Southeast Europe Journal of Soft Computing ISSN 2233 – 1859
Available online: http://scjournal.ius.edu.ba

Vol 13, No. 1, March 2024, pp. 01-08

Heart Attack Prediction Model Based on Feature Selection and Decision Tree Approaches

[1]Hussein Abdullah Jaber, [1]Mortada Sadoun Thabet, [1]Rabab Abdul Hussein Fahd, [1]Dalal Khatib Muhbis, and [1]Alaa Khalaf Hamoud

[1]Department of Computer Information Systems,
College of Computer Science and Information Technology,
University of Basrah, Basrah, Iraq.
**Corresponding Author*: alaa.hamoud@uobasrah.edu.iq

Article Info

Article history:

Article received on 23 12 2023

Received in revised form 20 01 2024

Keywords:

Heart Attack; Feature Selection; Machine Learning; Decision Tree, LMT, Weka.

ABSTRACT: The purpose of this study is creating a machine learning based model is to predict heart attacks is to improve the capacity to anticipate the occurrence of this dangerous medical condition. It is feasible to find significant and linked variables that may cause heart attacks by using the decision tree as a tool for medical data analysis. The system analyzes clinical data using artificial intelligence techniques to find patterns that might suggest the possibility of a heart attack. The advantage is early disease detection and prediction, which allows the medical staff to better plan treatment and take preventative action. This kind of system can aid in enhancing patient care and lowering the likelihood that. Throuought the study, two paths will be examined, the first one is applying machine learning algorithms without applying feature selection, and the second one with feature selection process. Three mainly feature selection algorithms will be examined to find the most correlated features that affect the heart attack. The model will examine six machine learning decision tree algorithms namely (decision stump, hoeffding tree, j48, LMT, random forest, and rep tree) to find the accurate algorithm in prediction. The results show that LMT have the accurate prediction accuracy with 82.5%.

1. INTRODUCTION

Myocardial infarctions, commonly referred to as heart attacks, are major causes of death in many countries. According to the report, about 805,000 Americans suffer a heart attack annually, of which 605,000 are first-time sufferers and the remaining 200,000 have experienced a heart attack before. The prevalence and seriousness of heart attacks are highlighted by this data, making them a serious public health issue, while that heart attacks are frequently caused by coronary heart disease [1]. The accumulation of fat or bad cholesterol in the artery walls causes a blockage that is the main cause of this condition. A heart attack damages the heart's muscle tissues because there is not enough blood flow to the heart. One of the most concerning industries when it comes to data collection and processing is the healthcare sector. A huge volume of multidimensional data in different clinical sectors had been created in the last few years. Data of hospital resources, clinical factors, diagnosis data, medical records of patients, besides, medical equipements data. Thesr complicated, and dense data need to be preprocessed, and enhanced to get the knowledge required for supporting decision-making process [2]. One promising use of medical data mining is to uncover hidden patterns in medical data sets. The human body's vital pumping organ needs to be properly cared for. Unexpected developments could encourage the confinement of sporadic illnesses. Proper datasets from the verified repository and the application of supervised techniques would help with monitoring and prediction in the event that an unfavorable event occurred. Machine learning techniques are one such method that may be preferred by the market. The foundation of machine learning is data testing and training, whereby the system uses experience-based data [3], [4]. It is very important to find the most correlated feature that affects the heart attack, and this process may assist both clinical staff and patients. This process can determine the factors that can reduce the number of heart attack cases [5]–[7]. Using machine learning for this aim can give accurate results in both finding the correlated features and predicting the attacks. Machine learning and decision tree algorithms are utilized in different sectors and prove their accurate results in prediction, such as education [8]–[11], healthcare [12]–[14], the industial sector [3], [13], and agriculture [15], [16].

The benefit of creating a system to predict heart attacks is to assist patients in early disease diagnosis and taking necessary preventive actions, promoting healthier lives, steering clear of habits that contribute to this condition, and ultimately reducing the incidence and fatalities associated with it. After the model has been trained, predictions are made using the test dataset. The term ”supervised learning” refers to learning that occurs while a teacher is present and includes methods like logistic regression, Naive Bayes, decision trees, SVM models, KNN, and random. For the development process, machine learning techniques are helpful in integrating large amounts of data.

The remaining sections are organized as follows: Section 2 reviews the background information about the problem statement; Section 3 demonstrates the methodology steps, starting with data preprocessing and ending with model evaluation; and Section 4 lists the concluded points, recommendations, and future works..

2. LITERATURE REVIEW

This section lists the related works that utilize data mining techniques in the field of heart attacks and diseases related to this field. In [17], Saboor et al. proposed a study that focused on accurate and timely identification utilizing effective and dependable machine learning algorithms to detect heart disease. The field of medical sciences underwent significant diversification, driven by computing capabilities and technique improvements, particularly in the process of diagnosing human heart diseases. Various classifiers, incorporating a variety of algorithms including logistic regression (LR), support vector machine (SVM), LDA, random forest (RF), and XG Boost (XGB), were employed to predict heart disease based on a standard dataset. Performance evaluation with metrics like sensitivity, specificity, F-measure, and classification accuracy was evaluated before and after hyperparameter tuning. Results indicated notable enhancements in the precision of prediction classifiers following the tuning of hyperparameters, highlighting the effectiveness of data standardization and tuning in machine learning classifiers.

In [18], Bhatt et al. explored the application of machine learning, specifically employing k-implementing modes clustering with Huang’s initial method to enhance classification precision. Models such as RF, decision tree classifier (DT), multilayer perceptron (MP), and XGB were employed and optimized using GridSearchCV. The suggested model, utilized on a practical dataset comprising 70,000 instances from Kaggle, demonstrated high accuracy: decision tree (achieving 86.37 accuracy with cross-validation), XGBoost (achieving 86.87 accuracy with cross-validation), random forest (reaching 87.05 accuracy with cross-validation), and multilayer perceptron (87.28 accuracy with cross-validation). the AUC values for the models were decision tree (score of 0.94), XGBoost (score of 0.95), random forest (score of 0.95), and multilayer perceptron (with a score of 0.95). The research determined that the multilayer perceptron, The accuracy was enhanced with a score of 87.27 when applying the alogrithms using cross-validation.

In [4] Gupta et al. proposed a study using supervised machine-learning techniques, including K-Nearest Neighbour, DT, LR, NB, and SVM, to predict cardiac disease. The dataset used was collected from the University of California, Irvine (UCI) repository. LR outperformed other supervised machine learning algorithms regarding performance while demonstrating a low false negative rate. The study suggested the potential for accuracy improvement through ensemble techniques. Jupyter Notebook, with its versatile libraries and header files, was deemed the optimal tool for the precise and accurate implementation of Python programming in this contex.

In [19], Jindal et al. proposed a model that focused on predicting the heart attack based on different attributes based on K-nearest neighbor and LR. The model demonstrated significant accuracy in predicting heart disease, alleviating the pressure associated with accurate identification. The system, implemented in the.pynb format, not only enhanced medical care but also reduced costs. The study provided valuable insights for predicting heart disease in individuals, contributing to proactive healthcare measures.

3. BACKGROUND

The heart is an essential component of the cardiovascular system and muscle that pumps blood throughout the body. The cardiac muscle, which is housed in the chest behind the ribs, is a powerful muscle that makes up the heart. A heart attack is an emergency condition that happens when blood flow to a part of the heart muscle is blocked, seriously damaging the heart muscle. A heart attack is one of the arteries supplying the many diseases that can affect the heart. A blockage in one of the arteries supplying the heart with blood is typically the cause. According to WHO, the deaths due to heart diseases in Iraq reached 18.50% of total deaths in 2017, and Iraq has been identified as having the 19th rank at the international level [20], [21] [22].

Between 2015 and 2016, 165 patients diagnosed with myocardial infarction (MI) were admitted to the Cardiac Specialty Hospital (Coronary Care Unit) in Sulaimani, Iraq. Different factors such as lifestyle, family history, socioeconomic status, and classical risk factors are investigated, and the results showed that the most common category of individuals who have been infected is the category of individuals aged ≥61 years. The findings showed that 98.2% did not consume fish, and 56.4% and 58.8% did not eat daily fruits and vegetables, respectively. Smoking correlated significantly with age, gender, education level, and marital status. In Sulaimani, predictors of heart attack included a sedentary lifestyle, high body mass index, positive family history, an unhealthy diet, smoking, high blood pressure, dyslipidemia, and diabetes mellitus [23]. In the field of machine learning, there are several tools available that aid in data mining and predictive modeling. One such tool is Weka, which stands out from the others due to its extensive collection of machine learning algorithms and its open-source nature [24][25], [26]. Weka tool is one of powerful tools that been utilized in different sectors and help the analysits to perform the knowledge discovery and help in supporting decisions such as healthcare [27], [28], and eduation [29], [30].

4. METHODOLOGY

The methodology framework that has been followed in this study is shown in Figure 1. The first step is performing data preprocessing, since it is a critical step that affects the overall model accuracy. The second step is the feature selection process, where the dataset is examined to find the correlated features that affect the heart attack. The third step includes implementing six decision tree approaches (namely decision stump (DS), hoeffding tree (HT), logistic model tree (LMT), RF, random tree (RT), and reduced error pruning tree (ReT)) in order to find the optimal algorithm to predict heart attacks. The final step includes examining the model before and after feature selection to find the optimal algorithm for prediction and to examine the effect of feature selection on the prediction process.

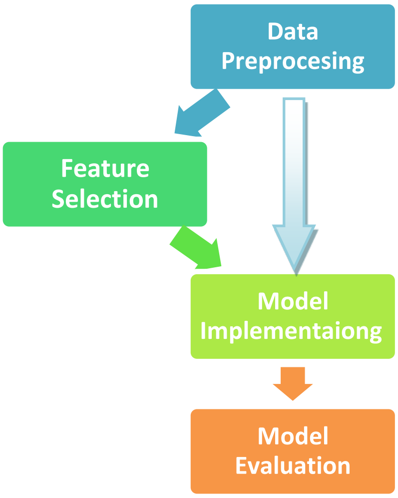

Figure 1: Model framework

4.1. Data Preprocessing

The dataset used in this study [31] consists of 303 records of patients diagnosed with heart attacks. The 1 holds the metadata of the dataset, the minimum and maximum values, and if the column holds an empty value, the age represents the patient's age, and the sex represents the patient's gender (male or female). The column cp represents the type of chest pain, while trtbps represents blood pressure at rest. Column chol represents the level of cholesterol in the blood, while fbs represents the blood sugar level while fasting. The column restecg represents the results of the electrocardiogram at rest, while thalachh is the maximum heart rate achieved during the physiological test. Column exng represents the presence or absence of chest pain during exercise, while oldpeak represents the relative decrease in the peak of the ST segment of the heart. Column slp represents the slope of the heart rate, caa represents the number of major vessels colored by fluoroscopy, thall represents the type of anemia, while the output represents the final class.

Table 1 Dataset Metadata

Seq	Field	Min	Max	Null
1	age	29	77	None
2	sex	0	1	None
3	cp	0	3	None
4	trtbps	94	200	None

5	chol	126	564	None
6	fbs	0	1	None
7	restecg	0	2	None
8	thalachh	71	202	None
9	exng	0	1	None
10	oldpeak	0	6.2	None
11	slp	0	2	None
12	caa	0	4	None
13	thall	0	3	None
14	output	0	1	None

The platform specifications utilized in this study are listed in Table 2. Since the data is cleaned, so the required data preprocessing steps, including data cleaning is not nessecirely.

Table 2: Platform Specification

Hardware	Specification
CPU	Intel Corei7-7
RAM	16GB
HARD	256G SSD
OS	Win 10 Home

4.2. Feature Selection

In this step, three main feature selection (FS) approaches will be examined. Info Gain (IG) is a method used in machine learning to evaluate the importance of a feature by measuring the reduction in uncertainty it brings to a dataset. It quantifies the amount of information gained when a feature is added to the model. Info Gain is frequently employed to rank features based on their relevance in classification tasks [32], [33]. Gain Ratio (GR) is a refinement of Info Gain, designed to account for the bias of Info Gain towards features with numerous values or categories. It normalizes information gain by dividing it by split information, which assesses data branching based on the selected feature. The gain ratio is utilized to mitigate the bias of Info Gain for high-cardinality features [34], [35]. The symmetric evaluator (SE), or symmetrical uncertainty, is a metric for evaluating the relationship between two features in a dataset. It assists in quantifying the degree of correlation between two candidate features. If a strong relationship exists between two features, removing one of them may be advisable to reduce redundancy and enhance model performance [36]–[38]. These feature selection techniques are crucial for dimensionality reduction and enhancing machine learning models' effectiveness. They enable the selection of a subset of relevant features with less redundancy and noise, ultimately leading to more accurate and efficient models. Please feel free to explore the provided resources for a deeper understanding of each algorithm's applications and implementations.

Table 3: Features' Correlation

FS Algorithm	Selected Features
IG	13, 3, 12, 10, 9, 8, 11, 1, 2
GR	3, 13, 12, 9, 8, 11, 10, 2, 1
SE	3, 13, 12, 9, 8, 10, 11, 2, 1

According to Table 3, according to the IG filter, the featrurer (13, 3, 12, 10, 9, 8, 11, 1, and 2) have the most effect on the heart attack, starting from feature (thall, cp, caa, oldpeak, thalachh, slp, and age), while other features have no correlation with the target class. Next, according to GR, features (3, 13, 12, 9, 8, 11, 10, 2, and 1) have the most effect on the target class, while there is no correlation between the remaining features and the target class. Finally, according to SE, features (3, 13, 12, 9, 8, 10, 11, 2, and 1) affect the heart attack, while other features have no correlation with the target class. The top three features that affect the heart attack according to these algorithms are (thall, cp, and caa); see Figure 2.

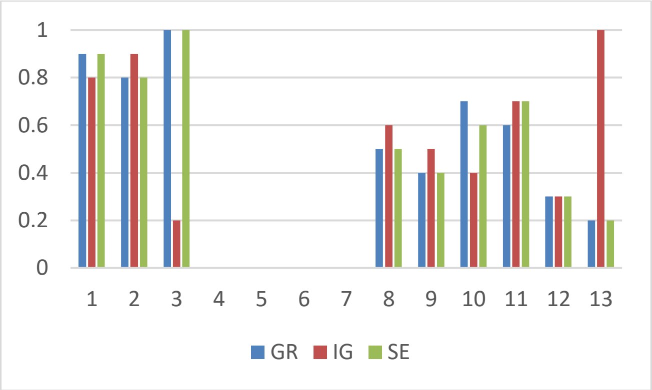

Figure 2: Attributes' correlations

4.3. Model Implementation

Six decision tree algorithms are utilized in this study. The first one is DS, which is a basic one-level decision tree used for binary classification. It makes decisions based on a single feature and is often employed in ensemble methods like AdaBoost [39]. The second one is HT which is designed for incremental learning from data streams. This approach updated the tree structure continuously [14], [40]. The third algorithm is J48, the widely used DT algorithm, where the dataset is recursively splitted to increase the gained information while the resultant decision tree is created based on the splitted dataset. J48 is a suitable choice for the divide-and-conquer and top-down approaches [13]. Next, the fourth approach is the LMT algorithm, where two approaches are combined (LR and DT). DT is utilized in the dataset partition process, while LR is applied to each partition of the dataset. This algorithm is optimal for nonlinear and complex relationships within datasets [41], [42]. The fifth algorithm is RF, one of the ensemble techniques that can enhance accuracy by combining multiple DTs. This algorithm utilized random feature selection and bootstrapped the samples of data to introduce randomness [43], [44]. Next, the sixth algorithm is RT, where the randomization is incorporated to construct the DT. To perform splitting, random features are selected to make it suitable for the high-dimension datasets and to overcome the overfitting problem [26], [44]. Finally, ReT is one of the DT algorithms that builds the structure of the tree and improves generalization by pruning it. It iteratively removes branches that do not significantly affect accuracy [45]. These machine learning algorithms have unique characteristics and applications, making them suitable for different types of classification tasks. By understanding their distinctions, you can choose the most appropriate algorithm based on the specific requirements and nature of your dataset.

4.4. Model Evaluation

After implementing the FS step and determining the most correlated features, the model evaluation process is performed with feature selection (WFS) and without feature selection (WOFS) to measure the data mining accuracy before and after the feature selection process. According to Figure 3, the accuracies of algorithms (Hoeffding tree and random tree) are enhanced by 1% and 5%, respectively, while the accuracies are decreased by 0.33%, 0.33%, 0.99%, and 3.02% for J48, LMT, RF, and ReT, respectively. The best accuracy score is LMT without feature selection (82.51%).

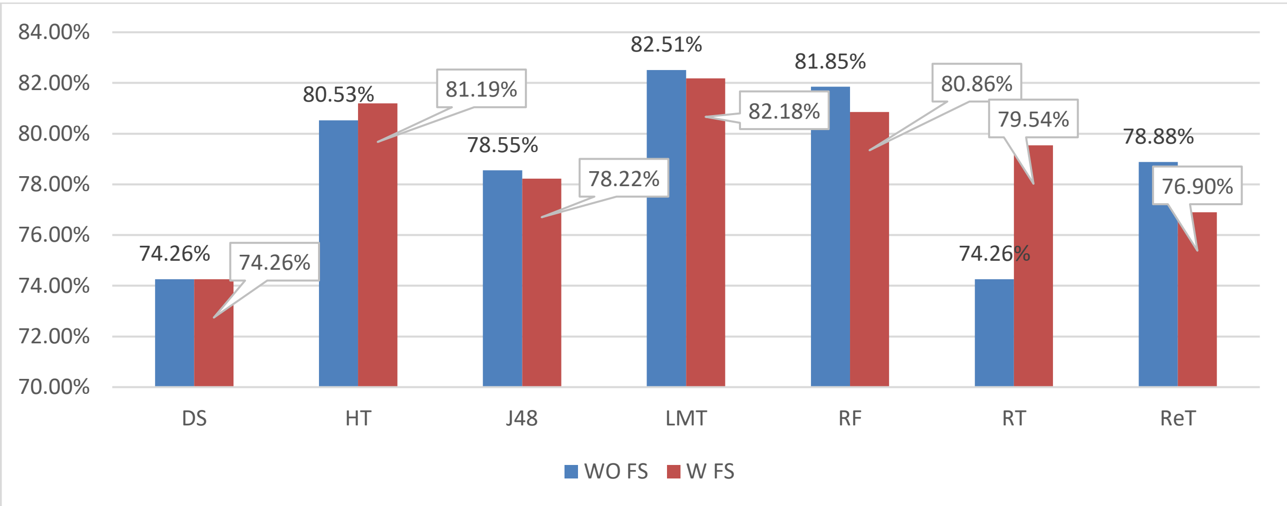

Figure 3: Model evaluation before/after feature selection process

5. CONCLUSION AND FUTURE WORKS

The purpose of creating an automated heart attack-based system to predict attacks is to improve the capacity to anticipate the occurrence of this dangerous medical condition. It is feasible to find significant and linked variables that may cause heart attacks by using the decision tree as a tool for medical data analysis. The system analyzes clinical data using artificial intelligence techniques to find patterns that might suggest the possibility of a heart attack. The advantage is early disease detection and prediction, which allows the medical staff to better plan treatment and take preventative action. This kind of system can aid in enhancing patient care and lowering the likelihood of that. A feature selection process can be adopted to find the correlation between the features in the dataset and the heart attack. According to the feature selection process in the paper, it is obvious that there are correlations between some features while others may have no effect on the heart attack. According to the results, the top three features that affect the heart attack are thall, cp, and caa. The effectiveness of machine learning algorithms is found in their ability to learn from the data and predict future trends. Six decision tree algorithms were utilized in this study, and the best algorithm for predicting was LMT. The model will be examined with more data, and it can be implemented to predict real-time cases to help the clinical staff. The proposed model can be utilized as an assistant tool to predict heart attack cases.

REFERENCES

[1] M. W. Eladham, A. B. Nassif, and M. A. AlShabi, "Heart attack prediction using machine learning," in *Smart Biomedical and Physiological Sensor Technology XX*, 2023, pp. 86–93.

[2] G. Thilagavathi, S. Priyanka, V. Roopa, and J. S. Shri, "Heart disease prediction using machine learning algorithms," in *2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, 2022, pp. 494–501.

[3] K. Arumugam, M. Naved, P. P. Shinde, O. Leiva-Chauca, A. Huaman-Osorio, and T. Gonzales-Yanac, "Multiple disease prediction using Machine learning algorithms," *Mater Today Proc*, vol. 80, 2023, doi: 10.1016/j.matpr.2021.07.361.

[4] C. Gupta, A. Saha, N. V. S. Reddy, and U. D. Acharya, "Cardiac Disease Prediction using Supervised Machine Learning Techniques," in *Journal of Physics: Conference Series*, 2022. doi: 10.1088/1742-6596/2161/1/012013.

[5] A. Purnomo, M. A. Barata, M. A. Soeleman, and F. Alzami, "Adding feature selection on Naïve Bayes to increase accuracy on classification heart attack disease," in *Journal of Physics: Conference Series*, 2020. doi: 10.1088/1742-6596/1511/1/012001.

[6] O. Shakir and I. Saleh, "Hybridization of Swarm for Features Selection to Modeling Heart Attack Data," *AL-Rafidain Journal of Computer Sciences and Mathematics*, vol. 16, no. 2, 2022, doi: 10.33899/csmj.2022.176587.

[7] H. Takci, "Improvement of heart attack prediction by the feature selection methods," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 26, no. 1, 2018, doi: 10.3906/elk-1611-235.

[8] Z. J. Kovacic, "Early Prediction of Student Success: Mining Students Enrolment Data," in *Proceedings of the 2010 InSITE Conference*, 2010. doi: 10.28945/1281.

[9] A. K. HAMOUD, "CLASSIFYING STUDENTS'ANSWERS USING CLUSTERING ALGORITHMS BASED ON PRINCIPLE COMPONENT ANALYSIS.," *J Theor Appl Inf Technol*, vol. 96, no. 7, 2018.

[10] A. Khalaf Hamoud and A. Majeed Humadi, "STUDENT'S SUCCESS PREDICTION MODEL BASED ON ARTIFICIAL NEURAL NETWORKS (ANN) AND A COMBINATION OF FEATURE SELECTION METHODS," *Journal of Southwest Jiaotong University*, vol. 54, no. 3, 2019, doi: 10.35741/issn.0258-2724.54.3.25.

[11] A. Hamoud, A. S. Hashim, and W. A. Awadh, "Predicting student performance in higher education institutions using decision tree analysis," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, pp. 26–31, 2018.

[12] A. K. Hamoud, "Applying Association Rules and Decision Tree Algorithms with Tumor Diagnosis Data," *International Research Journal of Engineering and Technology*, vol. 3, no. 8, pp. 27–31, 2016.

[13] U. Pehlivan, N. Baltaci, C. Acarturk, and N. Baykal, "The analysis of feature selection methods and classification algorithms in permission based Android malware detection," in *IEEE SSCI 2014: 2014 IEEE Symposium Series on Computational Intelligence - CICS 2014: 2014 IEEE Symposium on Computational Intelligence in Cyber Security, Proceedings*, 2014. doi: 10.1109/CICYBS.2014.7013371.

[14] T. E. Mathew, "Appositeness of Hoeffding tree models for breast cancer classification," *J Curr Sci Technol*, vol. 12, no. 3, 2022.

[15] S. Bishnoi and B. K. Hooda, "Decision Tree Algorithms and their Applicability in Agriculture for Classification," *Journal of Experimental Agriculture International*, 2022, doi: 10.9734/jeai/2022/v44i730833.

[16] O. Abrishambaf, P. Faria, Z. Vale, and J. M. Corchado, "Energy scheduling using decision trees and emulation: Agriculture irrigation with run-of-the-river hydroelectricity and a PV case study," *Energies (Basel)*, vol. 12, no. 20, 2019, doi: 10.3390/en12203987.

[17] A. Saboor *et al.*, "A method for improving prediction of human heart disease using machine learning algorithms," *Mobile Information Systems*, vol. 2022, 2022.

[18] C. M. Bhatt, P. Patel, T. Ghetia, and P. L. Mazzeo, "Effective heart disease prediction

using machine learning techniques," *Algorithms*, vol. 16, no. 2, p. 88, 2023.

[19] H. Jindal, S. Agrawal, R. Khera, R. Jain, and P. Nagrath, "Heart disease prediction using machine learning algorithms," in *IOP conference series: materials science and engineering*, 2021, p. 12072.

[20] B. J. Saleh, R. R. K. Al_Taie, and A. A. Mhawes, "Machine Learning Architecture for Heart Disease Detection: A Case Study in Iraq," *International journal of online and biomedical engineering*, vol. 18, no. 2, 2022, doi: 10.3991/ijoe.v18i02.27143.

[21] R. K. Abd, S. N. Abd, and V. Raman, "Tracing the risk factors of heart diseases at al-Nasiriyah heart center in Iraq," *J Cardiovasc Dis Res*, vol. 10, no. 1, 2019, doi: 10.5530/jcdr.2019.1.6.

[22] M. D. Ahmed, I. H. Hameed, and M. Q. Abd-Ali, "Prospective and retrospective study of the acute heart attack cases in marjan hospital-hillah city-Iraq," *Res J Pharm Technol*, vol. 10, no. 10, 2017, doi: 10.5958/0974-360X.2017.00606.0.

[23] B. O. Sharif and S. Y. Lafi, "Common Risk Factors of Myocardial Infarction and Some Socio Demographic Characteristics in Sulaimani City," *Kurdistan Journal of Applied Research*, 2022, doi: 10.24017/science.2021.2.13.

[24] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, 2009, doi: 10.1145/1656274.1656278.

[25] E. Frank, M. A. Hall, and I. H. Witten, "The WEKA Workbench Data Mining: Practical Machine Learning Tools and Techniques," *Morgan Kaufmann, Fourth Edition*, 2016.

[26] S. Kalmegh, "Analysis of WEKA Data Mining Algorithm REPTree , Simple Cart and RandomTree for Classification of Indian News," *International Journal of Innovative Science, Engineering & Technology*, vol. 2, no. 2, 2015.

[27] K. Rijhwani, V. R. Mohanty, Y. B. Aswini, V. Singh, and S. Hashmi, "Applicability of data mining and predictive analysis for tobacco cessation: An exploratory study," *Front Dent*, vol. 17, 2020, doi: 10.18502/fid.v17i24.4624.

[28] P. P. Shinde, K. S. Oza, and R. K. Kamat, "Systematic acuity of medicinal big data: need of health industry," *Pers Ubiquitous Comput*, vol. 27, no. 3, 2023, doi: 10.1007/s00779-022-01681-1.

[29] S. Alija, E. Beqiri, A. S. Gaafar, and A. K. Hamoud, "Predicting Students Performance Using Supervised Machine Learning Based on Imbalanced Dataset and Wrapper Feature Selection," *Informatica*, vol. 47, no. 1, 2023.

[30] L. Rashid, "Towards successful entrepreneurial outcomes amidst extreme fragility," 2020.

[31] Rashik Rahman, "Heart Attack Analysis & Prediction Dataset," *https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset/data*.

[32] C. Lee and G. G. Lee, "Information gain and divergence-based feature selection for machine learning-based text categorization," *Inf Process Manag*, vol. 42, no. 1, pp. 155–165, 2006.

[33] H. Uğuz, "A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm," *Knowl Based Syst*, vol. 24, no. 7, pp. 1024–1032, 2011.

[34] N. A. Suwadi *et al.*, "An Optimized Approach for Predicting Water Quality Features Based on Machine Learning," *Wirel Commun Mob Comput*, vol. 2022, 2022, doi: 10.1155/2022/3397972.

[35] H. H. Htun, M. Biehl, and N. Petkov, "Survey of feature selection and extraction techniques for stock market prediction," *Financial Innovation*, vol. 9, no. 1. 2023. doi: 10.1186/s40854-022-00441-7.

[36] W. Wang, M. Guo, T. Han, and S. Ning, "A novel feature selection method considering feature interaction in neighborhood rough set," *Intelligent Data Analysis*, vol. 27, no. 2, 2023, doi: 10.3233/IDA-216447.

[37] A. Sikri, N. P. Singh, and S. Dalal, "Analysis of Rank Aggregation Techniques for Rank Based on the Feature Selection Technique," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 11, 2023, doi: 10.17762/ijritcc.v11i3s.6160.

[38] B. Niu, J. Sun, and B. Yang, "Multisensory based tool wear monitoring for practical applications in milling of titanium alloy," in *Materials Today: Proceedings*, 2020. doi: 10.1016/j.matpr.2019.12.126.

[39] M. Kamaladevi and V. Venkatraman, "Tversky Similarity based Under Sampling with Gaussian Kernelized Decision Stump Adaboost Algorithm for Imbalanced Medical Data Classification," *International Journal of Computers, Communications and Control*, vol. 16, no. 6, 2021, doi: 10.15837/IJCCC.2021.6.4291.

[40] E. García-Martín, A. Bifet, and N. Lavesson, “Energy modeling of Hoeffding tree ensembles,” *Intelligent Data Analysis*, vol. 25, no. 1, 2021, doi: 10.3233/IDA-194890.

[41] W. Chen *et al.*, “A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility,” *Catena (Amst)*, vol. 151, 2017, doi: 10.1016/j.catena.2016.11.032.

[42] V. H. Nhu *et al.*, “Shallow landslide susceptibility mapping: A comparison between logistic model tree, logistic regression, naïve bayes tree, artificial neural network, and support vector machine algorithms,” *Int J Environ Res Public Health*, vol. 17, no. 8, 2020, doi: 10.3390/ijerph17082749.

[43] L. Breiman, “Random Forests,” *Mach Learn*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[44] J. F. Le Gall, “Random trees and applications,” *Probability Surveys*, vol. 2, no. 1, 2005, doi: 10.1214/154957805100000140.

[45] S. K. Jayanthi and S. Sasikala, “Reptree classifier for identifying link spam in web search engines,” *IJSC*, vol. 3, no. 2, pp. 498–505, 2013.