International Journal of Computer and Communication Technology

Volume 9
Issue 1 *Research on Computing and Communication Sciences.*

Article 7

July 2023

Comparative Analysis of Data Mining Techniques for Heart Disease Prediction: A Focus on Neural Networks and Decision Trees

Suman Kumari Panigrahi
Gandhi Institute for Education & Technology, sumankpanigrahi@gmail.com

Abantika Roy
KIIT University, abantikaroy.2001@gmail.com

Gargi Balabantaray
Siksha 'O'Anusandhan University, gargibalabantaray@gmail.com

Karishma Rana
Jaypee Institute of Information Technology, kar97na@gmail.com

Recommended Citation

Panigrahi, Suman Kumari; Roy, Abantika; Balabantaray, Gargi; and Rana, Karishma (2023) "Comparative Analysis of Data Mining Techniques for Heart Disease Prediction: A Focus on Neural Networks and Decision Trees," *International Journal of Computer and Communication Technology*: Vol. 9: Iss. 1, Article 7.
DOI: 10.47893/IJCCT.2023.1442
Available at: https://www.interscience.in/ijcct/vol9/iss1/7

Comparative Analysis of Data Mining Techniques for Heart Disease Prediction: A Focus on Neural Networks and Decision Trees

Suman Kumari Panigrahi[1], Abantika Roy[2], Gargi Balabantaray[3], Karishma Rana[4*]
[1]*Department of Computer Science & Engineering, Gandhi Institute for Education & Technology, Odisha, India*
[2]*Department of Biotechnology, KIIT University, Bhubaneswar, Odisha, India*
[3]*Department of Immunology& Rheumatology, Institute of Medical Sciences, Sum Hospital, Siksha 'O'Anusandhan University, Odisha, India*
[4]*Department of Biotechnology, Jaypee Institute of Information Technology, Noida, Uttar Pradesh, India.*
sumankpanigrahi@gmail.com, abantikaroy.2001@gmail.com, gargibalabantaray@gmail.com, kar97na@gmail.com

Abstract
Heart disease is a general term used to describe numerous medical conditions that directly affect the heart and its various components. It is a prevalent health concern in modern times. The focus of this paper is to evaluate different data mining techniques for the prediction of heart disease, which have been introduced in recent years. The findings indicate that neural networks using 15 attributes demonstrate the best performance among all other data mining techniques. Additionally, the analysis concludes that decision trees, with the assistance of genetic algorithms and feature subset selection, also exhibit high accuracy.The study concludes that data mining techniques can effectively predict heart disease and that the choice of technique depends on the specific context of the analysis. The study suggests that decision trees and artificial neural network models are suitable for heart disease prediction.The study also recommends further research to explore the use of other data mining techniques for heart disease prediction.

Keywords: heart disease,data mining techniques, neural networks, decision tree, genetic algorithm, prediction.

1. Introduction

Data mining involves the identification of previously unknown trends and patterns in databases, which can be utilized to develop predictive models. Within the healthcare industry, data mining is becoming increasingly indispensable as a means of handling vast quantities of complex data, such as electronic patient records, disease diagnoses, medical equipment, and information on hospital resources. The processing and analysis of this data are essential to extract knowledge that can support decision-making and cost savings. Data mining provides a range of techniques and tools that can be applied to this processed data, allowing hidden patterns to be discovered. This, in turn, provides healthcare professionals with additional sources of knowledge to aid in making informed decisions. The fundamental model for the data mining process is illustrated in Figure 1.

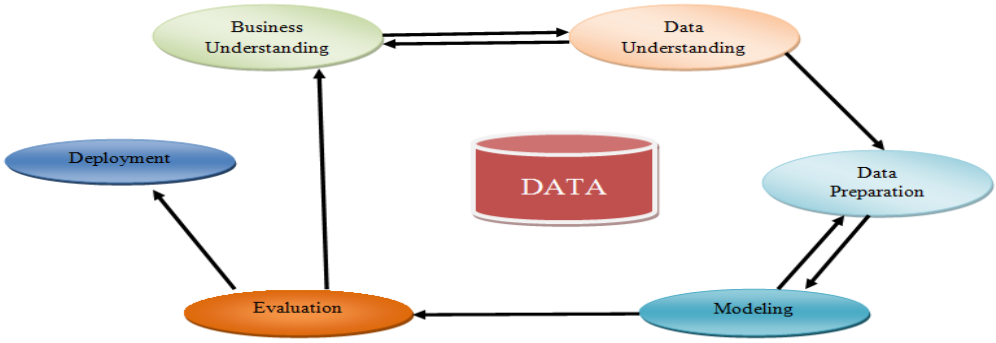

Figure 1: Processing Model of data mining

This diagram shows the different stages of the data mining process. The first stage is to define the business problem that needs to be solved. The next stage is data understanding, where the data that is available is explored and analyzed. The third stage is data preparation, where the data is cleaned, transformed, and formatted so that it can be used for modeling. The fourth stage is data modeling, where different models and algorithms are used to analyze the data and identify patterns. The fifth stage is model evaluation and validation, where the accuracy and effectiveness of the models are tested using a separate testing dataset. The final stage is deployment and maintenance, where the models are put into practice and continuously monitored and updated as needed.

According to the World Health Statistics 2019 report, one out of every three adults globally suffers from high blood pressure - a condition that contributes to approximately 50% of all fatalities due to heart disease and stroke. Heart disease, also known as cardiovascular disease (CVD), encompasses a variety of conditions that affect the heart, including irregular heart rhythms and heart valve malfunctions, in addition to heart attacks. These issues can cause heart failure and a range of other complications. The implementation of efficient and effective automated heart disease prediction systems can have significant benefits for the healthcare sector. Our research aims to provide an in-depth analysis of different data mining techniques that can be utilized in such automated systems, which, in turn, would reduce the number of tests required for a patient.

Therefore, these systems will not only save costs but also save time for both analysts and patients.

2. Methodology

This paper presents an analysis of different data mining techniques that can aid medical analysts or practitioners in accurately diagnosing heart disease. To conduct this study, we examined publications, journals, and reviews in the field of computer science and engineering, data mining, and cardiovascular disease that were published in recent times. [5]

Issues and Challenges of Disease Prediction Using Different Data Mining Techniques

There are several issues &challenges while analyzing heart disease prediction using different data mining techniques. Some of these challenges are:

1. Data quality: The accuracy of the prediction model depends heavily on the quality of the data used to train and test the model. Incomplete, inconsistent, or noisy data can lead to inaccurate predictions and affect the performance of the model.
2. Imbalanced data: The distribution of data in the training dataset may be skewed, with one class having significantly more instances than the other. This can result in biased models that perform well on the majority class but poorly on the minority class.
3. Overfitting: Overfitting occurs when a model is too complex and learns the noise in the data instead of the underlying patterns. This can lead to a model that performs well on the training data but poorly on new, unseen data.
4. Feature selection: The selection of relevant features is crucial for the accuracy of the prediction model. However, selecting the most informative features from a large dataset can be a challenging task.
5. Interpreting results: Different data mining techniques may produce different results, and it can be challenging to interpret and compare these results. Additionally, some techniques may not provide insight into the underlying patterns in the data, making it difficult to interpret the results and make informed decisions.
6. Generalizability: The performance of a model on a specific dataset may not generalize well to other datasets or real-world scenarios. It is important to evaluate the performance of the model on multiple datasets and in different settings to ensure its generalizability.

3. Experimental Analysis

3.1 Data Mining & Neural Networks

An artificial neural network (ANN), which is sometimes referred to as a "neural network" (NN), is a mathematical or computational model that imitates biological neural networks. This system is an emulation of the biological neural system. In this study, a heart disease prediction system was created utilizing 15 attributes [4]. In previous work, only 13 attributes were used for prediction, but this research incorporated two additional attributes, namely obesity, and smoking, to enable a more efficient diagnosis of heart disease.

To conduct the experiment, the researchers utilized the data mining tool Weka 3.6.6. Firstly, missing values in the dataset were identified and then replaced with appropriate values using the Replace Missing Values filter from 3.6.6 [4]. After this, various data mining techniques were applied and analyzed on the heart disease database. A confusion matrix was generated for each classifier. Table 1 displays the results of this research work, demonstrating that neural networks outperformed other data mining techniques

Table1: Comparison of Data mining techniques

Classifications Techniques	Accuracy
Naive Bayes	91.89%
Decision Tree	97.65%
Neural Network	99.75%

3.2 Fuzzy Logic& Genetic Algorithm

According to this research, the proposed method is an enhanced version of the model that utilizes genetic algorithms for feature selection and a fuzzy expert system for accurate classification. Fuzzy set theory and fuzzy logic are considered appropriate for developing knowledge-based systems in healthcare to diagnose diseases [2].

The experiments were carried out using Matlab and its fuzzy tool. A Mamdani model of the fuzzy system was employed, and the fuzzy rules were generated based on the experts' knowledge in the field. The researchers used a dataset from the UCI machine learning repository and identified that only 6 attributes were necessary and effective for predicting heart disease. The proposed system takes the selected features as input and produces an output value of either 0 or 1, indicating the absence or presence of heart disease in patients.

The fuzzy logic process involves several steps. Firstly, the input data is collected as a crisp set and is converted into a fuzzy set using fuzzy linguistic variables, terms, and membership functions through the fuzzification process. Then, an inference is made based on a set of rules using the fuzzy rules generated by the system. Finally, the defuzzification process is performed to obtain a crisp output [2]. In this system, the fuzzy rules are generated based on the support sets obtained, and Table 2 shows the support set.

Table 2: Values of the support set

Sl. No.	Attributes	Support set	
		Heart Patients	Non-heart Patients
1	Types of Chest pain	8	1,2,3,4,5,6,7
2	Exang	Yes	No
3	Rbps	138-159	146-160
4	Oldpeak	2.05-6.8	<2.05
5	Thalach	81-146	146-178
6	ca	1,2,3,4,5,6,7	0

3.3 Data Mining & Machine Learning Algorithms

The research work presented in this paper focuses on data classification using various supervised machine learning algorithms, namely Naive Bayes, Decision List, and KNN. The TANAGRA tool is used for classification, and the data is evaluated using 10-fold cross-validation. TANAGRA is a data mining tool designed for academic and research purposes. It offers several data mining methods from exploratory data analysis, statistical learning, machine learning, and database areas. It provides an easy-to-use interface, allowing users to analyze real or synthetic data. Moreover, it offers architecture for users to add their own data mining methods and compare their performances. Additionally, TANAGRA [20] provides access to a wide range of data sources, direct access to data warehouses and databases, data cleansing, and interactive utilization.

The experiments conducted in this research used a training dataset consisting of 3000 instances with 14 different attributes. The dataset was split into two parts, with 70% of the data used for training and the remaining 30% used for testing. The data was classified using three different supervised machine learning algorithms: Naive Bayes, Decision List, and KNN, with evaluation performed using 10-fold cross-validation. The TANAGRA tool was used for classification and evaluation, providing an easy-to-use interface for exploratory data analysis, statistical learning, machine learning, and database tasks [12]. Results showed that Naive Bayes algorithm outperformed the other two algorithms in terms of accuracy and evaluation time. Figure 2 presents a performance study of the various algorithms used.

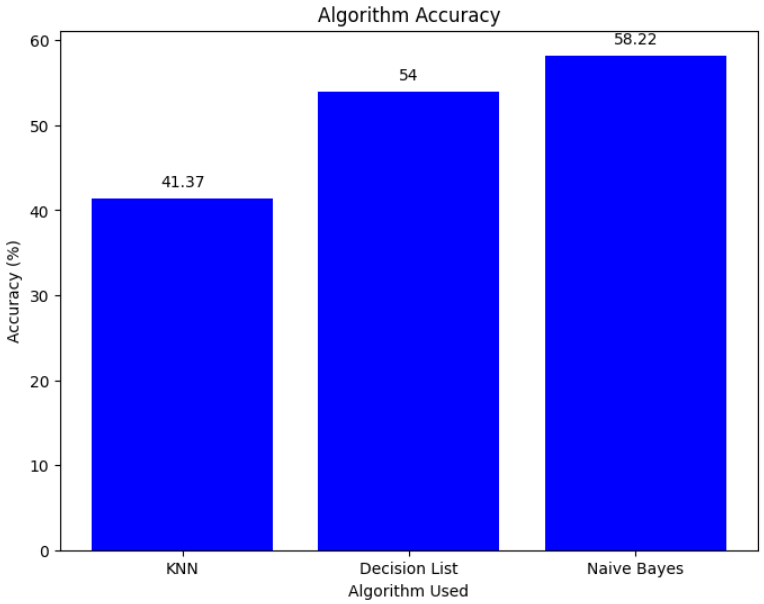

Figure 2: Performance analysis of various Algorithms

3.4 Data Mining & Genetic Algorithm

The use of a Genetic Algorithm and Feature Subset Selection in this research work aimed to reduce the number of attributes required for heart disease diagnosis. The process involved starting with zero attributes and generating an initial population with randomly generated rules. New populations were constructed based on the idea of survival of the fittest, where new rules were generated by applying genetic operators such as cross-over and mutation. The process continued until a population was obtained where every rule satisfied the fitness threshold. The genetic search resulted in reducing the number of attributes from 13 to 6.

In addition to the genetic algorithm, CFS Evaluator was also used. The observations were conducted using Weka 3.6.0 tool on a dataset of 907 records with 13 attributes. All attributes were made categorical, and inconsistencies were resolved for simplicity. After reducing the number of attributes to 6, various classifiers were used on the dataset for heart disease prediction. The performance analysis of these classifiers is shown in Figure 3, where it can be seen that the Decision Tree classifier had the highest accuracy and least mean absolute error.

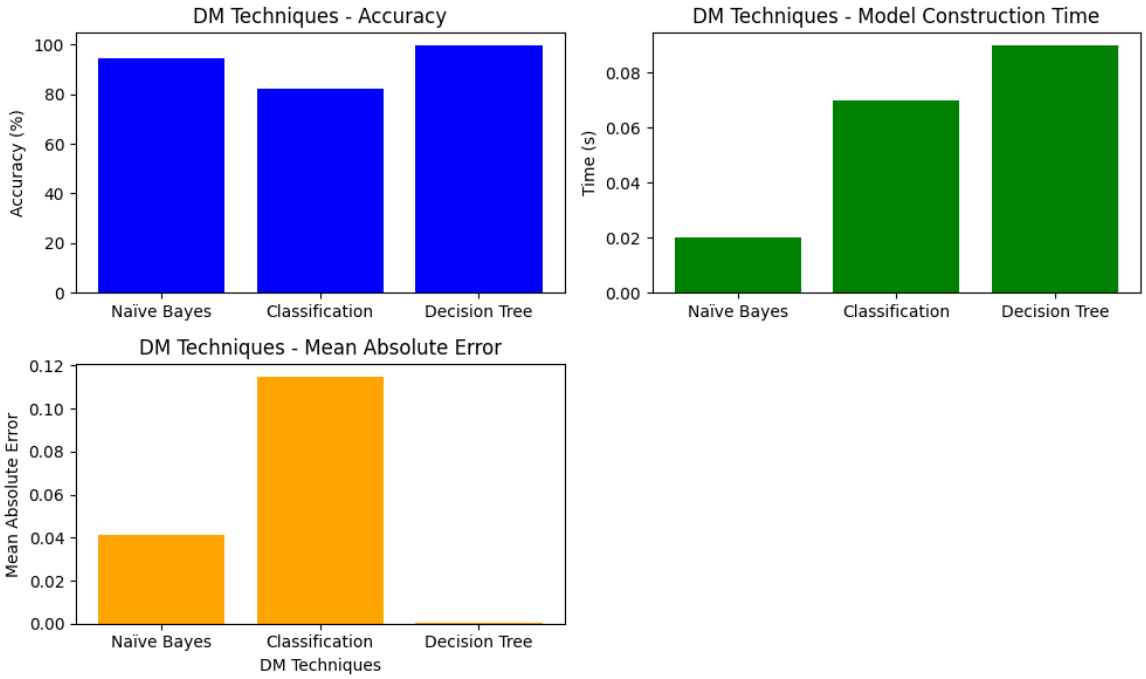

Figure 3: Comparison figure for 3 classifiers

3.5 IHDPS

The Intelligent Heart Disease Prediction System (IHDPS) is a web-based, user-friendly, and expandable system developed using data mining techniques such as Decision Trees, Naive Bayes, and Neural Networks. The system is built on the .NET platform and is capable of discovering and extracting hidden knowledge associated with heart disease from a historical heart disease database. The system can answer complex queries related to heart disease diagnosis, enabling healthcare analysts and practitioners to make intelligent clinical decisions that traditional decision support systems cannot. The IHDPS is based on 17 attributes and was developed using a dataset of 1010 records from the Cleveland Heart Disease database, which was equally divided into training and testing datasets. The system was found to be most effective in predicting heart disease in patients using Naive Bayes, which had the highest percentage of correct predictions (89.23%), followed by Neural Networks (87.83%) and

Decision Trees in Figure 4. However, Decision Trees were found to be most effective in predicting patients with no heart disease (94.45%) compared to the other two models. The IHDPS can display the results both in tabular and graphical forms, providing effective treatments and reducing treatment costs.

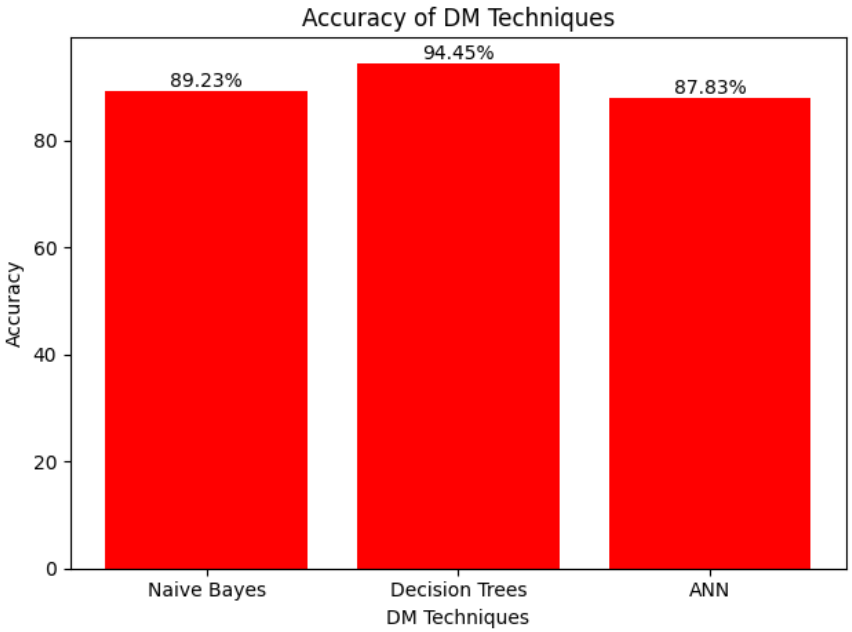

Figure 4: Performance analysis of IHDPS

4. Results

The accuracy of a classifier can be influenced by various factors such as the size and quality of the dataset, the pre-processing techniques used the choice of parameters for the classifier, and the specific characteristics of the data mining technique. Therefore, it is important to carefully select and evaluate different combinations of data mining techniques and classifiers to identify the most effective approach for a given problem.

5. Recommendation

Based on the challenges and issues discussed above, here are some recommendations for future studies on heart disease prediction using data mining techniques:

1. Use larger datasets: In order to improve the accuracy and generalizability of heart disease prediction models, it is recommended to use larger datasets that include more diverse patient populations.
2. Address class imbalance: Class imbalance is a common issue in heart disease prediction datasets, where the number of positive (with heart disease) cases is much lower than negative (without heart disease) cases. Researchers can employ techniques such as oversampling, undersampling, or SMOTE to address this issue.
3. Comparison of multiple classifiers: While many studies have compared different data mining techniques, it is recommended to compare multiple classifiers within each technique to identify the most accurate and efficient model for heart disease prediction.
4. Incorporate domain knowledge: Incorporating domain knowledge can help improve the accuracy and interpretability of heart disease prediction models. For example, incorporating medical knowledge about risk factors and symptoms of heart disease can improve the selection of relevant attributes and improve the performance of the models.
5. Use multiple evaluation metrics: Using multiple evaluation metrics such as accuracy, sensitivity, specificity, and

AUC can provide a more comprehensive understanding of the performance of heart disease prediction models.

6. Replication of studies: Replication of studies on heart disease prediction using different datasets and data mining techniques can help confirm the validity and generalizability of the results.

Overall, conducting rigorous and comprehensive studies on heart disease prediction using data mining techniques can lead to more accurate and efficient diagnosis and treatment of heart disease, ultimately improving patient outcomes and reducing healthcare costs.

6. Discussion

The analysis of heart disease prediction using different data mining techniques is a rapidly growing field of research, as it has the potential to improve the accuracy and efficiency of diagnosis and treatment. The studies reviewed in this discussion have shown that various data mining techniques, such as decision trees, naive Bayes, and neural networks, have been used to predict heart disease with high accuracy.One of the challenges identified in these studies is the lack of standardization in the selection of attributes and classifiers. This makes it difficult to compare the results of different studies and generalize the findings. Additionally, the small sample sizes used in some studies may limit the generalizability of their findings to other populations.Another challenge is the imbalance in the dataset, which is often skewed toward healthy individuals, leading to a bias toward negative predictions. The use of oversampling and undersampling techniques can help mitigate this issue, but care must be taken to ensure that the resulting dataset is representative of the population being studied.Despite these challenges, the studies reviewed have demonstrated that data mining techniques can improve the accuracy of heart disease prediction. Furthermore, the use of intelligent heart disease prediction systems, such as the Intelligent Heart Disease Prediction System (IHDPS), can provide valuable decision support to healthcare practitioners and help reduce treatment costs.However, the analysis of heart disease prediction using different data mining techniques is a promising field of research that can contribute to improved diagnosis and treatment of heart disease. However, further research is needed to standardize the selection of attributes and classifiers, address issues of dataset imbalance, and evaluate the effectiveness of these techniques in diverse populations.

7. Conclusion

This Paper shows different data mining techniques and classifiers have shown varying levels of accuracy in predicting heart disease. The use of Neural Networks and Decision Trees with a larger number of attributes has shown higher accuracy, while the use of Genetic Algorithms and feature subset selection has reduced the number of attributes needed while still maintaining high accuracy. These findings suggest that the choice of data mining technique and classifier should be carefully considered based on the specific needs and characteristics of the heart disease prediction system being developed.

8. References

[1] Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. *IEEE access*, *7*, 81542-81554.

[2] Shah, D., Patel, S., & Bharti, S. K. (2020). Heart disease prediction using machine learning techniques. *SN Computer Science*, *1*, 1-6.

[3] Repaka, A. N., Ravikanti, S. D., & Franklin, R. G. (2019, April). Design and implementing heart disease prediction using naives Bayesian. In *2019 3rd International conference on trends in electronics and informatics (ICOEI)* (pp. 292-297). IEEE.

[4] Khan, M. A. (2020). An IoT framework for heart disease prediction based on MDCNN classifier. *IEEE Access*, *8*, 34717-34727.

[5] Ayon, S. I., Islam, M. M., & Hossain, M. R. (2022). Coronary artery heart disease prediction: a comparative study of computational intelligence techniques. *IETE Journal of Research*, *68*(4), 2488-2507.

[6] Bashir, S., Khan, Z. S., Khan, F. H., Anjum, A., & Bashir, K. (2019, January). Improving heart disease

prediction using feature selection approaches. In *2019 16th international bhurban conference on applied sciences and technology (IBCAST)* (pp. 619-623). IEEE.

[7] Ritchie, R. H., & Abel, E. D. (2020). Basic mechanisms of diabetic heart disease. *Circulation Research*, *126*(11), 1501-1525.

[8] Stoltzfus, K. C., Zhang, Y., Sturgeon, K., Sinoway, L. I., Trifiletti, D. M., Chinchilli, V. M., & Zaorsky, N. G. (2020). Fatal heart disease among cancer patients. *Nature communications*, *11*(1), 2011.

[9] Iung, B., Delgado, V., Rosenhek, R., Price, S., Prendergast, B., Wendler, O., & Eorp Vhd Ii Investigators. (2019). Contemporary presentation and management of valvular heart disease: the EURObservational Research Programme Valvular Heart Disease II Survey. *Circulation*, *140*(14), 1156-1169.

[10] Reddy, G. T., Reddy, M. P. K., Lakshmanna, K., Rajput, D. S., Kaluri, R., & Srivastava, G. (2020). Hybrid genetic algorithm and a fuzzy logic classifier for heart disease diagnosis. *Evolutionary Intelligence*, *13*, 185-196.

[11] Orton, E. C. (2023). Valvular heart disease. *Small Animal Soft Tissue Surgery*, 936-943.

[12] Sekar, J., Aruchamy, P., Sulaima Lebbe Abdul, H., Mohammed, A. S., & Khamuruddeen, S. (2022). An efficient clinical support system for heart disease prediction using TANFIS classifier. *Computational Intelligence*, *38*(2), 610-640.