EARLY PREDICTION OF CARDIVASCULAR DISEASE USING ENSEMBLE MACHINE LEARNING TECHNIQUES

ABDULLAH SAAD AL-MALAISE AL-GHAMDI

Department of Information Systems, Faculty of Computing and IT, King Abdulaziz University, Jeddah, Saudi Arabia
Information Systems Department, HECI School, Dar Alhekma University, Jeddah 22246, Saudi Arabia
E-mail: aalmalaise@kau.edu.sa

Abstract - The most important organ in the human body is the heart, whose incorrect operation can lead to a variety of complications. The medical industry frequently sponsors campaigns to disseminate information, offer guidance, and discover a treatment for cardiovascular illnesses. The computing scholars are also doing their best to assist the healthcare industry by offering a variety of solutions utilizing cutting-edge technologies. Similar to that, this research proposed an ensemble machine learning framework to forecast a patient's likelihood of developing heart failure based on a variety of attributes. On a real-world dataset, we ran a number of tests using various machine learning algorithms to achieve this. In this research, individual machine learning algorithms like Random Forest, Decision Tree, and Logistics Regression were used as well as ensemble techniques. The findings indicated that, with an accuracy of 84.48%, the voting approach using all three algorithms outperformed. The results of this research may have a positive effect and provide medical professionals with guidance by highlighting the vital indicators that must be kept under close watch in order for them to make timely decisions. The findings of this research can aid the medical community in developing awareness and prevention campaigns to appropriately direct heart patients.

Keywords - Cardiovascular Disease, Machine Learning Algorithms, Heart Failure Factors, Ensemble Techniques.

I. INTRODUCTION

Cardiovascular disease is one of the leading causes of mortality in the world, accounting for nearly 17.9 million deaths each year. Heart disease can be caused by any condition that causes blood vessel blockages, including rheumatic heart disease, high blood pressure, and others[1]. However, if appropriate precautions and healthy lifestyle approaches are not implemented in our daily lives, the percentage of deaths can rise. According to research, even recent technological advancements can fail in dealing with heart disease issues if proper and healthy lifestyle choices are not made [2]. According to the World Health Organization (WHO), several factors are increasing the number of heart patients, including a lack of physical activity and an unhealthy diet, which increases abnormal blood pressure, high blood sugar, and obesity, all of which contribute to heart disease[3]. According to the earlier discussion, it is clear that a proper lifestyle should be adopted. From the perspective of a heart consultant, they can follow up and watch several factors in their patients, such as blood pressure, blood sugar, and other artery-related diseases, in order to make timely choices.A research [4]presented the effect of smoking that can cause heart problems. The research examined 41 chronic smokers' data using various attributes such as age, height, weight, and gender to determine the likelihood of heart disease. The findings indicated that arterial stiffness is more prevalent in smokers when compared to datasets from 116 nonsmokers. Furthermore, the hypertensive disease has been identified as a significant contributor to heart disease problems [5].The researcher discovered that, in addition to some common risk factors for heart disease, the detection of histological changes can be used to quickly identify a cardiovascular problem.

Several techniques and processes to treat heart disease have been suggested by researchers in recent years. In particular, the use of cutting-edge technology, prediction algorithms, and artificial intelligence methods can assist the health industry in dealing with deadly diseases. Machine Learning (ML), an important subfield of artificial intelligence methods, has been used extensively in the treatment of cardiovascular disease. In literature [6]A predictive ML algorithm used to forecast a patient's cardiac arrest. The researchers used a variety of ensemble approaches to enhance the model's accuracy and predict the likelihood of cardiac arrest in a patient.

The previous research showed that a number of risk factors can be used to calculate the likelihood of developing heart disease. While medical professionals can pay particular attention to those factors and guide patients appropriately. In literature, a variety of ML models, including Naive Bayes, Logistic Regression, Random Forest, Stacking, and others, have been used and applied to the potential risk data. The outcomes demonstrated the effective application of the stacking ensemble method, which gave patients with heart disease an 87.8% accurate prediction [6]. This demonstrates the effectiveness of ML in predicting a patient's level of risk for developing any heart-related disease. Additionally, numerous studies have used ML models to treat cardiovascular diseases. For instance, support vector regression for early cardiac disease prediction [7], risk in heart failure based on logistic regression model [8], ML algorithm based message alert system

for predicting heart problem [9], and an ensemble ML implementation for diagnosing artery disease [10].
Another study [11]discussed the heart problem associated with the pregnant women and called this disease as "Maternal Heart Disease" [6]. The research emphasized the lack of knowledge regarding maternal heart disease issues and suggested a framework to handle these types of patients. The research made use of the Korea NHI dataset, which contained data on nearly 174,926 patients. To determine the relationship between preterm birth and cardiovascular disease, a random forest technique built on machine learning was used. As 12234 of the 12701 women who gave birth to preterm babies had at least one heart condition, the research looked into the role that heart disease may play in preterm birth.There are several other studies who investigated that heart-related issues are commonly found in the pregnant women [12], [13]. In addition, ML algorithms are commonly used and applied for different industry-related dataset and provided a number of solutions for different problems [14]–[18].

Therefore, the primary goal of this research is to discuss the value of ML algorithms in cardiovascular disease prediction. Additionally, using real-world datasets and various ML algorithms, we applied the theory to predict risk linked with heart failure. The principal contributions of this study are highlighted in the list below:

- The experimental set-up for using various ML algorithms to forecast heart disease.
- The preprocessing of the collected dataset in order to prepare it for conducting the experiment.
- The use of ensemble approaches for improving the performance prediction and model accuracy.
- Experimental evaluation performed and comparison is presented using multiple assessments factors such as accuracy, precision, and recall.
- A comparison of the performance with and without ensemble techniques was used to highlight the performance.

The remainder of the paper is set as follows. The next part describes the precise experimental setup, a summary of the data, and an explanation of the machine learning techniques used in this study. Section 3 talks about the model's execution. Additionally, we review the findings and discussion in section 4. The paper ends with a discussion of the study's major accomplishments.

II. MATERIALS AND METHODS

2.1. Dataset Overview

The suggested model was built using a real-world dataset obtained from Kaggle and named the Heart Disease Dataset[19]. The dataset was collected initially by four different organizations [20]. The dataset includes a large number of attributes, but 14 distinct variables were chosen after modification, as shown in Table 1. This dataset contained 303 patient records with numerous attributes such as age, chest pain, blood pressure, blood sugar, and results of various tests such as thallium, fluoroscopy, ECG, and others.

Attribute	Description	Data Values
Age	The approximate age of the patient.	Different values from 29 to 77
CP	The chest pain values. The value identifies the chest pain severity and categorized in four different groups.	0, 1, 2, 3
Chol	The level of cholesterol in actual values.	Values between 126 to 564
RestECG	The test values of ECG, which categorized in different group based on its severity.	0, 1, 2
Exang	This parameter explains if the exercise can bring the angina pain or not. Yes=1, No=1	0, 1
Slope	To know the patient's health during the peak of the exercise.	1, 2, 3
Thal	This test is known as Thallium test. The test is useful for understanding the health-related issues.	0, 1, 2, 3
Gender	Gender of the patient. Male=1, Female=0	0, 1
RestBP	The level of blood pressure.	Values between 94 to 200
FBS	The level of blood sugar. Yes=1, No=0	0, 1
HeartBeat	The values of test for measuring heartbeat.	Values between 71 to 202
OldPeak	The level of depression in the patients.	Values between 0 to 6.2.
CA	The test value generated by fluoroscopy.	0, 1, 2, 3

Target	The label column in the dataset. This column categorized in two distinct values. This attribute describes the chances of heart failure. 0= No chances 1= Strong chances	0, 1

Table 1: Data Overview

2.2. Data Preprocessing

Preparing the dataset is an essential step before implementing the proposed framework. To review the data values in each attribute, a comprehensive method was used. We used the Rapid Miner tool for data preprocessing and model implementation, which is a popular tool for running machine learning, data science, and other algorithms[21]. The tools is commonly used by academic scholars for data preprocessing and ML model's execution [22]–[26]. First, the dataset was loaded into rapid miner using the "Read CSV" operator to examine the attributes and see if there were any missing values. As we got the cleaned version from the repository, the dataset found no missing values. Furthermore, an outlier detection algorithm is used to determine which factors have outlier values and to correct them. Finally, a data transformation method is used to convert values from numerical to binominal in order to make them suitable for ML model implementation.

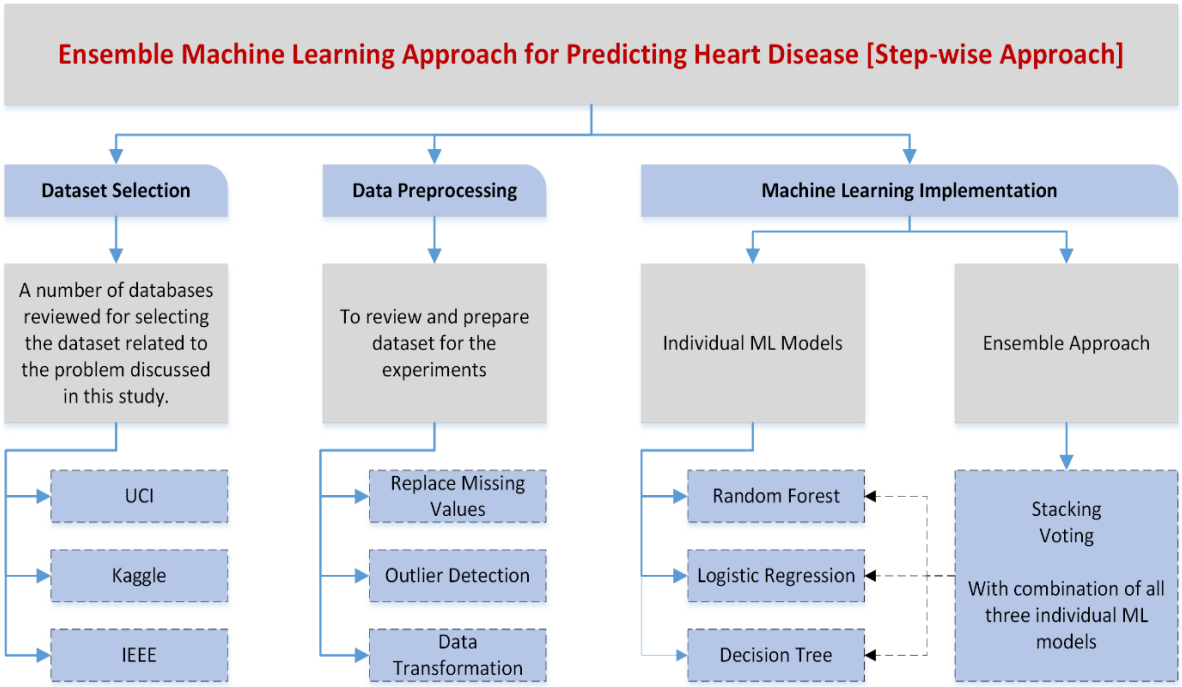


Fig. 1. Proposed Framework for Heart Disease Prediction

2.3. Methodology

According the proposed model as shown in Fig. 1 the number of ML algorithms used in this study for predicting the heart failure based on the given dataset. Rapid miner tool selected in this study, which is widelyused for training machine using health industry dataset [27], [28]. The suggested methodology included several stages, which are depicted in the proposed framework: data selection, data preprocessing, ML implementation, results, and discussion. This study chose three ML algorithms to forecast heart failure in a patient: Decision Tree (DT), Random Forest (RF), and Logistic Regression (LR). In addition, a pair of ensemble approaches (Stacking and Voting) were used to improve the model's accuracy. The findings indicated that using ensemble approaches resulted in a significant improvement in the model's performance.

III. MODEL IMPLEMENTATION

We ran nine different experiments in this research, each with a different ML algorithm. Specifically, the model was implemented in two stages: (i) using individual ML methods and (ii) using ensemble approaches. The rapid miner tool was used to carry out the tests. First, we used separate machine learning algorithms. (DT, RF, and LR). The dataset loaded, as shown in Fig. 2, was earlier loaded in the tool and retrieved from the local repository. The target variable, also known as the class variable, is then converted from "numerical to binominal".The primary goal of this conversion is to make it more useful for classification.

Initially, the target variable had two different values "0" and "1," which represented the dataset's classification into "No Heart Failure" and "Heart

Failure" respectively. Furthermore, the dataset was multiplied with the "Multiply" operator in order to pass it through three distinct algorithms. Finally, the "Cross Validation" approach was used for training and testing the model, which enables the execution of training and testing phases using the K-fold validation approach. We used a 10-fold cross validation approach to include as many transactions as possible in training and testing the model.The subsequent section discusses the outcomes of each experiment.

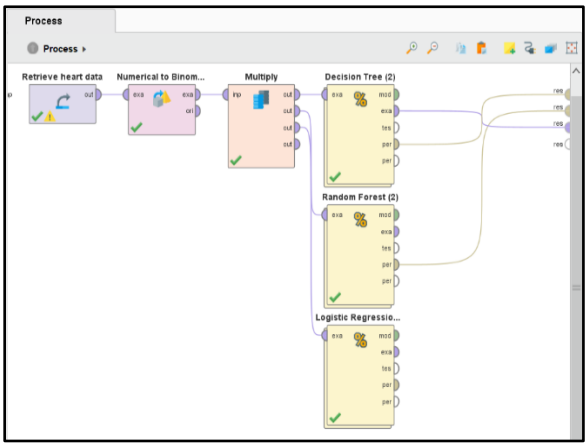


Fig.2. ML Implementation without Ensemble Approach

The experiment was carried out in the second stage using the first ensemble method known as stacking. Stacking is a machine learning technique that uses multiple base learners to train the model, as shown in Fig. 3. Following the execution of the multiple learners on the left side of the process, the right side of the process requires a unique stacking learner model. We ran three stacking experiments, with all three algorithms serving as base learners and each ML model serving as a stacking learner model in every experiment.

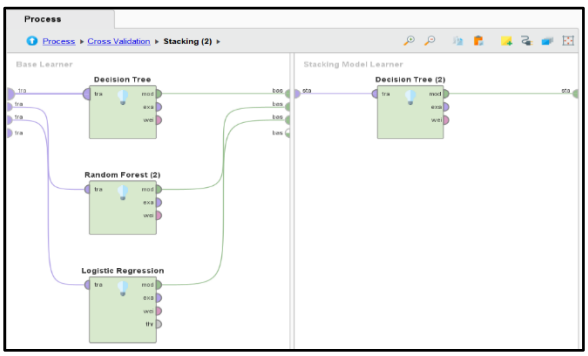


Fig.3. ML Implementation with Stacking Ensemble Approach

Finally, the voting approach was applied in the final phase of this research. This method, also known as the ensemble technique, employs a majority vote plan with multiple algorithms. The model can be applied to unseen data for testing purposes based on the results produced by inner sub-processes (ML algorithms). Figure 4 depicts the ensemble voting method in action. In this step, we tried three different ML algorithm combinations to assess and attain maximum accuracy.

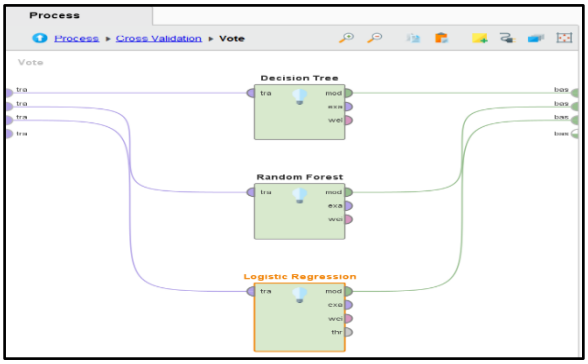


Fig.4. ML Implementation with Voting Ensemble Approach

IV. RESULTS AND DISCUSSION

This section describes the outcomes of all experiments carried out in this research. Rapid Miner includes a number of operators for measuring model performance using various evaluation matrices. The classification method was used to evaluate the model's performance on the target variable. The percentage of accurate predictions for each class is shown by the accuracy percentage. As previously discussed, the target variable represents the likelihood of heart failure in a patient using numerous independent variables in the dataset. Table 2 shows the performance of the various models. A variety of assessment metrics, such as precision, recall, and overall accuracy, are used to assess performance. Overall, the LR model did well in all three algorithms, with an accuracy of 82.15%.

Classifier	Precision	Recall	Overall Accuracy
Decision Tree	77.14%	73.40%	73.29%
Random Forest	81.95%	83.15%	80.53%
Logistic Regression	82.84%	84.85%	82.15%

Table 2: The Performance of ML Models without Ensemble

Table 3 depicts the performance of ML models using the stacking ensemble method. Multiple algorithms can be used as base learners in this method. It offers a comprehensive approach to training the machine that employs numerous techniques. As a stacking learner, we used three classifier combinations. We conducted three experiments for this, and the findings are shown in the table below. Stacking with random forest as model learner outperformed with 84.16% accuracy, as shown in the table. The precision and recall values were also higher than in other experiments, at 83.62% and 88.63%, respectively.

Classifier	Precision	Recall	Overall Accuracy
Stacking with Decision Tree as Model Learner	81.68%	81.90%	79.86%
Stacking with Random Forest as Model Learner	83.62%	88.63%	84.16%
Stacking with Logistic Regression as Model Learner	82.13%	77.67%	77.90%

Table 3: The Performance of ML Models with Stacking Ensemble Approach

Finally, in the final stage of this research, we used a voting ensemble method. We used different combinations of ML algorithms in this approach as well to discover the best model accuracy. Table 3 displays the precision, recall, and accuracy of each of the three voting models created in this study. Voting model-3 with decision tree and logistic regression did better than others in this study, with an accuracy of 84.48%. This result backs up the performance of the LR model, which had the best accuracy of any ML algorithm, as shown in Table 2. Finally, in all experiments, the voting ensemble approach gave the best solution and the greatest accuracy in predicting heart failure in patients based on various attributes



Fig. 5. Performance Comparison among all Experiments

Classifier	Precision	Recall	Overall Accuracy
Voting Model-1 [Decision Tree and Random Forest]	85.88%	70.42%	77.58%
Voting Model-2 [Decision Tree and Logistic Regression]	87.24%	67.37%	76.91%
Voting Model-3 [Decision Tree, Logistic Regression, and Random Forest]	84.24%	88.63%	84.48%

Table 3: The Performance of ML Models with Voting Ensemble Approach

Fig. 5 depicts the outcomes and comparisons of all experiments performed in this research. The diagram depicts the precision, recall, and total accuracy values attained by all models. To begin with, it is clear from the results that all of the models obtained accuracy greater than 70%. According to prior research, accuracy greater than 70% is appropriate for machine learning predictive models. We also used ensemble methods to improve the model's performance, and both approaches helped to improve predicting accuracy, as shown in the figure below. Altogether, stacking with RF and voting with DT, LR, and RF are among the best classifiers in this study. In which, voting with DT, LR, and RF outperformed among all models with 84.48%.

V. CONCLUSIONS

Due to an increasing number of patients, cardiovascular disease is one of the main concerns for the health industries. Medical specialists are constantly attempting to conduct awareness campaigns in order to guide patients in avoiding heart-related diseases. In this research, we offered an optimal solution for the healthcare industry by utilizing machine learning approaches to predict the likelihood of heart failure in a patient based on various demographic and medical test attributes. The research employed a number of individual ML models as well as an ensemble strategy to improve model performance in predicting heart disease. Following the implementation of numerous methods, the voting approach with DT and LR outperformed with an accuracy of 84.48%. Using the dataset's various attributes, the research can help medical experts treat heart patients. In the future, the author would like to add more diversity to the dataset by including various attributes that can better support the data. Furthermore, incorporating deep learning methods can aid in providing better solutions for heart consultants.

REFERENCES

[1] "Cardiovascular diseases," World Health Organization. https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1 (accessed Mar. 20, 2023).

[2] M. Trigka and E. Dritsas, "Long-Term Coronary Artery Disease Risk Prediction with Machine Learning Models," Sensors, vol. 23, no. 3, p. 1193, 2023.

[3] "Cardiovascular diseases symptoms," World Health Organization. https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_2 (accessed Mar. 20, 2023).

[4] A. Mahmud and J. Feely, "Effect of smoking on arterial stiffness and pulse pressure amplification," Hypertension, vol. 41, no. 1, pp. 183–187, 2003, doi: 10.1161/01.HYP.0000047464.66901.60.

[5] J. Díez and J. Butler, "Growing heart failure burden of hypertensive heart disease: a call to action," Hypertension, vol. 80, no. 1, pp. 13–21, 2023.

[6] N. Mishra, N. P. Desai, A. Wadhwani, and M. F. Baluch, "Visual Analysis of Cardiac Arrest Prediction Using Machine Learning Algorithms: A Health Education Awareness Initiative," in Handbook of Research on Instructional Technologies in Health Education and Allied Disciplines, IGI Global, 2023, pp. 331–363.

[7] O. Taylan, A. S. Alkabaa, H. S. Alqabbaa, E. Pamukçu, and V. Leiva, "Early Prediction in Classification of Cardiovascular Diseases with Machine Learning, Neuro-Fuzzy and Statistical Methods," Biology (Basel)., vol. 12, no. 1, p. 117, 2023.

[8] S. Chen et al., "Predicting Six-Month Re-Admission Risk in Heart Failure Patients Using Multiple Machine Learning Methods: A Study Based on the Chinese Heart Failure Population Database," J. Clin. Med., vol. 12, no. 3, p. 870, 2023.

[9] C. Dhanamjayulu, G. V. Suraj, M. Nikhil, R. Kaluri, and S. Koppu, "A Machine Learning Algorithm-Based IoT-Based Message Alert System for Predicting Coronary Heart Disease," in Advancements in Smart Computing and Information Security: First International Conference, ASCIS 2022, Rajkot, India, November 24–26, 2022, Revised Selected Papers, Part I, 2023, pp. 362–376.

[10] B. Kolukisa and B. Bakir-Gungor, "Ensemble feature selection and classification methods for machine learning-based coronary artery disease diagnosis," Comput. Stand. Interfaces, vol. 84, p. 103706, 2023.

[11] J. S. Lee, E.-S. Choi, Y. Hwang, K.-S. Lee, and K. H. Ahn, "Preterm birth and maternal heart disease: A machine learning analysis using the Korean national health insurance database," PLoS One, vol. 18, no. 3, p. e0283959, 2023.

[12] M. Ahmed, M. A. Kashem, M. Rahman, and S. Khatun, "Review and analysis of risk factor of maternal health in remote area using the internet of things (IoT)," in InECCE2019: Proceedings of the 5th International Conference on Electrical, Control & Computer Engineering, Kuantan, Pahang, Malaysia, 29th July 2019, 2020, pp. 357–365.

[13] G. Özsezer and G. Mermer, "Prevention of Maternal Mortality: Prediction of Health Risks of Pregnancy with Machine Learning Models," Available SSRN 4355295.

[14] Z. Ullah, N. Alsubaie, M. Jamjoom, S. H. Alajmani, and F. Saleem, "EffiMob-Net: A Deep Learning-Based Hybrid Model for Detection and Identification of Tomato Diseases Using Leaf Images," Agriculture, vol. 13, no. 3, p. 737, 2023.

[15] Z. Ullah et al., "Detecting High-Risk Factors and Early Diagnosis of Diabetes Using Machine Learning Methods," Comput. Intell. Neurosci., vol. 2022, 2022.

[16] F. Alsolami et al., "A unified decision-making technique for analysing treatments in pandemic context," Comput. Mater. Contin., pp. 2591–2618, 2022.

[17] Z. Ullah, F. Saleem, M. Jamjoom, and B. Fakieh, "Reliable Prediction Models Based on Enriched Data for Identifying the Mode of Childbirth by Using Machine Learning Methods: Development Study," J. Med. Internet Res., vol. 23, no. 6, p. e28856, 2021.

[18] F. Saleem, A. S. A.-M. Al-Ghamdi, M. O. Alassafi, and S. A. AlGhamdi, "Machine Learning, Deep Learning, and Mathematical Models to Analyze Forecasting and Epidemiology of COVID-19: A Systematic Literature Review," Int. J. Environ. Res. Public Health, vol. 19, no. 9, p. 5099, 2022.

[19] UCI, "Heart Disease Data Set." https://www.kaggle.com/ronitf/heart-disease-uci (accessed Mar. 20, 2023).

[20] R. Detrano et al., "International application of a new probability algorithm for the diagnosis of coronary artery disease," Am. J. Cardiol., vol. 64, no. 5, pp. 304–310, 1989.

[21] R. M. Team, "Rapid Miner." https://rapidminer.com/ (accessed Mar. 01, 2023).

[22] F. J. Alsolami, F. Saleem, and A. L. Abdullah, "Predicting the Accuracy for Telemarketing Process in Banks Using Data Mining."

[23] F. Saleem and A. Malibari, "DATA MINING COURSE IN INFORMATION SYSTEM DEPARTMENT–CASE STUDY OF KING ABDULAZIZ UNIVERSITY," 2011.

[24] S. Angra and S. Ahuja, "Implementation of Data Mining Algorithms on Student's Data using Rapid Miner," in International Conference On Big Data Analytics and computational Intelligence (ICBDACI), 2017, pp. 387–391.

[25] F. Saleem and A. AL-Malaise AL-Ghamdi, "Implementation of data mining approach for building automated decision support systems," in International Conference on Information Society, i-Society 2012, 2012, pp. 127–130.

[26] F. Saleem, Z. Ullah, B. Fakieh, and F. Kateb, "Intelligent Decision Support System for Predicting Student's E-Learning Performance Using Ensemble Machine Learning," Mathematics, vol. 9, no. 17, p. 2078, 2021.

[27] F. Khounraz et al., "Prognosis of COVID-19 patients using lab tests: A data mining approach," Heal. Sci. Reports, vol. 6, no. 1, p. e1049, 2023.

[28] C. Kumbhar and A. Hussain, "Prediction of Diabetics in the Early Stages Using Machine-Learning Tools and Microsoft Azure AI Services," in Machine Learning, Blockchain, and Cyber Security in Smart Environments, Chapman and Hall/CRC, 2023, pp. 59–80.

★★★