IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 9, September 2022

DOI: 10.17148/IJARCCE.2022.11920

Cardiovascular Disease Prediction using Machine Learning Methods

Manoj M[1], Yogeshwar K[2], Mangala Madhan Kumar [3]

Department of Electrical and Electronics Engineering, BMS Institute of Technology and Management, Bengaluru, India[1]

Department of Electrical and Electronics Engineering, BMS Institute of Technology and Management, Bengaluru, India[2]

Department of Electrical and Electronics Engineering, BMS Institute of Technology and Management, Bengaluru, India[3]

Abstract: Heart is the most important organ for all living organisms. The heart related diseases caused numerous numbers of deaths worldwide from past few decades. Prediction and diagnosis of the heart diseases with very high precision and correctness is required for early diagnosis. Machine Learning helps in making predictions and decisions from large data sets of data consisting medical parameters. The paper demonstrated many Machine Learning algorithms such as Decision Tree Classifier, KNeighbors Classifier, Naive bayes, Random Forest Classifier, Grid Search CV, SVM for predicting the heart disease using Erbil heart disease dataset from kaggle having 22 different medical and non-medical attributes. The precision, accuracy, F1-score, and recall of all algorithms used for predicting the heart disease is evaluated. Decision Tree Classifier algorithm provided a good accuracy of 98% among all other algorithms.

Keywords: Machine learning, Naive bayes, KNeighbors Classifier, Random Forest Classifier, Grid Search CV, SVM, Decision Tree Classifier, F1-score, Confusion matrix.

I. Introduction

For the past few decades, heart-related diseases have been one of the major reasons for death. Blood pressure, pulse rate, and cholesterol are the main reasons for heart disease. Heart diseases are also caused by non-modifiable factors such as Age, Family history and other habits like Poor diet, smoking, and drinking. When blood vessels are overstretched it leads to a change in blood pressure. Blood pressure is normally measured in terms of 2 phases of the cardiac cycle - diastole and systole. Systolic blood pressure measures the pressure present in the arteries when the heart beats whereas diastolic blood pressure measures the pressure present in the arteries when the heart is in a resting state. The increase in fat or level of lipids in the blood may also cause heart diseases. The lipids are present in the arteries and reduce the blood flow as they narrow the arteries. Age is a non-modifiable factor, with an increase in age, there are changes in heart and blood vessels. Smoking is a primary reason of heart disease, and it causes 40% of death as it limits the oxygen level in the blood, tightens and causes damage to the blood vessels [1].

Data mining is the process of retrieving essential knowledge and information from vast volumes of data. Data mining is mostly used to uncover patterns and extract hidden information from massive amounts of data. The four primary approaches used in data mining are classification, association rules, clustering, and regression. In many fields, data mining is primarily required to quickly analyse and extract valuable information from a massive volume of data. As there is a vast amount of information in fields like medicine, education, and business, this data can be mined using certain methods to discover relevant information. Algorithms of machine learning can be used to implement all these data mining techniques [2]. One of the areas of artificial intelligence that is constantly expanding is machine learning. Machine learning algorithms can analyse data from a variety of domains and create trained models using their findings.

For the analysis of heart disease attribute variables and the prediction of heart diseases, various types of data mining approaches are available. Classification is one of the popular types of data mining techniques and it is used to classify the data into different classes or groups. Feature selection is another technique that includes the process of choosing a subset of similar features that helps in model construction. Association rule mining involves identifying associations in the huge database and their values [3]. Classification techniques like Naive Bayes, KNeighbors Classifier, Random Forest Classifier, GridSearchCV, SVM and Decision Tree Classifier are used in the classification and prediction of heart diseases. To predicting heart disease using trained data, the classification model is created using classification methods.

We used the Erbil heart disease dataset from Kaggle for this study and we compared the effectiveness of different heart disease prediction systems.

II. Related Work

Multiple studies that use data mining or machine learning techniques to predict heart diseases have been conducted. Various heart disease-related variables and a model for prediction are presented in a research paper [4] by Devansh Shah, Samir Patel, and Santosh Kumar Bharti. The model is based on supervised learning methods. It makes use of the already-existing dataset of heart disease patients from the UCI repository's Cleveland database. The 303 instances and 76 attributes in the dataset were used for analysis. Out of the 76 qualities, only 14 are tested to demonstrate how well various algorithms perform. Naïve Bayes, decision tree, K-nearest Neighbor and random forest algorithms were employed. The probability that patients may develop heart disease is the main topic of this research paper. The findings show that the K-nearest neighbour method with a value of K=7 achieves the greatest accuracy score of 90.789%.

By mining data comprising previous health records, Kumari Deepika and Dr. S. Seema provided effective mechanisms that have been utilized for chronic disease prediction in their research paper [5]. For the diagnosis of diabetes and heart disease, they employed the classifier algorithms such as Decision Tree, Support Vector Machine (SVM), Naive Bayes and Artificial Neural Networks (ANN). To diagnose diabetes and heart disease, they used two different datasets from the UCI machine learning repository. To evaluate the effectiveness of the algorithm's classification based on accuracy rate, the research provided a comparative analysis of various classifiers. For predicting diabetes, the Naive Bayes classifier had the highest accuracy (73.588%), whereas the SVM had the highest accuracy (95.556%) for predicting heart disease.

In a different research paper [6], Mr. Santhana Krishnan J and Dr. Geetha S offer a system that predicts the likelihood that a patient would have cardiac disease using a classification model to examine the data. The clinical variables in the dataset are indicative of heart conditions, including the nature of chest discomfort, blood pressure level, electrocardiographic findings, and others. The Naive Bayes Classifier Algorithm and the Decision Tree Classification Algorithm were the two primary algorithms used in this system. With an accuracy level of 91%, the Naive Bayes classifier predicted the heart disease patient, whereas the Decision tree model predicted with an accuracy level of 87%.

In order to prevent biased algorithm performance, D.P. Yadav, Prabhav Saini, and Pragya Mittal offered the model using 3-fold cross-validation in their research paper [7]. The dataset was subjected to the use of well-known machine learning techniques Random Forest, Naive Bayes, Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) for the prediction of heart disease. The Naive Bayes method had the highest average accuracy, at 87.78%. Additionally, an accuracy of 96% was attained after using a genetic algorithm to improve the features in the dataset.

The Research paper [8] by Ghadekar Premanand Pralhad, Anshul Joshi and others presented the model using the Grid Search CV algorithm for Dyslexia. The techniques used were SVM and Grid search CV with an accuracy of 97.42%. They improved the accuracy using the model for predicting dyslexia by using conventional methodologies.

An automated system for estimating the risk of heart disease was offered in another research paper [9] by Farzana Tasnim and Sultana Umme Habiba. The dataset from the UCI machine learning repository was used to assess the prediction of cardiac disease. When compared to how machine learning algorithms typically operate, the feature selection strategy improved algorithm performance. The Random Forest algorithm with Principal Component Analysis (PCA) provided the best accuracy of 92.85% for classifying heart disease among the various classification techniques.

III. About Dataset

For this study, we have made use of the Erbil heart disease dataset [11] from Kaggle. It contains records of 333 patients with 22 attributes (TABLE I). The information in this dataset was manually and directly gathered from hospital patients. All this information was gathered from Medical Help Facility, a private heart centre and hospital in Erbil, Iraq. Unlike the Cleveland database of the UCI Machine learning repository [10], this data has five categories which have been used to group the collected data. Some of these factors include the patient's demographics, medical history, physical exams and symptomatology, medical lab testing, and diagnostic characteristics.

Out of 333 records, 118 have confirmed the presence of heart disease, amongst the confirmed cases 90 patients are alive while 28 are dead, 56 patients are female and the rest 62 are male, age of patients varies from 20 years to 90 years.

The prepossessing of the data is done to remove any missing or noisy records. To have accurate results the data is cleaned for noisy records and checked for missing values and found to be none.

Feature scaling is an essential step in modelling algorithms with data sets. This allows the resulting data to contain features of varying dimensions and scales throughout. Data characteristics at different scales have a negative impact on modelling data sets. This skews the prediction results in terms of misclassification error and accuracy rate. Therefore, we need to scale the data before modelling. The 30% of the dataset was used for testing while 70% for training.

TABLE I Attributes and details of the data set.

Attributes	**Details**
Age	Age of patient in years
Sex	Gender of patient female=1, male=0
Smoking	If patient is smoking Yes=1, No=0
Years	Number of years of smoking if smoker
LDL	Low-density lipoprotein ratio of patient
CHP	Chest pain type 1=Typical angina, 2=Atypical angina,3=non-anginal pain, 4=Asymptomatic
Height	The height of the patient in cm
Weight	The weight of patients in kg
FH	Family history of heart disease Yes=1, No=0
Active	If the patient is alive or dead Yes=1, No=0
Lifestyle	The place of living 1=City, 2=Town, 3=Village
IHD	Does the patient have ischemic heart disease 0=No, 1=Yes
HR	Heart Rate ratio
DM	Does the patient have diabetes 0=No, 1=Yes
Bpsys	The ratio of Systolic Blood Pressure
Bpdias	The Diastolic ratio of Blood Pressure
HTN	Does the patient suffer from hypertension 0=No, 1=Yes
IVSD	It is a measurement that is used to determine Left Ventricular Hypertrophy (LVH)
ECGpatt	Contains four categories of an ECG test 1=ST-Elevation, 2=ST-Depression, 3=T-Inversion, 4=Normal
Qwave	The presence of the Q wave 0=No, 1=Yes
Target	If the patient suffers from heart disease or not 0=without heart disease, 1=with heart disease

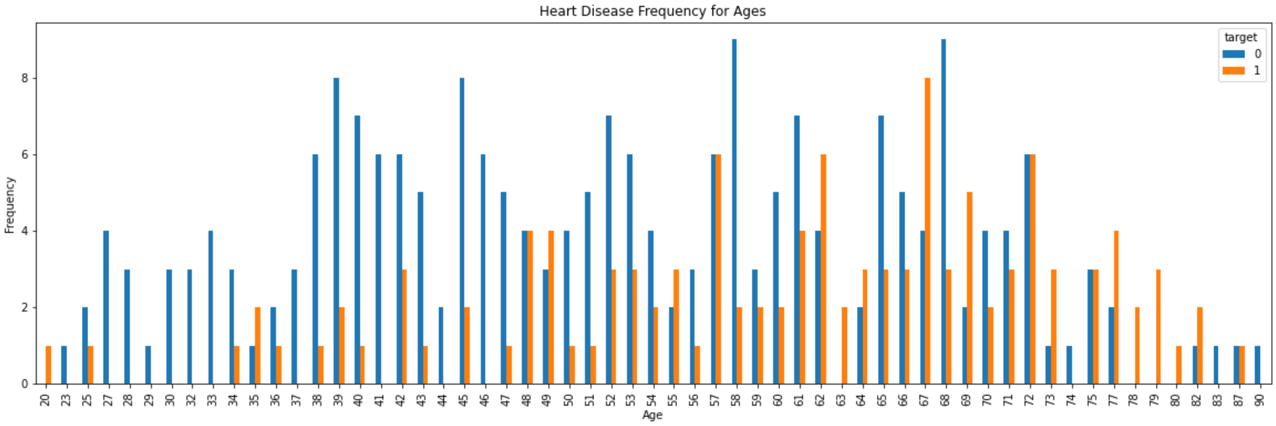

Fig. 1 Frequency of disease records over ages

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 9, September 2022

DOI: 10.17148/IJARCCE.2022.11920

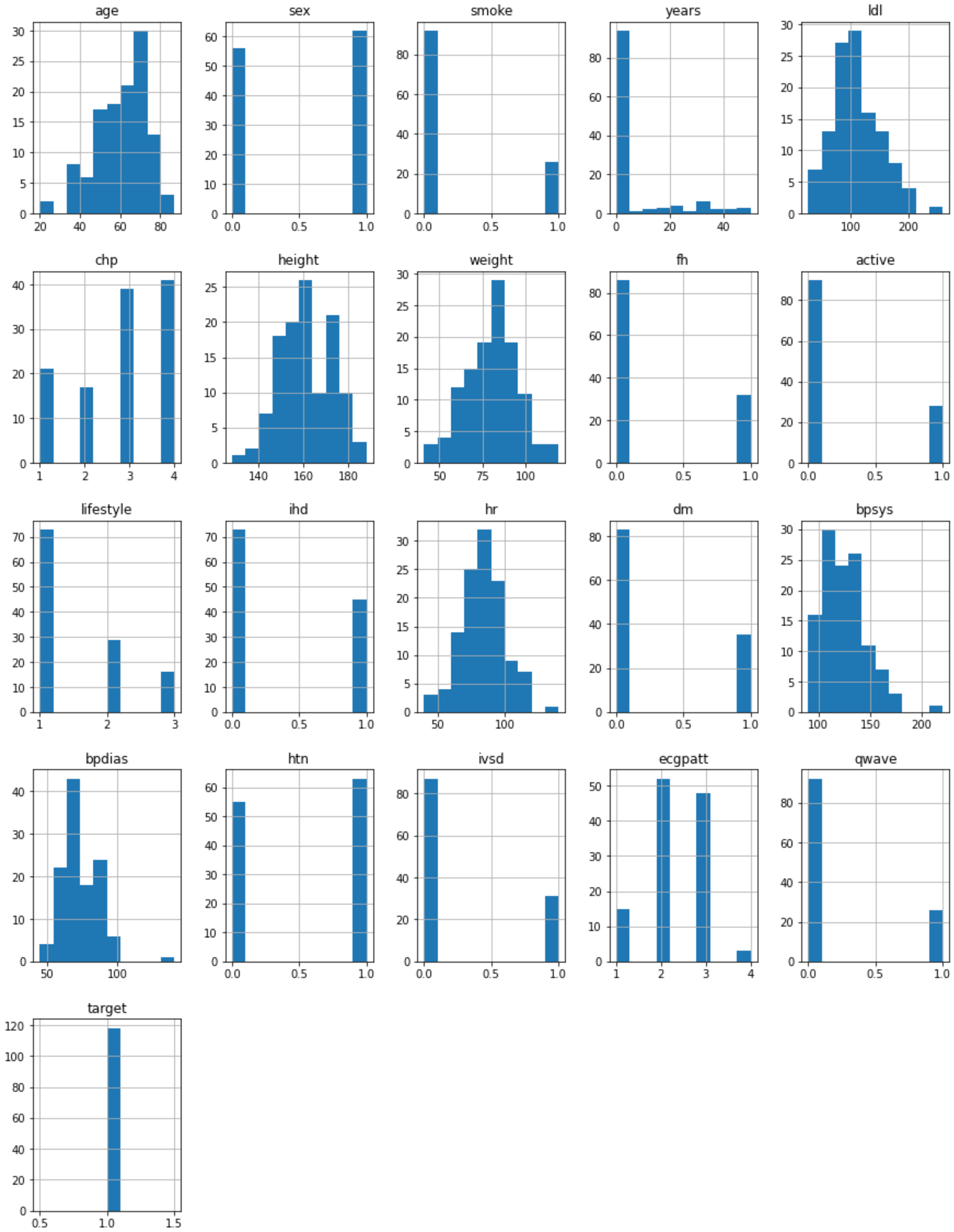

Fig. 2 Plot of different attributes versus confirmed heart disease records.

IV. ALGORITHMS USED

A. Naïve Bayes Classifier

The Bayes Theorem is the foundation of the probabilistic machine learning algorithm known as the Naive Bayes Classifier, which is employed in a wide range of classification problems. By applying Bayes' theorem, we can determine the likelihood that A will occur given that B has already occurred. A is the hypothesis in this instance, and B is the proof.

DOI: 10.17148/IJARCCE.2022.11920

The predictors and attributes are assumed to be independent in this scenario. It is when one quality is present but does not affect the other. That is why it is referred to as naive. D.P. Yadav, Prabhav Saini and Pragya Mittal [7] have achieved an accuracy of 87.78%, while we achieved a low accuracy of 65% using the Erbil heart disease dataset.

B. K-Nearest Neighbor (KNN)

The K-Nearest Neighbor algorithm classifies the records based on the nearest neighbor and uses feature similarity to forecast the values of new data points. This further indicates that the new data point will be assigned a value depending on how closely it resembles the points in the training set. Euclidean distance is used to calculate how far away an attribute is from its neighbors. Devansh Shah, Samir Patel and Santosh Kumar Bharti [4] made use of KNN and achieved 90.78% accuracy for k=7, in our study we achieved an accuracy of 76% for k=3.

C. Decision Tree Classifier

A decision tree is a tree-structured classification model in which core nodes represent the attributes of a dataset, branches represent the actions taken, and leaf nodes represent the outcomes. Decision nodes are used to make decisions and have many branches, whereas Leaf nodes are the outcomes of decisions and do not have any more branches.

- Root Node: At the root node, the decision tree is initiated. The complete dataset is represented, and it is then divided into two or more homogeneous sets.
- Leaf Node: The tree cannot be divided further after receiving a leaf node; leaf nodes are the final output nodes.

In a decision tree, the method begins at the root node of the tree to forecast the class of a given data set. Based on a comparison that follows the branch and moves to the next node, this algorithm compares the values of the original property with the record attribute (the real data set). The procedure continues after comparing the attribute value for the subsequent node with those of its child nodes once more. The procedure is carried out until the tree's leaf node is reached. Our study found the decision tree has the highest accuracy of 98%.

D. Random Forest Classifier

It is a set of multiple Decision trees, used to prevent overfitting by forming the trees on random subsets, though computation is lower than a decision tree, a Random Forest is more accurate than a decision tree classifier. The main idea behind the random forest method entails building a lot of "simple" decision trees during training and using a majority vote (mode) across them for classification. An accuracy of 93% was found in our study, while Farzana Tasnim and Sultana Umme Habiba [9] achieved 92.85% accuracy.

E. Grid Search CV

GridSearchCV is the process of tuning hyperparameters to determine the optimal value for a given model. The performance of a model depends significantly on the value of the hyperparameter, there is no way to know the best value for the hyperparameter in advance, so ideally one should try all possible values to know the optimal value, hence GridSearchCV is used to auto-tune hyperparameter. After going through the work of Ghadekar Premanand Pralhad and Anshul Joshi [8] we used Random Forest Classifier using GridSearchCV and found an accuracy of 93%.

F. Support Vector Machines

Support vector machines are supervised machine learning methods used for classification, regression, and outliers' detection. It generates the hyperplanes iteratively that separate the classes in the best way then the optimal hyperplane is selected in such a way that it segregates the dataset into several different classes to find a maximum marginal hyperplane. The SVM classifier provides good accuracy and faster prediction performance than the Naïve Bayes algorithm. SVM uses less memory comparatively because they use a subset of learning points in the decision phase. The algorithm works well with high dimensional space and with a clear separation margin.

A paper published by Kumari Deepika and Dr S. Seema [5] had an accuracy of 95.55% while we got an accuracy of 94% using SVM.

V. EVALUATION

The model's accuracy on the data set is gauged by the F1 score. It's employed to assess binary categorization schemes that label examples as "positive" or "negative." Model recall and accuracy are also factors. We used this method to assess

DOI: 10.17148/IJARCCE.2022.11920

the outcomes since it is defined as the harmonized mean of the model's accuracy and recall. A classification model's performance on a set of test data for which the true values were known is typically described using the confusion matrix.

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

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- True positives (TP): These are cases in which we correctly predicted that the patient has the disease and indeed they do.
- True negatives (TN): They don't have the disease, as we predicted.
- False positives (FP): They don't have the disease, but our prediction was that they would have it.
- False negatives (FN): They have the disease, but our prediction was that they would not have it.

VI. Results

This research aims to predict whether a patient will develop heart disease or not by using the Erbil heart disease dataset. The dataset includes patient's demographics, their medical history, physical exams and symptomatology, medical lab testing, and diagnostic characteristics which have different attributes compared to the Cleveland database of the UCI Machine learning repository. Our study was based on supervised machine learning techniques using Naive Bayes Classifier, Random Forest Classifier, Grid Search CV, Support Vector Machines and Decision Tree Classifier. The confusion matrix obtained after applying the above techniques is listed in Table III and the accuracy of different models is shown in TABLE II.

TABLE II Model and corresponding Accuracy

Model	Accuracy
Naive Bayes Classifier	65%
KNeighbors	76%
Random Forest Classifier	93%
Grid Search CV	93%
Support Vector Machines	94%
Decision Tree Classifier	98%

TABLE IIIIII Confusion values of various models

Model	True Positive	True Negative	False Negative	False Positive
Naïve Bayes Classifier	52	4	31	13
KNeighbors	51	5	19	25
Random Forest Classifier	50	6	1	43
Grid Search CV	50	6	1	43
Support Vector Machines	50	6	0	44
Decision Tree Classifier	54	2	0	44

VII. Conclusion

One of the most important and critical organs in the human body is the heart, and heart disease prevention is a major concern for people. One of the crucial factors for evaluating an algorithm's performance throughout a prediction process is its accuracy. A suitable machine learning model can not only predict a disease rapidly, but also predict it accurately enough to improve treatment, decrease the need for human work, and eliminate medical lab tests. This study aims to forecast whether a patient will experience heart disease or not. The Erbil heart disease dataset was used in this study's supervised machine-learning classification work. The Decision Tree Classifier used in this study achieved the maximum classification accuracy, which was 98%.

DOI: 10.17148/IJARCCE.2022.11920

In the future, this research can be carried out using a variety of machine-learning technique combinations to predict heart problems with higher accuracy. Additionally, we may create novel feature selection techniques to get a more comprehensive understanding of the important features in the dataset and improve the accuracy of heart disease prediction.

REFERENCES

[1]. J. Thomas and R. T. Princy, "Human heart disease prediction system using data mining techniques," 2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT), 2016, pp. 1-5, doi: 10.1109/ICCPCT.2016.7530265.

[2]. S. K. J. and G. S., "Prediction of Heart Disease Using Machine Learning Algorithms," 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT), 2019, pp. 1-5, doi: 10.1109/ICIICT1.2019.8741465.

[3]. C. Sowmiya and P. Sumitra, "Analytical study of heart disease diagnosis using classification techniques," 2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), 2017, pp. 1-5, doi: 10.1109/ITCOSP.2017.8303115.

[4]. Shah, D., Patel, S. and Bharti, S.K., 2020. Heart disease prediction using machine learning techniques. SN Computer Science, 1(6), pp.1-6. https://doi.org/10.1007/s42979-020-00365-y.

[5]. K. Deepika and S. Seema, "Predictive analytics to prevent and control chronic diseases," 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), 2016, pp. 381-386, doi: 10.1109/ICATCCT.2016.7912028.

[6]. S. K. J. and G. S., "Prediction of Heart Disease Using Machine Learning Algorithms," 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT), 2019, pp. 1-5, doi: 10.1109/ICIICT1.2019.8741465.

[7]. D. P. Yadav, P. Saini and P. Mittal, "Feature Optimization Based Heart Disease Prediction using Machine Learning," 2021 5th International Conference on Information Systems and Computer Networks (ISCON), 2021, pp. 1-5, doi: 10.1109/ISCON52037.2021.9702410.

[8]. G. P. Pralhad, A. Joshi, M. Chhipa, S. Kumar, G. Mishra and M. Vishwakarma, "Dyslexia Prediction Using Machine Learning," 2021 International Conference on Artificial Intelligence and Machine Vision (AIMV), 2021, pp. 1-6, doi: 10.1109/AIMV53313.2021.9671004.

[9]. F. Tasnim and S. U. Habiba, "A Comparative Study on Heart Disease Prediction Using Data Mining Techniques and Feature Selection," 2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), 2021, pp. 338-341, doi: 10.1109/ICREST51555.2021.9331158.

[10]. Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[11]. Hangaw Qadir Ahmed, Shwan Othman Amen, Banan Qasim Rassol, & Ibrahim Ismael Hamad. (2022). Erbil Heart Disease Dataset [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/3989065.