Physical Sciences, Life Science and Engineering Volume: 2, Number 3, 2025, *Page*: 1-9

Classification Of Heart Disease Using Feature Selection and Machine Learning Techniques

Sondos Jameel Mukhyber

General Directorate for Education of Diyala, Iraq

DOI: https://doi.org/10.47134/pslse.v2i3.386
*Correspondence: Sondos Jameel Mukhyber
Email: shabaadameeriraqi@gmail.com

Received: 07-02-2025
Accepted: 13-03-2025
Published: 07-04-2025

Abstract: Heart disease is a complex disease that affects a large number of people worldwide. The timely and accurate detection of heart disease is critical in healthcare, particularly in the field of cardiology. In various fields around the world, machine learning is used. There are no exceptions in the healthcare sector. Machine learning can be crucial in determining whether or not there will be locomotor abnormalities, heart ailments, and other conditions. If foreseen far in advance, such information can offer crucial intuitions to doctors, who can then modify their diagnosis and approach per patient. in this paper it has been used a variety of machine learning techniques and used the heart disease dataset to evaluate its performance using different metrics for evaluation, such as accuracy, precision, recall ,and F-measure. For this purpose, it has been used five classifiers of machine learning such as Support Vector Machine, Gaussian Naïve Bayes, Decision Trees, Artificial Neural Network, and Logistic Regression. Furthermore, it has been check their accuracy on the standard heart disease dataset by performing certain pre-processing of dataset, and feature section. Finally, the experimental result indicated that the accuracy of the prediction classifiers.

Keywords: Heart Disease, Support Vector Machine, Logistic Regression, Decision Trees, Artificial Neural Network.

Introduction

Heart Failure (HF) is the failure of heart to pump sufficient amount of blood to meet the needs of the body. Narrowing or blockage of the coronary arteries is considered to be the main cause of HF. Coronary arteries are those arteries which are responsible for carrying blood to the heart itself. About half of the people who develop heart failure (HF) die within five years of diagnosis. The common symptoms of HF include shortness of breath, swollen feet and weakness of the body.

Heart disease has created a lot of serious concerned among researches; one of the major challenges in heart disease is correct detection and finding presence of it inside a human. Early techniques have not been so much efficient in finding it even medical professor are not so much efficient enough in predicating the heart disease. There are various medical instruments available in the market for predicting heart disease there are two major problems in them, the first one is that they are very much expensive and second one is that they are not efficiently able to calculate the chance of heart disease in human.

With advancement in computer science has brought vast opportunities in different areas, medical science is one of the fields where the instrument of computer science can be used. Machine Learning is one such tool which is widely utilized in different domains because it doesn't require different algorithm for different dataset. Reprogrammable capacities of machine learning bring a lot of strength and opens new doors of opportunities for area like medical science. In medical science heart disease is one of the major challenges; because a lot of parameters and technicality is involve for accurately predicating this disease.

Machine learning could be a better choice for achieving high accuracy for predicating not only heart disease but also another diseases because this vary tool utilizes feature vector and its various data types under various condition for predicating the heart disease, algorithms such as Naive Bayes, Decision Tree, Neural Network, are used to predicate risk of heart diseases. All these techniques are using old patient record for getting predication about new patient. This predication system for heart disease helps doctors to predict heart disease in the early stage of disease resulting in saving millions of life.

In this paper, it has been proposed a machine learning classifier that includes Support Vector Machine , Gaussian Naïve Bayes, Decision Trees, Artificial Neural Network, and Logistic Regression for heart disease prediction. the normalization and reduce the features for the best machine learning classifier. Apart from that, various performance evaluation parameters, such as accuracy, precision, recall, and F-measures, are used for the machine learning classifier's performance. The proposed method has been tested on the Cleveland heart disease dataset.

Literature Review

Mr.Valle Harsha Vardhan et al employed seven machine learning algorithms in this study to gauge performance are SVM, applied to the dataset along with Decision Tree, Random Forest, Naive Bayes, Logistic Regression, Adaptive Boosting, and Extreme Gradient Boosting. A prediction model is created after comparing the accuracy of each of the seven machine learning techniques. the objective is to employ a variety of evaluation metrics, such as the confusion matrix, accuracy, precision, recall, and f1-score, which accurately predicts the disease. The extreme gradient boosting classifier has the highest accuracy (81%), when all seven are compared.

K. Karthick et al employed six ML classification algorithms for developing heart disease risk prediction model, namely, SVM with RBF kernel, Gaussian Naive Bayes, logistic regression, LightGBM, XGBoost, and random forest, were applied to UCI ML repository's Cleveland HD dataset, and obtained the accuracy as 80.32%, 78.68%, 80.32%, 77.04%, 73.77%, and 88.5%, respectively. Random Forest achieves 88.5% accuracy during validation for 303 data instances with 13 selected features of Cleveland HD dataset.

Rohit Bharti et al presented a paper for predicting heart disease. In this paper different machine learning algorithms and deep learning are applied to compare the results and analysis of the UCI Machine Learning Heart Disease dataset. Various results are

achieved and are validated using accuracy and confusion matrix. They used algorithms like LR, KNN, SVM, RF,and DT and obtained the accuracy as 83.3%, 84.8%, 83.2%, 80.3%,and 82.3%, respectively.

Yamala Sandhya presented a research paper for predicting heart disease using machine learning algorithms on Cleveland HD dataset. She used the Support Vector Machine to predict and identify the heart diseases of patients. she compared the result of the Support Vector Machine with the other machine algorithms like Artificial Neural Network, Naive Bayes, KNN and Decision Support. The Support Vector Machine algorithm gives the better accuracy, specificity and sensitivity.

Senthilkumar mohan et al proposed a novel method that aims at finding significant features by applying machine learning techniques resulting in improving the accuracy in the prediction of cardiovascular disease. The prediction model is introduced with different combinations of features and several known classification techniques such as Naive Bayes, Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine.

Methodology

For evaluation of risk of heart disease using a combination of models, this paper proposes the framework shown in Figure 1. This approach is divided into six modules involving pre-processing, feature section, training, testing, application of classifies, and finally, performance evaluation of the classifiers . The modules have been described below

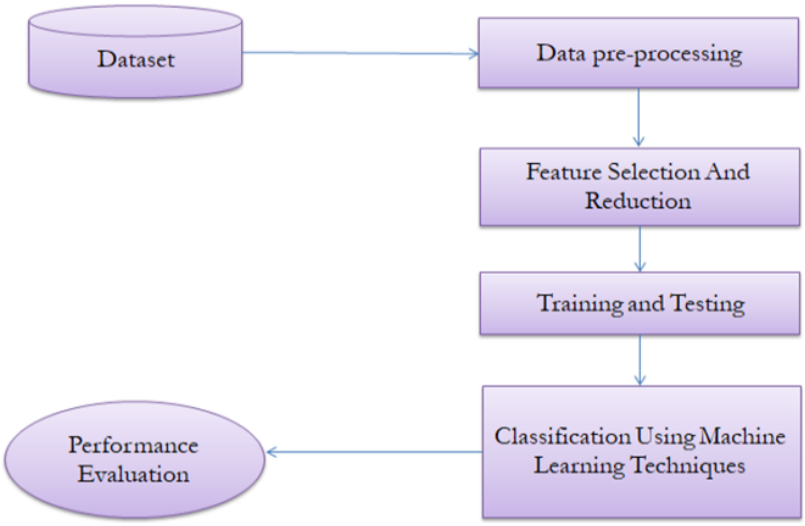


Figure 1. Block diagram of the proposed model.

Data Collection

The Cleveland heart disease dataset, available from the University of California, Irvine (UCI) online repository for machine learning, is the most prominent dataset used by the researchers. There are 303 records, where 6 records are with some missing values. Those 6 records have been removed from the dataset and the remaining 297 patient records are used in this study. The dataset indicate that 137 records show the value of 1 establishing

the presence of heart disease while the remaining 160 reflected the value of 0 indicating the absence of heart disease.

Table 1. UCI dataset attributes detailed information

Attributes	Description	Type
Age	Patient's age	Numeric
Sex	Patient's gender (1 for male and 0 for female)	Nominal
Cp	Type of chest pain categorized into 4 value: 1. typical angina 2.atypical angina 3.nonanginal pain and 4.asymptomatic	Nominal
Treastbps	Level of blood pressure at resting mode	Numeric
Chol	Serum cholesterol	Numeric
Fbs	Blood sugar levels (1 in case of true and 0 in case of false)	Nominal
Resting	Results of electrocardiogram (0 for normal state, 1 for abnormality in ST-T wave, and 2 for probability or certainty of LV hypertrophy by Estes)	Nominal
Thalach	Maximum heart rate achieved	
Exang	Exercise induced angina (1 for yes and 0 for no)	Nominal
Oldpeak	ST depression induced by exercise relative to rest	
Slope	The slope of the peak exercise ST segment (0 for unsloping, 1 for flat, and 2 for downsloping)	Nominal
Ca	Number of major vessels (0–3) colored by fluoroscopy	Nominal
Thal	Probably thalassemia (0 for normal, 1 for fixed defects, and 2 for reversible defects)	Nominal
condition	0 = no disease and 1 = disease	Nominal

Data pre-processing

Normalization–is one of the most general data pre-processing techniques in machine learning. A dataset's attribute is normalized as its values are scaled to fall into a narrow range, such as 0.0 to 1.0.There are various techniques of normalization are available such as minmax, z-score, and decimal normalization. In this paper used max-min normalization.

$$X_{new} = (X_i - X_{max})/(X_{max} - X_{min}) \quad (1)$$

Where X_{max} is greater value in the column and X_{min} is smaller value in the column.

Feature Selection and Reduction

From among the 13 attributes of the data set, two attributes pertaining to age and sex are used to identify the personal information of the patient. The remaining 11 attributes are considered important as they contain vital clinical records. Clinical records are vital to diagnosis and learning the severity of heart disease. Those two attributes have been removed and the remaining 11 attributes vital clinical records are used in this study.

Training and Testing

Train/Test is a method to measure the accuracy of a algorithm. It is called Train/Test because it split the data set into two sets: a training set and a testing set. In this paper the data has been divided into 76% for training, and 24% for testing.

Classification Using Machine Learning Techniques

- **ANN (Artificial Neural Network)** – ANN is a classification model which is collected by interconnected nodes. It can be viewed as a circular node which is represented as an artificial neuron that reveals the output of one neuron to the input of other. The ANN model is useful in revealing the hidden relationships in the historical data, thus facilitating the prediction and forecasting of diseases of patients. ANN model is accurate enough to make significant and relevant decisions regarding data usag .
- **LR (Logistic Regression)** – Logistic regression has been widely used in artificial intelligence and machine learning due to its deep theoretical basis and good practical performance. The gradient descent approach is the most commonly used and it used in this paper**.**
- **Gaussian NB (Gaussian Naive Bayes)** – This technique is based on Bayes' Theorem. It is a type of classification algorithm working on continuous normally distributed features that is based on the Naive Bayes algorithm. This classification algorithm does really well in predicting the correct class the present features belong to. The 'naive' in the name of this algorithm is used based on the assumption that this algorithm considers while predicting the label of the features. The assumption here is that all the features are independent of each other, though this might not be true in a real-world scenario. Still, the algorithm works fine. In the training process of a Naive Bayes classifier.
- **DT (Decision Tree)** – Decision trees are a frequently used classification and regression technique because of their easy interpretation compared to other classification methods, their realization at lower costs, the ease of integration with databases and a good level of reliability .In addition, leaves shown as decision rules in decision trees can be easily interpreted by people working in this field and this method is used effectively in high dimensional data.
- **SVM (Support Vector Machine)** – Support Vector Machine (SVM) algorithm was introduced and presented by Vapnik in 1995 as a supervisory algorithm. The instruction of this algorithm is using accuracy for generalization of errors. The algorithm operates by forming a hyper plane and divides the data into classes. The method of this division is in this way: all the samples that are belonged to a class are placed in one side and other classes are placed in the other side. For performing the SVM classifier operation, a linear classification of the data is defined and in division process, it tries to choose a line that has the greatest margin of safety.

Assessment Criteria

Confusion Matrix: is a table that displays actual or predicted values.

- **Accuracy**: This performance measure is calculated by performing ratio of total number of correctly diagnosed cases to the total number of cases.

$$\text{Accuracy} = (TP + TN) / (TP + FP + TN + FN) \quad (2)$$

- **Precision**: This is the ratio of number of correctly classified instances to the total number of instances fetched.

$$\text{Precision} = TP / (TP + FP) \quad (3)$$

- **Recall**: This is a ratio of number of correctly classified instances to the total number of instances in the dataset.

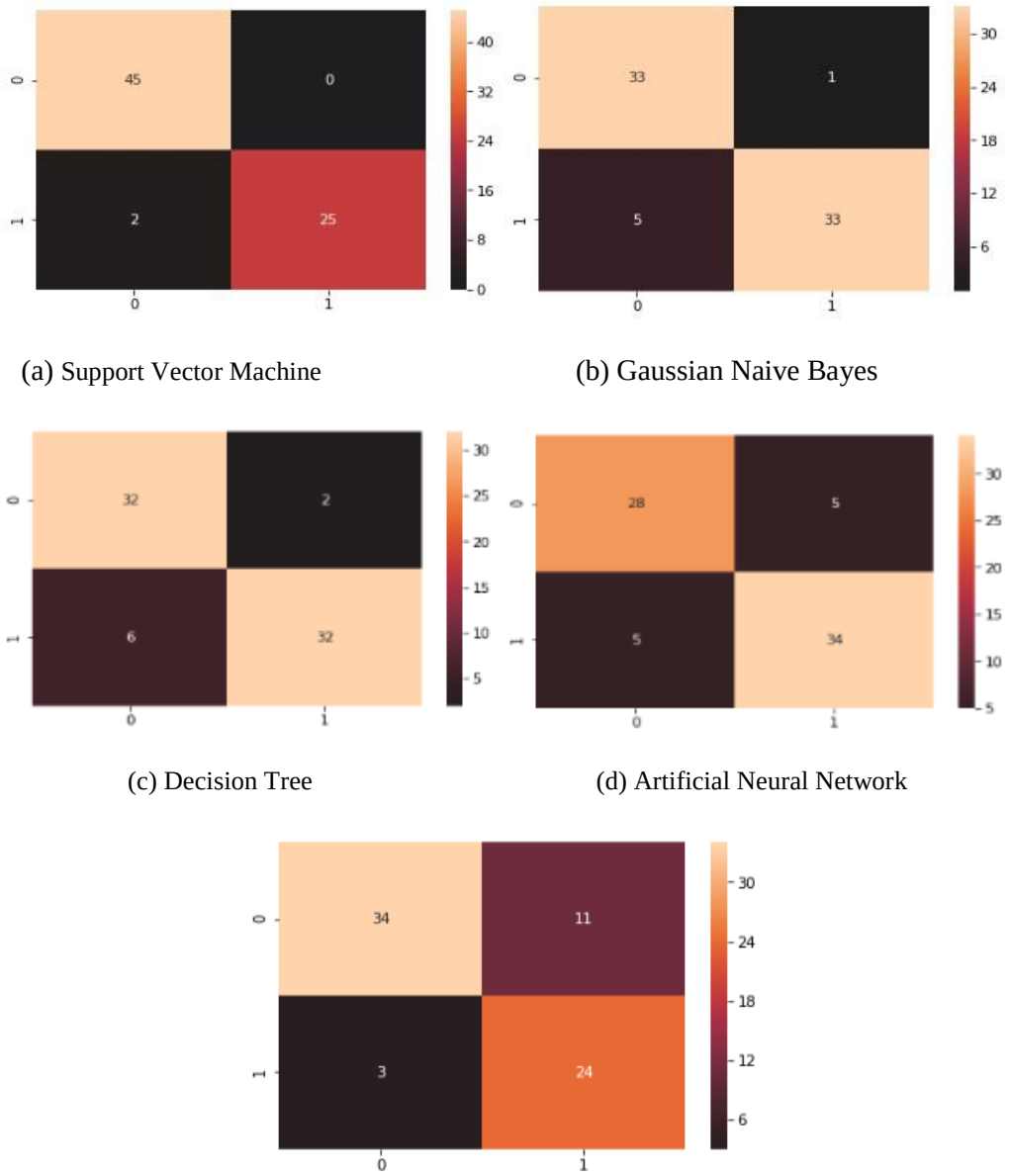
$$\text{Recall} = TP / (TP + FN) \quad (4)$$

- **F – Measure**: This measure computes some average of the information retrieval precision and recall metrics.

$$F - \text{Measure} = (2 * \text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (5)$$

Result and Discussion

All the five classification algorithm is been tested for The Cleveland heart disease dataset



(a) Support Vector Machine (b) Gaussian Naive Bayes

(c) Decision Tree (d) Artificial Neural Network

(e) Logistic Regression

Figure 2. Confusion matrix

Table 2. Classifiers evaluation in terms of precision, recall, F-measure, and accuracy

S. No	Accuracy (%)	Precision (%)	Recall (%)	F – Measure (%)
SVM	97.22	100	92.59	96.15
Gaussian NB	91.66	97.05	86.84	91.66
DT	88.88	94.11	84.21	88.88
ANN	86.11	87.17	87.17	87.17
LR	80.55	68.57	88.88	77.41

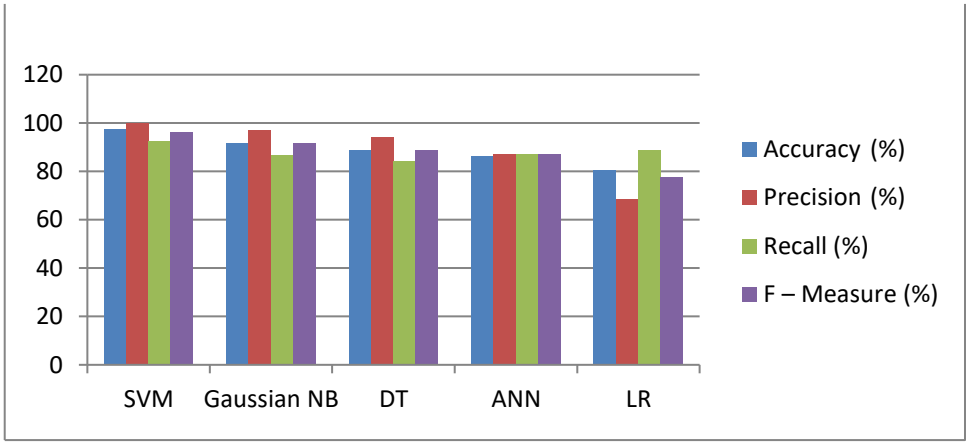


Figure 3. Graphical representation of evaluation of Classifiers performance

Conclusion

The five ML classification algorithms, namely, Support Vector Machine, Gaussian Naive Bayes, Decision Trees, Artificial Neural Network, and logistic regression, were applied to UCI Cleveland heart disease dataset, and obtained the accuracy as 97.22%, 91.66%, 88.88%, 86.11%, and 80.55%, respectively for the selected 11 attributes. The SVM algorithm provides better accuracy as 97.22% . In this classification model, totally 297 data instances have been used. In future, various heart disease datasets from health data repository can be combined, and the best performing classification model using contemporary machine learning models can be outlined.

References

Abdar, M. (2015). A survey and compare the performance of IBM SPSS modeler and rapid miner software for predicting liver disease by using various data mining algorithms. Cumhuriyet Üniversitesi Fen Edebiyat Fakültesi Fen Bilimleri Dergisi, 36(3), 3230-3241.

Ahmad, G.N. (2022). Efficient Medical Diagnosis of Human Heart Diseases Using Machine Learning Techniques with and Without GridSearchCV. *IEEE Access*, *10*, 80151-80173, ISSN 2169-3536, https://doi.org/10.1109/ACCESS.2022.3165792

Ali, L., Niamat, A., Khan, J. A., Golilarz, N. A., Xingzhong, X., Noor, A., ... & Bukhari, S. A. C. (2019). An optimized stacked support vector machines based expert system for the effective prediction of heart failure. *IEEE Access*, *7*, 54007-54014.

Ayon, S.I. (2022). Coronary Artery Heart Disease Prediction: A Comparative Study of Computational Intelligence Techniques. *IETE Journal of Research*, *68*(4), 2488-2507, ISSN 0377-2063, https://doi.org/10.1080/03772063.2020.1713916

Azmi, J. (2022). A systematic review on machine learning approaches for cardiovascular disease prediction using medical big data. *Medical Engineering and Physics*, *105*, ISSN 1350-4533, https://doi.org/10.1016/j.medengphy.2022.103825

Bharti, R., Khamparia, A., Shabaz, M., Dhiman, G., Pande, S., & Singh, P. (2021). Prediction of heart disease using a combination of machine learning and deep learning. Computational intelligence and neuroscience, 2021.

Bhat, S.S. (2022). Prevalence and Early Prediction of Diabetes Using Machine Learning in North Kashmir: A Case Study of District Bandipora. *Computational Intelligence and Neuroscience*, *2022*, ISSN 1687-5265, https://doi.org/10.1155/2022/2789760

Çetinkaya, Z., & Horasan, F. (2021). Decision trees in large data sets. International Journal of Engineering Research and Development, 13(1), 140-151.

Chandra, P., & Deekshatulu, B. L. (2012, November). Prediction of risk score for heart disease using associative classification and hybrid feature subset selection. In 2012 12th international conference on intelligent systems design and applications (ISDA) (pp. 628-634). IEEE..

Desai, F. (2022). HealthCloud: A system for monitoring health status of heart patients using machine learning and cloud computing. *Internet of Things (Netherlands)*, *17*, ISSN 2542-6605, https://doi.org/10.1016/j.iot.2021.100485

Dileep, P. (2023). An automatic heart disease prediction using cluster-based bi-directional LSTM (C-BiLSTM) algorithm. *Neural Computing and Applications*, *35*(10), 7253-7266, ISSN 0941-0643, https://doi.org/10.1007/s00521-022-07064-0

Galal, A. (2022). Applications of machine learning in metabolomics: Disease modeling and classification. *Frontiers in Genetics*, *13*, ISSN 1664-8021, https://doi.org/10.3389/fgene.2022.1017340

Guleria, P. (2022). XAI Framework for Cardiovascular Disease Prediction Using Classification Techniques. *Electronics (Switzerland)*, *11*(24), ISSN 2079-9292, https://doi.org/10.3390/electronics11244086

Kamel, H., Abdulah, D., & Al-Tuwaijari, J. M. (2019, June). Cancer classification using gaussian naive bayes algorithm. In 2019 international engineering conference (IEC) (pp. 165-170). IEEE.

Krishnamoorthi, R. (2022). A Novel Diabetes Healthcare Disease Prediction Framework Using Machine Learning Techniques. *Journal of Healthcare Engineering*, *2022*, ISSN 2040-2295, https://doi.org/10.1155/2022/1684017

Kumar, V. (2022). Addressing Binary Classification over Class Imbalanced Clinical Datasets Using Computationally Intelligent Techniques. *Healthcare (Switzerland)*, *10*(7), ISSN 2227-9032, https://doi.org/10.3390/healthcare10071293

Methods in Medicine, C. A. M. (2023). Retracted: Implementation of a Heart Disease Risk Prediction Model Using Machine Learning.

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

Mudawi, N. Al (2022). A Model for Predicting Cervical Cancer Using Machine Learning Algorithms. *Sensors*, *22*(11), ISSN 1424-8220, https://doi.org/10.3390/s22114132

Mukhyber, S. J., Abdulah, D. A., & Majeed, A. D. (2023, March). Classification of liver dataset using data mining algorithms. In AIP Conference Proceedings (Vol. 2475, No. 1). AIP Publishing. [10] Song, Y., Kong, X., Huang, S., & Zhang, C. (2021). Fast training logistic regression via adaptive sampling. Scientific Programming, 2021, 1-11.

Rastogi, R. (2023). Diabetes prediction model using data mining techniques. *Measurement: Sensors*, *25*, ISSN 2665-9174, https://doi.org/10.1016/j.measen.2022.100605

Sandhya, Y. (2020). Prediction of heart diseases using support vector machine. International Journal for Research in Applied Science & Engineering Technology (IJRASET)(ISSN: 2321-9653) Volume, 8.

Sharma, H., & Rizvi, M. A. (2017). Prediction of heart disease using machine learning algorithms: A survey. International Journal on Recent and Innovation Trends in Computing and Communication, 5(8), 99-104.

Tasin, I. (2023). Diabetes prediction using machine learning and explainable AI techniques. *Healthcare Technology Letters*, *10*(1), 1-10, ISSN 2053-3713, https://doi.org/10.1049/htl2.12039

Vardhan, M. V., Kumar, M., U., R., Vardhini, M., V., Varalakshmi, M., S., and Kumar, M., A., S.(2023). HEART DISEASE PREDICTION USING MACHINE LEARNING. Journal of Engineering Sciences. Issue 04 Vol 14.