Diagnosis of Cardiovascular Diseases Using Classification Algorithms

Mehmet Akif Tanisik
[1]Department of Computer Science and Engineering
International University of Sarajevo
Bosnia & Herzegovina
matnsk@outlook.com

Emine Yaman
[1]Department of Computer Science and Engineering
International University of Sarajevo
Bosnia & Herzegovina,
eyaman@ius.edu.ba

Ali Almisreb
[2]Faculty of Engineering and Natural Sciences
International University of Sarajevo
Bosnia and Herzegovina
aalmisreb@ius.edu.ba

[3,4,5,*]Nooritawati Md Tahir
[3]College of Engineering
Universiti Teknologi MARA, Malaysia
[4]Institute for Big Data Analytics and Artificial Intelligence (IBDAAI)
Universiti Teknologi MARA,
Selangor, Malaysia.
[5]Integrative Pharmacogenomics Institute (iPROMISE)
Universiti Teknologi MARA,
Malaysia.
*nooritawati@ieee.org

2022 IEEE Symposium on Wireless Technology & Applications (ISWTA) | 978-1-6654-8482-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISWTA55313.2022.9942793

***Abstract*—Heart diseases are the most common diseases in the world and will continue to be the number one cause of death for a long time. Each year 17.9 million people die due to cardiovascular diseases (CVDs), an estimated 32% of all deaths worldwide. However, many heart disease factors are preventable or treatable. If these factors are prevented or treated, it is an excellent opportunity to reduce the loss of life due to heart diseases. Nowadays, data science is actively used by people, and the importance of data science is increasing daily. It is vital for humanity that heart diseases and similar medical problems can be predicted using data science. For this reason, early disease detection aims to apply statistical methods in medicine. This research determines the relation between heart diseases and other human body characteristics to early diagnosis of heart diseases. In this research, data mining approaches specifically using different data science algorithms were applied to predict patients' heart diseases, namely Naïve Bayes, Logistic Regression, Multilayer Perceptron, and Random Forest algorithms for classification and diagnosis of cardiovascular diseases prediction. Results showed that the Naïve Bayes algorithm obtained an accuracy of 88.5% and was the best among all other algorithm.**

Keywords— Heart Disease Prediction, Naïve Bayes, Logistic Regression, Random Forest, Machine Learning, Classification

I. INTRODUCTION

The heart is one of the human body's essential components of the circulatory system. Also, despite its small size, it is the strongest muscle in the body. The heart begins to beat while a fetus in the mother's womb performs an average of a hundred thousand beats per day in a healthy adult body. The heart beats approximately two and a half billion times in an average human life and pumps the necessary clean blood to every body part with each beat [1]. About five litres of blood per minute passes through the heart and is distributed throughout the body. However, the heart is not fed by the blood passing through it. Some veins feed the heart, which is intense and pegs away at its job. These vessels are called coronary vessels. The disease that occurs due to narrowing or occlusion of the coronary vessels is called coronary artery disease [2]. Coronary artery disease is the root of most cardiovascular diseases [3]. According to the World Health Organization (WHO) data, deaths as a result of cardiovascular diseases take the first place among all deaths with 32 percent [4].

Myocardial infarction (MI) is one of the coronary artery diseases and is popularly called a heart attack. Myocardial infarction (MI) occurs due to blood building up in a Heart or impaired oxygenation in the heart. In other words, it is due to the inability of sufficient pumping blood to the heart and thus adequate oxygen to the heart. Damage occurs in a heart when there is insufficient oxygen, and death occurs in a heart when there is no sufficient oxygen for a long time [5]. The leading causes of this disease, which pose a risk of death or adversely affect human health [6] includes obesity, hypertension, diabetes, cholesterol, gender, age, tobacco use, alcohol use and family history.

In addition to these substances, drug use, stress, and intense lifestyle are among the factors that trigger the disease [7]. The diagnosis of heart attack can be made according to the results of physical examination and tests such as creatinine kinase, troponin, myoglobin, and electrocardiography (ECG) [8]. The diagnosis of the disease is made only by doctors today. But we can expect data scientists to help doctors in this regard with the power of data mining approaches. The medical world is very rich in data. However, the data usage and studies about medical data are insufficient, as expected. In recent years, many algorithmic and statistical analyses have been conducted on CVDs. The presented studies can be shown as examples of how data processing and information extraction can yield valuable results that can be used in medicine. As the number of studies on cardiovascular diseases increases, it will provide more support for preventing heart diseases.

Therefore, this research aimed to determine the relationship between age, gender, chest pain type, resting blood pressure, cholesterol value, fasting blood glucose, resting electrocardiographic result, maximum heart rate, angina, ST (Sinus Tachycardia), depression, ST segment slope and diagnosis of a heart attack. In this study, the data

were taken from the UCI database for pre-processing data with data mining for diagnosis prediction. In the second section, related works about heart disease predictions are explained. The third section describes the materials and methods used in this study. The analysis results are presented in the fourth section, along with discussions, conclusions, and future work.

II. Relatad Works

In recent years, Data Mining and Machine Learning techniques have been applied to many areas of health informatics and several industries since valuable information can be identified using data mining techniques. The applications of artificial decision support systems in the field of health informatics are increasing daily. Artificial intelligence-based decision support systems are also becoming widespread to help diagnoses and treatments made by data analysts to prevent human-induced errors. In addition to algorithmic studies aimed at detecting cardiovascular diseases in advance, various statistical analyses have also been used. Many studies are conducted under the discipline of artificial intelligence (AI) and pattern recognition for CVDs predictions.

R. Perumal [9] applied a prediction model to the UCI-Cleveland heart disease dataset of 303 instances and 14 attributes. The proposed heart disease prediction model used principal component analysis (PCA) as a feature reduction and standardization technique. Results showed that the Logistic Regression classifier accuracy was 87%, Support Vector Machine (SVM) was 85%, and the lowest was the K-Nearest Neighbor classifier with 69% accuracy. Next, C. B. C. Latha [10] applied comparative analysis of various classification techniques to increase the predictive accuracy of MI risk prediction using ensemble algorithms and procedures on the UCI-Cleveland heart disease dataset with 303 instances and 14 attributes, specifically Bayes Net, Naïve Bayes, Random Forest, C4.5, Multilayer Perceptron, and PART. In addition, Boosting, Bagging, Stacking, and Majority Vote were also used to increase the predictive accuracy of the classifier algorithms. It was found that the accuracy of the weak classifier was increased by 7.26%, namely 85.48% accuracy, using nine attributes with Majority Vote for Naive Bayes, Random Forest, Bayes Net, and Multilayer Perceptron algorithms. Further, D. Ananey-Obiriet [11] applied three classification algorithms, namely Logistic Regression, CART, and Gaussian Naïve Bayes Model, to the UCI-Cleveland heart disease dataset. Results showed that both Logistic Regression and Gaussian Naïve Bayes Model obtained 82.75% accuracy, and the CART algorithm accuracy was 79.31%. However, the AUC ROC value for the Gaussian Naïve Bayes Model was higher than the Logistic Regression algorithm. CART algorithm accuracy was less than the other two algorithms due to the sample size of the dataset.

Conversely, A. Gupta [12] preprocessed the dataset with Data Imputation, Data Standardization, and Data Stratification to replace the missing values for the UCI-Cleveland heart disease dataset to develop a model for classifying the MI risk prediction model and trained using Logistic Regression, K-Nearest Neighbor, Support Vector Machine, CART, and Random Forest classifiers based on FAMD method. Random Forest achieved the best accuracy of 93.44% compared to other classifiers. S. Mohan [13] applied HRFLM, a hybrid Random Forest algorithm with Linear Model, to increase the accuracy of MI risk prediction using the UCI-Cleveland heart disease dataset based on eight different classifier algorithms. Results showed that Random Forest and Linear Model are the best among the eight classifier algorithms. The Linear Model method was the best compared to the CART and Random Forest methods and combined Random Forest and Linear Model methods to propose the HRFLM method to improve the results. S. Kodati [14] applied various data mining classifying algorithms to develop an MI risk prediction system using the UCI-Cleveland dataset based on Orange and Weka data mining tools. The main idea was to compare different data mining tools according to their classification precision and recall. Naïve Bayes precision is the highest compared to SVM, Random Forest, and K-Nearest Neighbor algorithms. Naïve Bayes precision was 0.824 with Orange and 0.837 with Weka, and the recall was 0.806 with Orange and 0.837 with the Weka. Weka showed higher precision and recall than the Orange data mining tool.

C. Gazeloglu [15] applied 18 machine learning algorithms using all these datasets specifically the UCI-Cleveland with 303 instances and 14 attributes, UCI-Hungarian with 294 instances and 14 attributes, Statlog with 261 instances and 14 attributes and Z-Alizadeh Sani with 303 instances and 55 attributes. The proposed system has a two-tier ensemble built upon three different classifier ensembles, Random Forest, Gradient Boosting Machine, and Extreme Gradient Boosting Machine, in a stacked manner. The best prediction performance was the Z-Alizadeh Sani dataset, with an accuracy of 83.91% using different numbers of particles in PSO.

III. Materials and Methods

This part discusses the methodology used. The data mining techniques e used in this study are also elaborated.

A. WEKA Data Mining Tool

Weka was developed for machine learning and data mining at the University of Waikato, including machine learning algorithms and methods for data mining. The software is produced in Java language and can be integrated into a different application written in Java. Weka can operate Classification, Clustering, and Association Rule Mining with its pre-written algorithms inside and can also be used for data pre-processing and visualization purposes [18].

B. Dataset

In this work, the dataset used is the UCI Machine Learning repository, namely the Heart Disease Dataset, which consists of four different datasets. Each dataset contains information from several patients. The patients' information was obtained from four different hospitals namely Cleveland Clinic Foundation, Hungarian Institute of Cardiology, Budapest, V.A. Medical Centre, Long Beach, CA, and University Hospital, Zurich, Switzerland. Each dataset has the same instance format, precisely 76 attributes such as age, sex, cp (chest pain type), and many more. Table 1 shows the selected fourteen (14) attributes used in this study.

Table 1: Heart Disease Features of Cleveland Dataset

Attribute & Description	Type of Attribute	Value & Description
Age in years	Numeric	29-77 In range between
Sex Gender	Nominal	0,1 0=female 1=male
Cp Chest paint type	Nominal	1,2,3,4 1= typical angina 2= atypical angina 3= non-angina pain 4=asymptomatic
Trestbps Resting blood pressure (in mm Hg on admission to hospital)	Numeric	94-200 In range between
Chol Serum cholesterol in mg/dl	Numeric	126-564 In range between
Fbs Fasting blood sugar > 120 mg/dl	Nominal	0,1 0=false 1=true
Restecg Resting electrocardiographic results	Nominal	0,1,2 0=normal 1=ST-T wave Abnormality 2=definite left ventricular hypertrophy
Thalach Maximum heart rate achieved	Numeric	71-202 In range between
Exang Exercise induced angina	Nominal	0,1 0=false 1=true
Oldpeak ST depression induced by exercise relative to rest	Numeric	0-6.2 In range between
Slope The slope of the peak exercise ST segment	Nominal	1,2,3 1=upsloping 2=flat 3=down sloping
Ca Number of major vessels (0-3) colored by fluoroscopy	Nominal	0-3 In range between
Thal Heart status	Nominal	3,6,7 3=normal 6=fixed defect 7=reversible defect
Class Diagnosis	Nominal	0,1 0=false 1=true

C. Techniques and Methods

This study involved two stages, specifically data-preprocessing and application of different data mining models to train and test the proposed models, namely Naïve Bayes, Logistic Regression, Multilayer Perceptron, and Random Forest. During preprocessing, some of the data mining techniques were applied to get a better and workable dataset and minimize the outliers and missing values.

i. Data Preprocessing

Recall that data preprocessing is related to handling missing values and outliers. This process should be done before any methods are applied. The process contains techniques such as completing missing data, eliminating inconsistencies, and removing noise to detect outliers. Some of the methods that can be used for missing data are instances with missing values detected, missing values replaced with the average or the median of the attribute instead of missing values, missing values replaced with the average of their classes instead of missing values or missing values replaced with the appropriate value produced using methods such as regression instead of missing values [16]. As for outlier's detection, the techniques that can be used are binning, clustering, or regression methods. For instance, some of the missing values in the UCI-Cleveland heart disease dataset [17] are "ca" and "thal" nominal attributes. These missing values are replaced based on the majority mark. The value of the "ca" was '0' as the majority mark and had four missing values. This value was gathered from 172 observations out of 303. The value of the "thal" was '3' as the majority mark and had two missing values. This value was gathered from 168 observations out of 303. In the end, the missing Values of "ca" and "thal" were replaced with the values of '0' and '3' according to the majority mark.

ii. Naïve Bayes Algorithm

Naive Bayes is used as a classification technique for data mining approaches. Naïve Bayes is predicated on Bayes Theorem. This statistical algorithm assumes no dependency between the dataset classes, which are attributes, and works well with big datasets. Equation (1) is the Naïve Bayes algorithm formula:

$$P(A\backslash B) = (P(B\backslash A) * P(A))/P(B) \quad (1)$$

where $P(A)$ is the independent probability (class prior probability) of the class A, $P(B)$ is the independent probability (predictor prior probability) of the class B, $P(A\backslash B)$ is if P(B) satisfies the situation, it is the probability (posterior probability) of class A and (P(B\A) is the likelihood [19].

iii. Logistic Regression Algorithm

Logistic Regression (LR) is used for machine learning approaches for discriminative models and multi-class classification models such as multinomial Regression. However, the limitation of this algorithm is that the algorithm learns only in linear decision boundaries. Logistic Regression cannot handle missing values like Naive Bayesian, which sometimes fails for the prediction of class labels [20]. Equation (2) is the Logistic Regression algorithm formula.

$$m = \frac{1}{1 + e^{(\beta 0 + \beta 1 * X1 + \cdots + \beta t * Xt)}} \quad (2)$$

iv. Multilayer Perceptron Algorithm

Multilayer Perceptron (MLP) is based on the complex architecture of nodes that can learn in nonlinear decision boundaries. MLP contains an input layer, an output layer, and a hidden layer. MLP uses backpropagation for training the data [21]. Fig. 1 shows an example of a MLP artificial

neural network (ANN).

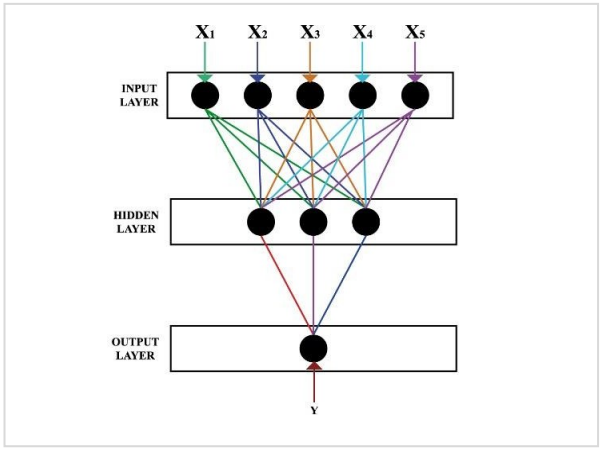

Fig 1: Example of MLP

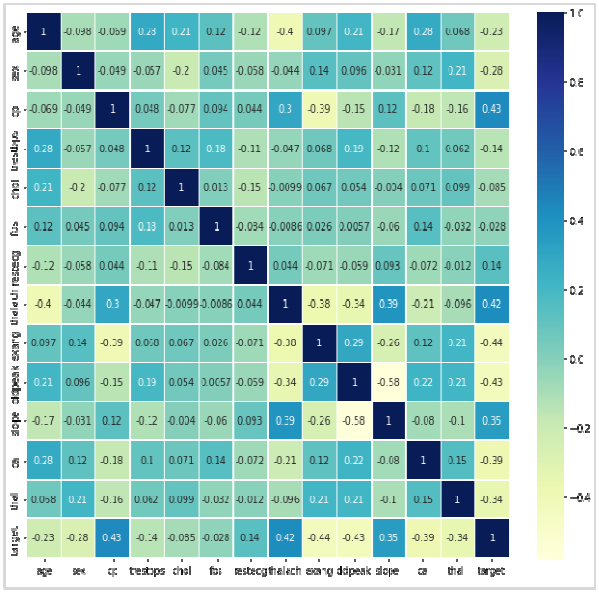

Fig. 2. Correlation Matrix of dataset

v. Random Forest

The Random Forest algorithm is used for classification and regression tasks in data mining based on tree decisions trained in different training sets rather than one single decision tree [22]. The standard is set by creating some calculations altogether trees. Then, the most used feature is selected by combining the features in other trees. The selected attribute is included in the tree, and the process is repeated at all levels [23].

IV. Results and Discussion

As mentioned earlier, this study used four classification algorithms to diagnose cardiovascular diseases. The performance measures to evaluate the proposed models' effectiveness are accuracy (Acc), true positive (TP), and false positive (FP) [24]. The dataset is divided into a ratio of 80:20, namely for training and testing.

Naïve Bayes algorithm predicted 54 correct predictions and 7 wrong predictions whilst Logistic Regression predicted 53 accurate predictions and 8 wrong predictions. As for MLP, this model predicted 52 correct predictions and 9 wrong predictions, and Random Forest predicted 51 correct predictions and 11 wrong predictions. Results showed that the Naïve Bayes algorithm predictions were the highest among all algorithms in this study. Further, Correlation analysis has been done to determine the correlation between the characteristics of the dataset and the diagnosis. The correlation matrix is shown in Fig. 2. As observed in Fig. 2, there is a high positive correlation between chest pain type, exercise-induced angina, and exercise-induced ST depression in the correlation of the characteristics with the diagnosis status. Hence it can be summarized that the chest pain type, exercise-induced angina, and exercise-induced ST depression affect the diagnosis. The detailed accuracies upon completing the Weka algorithm are as in Table 2.

Table 2: Performance Measure for each classifier

Classifier	Performance Measure (%)				
	Correctly Classified	Wrongly Classified	*Acc*	*TP*	*FP*
Naïve Bayes	**54**	**7**	**88.5**	**88.5**	**11.5**
Logistic Regression	53	8	86.9	86.9	13.1
Multilayer Perceptron	52	9	85.2	85.2	14.8
Random Forest	50	11	82.0	82.0	18.0

From the results obtained, Naïve Bayes is the best data mining algorithm with an accuracy of 88.5% to predict heart diseases for the UCI-Cleveland dataset. However, Logistic Regression (86.9%) and Multilayer Perceptron's (85.2%) accuracies were close to Naïve Bayes too. The naïve Bayes algorithm showed the best accuracy due to its learning mechanism. The dataset's attributes were not much dependent on each other, but the Naïve Bayes algorithm expects the features to be independent. As a result, the Naïve Bayes algorithm performed the best accuracy due to these independent attributes.

V. Conclusion

This study aims to predict heart diseases by determining the relationship between different characteristics of patients and heart diseases. Data mining classification algorithms such as Naïve Bayes, Logistic Regression, Multilayer Perceptron, and Random Forest have been used to determine the relationship between patient characteristics and cardiovascular diseases based on UCI- Cleveland heart disease dataset. In addition, the WEKA tool was used to run

algorithms for the classification tasks. Performance measures were used to evaluate the effectiveness of the classification algorithms. Originally, the UCI-Cleveland dataset had 76 attributes. However, only 14 attributes have been used. Results attained showed that Naïve Bayes had the highest accuracy of 88.5% to predict heart diseases for the UCI-Cleveland dataset. Future work includes utilizing several other classification algorithms like deep learning that could assist the medical practitioner in the early diagnosis of heart diseases.

ACKNOWLEDGMENT

This study is collaboration research between IUS and UiTM. The authors would like to thank the Research & Development Centre at IUS for all the support during this study. The conference registration fees are funded by Pembiayaan Yuran Prosiding Berindeks (PYPB), Tabung Dana Kecemerlangan Pendidikan (DKP), Universiti Teknologi MARA (UiTM), Malaysia.

REFERENCES

[1]"Body basics," Rady Children's Hospital-San Diego. [Online]. Available: https://www.rchsd.org/health-articles/heart-and-circulatory-system/. [Accessed: 13-Dec-2021].

[2] Quertermous, T., & Ingelsson, E. (2016). Coronary artery disease and its risk factors. Circulation Research, 118(1), 14–16. https://doi.org/10.1161/circresaha.115.307937

[3] Sweis, R. N., & Jivan, A. (2021, December 1). Overview of coronary artery disease - cardiovascular disorders. MSD Manual Professional Edition. Retrieved December 13, 2021, from https://www.msdmanuals.com/professional/cardiovascular-disorders/coronary-artery-disease/overview-of-coronary-artery-disease.

[4] World Health Organization. (n.d.). Cardiovascular diseases (cvds). World Health Organization. Retrieved December 13, 2021, from https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

[5] Storrow, A., & GIBLER, W. (2000). Chest pain centers: Diagnosis of acute coronary syndromes. Annals of Emergency Medicine, 35(5), 449–461. https://doi.org/10.1016/s0196- 0644(00)70006-0

[6] Hajar, R. (2017). Risk factors for coronary artery disease: Historical perspectives. Heart Views, 18(3), 109. https://doi.org/10.4103/heartviews.heartviews_106_17

[7] Vassalle, C., Petrozzi, L., Botto, N., Andreassi, M. G., & Zucchelli, G. C. (2004). Oxidative stress and its association with coronary artery disease and different atherogenic risk factors. Journal of Internal Medicine, 256(4), 308–315. https://doi.org/10.1111/j.1365-2796.2004.01373.x

[8] Al-Hadi HA, Fox KA. Cardiac markers in the early diagnosis and management of patients with acute coronary syndrome. Sultan Qaboos Univ Med J. 2009 Dec;9(3):231-46. Epub 2009 Dec 19. PMID: 21509305; PMCID: PMC3074795.

[9] Ramya Perumal, Kaladevi AC, "Early Prediction of Coronary Heart Disease from Cleveland Dataset using Machine Learning Techniques", IJAST, vol. 29, no. 06, pp. 4225 - 4234, May 2020.

[10] Latha, C. B., & Jeeva, S. C. (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. Informatics in Medicine Unlocked, 16, 100203. https://doi.org/10.1016/j.imu.2019.100203

[11] Ananey-Obiri, Daniel & Sarku, Enoch. (2020). Predicting the Presence of Heart Diseases using Comparative Data Mining and Machine Learning Algorithms. International Journal of Computer Applications. 176. 975-8887. 10.5120/ijca2020920034.

[12] A. Gupta, R. Kumar, H. Singh Arora and B. Raman, "MIFH: A Machine Intelligence Framework for Heart Disease Diagnosis," in IEEE Access, vol. 8, pp. 14659-14674, 2020, doi: 10.1109/ACCESS.2019.2962755

[13] S. Mohan, C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in IEEE Access, vol. 7, pp. 81542-81554, 2019, doi: 10.1109/ACCESS.2019.2923707

[14] Kodati, S., Vivekanandam, R., & Ravi, G. (2019). Comparative analysis of clustering algorithms with heart disease datasets using data mining weka tool. Advances in Intelligent Systems and Computing, 111–117. https://doi.org/10.1007/978-981-13-3600-3_11

[15] C. Gazeloglu, "Prediction of heart disease by classifying with feature selection and machine learning methods ", Progr Nutr, vol. 22, no. 2, pp. 660–670, Jun. 2020

[16] Tama, B. A., Im, S., & Lee, S. (2020). Improving an intelligent detection system for coronary heart disease using a two-tier classifier ensemble. BioMed Research International, 2020, 1–10. https://doi.org/10.1155/2020/9816142

[17] UCI. (n.d.). UCI Machine Learning Repository: Heart disease data set. Retrieved December 14, 2021, from https://archive.ics.uci.edu/ml/datasets/heart+disease.

[18] The University of Waikato. (n.d.). Weka Wiki. Retrieved December 14, 2021, from https://waikato.github.io/weka-wiki/.

[19] Marathe, N., Gawade, S., & Kanekar, A. (2021). Prediction of heart disease and diabetes using naive Bayes algorithm. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 447–453. https://doi.org/10.32628/cseit217399

[20] T., M., Mukherji, D., Padalia, N., & Naidu, A. (2013). A heart disease prediction model using SVM-decision trees-logistic regression (SDL). International Journal of Computer Applications, 68(16), 11–15. https://doi.org/10.5120/11662-7250

[21] Yan, H., Jiang, Y., Zheng, J., Peng, C., & Amp; LI, Q. (2006). A multilayer perceptron-based medical decision support system for heart disease diagnosis. Expert Systems with Applications, 30(2), 272–281. https://doi.org/10.1016/j.eswa.2005.07.022

[22] Fawagreh, K., Gaber, M. M., & Elyan, E. (2014). Random forests: From early developments to recent advancements. Systems Science & Control Engineering, 2(1), 602–609. https://doi.org/10.1080/21642583.2014.956265

[23] Jabbar, M. A., Deekshatulu, B. L., & Chandra, P. (2015). Prediction of heart disease using random forest and feature subset selection. Advances in Intelligent Systems and Computing, 187–196. https://doi.org/10.1007/978-3-319-28031-8_16

[24] Tan, P. N., Steinbach, M., Karpatne, A., & Kumar, V. (2019). 4. In Introduction to data mining (Second, pp. 313–330). essay, Pearson Education.