ISSN: 2633-4828 Vol. 5 No.3, September, 2023

International Journal of Applied Engineering & Technology

ROLE OF HYBRID MACHINE LEARNING ALGORITHM IN ENHANCING PREDICTION OF HEART DISEASES

Inam Ullah[1], Zafar Iqbal[2] and Nayab Khalil[3]
[1]Chairperson Department of Computer Sciecne, Mir Chakar Khan Rind University Sibi, Balochistan
[2,3]M.S Scholar, Mir Chakar Khan Rind University Sibi, Balochistan

ABSTRACT

Globally, blockage of blood vessels is considered main cause of heart failure that leads to death, pain (Angina), and paralysis. World Health Organization (WHO) reports that yearly, 17.9 million people die due to heart disease across the world. Secondly, the number of cases is increasing specifically in teenagers because of unawareness of techniques available for identification of heart disease. Insite of various studies have been conducted worldwide, still the identification of heart disease is missing element in existing body of literature especially in the county of Pakistan. Some past studies report high accuracy by using some kind of organizers to magnify accuracy which may affect the real identification of disease. In this research, a machine learning algorithm are used by combining eXtreme Gradient Boosting (XGB), Gradient Boost (GB), and Random Forest algorithms, known as hybrid approach. The proposed method outperforms existing research in terms of accuracy by comparing with state-of-the-art methods, achieving a better accuracy rate of approximately 90.21%.

Keywords: Blood vessels, heart disease, Machine Learning algorithm, hybrid, Gradient Boosting

INTRODUCTION

Machine learning facilitates researchers in deploying and extracting known and unknow information from previously available data (Sini, 2011). It is a broader filed as its scope and implementation status gaining importance day by day. It boosts numerous supervised, unsupervised, and ensemble learning classifiers which are used to predict accuracy of available data with high efficiency. We can use this advanced technology for the identification of heart diseases at the early stage, that will help number of people because cardiac disease is one of the challenging illnesses that severely affect the human heart with described range of conditions.

Health is one of the global challenges for humanity (Haldane, 2023; Saxena, 2023). The World Health Organization (WHO) says good health is a basic human right (AMA J Ethics. 2015). WHO estimates that 17.9 million global deaths from heart diseases (Dangare 2012). Therefore, it is important to provide them with good health care to stay healthy and fit. Worldwide, 31% of deaths are due to heart problems (Ramkumar & Sathyashree, 2021). Diagnosis and treatment of heart disease can be very difficult due to lack of availability of diagnostic tools and such unavailability makes physicians and other professionals uncomfortable to provide accurate diagnosis and treatment to heart patients especially in developing countries of the world. In this regard machine learning and other computer technologies have been emerged as beneficial toolkit to create programs that help physicians make decisions with information related to cardiology. Early diagnosis can reduce the risk of heart disease. The technology namely Medical Data Mining (MDM) supports to get valuable knowledge, information, and patterns from medical data. As literature suggests time is considered main gadget in medical information and could affect the efficacy in case of misdiagnosis. Similarly, it is also a big challenge to deal large amount of data in healthcare centers. In order to overcome these challenges Data Mining helps to cover these with actionable approach by using techniques and methods. Such actionable information may help cardiologists facilitate larger amount of patients in shorter period with quick decisions (Manna, 1990).

Heart disease is a condition affecting the heart, encompassing issues like narrowed or blocked blood vessels that may result in heart attacks, chest pain (angina) due to reduced blood flow, or strokes. Variants of heart disease also include conditions impacting the heart muscle, valves, or rhythm. Symptoms vary depending on the specific type of heart disease a patient has. There are distinctions in how heart disease manifests in men and women; for instance, men are more prone to chest pain, while women may experience symptoms like shortness of breath, nausea, or extreme fatigue. Narrowed blood vessels in these areas can lead to symptoms such as pain, numbness, weakness, or coldness in the hands or feet.

The heart is an important part of the body. It pumps blood, providing oxygen, nutrients and other resources to different parts of the body. Therefore, this muscular system plays an important role in the body. Heart problems also affect the normal functioning of other parts of the body. They are responsible for a third of the world's human deaths. Thus, it is very important to diagnose the disease at an early stage. Effective diagnosis and treatment of the patient play an important role in health care. Improper diagnosis leads to unacceptably serious consequences (Arghandabi, 2020).

Heart attacks are currently on the rise. An estimated 12 million people worldwide die each year, according to WHO research. The early symptoms of heart disease can be identified by the following symptoms: headache, dizziness, frequent fainting, difficulty eating and breathing, and indescribable fatigue and chest pain and heart palpitations. Causes of heart disease include: high blood pressure, diabetes, smoking, high cholesterol and obesity. (Purnomo, 2021).

This problem accounts for about half of all deaths caused by stroke and heart disease. Cardiovascular disease, also known as coronary heart disease (CVD), refers to a group of heart-related diseases that include more than one heart attack. Cardiovascular disease is divided into several types, including cardiovascular disease, cardiomyopathy and cardiomyopathy. Cardiovascular covers groups of disordering that may affect the heart functioning, blood vessels, and flow of blood delivered to throughout the body. This case is difficult to diagnose but important that could be completed in effective manner. Various monitoring systems use methods to identify stored data, such as ocean neural network procedures and weight measurement systems. Weka 3.6.6 is used to measure data processes (Krishna, 2018).

Heart disease prediction, the most important question in this area of research is to find the best system for it. There is a lot of work in this area, where a variety of techniques have been used to predict heart disease, but only because of certain conditions. The system still needs to be made more precise and efficient. Therefore, it is important to develop an intelligent prognostic model that accurately and effectively predicts heart disease.

LITERATURE REVIEW

Machine Learning Methods and Data Mining Approach are commonly used to predict cardiac attack. The body of literature suggests numerous machine learning algorithms including Decision Tree, Naïve Bayes, Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Blurred Logic, and Artificial Neural Network (ANN) are used to analyze medical data for desirable results (Pravesjit, 2022). Katarya et al. (2020) proposed two algorithms for detecting heart disease, a decision tree and a random forest. Accuracy of random forest and decision tree is 71.50% and 75%.

Al-Yarimi (2021) the author focuses on the most used machine learning algorithms for logistic regression, immediate K neighborhood, artificial neural network, SVM, Naive Bayes, critical tree, and random forest. Accuracy achieved logistic regression 84, near neighborhood K 76, artificial neural network 74, SVM 75, new rule 83, decision tree 74, and random forest 83. Logistic regression accuracy is higher than others.

Arghandabi (2020) the author focuses on machine learning methods. The author used hybrid methods, the accuracy of the analysis was 88.7%. Bhatt and Chouhan (2021) used two ML algorithms, Decision Tree and Naïve Bayes. The authors claims decision tree works better than Naive Bayes by getting the accuracy decision tree is 91% and naïve bayes is 87%.

Krishna (2018) the author proposes hybrid machine learning methods for heart disease prediction using various algorithms. Gaussian NB 93.44%, Vector Machine Support 90.16%, Random forest 95.08%, Hoffing tree 81.24%, logistic model tree 80.69%

Jindal, Agrawa and Jain (2021) Use different algorithms to analyze accuracy, possibly the nearest neighbor, decision tree, linear regression, and support vector machine (SVM) by getting the accuracy a support vector machine 83%, a resolution tree of 79%, a linear regression of 78%, and a K nearest neighbor 87%.

Kirubakaran, Kumar, Rajeswari, and Daniya (2021) used Machine Learning Algorithms including Support Vector Machine, Nio 22, MLP Classifier, Radmom Forest, Regression Logistics, and Decision Tree in order to predict heart disease with high accruacy. After implementing the algorithm, the author obtained a resolution estimate without PCA decision tree 59, random forest 58, logistic regression 86, support vector machine (SVM) 53, MLP rating 62 and naive Bayesian 79 and solution level accuracy PCA tree 70%, random forest 84%, logistic regression 68%, support vector machine (SVM) 55%, estimated MLP 69% and naive Bayesian 68%.

Ramotra and Mansotra (2021) the author seeks to increase the accuracy of heart disease prediction by using and identifying key features and data mining methods. Various combinations of features and seven classification technologies for developing forecasting models. Hybrid techniques with Bayesian and logistic regression were used. The most popular Cleveland database was used in the study. After experimenting with a heart disease prediction model, after applying this technique to get accuracy of 87.4%.

Jindal, Agrawal, Khera, and Jain (2021) the authors used various patient characteristics to predict the patient's heart disease, such as: age, blood pressure, cholesterol, gender, blood sugar, and so on. League. The accuracy achieved are: Minimum Continuous Improvement (SMO) 84.07, Bayes Net (BN) 81.11, Multilayer Perception (MLP) 77.4, Navies Bayesian (NB) 89.77. Similarly, Tarawneh and Embarak (2019) in this study, the authors combine various machine learning algorithms. Using the hybrid method, it achieves an accuracy level of 89.2%.

Dewan, Sharma, and Meghna, (2015) conducted a research study with the key objective of to create a model than can support to identify key pattenrs, information, and relationship realted to heart disease. It may detect quires related to heart disease and support medical consultants to make better clinical decisions as compared to past medical decision system. It can help to reduce treatment costs by delivering effective treatments. The proposed system was implemented in MATLAB.

Pahwa, (2017) proposed to apply Naïve Bayes and Random Forest in order to predict cardiac disease. A proposed approach involves feature selection prior to classification to enhance model quality. The SVM-RFE and gain ratio algorithms are utilized on the dataset for feature selection, assigning weights to each feature in the process. This methodology contributes to heightened reliability and decreased processing time. Experimental findings demonstrate that this selection method significantly enhances the accuracy of both models.

Yekkala, Dixit, and Sunanda (2017) this paper analyzes different ensemble methods (Bagged Tree, Random Forest, and AdaBoost) together with Feature subset selection method- Particle Swarm Optimization (PSO) to accurately predict a patient’s incidence of heart disease. Experimental results show the highest accuracy obtained by Bagged Tree and PSO.

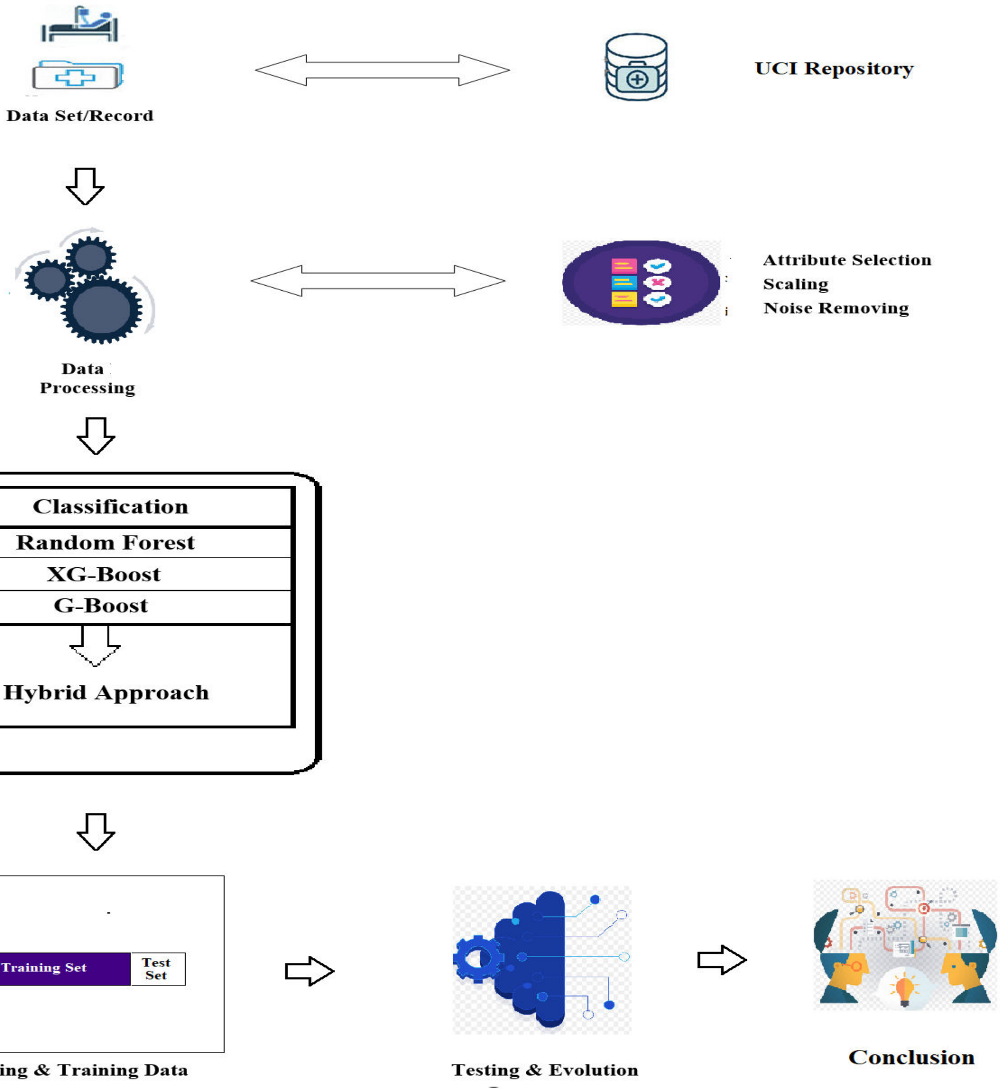
Bhatla (2012) this paper aims to analyze the different techniques evolved in recent years for the prediction of heart disease. The results show that neural networks with 15 different attributes have surpassed all other data mining techniques.

Purushottam and Richa (2015 developed a system in this analysis that can effectively detect rules for predicting patients ' risk levels based on the defined safety parameter. The laws may be prioritized according to the demand of the consumer. The system's efficiency is evaluated for diagnosis precision and the results show that it has a strong potential to better predict the risk of heart disease.

PROPOSED METHOD

This part of the study covers methods for the prediction of heart disease with a mathematical background. Figure 1 shows detailed information regarding the steps used in this study. The methodology is a systematic process that converts raw information into processed information which may provide real picture of data to readers. The flowchart covers steps including data collection, extraction of important values from data, and procession step of data. Preprocessing helps to clean up the data. After that classifier was used to classify the data. The proposed model is then implemented to assess its accuracy and performance using a range of performance metrics. This model utilizes 13 medical parameters, including chest pain, fasting sugar levels, blood pressure, cholesterol levels, age, and gender, among others, to make predictions.

Figure 1 Proposed Method



Data Collection

Data collection is the process of obtaining, measuring, and assessing an accurate data set for research using conventional verification procedures. The researcher can evaluate his hypothesis based on the data sets gathered. For our study, for this research collect the data on heart disease from online platform namely UCI Repository.

Data Preprocessing

After data collection the next phase is data preprocessing. Preprocessing is step of machine learning in which data preparation and transformation of the dataset and seeks at the same time to make knowledge discovery more efficient. In data preprocessing the first step is that we will collect only related dataset parts which are helpful in this research work and those dataset parts which are not related to our research work like inconsistent and or deficient in definite behaviors are trends and is likely to contain errors are discarded and then stored.

Classification and Prediction

After Data Processing the next phase is classification. Classification is the method of arranging data into different categories in order to give a specific form and coherent structure to the collected data, which facilitates its application in the most systematic and efficient way. It is the process of gathering statistical data under different homogeneous groups that is understandable for a comfortable interpretation. In Classification first we will apply algorithms on data sets after applying algorithms we will apply Hybrid approach.

Testing and Evaluation

After classification, the next step is testing and evaluation. Testing is one of the ways to assess potential behavior, while assessment is the process of examining a problem or condition so that it can be understood and diagnosed. In Testing & Evaluation we will produce more accurate results then the previous one. After this we will visualize the results in graphs and then conclusion. It is described that for heart disease prediction data mining techniques are used. Pre-defined datasets are cleaned at first stage by removing all irrelevant attributes. Then only relevant values from the selected attributes, which are needed can be separated for testing and training of the data. After that the researcher runs our datasets in PYTHON 3.9 GOOGLE COLAB and extract the prediction values. Four different algorithms are applied on the dataset and results are generated.

G Boost Algorithm

Gradient Boosting Algorithms are a set of machine learning techniques that blend numerous weak learning models to form a robust predictive model. Typically, decision trees are employed in gradient boosting processes. These models are gaining traction due to their efficacy in classifying intricate datasets and have notably excelled in winning various data science competitions on platforms like KAGGLE and the UCI repository. The popular Python machine learning library, Scikit-Learn, offers various implementations of gradient boosting classifiers, such as XG Boost.

XG Boost

Extreme Gradient Boosting and Distributed Gradient Bossted Decision Tree (GBDT) are scalable gadgets in emerging machine larning library. They provde similar tree boosting mechnasim in machine learning to cover, regression, ranking, classifciaiton related problems. The XG Boost is connceted with machine larning therefore to understand the boost first need to fully grap the cocecpt of machine learning in terms of decision tree, ensemble learning, supervised machine learning and gradient boosting. Supervised machine learning helps to train a model that may identify patterns with feactures and proper lables in a dataset. Later on this trained model may also help to predict the lables on a new dataset features.

Random Forest

Random Forest is a supervised learning algorithm. The 'forest' you create is a series of thin trees, usually trained in a 'packaging mode'. The general idea of the mobilization process is to combine the learning units with the total number of steps.

Hybrid Approach

A hybrid approach is employed for both classification and regression tasks, utilizing multiple sub-models. Each sub-model generates predictions, which are then combined using aggregation methods such as taking the mean or mode of the predictions. This aggregation process involves allowing each sub-model to contribute its prediction, effectively "voting" on the final outcome. It is effective & use to assemble algorithms. It uses parameters to configure the prediction of combined sub models. Voting approach is controlled by rule parameter which is used to take average of the probabilities by default.

EXPERIMENTAL RESUTS AND DISCUSSION

Basically, this proposed method is for prediction of heart disease using hybrid ML algorithm, which method was evaluated using various metrices, including accuracy, precision, recall and F1-score. The proposed method give high accuracy on the test dataset, indicates its effectiveness to identify the heart disease in early stage.

Confusion Matrix

In machine learning confusion matrix is a performance evaluation tool to represent the accuracy of a classification model by displaying the number of true positives, true negatives, false positives and false negatives.

We may generate several metrics to assess the correctness of our model using the confusion matrix.

1. Accuracy (all correct / all) = TP + TN / TP + TN + FP + FN
2. Misclassification (all incorrect / all) = FP + FN / TP + TN + FP + FN
3. Precision (true positives / predicted positives) = TP / TP + FP
4. Recall (true positives / all actual positives) = TP / TP + FN
5. Specificity (true negatives / all actual negatives) =TN / TN + FP

Accuracy

Accuracy is the ratio between correct and wrong prediction. Data quality and errors are important factors in measuring accuracy. It is measured by taking the appropriate image and displaying it as a percentage. Accuracy is measured using the following formula.

$$\text{Accuracy} = \frac{(TN+TP)}{(TN+TP+FN+FP)}$$

Were, TN-True Negative, TP-True Positive, FP-False Positive & FN-False Negative. Below is the description of TP rate, FP rate, Precision, Recall, F- Measure, MMC, ROC Area PRC Area and Class.

TP rate: True positive rate (correctly classified as an instances for a given class)

FP Rate: False positive rate (incorrectly classified as an instances for a given class)

Precision: The proportion of class that actually belong to a partition is divided by the total instance of the class.

Recall: The proportion of cases/instances categorized as a specific category divided by the actual total in that class (i.e. equal to TP rate) F-Measure: A combination of accuracy and retrieval, calculated as 2 * Accuracy * Call / (precision + Re Call)

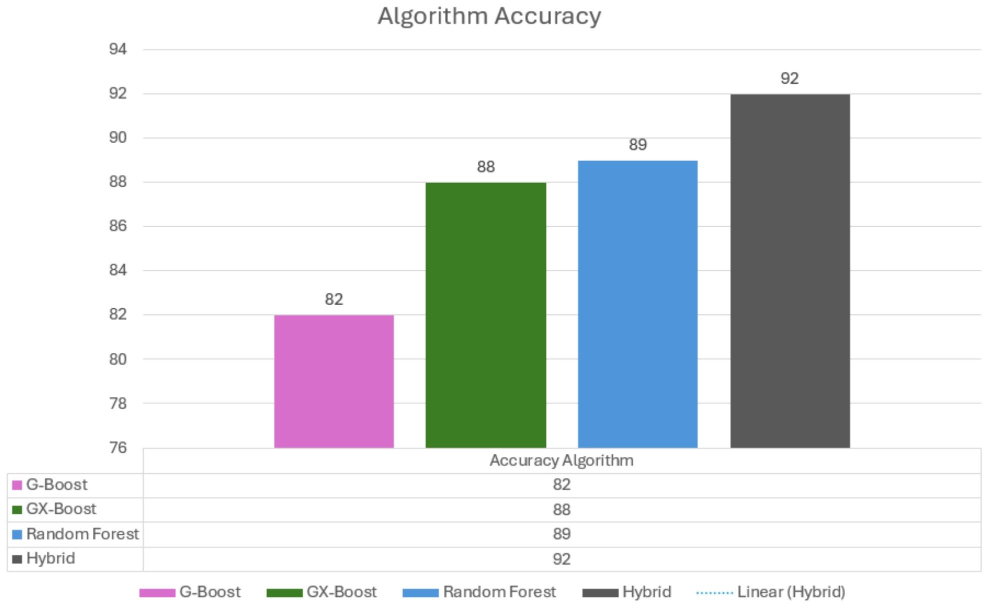
ROC (Receiver Operating Characteristics) Measurement Area: One of the most important WEKA output values. They let you know the overall performance of the work.

Table 1: Accuracy of Algorithms

Algorithms	Accuracy
G Boost	82.00%
XG Boost	88.00%
Random Forest	89.00%
Hybrid Approach	90.21%

Table 1 shows the accuracy of each algorithm, this means that each algorithm has correctly classified instances of class in percentage. Accuracy of each algorithm that G Boost algorithm gives accuracy of 82.00 percent, XG Boost gives 88.00, Random Forest gives accuracy of 89.00 and Hybrid approach gives the accuracy of 90.2128 percent.

Figure 2: Comparison of Algorithms Accuracy



In Figure 2 shows detailed comparison of Algorithms which shows accuracy of both techniques which is used in previous research work and in my research work. The graph shows that the technique used for this research work is better in predicting the heart disease.

CONCLUSION

It is concluded that the World Health Organization declared cardiac disease the most common syndrome in this era that leads to death. The study further concludes that today's modern technologies enables medical professionals to utilize their full potential in order to reduce the rate of death ratio. It would be a social injustice if medical practitioners did not utilize these technologies where every human error leads to a possible loss of life.

FUTURE WORK

The accuracy achieved by the hybrid algorithm can be enhanced further to minimize prediction errors. Future research efforts could focus on eliminating false positives in current prediction models. Additionally, these algorithms hold potential for application in predicting the progression patterns of various diseases beyond their current scope. There's also the opportunity to develop web and mobile applications where users can input their personal medical data and receive predictions based on their health conditions.

REFERENCES

Al-Yarimi, F. A. (2021). Feature optimization by discrete weights for heart disease prediction using supervised learning. Soft Computing, 25(3), 1821--1831.

AMA J Ethics. 2015;17(10):958-965. doi: 10.1001/journalofethics.2015.17.10.msoc1-1510.

Arghandabi, H. a. (2020). A Comparative Study of Machine Learning Algorithms for the Prediction of Heart Disease. International Journal for Research in Applied Science & Engineering Technology (IJRASET), 8(7), 1--9.

Bhatla, N. a. (2012). An analysis of heart disease prediction using different data mining techniques. International Journal of Engineering, 1(8), 1-4.

Dangare C S & Apte S S (2012). Improved study of heart disease prediction system using data mining classification techniques. International Journal of Computer Applications, 47(10), 44-8

Dewan, Sharma, A., & Meghna. (2015). Prediction of heart disease using a hybrid technique in data mining classification. In 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom (pp. 704--706).

Haldane, V., Ariyarajah, A., Berry, I., Loutet, M., Salamanca-Buentello, F., & Upshur, R. E. (2023). Global inequity creates local insufficiency: A qualitative study of COVID-19 vaccine implementation challenges in low-and-middle-income countries. PloS one, 18(2), e0281358.

Jindal, H., Agrawal, S., Khera, R., & Jain, R. (2021). Heart disease prediction using machine learning algorithms. Materials Science and Engineering, x(x), 1-11

Katarya, R., & Meena, S. K. (2020). Machine Learning Techniques for Heart Disease Prediction:. Health and Technology, x(x). doi:https://doi.org/10.1007/s12553-020-00505-7

Kirubakaran, S. S., Kumar, B. S., Rajeswari, R., & Daniya, T. (2021). Heart disease diagnosis systematic research using data mining and soft. Materials Today: Proceedings, xx(xx), 1-7.

Krishna, A. a. (2018). An Efficient Heart Disease Prediction using Various Data Mining Techniques. International Journal of Computer science engineering Techniques –, 3(2), 1--4.

Krishna, A. a. (2018). An Efficient Heart Disease Prediction using Various Data Mining Techniques. International Journal of Computer science engineering Techniques –, 3(2), 1--4

Manna, A. L. (1990). Promoting student health through children's literature. Educational Horizons, 69(1), 37--44.

Pahwa, K. a. (2017). Prediction of heart disease using hybrid technique for selecting features. In 2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON) (Vol. 1, pp. 500--504). IEEE

Pravesjit, S., Kantawong, K., That, V., & Longpradit, P. (2022). Segmentation of Broken Khmer Characters. In 2022 3rd International Conference on Big Data Analytics and Practices, IBDAP 2022 (pp. 65-68).

Purnomo, A. a. (2021, xx xx). Adding feature selection on Na{\"\i}ve Bayes to increase accuracy classification heart attack disease. (xx, Ed.) Journal of Physics: Conference Series, 1511, xx.

Purushottam, Prof. (Dr.) , K., & Richa , S. (2015). Efficient heart disease prediction system using decision tree.

Ramkumar, S., & Sathyashree, M. (2021). An Undertaken Report For Heart Disease Prediction And Identification Using Machine Learning Methods. Annals of the Romanian Society for Cell Biology, 25(4), 18313--18322.

Ramotra, A. K., & Mansotra, V. (2021). A Hybrid Cluster and PCA-Based. In V. S. ·, Rising Threats in Expert Applications and Solutions (pp. 111--117). Springer.

Saxena, A., Baker, B. K., Banda, A., Herlitz, A., Miller, J., Karrar, K., ... & Hassoun, N. (2023). Pandemic preparedness and response: beyond the Access to COVID-19 Tools Accelerator. BMJ global health, 8(1), e010615.

Soni J, Ansari U, Sharma D & Soni S (2011). Predictive data mining for medical diagnosis: an overview of heart disease prediction. International Journal of Computer Applications, 17(8), 43-8

Tarawneh, M., & Embarak, O. (2019). Hybrid approach for heart disease prediction using data mining techniques. 2016 International Conference on Circuit, Power and Computing Technologies [ICCPCT], 447--454

Yekkala, Dixit, I., & Sunanda , M. J. (2017). Prediction of heart disease using ensemble learning and Particle Swarm Optimization. In 2017 International Conference On Smart Technologies For Smart Nation (SmartTechCon) (Vol. 1, pp. 691-698). IEEE