PRINT ISSN 1119-8362
Electronic ISSN 2659-1499

Full-text Available Online at
https://www.ajol.info/index.php/jasem
https://www.bioline.org.br/ja

J. Appl. Sci. Environ. Manage.
Vol. 29 (1) 255-261 Jan. 2025

Predicting Heart Diseases by Selective Machine Learning Algorithms

[1]UMAR, N; [1]HASSAN, SK; [1]UMAR, A; [2]AHMED, SS

*[1]*Department of Computer Science; [2]Department of Statistics, Niger State Polytechnic, Zungeru, Niger State, Nigeria*

**Corresponding Author Email: nasirmaali67@gmail.com*
**ORCID: https://orcid.org/0009-0008-2076-1903*
**Tel: +2348035958967*

Co-Author Email: hsalihukuta@yahoo.com; umargud@gmail.com; ahmedsule710@gmail.com

ABSTRACT: Heart disease is among the leading causes of mortality worldwide. As a result, it's critical to diagnose patients appropriately and promptly. Consequently, the objective of this paper was to predict heart diseases using selective machine learning algorithms. The leverage technique was evaluated using the Cleveland heart disease dataset. In this study five classifiers were trained and tested with the unsmooth Cleveland dataset and the smooth Cleveland dataset. The results obtained showed all the classifiers performed better when tested with the smooth dataset with an accuracy of 98.11% than when tested with the unsmooth dataset with an accuracy of 89.71% The leverage technique performed better than works found in literature reviewed. These results show that feature engineering using data smoothing is effective for improved heart disease prediction.

DOI: https://dx.doi.org/10.4314/jasem.v29i1.32

Cite this Article as: UMAR, N; HASSAN, S. K; UMAR, A; AHMED, S. S. (2025). Predicting Heart Diseases by Selective Machine Learning Algorithms. *J. Appl. Sci. Environ. Manage.* 29 (1) 255-261

Dates: Received: 22 October 2024; Revised: 20 November 2024; Accepted: 28 December 2024; Published: 31 January 2025

Keywords: Heart Disease; Feature Improvement; Prediction; Data Smoothing; Feature Engineering.

Heart disease is among the leading causes of mortality worldwide (Sharma *et al.*, 2020). Any problem that affects the heart's ability to function normally is referred to as heart disease (Zhenya and Zhang, 2021). In heart disease, the heart generally fails to deliver enough blood to other regions of the body to allow them to operate normally. The narrowing and occlusion of coronary arteries causes heart failure (Muhammad *et al.*, 2020). Every year, an estimated 17 million individuals die from cardiovascular diseases, such as heart attacks and strokes, accounting for 31% of all fatalities worldwide (Dutta *et al.*, 2019). Heart disease is caused by a variety of variables, including personal and professional behaviors, as well as hereditary predisposition (Dutta *et al.*, 2019). Heart disease care and treatment is highly challenging, especially in poor nations, due to a lack of diagnostic instruments, physicians, and other resources, impairing appropriate prediction and treatment of cardiac patients (Pattnaik *et al.*, 2021; Sharma *et al.*, 2020). With these worries about cost and time, computer technology and machine learning approaches have recently been employed to produce software to assist clinicians in making preliminary decisions about heart disease (Muhammad, 2020). In line with this recent technology usage many academic and scientific researchers have worked on heart disease prediction using data mining models, however these researches have not paid much attention on the heart disease data attributes this is because much attention is paid to the classification models. Nevertheless, the performance of a classification model is dependent on the quality of features or attributes it is trained and tested on. A popular dataset used for heart disease prediction is the Cleveland heart disease dataset. Prediction using the Cleveland heart disease dataset has yielded low performance with accuracies ranging

**Corresponding Author Email: nasirmaali67@gmail.com*
**ORCID: https://orcid.org/0009-0008-2076-1903*
**Tel: +2348035958967*

from 76% to 89% for majority of the studies (Miao and Miao, 2018; Nandhini *et al.,* 2018) and 90% − 92% for few studies (Ali *et al.,* 2019). This low performance is because most of these studies focused on the classification without focusing on the feature set. Feature engineering has been identified by Uddin *et al.* (2018) to improve the quality of features which in turn improve classification model performance. Hence this research leverages a feature engineering method using the weighted moving average data smoothing technique to remove noise and capture important patterns in the dataset. Using the data mining tools Orange and Weka, Kodati and Vivekanandam (2018) investigated heart disease. In the Weka and Orange tools, Naive Bayes (NB), Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbor (KNN) classifiers were utilized for prediction. The models were also trained and tested using the Cleveland data set. Each of the data mining tools provided different prediction results for each classifier. However, the support vector machine produced the highest recall and precision of 83.7% and 84% respectively using the Weka tool. A drawback of this study is that the model accuracy was not measured which is critical for determining the system's performance. Hussein (2021) also performed heart disease prediction using Weka data mining tool. In this study two algorithms namely decision tree J84 and Nave Bayesian were analyzed and compared for heart disease prediction. The performance evaluation of the two algorithms were carried out using the Cleveland dataset from UCI repository. Before classification the dataset was preprocessed to remove missing data, and to remove outliers. Feature selection was also performed. Nave Bayes and J48 classier produced an accuracy of 83.70% and 76.66% respectively. This low performance could be as a result of important features been removed as outliers, or as a result of selecting features that are not optimal. Consequently, the objective of this paper was to predict heart diseases using selective machine learning algorithms.

MATERIALS AND METHODS

The processes involved in diagnosis of heart disease starts with dataset (Figure 1) followed by feature engineering. Under feature engineering is the data imputation and data smoothing. The data imputation come before the data smoothing. The imputed dataset is sent directly to the classifier for prediction and the smooth dataset is also feed to the same classifier for prediction.

Classification was done using five classifiers and the performance of each of these classifiers was evaluated.

Dataset: The Cleveland heart disease data set was used in this investigation. The data set can be found via the UCI machine learning repository and the UCI Kaggle repository, both of which are open to the public (Janosi *et al.,* 2018). The Cleveland dataset consist of 303 instances. Each instance has distinct 13 attributes along with its target labels. The dataset is composed of two classes which are: present (1) or absent (0) of heart disease.

Data Imputation: The process of replacing missing data with statistical approximations of the missing values is known as data imputation. In this paper the Nearest Neighbor (NN) technique was used to carry out the imputation. NN imputation is a method for imputing a new observation by locating the observations in the training set that are nearest to it and averaging them to fill in the value (Beretta and Santaniello, 2016). The outcomes of NN algorithms, are dependent on the distance metrics used. The Minkowski distance was utilized as the distance metric in this study. The Minkowski distance of order b between two points $X = (x_1,x_2,...,x_n)$ and $Y = (y_1,y_2,...,y_n)$ is given in Equation 1 (Merigo, 2012; Olakoslu, 2020).

$$MD\ (X,Y) = \left(\sum_{i-1}^{n} |\ x_i - y_i|^b\right)^{\frac{1}{b}} \quad (1)$$

The benefits of utilizing NN imputation are its ease of implementation and speed in acquiring a complete data set (Kim *et al.,* 2019).

Data Smoothing: After performing data imputation on the dataset, the next preprocessing step was data smoothing. Data smoothing refers to a statistical approach of eliminating noise or outliers from datasets to make the patterns more noticeable (Adebola and Timothy, 2017). Data smoothing is known to allow the important patterns of data to stand out. In this study the weighted moving average technique was using for data smoothing.

Moving Average Smoothing Technique: Moving averages are used to examine data points by calculating the averages of various subsets of the entire data set (Adebola and Timothy, 2017; Shastri *et al.,* 2018). A moving average is a sort of finite impulse response filter. Given a sequence of values and a fixed subset size, the first component of the moving average is generated by taking the mean of the initial fixed subset of the number series. The subset is then changed by shifting ahead that is, removing the first number in the sequence and adding the next value to the subset. The weighted moving average approach was used in this study.

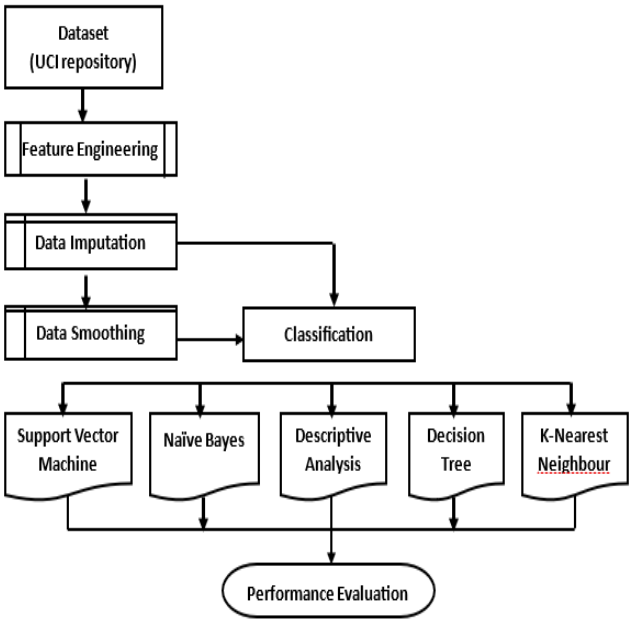

Fig. 1: Block diagram of all the processes involved in diagnosis of heart disease.

The weighted moving average is an average that uses multiplying parameters to give data at different points in the sample window distinct weights (Ekhosuehi, 2016). The weighted moving average formula is expressed in Equation 2.

$$M = \frac{\sum_{t=1}^{n} W_t * V_t}{\sum_{t=1}^{n} W_t} \quad (2)$$

Where *M* denotes the average value, *V* denotes the actual value, *W* denotes the weighting parameter, and *n* is the number of periods in the weighting group.

Data Classification: Machine learning capability lies in its ability to generalize by correctly classifying unknown information based on models developed using the training dataset (England, 2016).

Support Vector Machine (SVM), Decision Tree (DT), Naive Bayes (NB), K-Nearest Neighbor (KNN) and Discriminate Analysis were among the machine learning classification models utilized for training and testing. Patients were divided into two groups in this study: those who had no heart disease (0) and those who had heart disease (1).

Support Vector Machine (SVM): SVM is based on the structural risk minimization principle, which allows it to compress an array of raw data into a support vector set and learn how to achieve a classification decision function (Cao *et al.*, 2020).

The decision function of the SVM in the input space is expressed in Equation 3.

$$\gamma = h(x) = sign\left(\sum_{j=1}^{n} u_j y_j K(x, x_j) + v\right) \quad (3)$$

K-Nearest Neighbor (KNN): In KNN an item is classified based on its distance from its neighbors, and it is allocated to the most common class of its *k* closest neighbors (Kataria and Singh, 2013). The Euclidean distance is used to calculate the linear distance between two points in KNN technique (Greene and Cunningham, 2008). If two vectors x_i) and x_j are given where $x_i = (x_{i1}, x_{i2}, \dots, x_{in})$ and $x_j = (x_{j1}, x_{j2}, \dots, x_{jn})$, Then the Euclidean distance between x_i and x_j is given in Equation 4:

$$ED(x_i, x_j) = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2} \quad (4)$$

Naive Bayes (NB): The NB model is a probabilistic learning technique based on the Bayes theorem and the assumption of great feature independence. Learning with NB classifiers involves a large number of linear parameters in the number of problem functions (Chen and Fu, 2018). The Bayes theorem offers a way to calculate the posterior probability $P(x|y)$ from $P(x)$, $P(y)$ and $P(y|x)$ NB. Equations 5 and 6 presents the equation for posterior probability $P(x|y)$.

$$P(x|y) = \frac{P(y|x) \times P(x)}{P(y)} \quad (5)$$

$$P(x|y) = \frac{P(y_1|x) \times P(y_2|x) \times \ldots \times P(y_n|x) \times P(x)}{P(y_1, \ldots, y_n)} \quad (7)$$

Decision Tree: The decision tree employs a tree-like structure to progress from observations about an item (represented by the branches) to inferences about the item's target value (defined in the leaves) (Alsagheer *et al.*, 2017). Entropy is a popular techniques used in determining which attribute to position at the root or the different levels of the tree (Rokach and Maimon, 2005). Entropy is a measure of randomness in processed information (Bhanushali *et al.*, 2015; Li and Song, 2007). In Equation 7, entropy for a single attribute is expressed.

$$E(S) = \sum_{i=1}^{n} -p_i log_2 p_i \quad (7)$$

Where S represents the present state, p_i is the probability of an event i of state S.

Discriminant Analysis: Discriminate analysis is a multivariate statistical method for producing a model for predicting group membership. Centered on a linear combination of predictive variables, the model includes discriminating functions that arise to provide the best group segregation.

These functions are based on a survey of available group memberships (Büyüköztürk and Çokluk Bökeoğlu, 2008). Discriminate analysis is enabled when continuous quantities are measured on independent variables for each calculation (Ghojogh and Crowley, 2019).

Performance Metric: Accuracy: This is the number of correct guesses divided by the total number of right forecasts. The exact formula is given in Equation 8:

$$\text{Accuracy} = \frac{\text{True Positive + True negative}}{\text{True Positive + True negative + False Positive + False negative}} \quad (8)$$

Precision: Precision is a metric used to calculate how many positive predictions are accurately made. The formula in equation is used to dene precision 9:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (9)$$

Recall: The amount of correct positive predictions that could have been made from all positive predictions is calculated by recall. The recall is calculated using the formula in equation 10.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (10)$$

F-Score: The f-score of a model is defined as the harmonic average of recall and precision. F-Score is represented in equation 11.

$$\text{F} - \text{Score} = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} \quad (11)$$

RESULTS AND DISCUSSION

This section presents the results obtained using the unsmooth and smooth Cleveland dataset. These unsmooth and smooth datasets were feed as input to five classification models: SVM, DA, DT, K-NN and NB for heart disease prediction task.

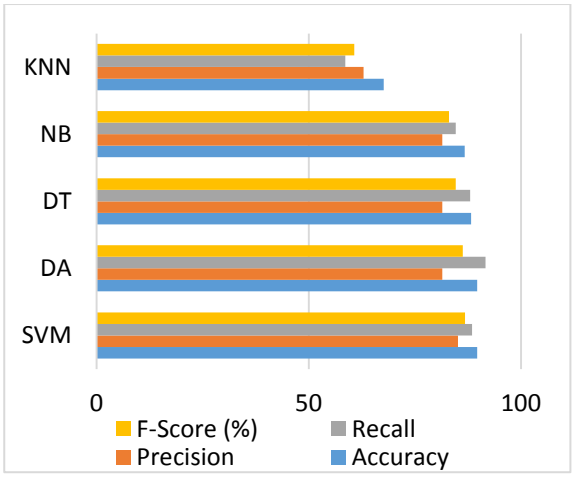

Fig. 2: Classification of the unsmoothed Cleveland dataset classification results

In terms of accuracy, precision, recall and f-score SVM and DA performed better than the other classifiers based on the results in Figure 1. After the SVM and DA with an accuracy of 89.71 % each, the next classifier in performance is the DT with an accuracy of 88.24%, followed by Nave Bayes with an accuracy of 86.76%. KNN had the least performance with an accuracy of 67.65%.

From the results shown in Figure 2 SVM achieved the best performance with an accuracy of 9811%., precision, of 100%, recall of 96.43% and F-score of 98.18% each. NB achieved the second best performance after the SVM is DA with an accuracy of 97.53%, precision of 100 %, recall of 96.43% and f-score of 98.18%. DT got accuracy of 89.71,

precision of 88.9%, recall of 85.71%, F-score of 87.27%, NB achieved accuracy 97.94%, precision of 96.3%, Recall of 96.3%, F-score of 74.58%. KNN attained the least performance in comparison with the other four classifiers. KNN achieved an accuracy of 77.94%, precision of 81.48%, recall of 68.75% and f-score of 74.58% which is lower than the values obtained by DA, DT, SVM and NB.

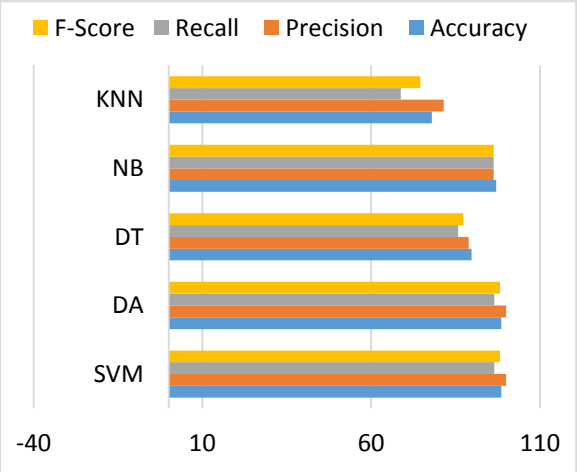

Fig. 3: Classification results using the Cleveland smooth preprocessed dataset

Table 1: Comparison of proposed system with related works

Algorithm	Dataset	Accuracy (%)
Hybrid Technique (Tarawneh and O. Embarak, 2019)	Cleveland	89.3
Random Forest (Pal and Parija, 2021)	Cleveland	86.9
SVM (Ware *et al.*, 2020)	Cleveland	86.07
Feature Selection + NB (Hussein, 2021)	Cleveland	83.7
SVM (Kodati and Vivekanandam, 2018)	Cleveland	83.7
Ensemble Classier (Latha and Jeeva, 2019)	Cleveland	85.48
Random Forest (Pattnaik *et al.*, 2021)	Cleveland	90.16
Data Smoothing + SVM	Cleveland	98.11

The data in Figure 3 show that data smoothing improved the performance of all the classifiers. For instance, the performance of SVM and DA were increased by 8.82% when tested with the smooth preprocessed dataset. DT improved in performance by 1.47% when tested with the smooth preprocessed dataset. NB improved in accuracy by 10.3% when tested with the smooth preprocessed dataset and KNN improved in performance by 10.29% when tested with the smooth preprocessed dataset. Table 1 is a comparison of the leverage Data smoothing +SVM with works found in literature reviewed on heart disease prediction using the Cleveland dataset.

Looking at the values of the performance measures in Table 3 the leverage technique achieved the highest accuracy of 98.11% than related works based on the Cleveland dataset. Hence the leverage technique performed better than previous works that used the Cleveland dataset for heart disease prediction. The results of the leverage technique is highlighted in bold. From the results shown in Table 1, 2, and 3 it can be seen that the data smoothing technique improved the dataset by filter noise and making important patterns to stand out which in turn improved the performance of all the classification models.

Conclusion: This study used the weighted moving average technique to smooth and improve the Cleveland features by removing noise from the dataset to render the patterns more visible. For future work additional data smoothing techniques such as random walk, exponential moving average, and simple exponential can be used. More heart disease dataset can be used to evaluate leverage technique.
Declaration of Conflict of Interest: The authors declare no conflict of interest.

Data Availability: Data are available upon request from the corresponding author.

REFERENCES

Adebola, FB; Timothy, EOT (2017). A new approach to smoothing time series data, *Int. J. Adv. Sci. Tech. Res.*, 5(1), 162-182.

Ali, L; Niamat, A; Khan, J; Amiri, GN; Xingzhong, X; Noor, A; Nour, R; Bukhari, SAC (2019). An Optimized Stacked Support Vector Machines Based Expert System for the Effective Prediction of Heart Failure. *IEEE Access*, 7(1), 54007-54014. Doi: 10.1109/ACCESS. 2019.2909969.

Alsagheer, RHA; Alharan, AFH; Al-haboobi, ASA (2017). Popular Decision Tree Algorithms of Data Mining Techniques A Review. *Int. J. Comput. Sci. Mob. Comput.*, 6(6), 133142.

Beretta, L; Santaniello, A (2016). Nearest neighbour imputation algorithms a critical evaluation. *BMC Med. Inform. Decis. Mak.*, 16(Suppl 3), 3-18. Doi: 10.1186/s12911-0160318- z.

Bhanushali, A; Mange, B; Vyas, H; Bhanushali, H; Bhogle, P. (2015). Comparison of Graphical Password Authentication Techniques. *Int. J. Comp. Appl.*, 116(1), 1114. Doi: 10.5120/20299-2332.

Buckberg, G; Nanda, N; Nguyen, C; Kocica, M (2018). What Is the Heart? Anatomy, Function, Pathophysiology, and Misconceptions. *J. Card. Dev. Dis.,* 5(2), 33. Doi: 10.3390 /jcdd 5020033.

Büyüköztürk, Ş; Çokluk, BO (2008). Discriminant function analysis: Concept and application, Egit. Arastirmalari. *Eurasian J. Educ. Res.,* 33(1), 7392.

Cao, J; Wang, M; Li, Y; Zhang, Q (2020). Improved support vector machine classification algorithm based on adaptive feature weight updating in the Hadoop cluster environment (pp. 112).

Chen, H; Fu, D (2018). An Improved Nave Bayes Classier for Large Scale Text. *Int. Conf. Art. Intel. Tech. Appl. (ICAITA),* 146, 33-36.

Dutta, A; Batabyal, T; Basu, M; Acton, ST (2019). Ancient convolutional neural network for coronary heart disease prediction. ArXiv. Doi: 10.2139 /ssrn.3514078.

Ekhosuehi, N (2016). On Forecast Performance Using a Class of Weighted Moving Average Processes for Time Series, *J. Nat. Sci. Res.* 6(13), 8792.

Englund, R (2016). Machine Learning for Technical Information. Quality Assessment.

Ghojogh, B; Crowley, M (2019). Linear and Quadratic Discriminant Analysis. *Tutorial*, 4(1), 116. Doi: http://arxiv.org/abs/1906.02590.

Greene, D; Cunningham, P (2017). Unsupervised Learning and Clustering.

Hussein, MTS (2021). Heart Diseases Prediction Using WEKA. *Baghdad Coll. Econ. Sci.,* 58(1), 23-41.

Janosi, A; Steinbrunn, M; William, P; Detrano, R (2018). *Heart Disease.* Kaggle: UCI. Doi: https://www.kaggle.com/ronitf/heart-disease-uci.

Kataria, A; Singh, MD (2013). A Review of Data Classification Using K-Nearest Neighbour Algorithm. *Int. J. Emer. Tech. Adv. Eng.,* 3(6), 354-360.

Kim, T; Ko, W; Kim, J (2019). Analysis and impact evaluation of missing data imputation in day-ahead PV generation forecasting. *Appl. Sci.,* 9(1), 118. Doi: 10.3390 /app 9010204.

Kodati, S; Vivekanandam, R (2018). Analysis of Heart Disease using in Data Mining Tools Orange and Weka. *Glob. J. Comput. Sci. Technol. C Softw. Data Eng.,* 18(1), 23-31.

Latha, CBC; Jeeva, SC (2019). Informatics in Medicine Unlocked Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. *Info. Med.,* 16(1), 100203. Doi: 10.1016 /j.imu.2019.100203.

Li, T; Song, J (2007). Construction of Decision Trees based Entropy and Rough Sets under Tolerance Relation. *Int. J. Comput. Intell. Syst.,* 1(1), 4-11. Doi: 10.2991 /iske. 2007.258.

Merigo, JM (2010). A New Minkowski Distance Based on Induced Aggregation Operators a New Minkowski Distance Based on Induced Aggregation Operators. *Int. J. Comp. Int. Sys.,* 4(2), 10-12. Doi: 10.1080/18756891.2011.9727769.

Miao, KH; Miao, JH (2018). Coronary heart disease diagnosis using deep neural networks. *Int. J. Adv. Comp. Sci. Appl.,* 9(10), 18. Doi: 10.14569 /IJACSA. 2018.091001.

Miao, KH; Miao, JH; Miao, GJ (2016). Diagnosing Coronary Heart Disease Using Ensemble Machine Learning, 7(10), 3039.

Muhammad, Y; Tahir, M; Hayat, M; Chong, KT (2020). Early and accurate detection and diagnosis of heart disease using intelligent computational model. *Sci. Rep.,* 10(1), 117. Doi: 10.1038/s41598-020-76635-9.

Nandhini, S; Debnath, M; Sharma, A; Pushkar, M (2018). Heart Disease Prediction Using Machine Learning. *Int. J. Recent Eng. Res. Dev.,* 3(10), 3946. Doi: 10.2139 /ssrn. 3167431.

Olakoslu, HB (2020). A generalization of the Minkowski distance and new denitions. *Turk. J. Math.,* 44(20), 319-333. Doi: 10.3906 /mat -1904-56.

Pal, M; Parija, S (2021). Prediction of Heart Diseases using Random Forest. *J. Phys. Conf. Series,* 1817(1), 1. Doi: 10.1088/1742-6596/1817/1/012009.

Pattnaik, PP; Padhy, SR; Mishra, BSP; Mishra, S; Mallick, PK (2021). Heart Disease Prediction Using Machine Learning Technique. *Lect. Notes*

Elect. Eng. (LNEE), 709(4), 193-201. Doi: 10.1007/978-981-15-8752-8_20.

Rokach, L; Maimon, O (2015). Decision Tree, in Data Mining and Knowledge Discovery Handbook. no. 2005, pp. 165192.

Sharma, V; Yadav, S; Gupta, M (2020). Heart Disease Prediction using Machine Learning Techniques. Proceedings of IEEE 2020 2nd Int. Conf. Adv. Comput. Commun. Control Networking, ICACCCN 2020, 1(6), 177181. Doi: 10.1109/ICACCCN51052.2020.9362842.

Shastri, S; Sharma, A; Mansotra, V; Sharma, A (2018). A Study on Exponential Smoothing Method for Forecasting. *Int. J. Comput. Sci. Eng.*, 6(4), 482485. Doi: 10.26438 /ijcse/v 6i4.482485.

Tarawneh, M; Embarak, O (2019). Hybrid Approach for Heart Disease Prediction Using Data Mining Techniques. *Lect. Notes Data Eng. Comm. Tech.*, 29(7), 447-454. Doi: 10.1007/978-3-030-12839-5_41.

Uddin, MF; Lee, J; Rizvi, S; Hamada, S (2018). Applied Sciences Proposing Enhanced Feature Engineering and a Selection Model for Machine Learning Processes. *Appl. Sci.* 8(1), 646, 132. Doi: 10.3390 /app 8040646.

Ware, S; Rakesh, S; Choudhary, B (2020). Heart Attack Prediction by using Machine Learning Techniques. *Int. J. Recent Technol. Eng.,* 8(5), 15771580. Doi: 10.35940 /ijrte.d 9439.018520.

Zhenya, Q; Zhang, Z (2021). A hybrid cost-sensitive ensemble for heart disease prediction. *BMC Med. Inform. Decis. Mak.*, 21(1), 122. Doi: 10.1186/s12911-021-01436-7.