100657. https://doi.org/10.1016/j.measen.2022.100657

[33] Doppala, B.P., Bhattacharyya, D., Chakkravarthy, M. and Kim, T. (2021) A Hybrid Machine Learning Approach to Identify Coronary Diseases Using Feature Selection Mechanism on Heart Disease Dataset. *Distributed and Parallel Databases*, **41**, 1-20. https://doi.org/10.1007/s10619-021-07329-y