HealthPulse Prognosis: Predicting Heart Disease through Intelligent Data Mining Strategies

D. PRABHU

Assistant professor

Department of computer science and Engineering, school of computing,

Vel Tech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology, (Deemed to be university) Avadi, Chennai, India.

[prabhu.klnce@gmail.com](mailto:prabhu.klnce@gmail.com)

Dr.D. Poornima

Assistant Professor,

Department of Computer Science and Engineering,

Sathyabama Institute of Science and Technology, Chennai - 600119, Tamil Nadu, India

[poornima.d.cse@sathyabama.ac.in](mailto:poornima.d.cse@sathyabama.ac.in)

T. ANAND

Assistant professor

Department of computer science and Engineering, school of computing,

Vel Tech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology, (Deemed to be university)Avadi, Chennai, India.

[t.anandme@gmail.com](mailto:t.anandme@gmail.com)

Ms. M.P.BOBBY,

Assistant Professor,

Department of Computer Science and Engineering,

Sathyabama Institute of Science and Technology, Chennai - 600119, Tamil Nadu, India

[bobby.m.p.cse@sathyabama.ac.in](mailto:bobby.m.p.cse@sathyabama.ac.in)

ABSTRACT

In the modern day, fatalities that are caused by heart disease have become a significant problem. It is estimated that nearly one person passes away due to heart disease every minute. In the majority of situations, the diagnosis of cardiac disease is dependent on a complicated mix of clinical and pathological data. This complexity results in exorbitant medical expenses, which in turn affects the quality of medical care that is provided. The objective of the research is to evaluate medical information in order to make a prediction regarding heart disease. In order to discover information contained within a dataset pertaining to heart illness, which was retrieved from the UCI Machine Learning repository, the suggested technique makes use of data mining classification algorithms such as KNN, SVM, and Decision Tree. In all, there are 282 observations and 75 characteristics that make up the input dataset. The feature selection of the input dataset is accomplished through the utilisation auto-correlation in order to locate the most optimised collection of features and enhance the overall performance of the classifiers. When it comes to accuracy, the results are achieved through the use of classification techniques, both with and without the utilization of feature selection. The final step is to make a prediction about the class labels by employing the Majority Voting method.

Keywords—heart disease, auto correlation, data mining, disease diagonsis, feature selection, classification and prediction;

# INTRODUCTION

Our works mainly focuses on various data mining techniques that are valuable in heart disease forecast. If the heart doesn’t function properly, this will distress the other parts of the human body such as brain, kidney etc. Heart disease is a kind of disease which effects the functioning of the heart. In today’s era heart disease is the primary reason for deaths. WHO (World Health Organization) has anticipated that 12 million people die every year because of heart diseases.

Different person body can show different symptoms of heart disease which may vary accordingly. Though, they frequently include back pain, jaw pain, neck pain, stomach disorders and tininess of breath, chest pain, arms and shoulders pains. There are a variety of different heart diseases which includes heart failure and stroke and coronary artery disease. A healthy way of life (main prevention) and timely analysis (inferior prevention) are the two major origins of heart disease director. Conducting steady check-ups (inferior prevention) shows outstanding role in the judgment and early prevention of heart disease difficulties. Several tests comprising of angiography, chest Xrays, echocardiography and exercise tolerance test support to this significant issue. Nevertheless, these tests are expensive and involve availability of accurate medical equipment. Heart expert’s create a good and huge record of patient’s database and store them. It also delivers a great prospect for mining a valued knowledge from such sort of datasets. The data used in dataset should be complete and consistent. So, ensure that data set is void of noisy, incomplete and inconsistent data. Researchers make use of several data mining techniques that are accessible to help the specialists or physicians identify the heart disease. Commonly used procedures are decision tree, k-nearest neighbor (KNN) and SVM (Support Vector Machine). Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. The data mining techniques have been developed in recent years and include generalization, characterization, classification, clustering, association mining, pattern matching, data visualization and meta-rule guided mining. Data mining can identify patterns, relationships and models, which support predictive and decision-making processes for diagnosis and treatment planning. The discovered models can be called predictive models, which can be integrated in hospitals’ information systems as models that are able to decrease the time of decision-making. In our proposed method we have used three classifiers namely SVM, Decision Tree and KNN. Among the three SVM gives the highest accuracy. This study proposes a method for predicting the heart disease. Different classification methods are employed to predict the heart disease based on the medical records. The remainder of this project is organized as follows: Chapter 2 provides a review on relevant studies. Chapter 3 explains the proposed method. The implementation and work done is described in Chapter 4.Conclusion is narrated in Chapter 5.

# LITERARTURE SURVEY

Over the course of the past several years, numerous hospitals have conducted a multitude of research on the anticipation of a variety of diseases. A variety of data mining strategies were utilised for the purpose of diagnosis, with variable degrees of success. The section that follows contains a concise literature overview of the various methods of data mining and assessment that are utilised in the healthcare industry. In order to accomplish the goal of this project, which is to create an automated medical heart disease prediction system, the obtained database will be carefully analysed. A medical examination of cardiac disorders can be carried out in two stages, according to the patient's condition. There are a number of different algorithms that have been developed for the pre-processing stage, which is the first one. The pre-processing step's purpose is to choose just the necessary features from the entire collection of database. Both the filter approach and the wrapper method are the two feature selection strategies that are utilised the most often. The methods for constructing the prediction model by using classifier are included in the second phase of the analysis section of the process. In general, the classifier is divided into two categories: those that are based on machine learning and those that are based on fuzzy systems according to the technology that is utilised.

A number of different machine learning strategies for the prediction of cardiac illnesses have been presented, all of which are based on a single classifier. For the purpose of prediction, single classifiers such as Support Vector Machine (SVM), Decision Tree, Neural Network, Naives Bayes, and Regression are utilised. For the purpose of illness prediction, a machine learning model that is based on supervised learning techniques [7] has been presented. A number of different algorithms, including Naïve Bayes, decision tree, K-nearest neighbour, and random forest algorithms, have been utilised in the development of the prediction model. The Cleveland database, which is a repository of heart disease patients at the University of California, Irvine, is used to evaluate the performance of each algorithm. Through the process of evaluating the model against 303 different instances of records and 76 different qualities, they arrived at the conclusion that the k-neatest neighbour model generates a high accuracy score data in comparison to other methods. Despite the fact that it was developed using a single classifier, the prediction model does not have sufficient accuracy when it comes to managing large amounts of data in a business context. It has been suggested that hybrid techniques be utilised in order to deal with the challenge.

For the purpose of making more accurate predictions regarding illnesses, the ensemble classifier is the one that will be employed. This classifier will consist of more than one classifier. For the purpose of identifying earlier stages of cardiac disease, the Ensemble learning approaches using Particle Swarm Optimisation (PSO) [1] have been proposed. In this article, a strategy for selecting feature subsets that is based on both filter and wrapper approaches is described and then implemented using the PSO technique. Principal component analysis (PSO) is mostly utilised for the purpose of selecting a subset of features from entire amounts of data by removing noisy and irrelevant data from a large volume of data. Using PSO, not only does it enhance the efficiency of learning by picking only the subset that is necessary for prediction, but it also reduces the complexity of the learning results [2], which allows for an improvement in the accuracy of the forecast.

In order to make accurate predictions regarding cardiac disorders, ensemble approaches like as Bagging, Random Forest, and AdaBoost are utilised. When it comes to fitting a model to bins, the signal-to-noise ratio is an extremely important factor. It is possible for even a slight variation in the data of the model to lessen the impact of fitting the model. In order to increase the signal-to-noise ratio and also to categorise the subset characteristics that are required for accurate prediction from a large quantity of data, a data reduction process that is based on discretization [3] has been developed. Once this is done, the ensemble classifier, which consists of Bagging, Random Forest, and Voting ensemble, is utilised in order to construct the most accurate prediction model possible. Eighty percent of the data that was gathered is used as a training dataset, and twenty percent of the data is regarded to be a test dataset for the purpose of evaluating the prediction model.

It is recommended that an ensemble classifier be developed using a machine learning technique that is based on the Adaptive Boosting method [4]. The method is applied to four distinct datasets, and the results demonstrate that the ensemble classifier that is based on adaptive boosting algorithms has reached higher levels of dependability and delivers levels of accuracy that are superior to those produced by earlier studies. One of the models that is utilised in clinical settings all around the world by professionals in the field of cardiac illnesses is the prediction model that is utilised in this technique [5]. The ensemble classifier that is based on adaptive boosting provides for a great deal of flexibility in terms of weight modification, which ultimately results in a robust and single composite ensemble classification and prediction model. The bagging ensemble learning approach with decision tree [6, 7] is provided here with the purpose of enhancing the accuracy of the prediction of heart disease.

Improvements in the performance of the prediction mechanism may be achieved by the extraction of features through the use of Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA) techniques [8]. In the early days of algorithm development, the ensemble classifier algorithms that were utilised the most frequently were bootstrapping aggregation [9], boosting [10], and random decision forests [11]. When these ensemble classifiers are utilised, there has been an improvement in the performance of the classification efficiency. It has been suggested that in order to accomplish ensemble learning, a system that integrates a large number of classifiers or partitions based on voting techniques should be utilized [12].

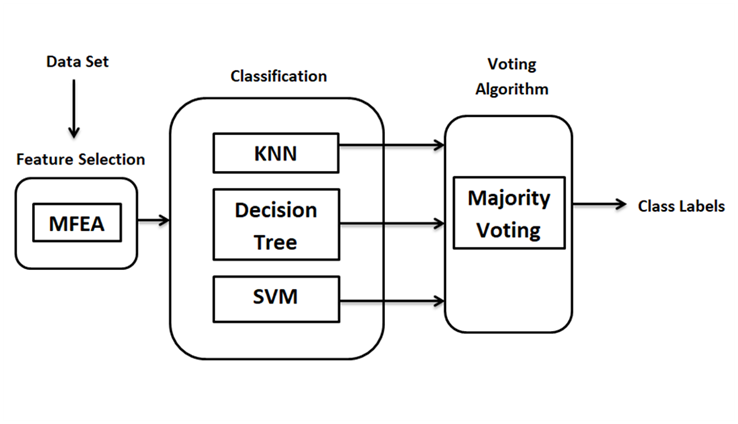
Voting is a term that refers to the process of picking the best answer from among the available options or developing views that result in a consensus among the participants. Through the process of voting, every voter has the opportunity to express their preference for one or more candidates [13]. The candidate who is found to be the most desirable is chosen to be the best classifier. Various methodologies, including parametric, nonparametric, heuristics, logical, and probabilistic approaches, can be utilised in the construction of voting systems. Voting systems may be used for a variety of purposes, including selecting the most effective classifier [14]. One innovative method involves combining a voting system with a number of different machine learning algorithms and then choosing the most effective one. A number of different classifiers, including decision trees, k-nearest neighbours, and multilayer perception, are selected by the machine learning algorithms, and they are all trained using the same data set [15]. In a different manner, the voting system may be utilised with a single classifier, but it can be trained using a separate data set.

It is decided to conduct a survey in order to investigate the impact that voting methods [16] have on categorization tasks. To investigate the influence of voting approaches on the performance of the k-Nearest Neighbour and Naïve Bayes algorithms, a variety of datasets are utilised for the purpose of conducting statistical analysis. It is decided to do a comparative analysis of the ensemble decision tree approaches [17]. In order to carry out the studies, ensemble decision tree methods, such as bootstrap aggregating and server randomization, were utilised. After careful analysis, it was shown that random forests, boosting, and randomised tree performed better than bagging. There are a variety of basic classifiers that may be utilised with the ensemble learning approach. Ensembles may be built by first splitting the dataset into subgroups and then mixing those subsets together [18, 19].

The formation of the additional datasets was accomplished by the use of random permutation [20] and dividing of the first dataset. The resultant ensemble achieved a high level of generalisation as a consequence of the overall compatibility of the various models being greatly improved. [21, 22]. The use of ensemble learning has been utilised in the prediction of a number of medicinal treatments. An ensemble model for the prediction of cardiac illness is suggested in this study. The model makes use of machine learning classifier methods, namely support vector machines (SVM), decision trees, and random forests [23]. The research efforts are being done with the intention of selecting the relevant characteristics in order to improve the accuracy of the prediction while simultaneously reducing the amount of computer resources required. This is accomplished by the use of feature selection methods, auto correlation, which are mostly employed for the purpose of feature reduction. The output class is predicted by the voting classifier based on the one that received the biggest majority of votes. The predictions of each base classifier are then sent into the voting classifier. Creating a single model by training several base classifiers and predicting the output based on the cumulative majority of votes for each output class is the primary objective of this. This involves training multiple base classifiers.

# PROPOSED METHODOLOGIES

The following section outlines the development of the Heart Disease Prediction System (HDPS), which utilises auto correlation feature selection algorithms in combination with voting ensemble classifiers. The Cleveland dataset comprises 82 observations and 75 variables sourced from the UCI Machine Learning Repository, namely the heart disease dataset. The initial dataset contains 75 characteristics, and a auto correlation feature selection technique is used to identify and choose 10 features. The altered dataset is subsequently inputted into classification techniques such as Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbours (KNN). The output obtained is analysed utilising voting ensemble classification algorithms. Figure 1 illustrates the architecture of the heart disease prediction system.



Auto correlation

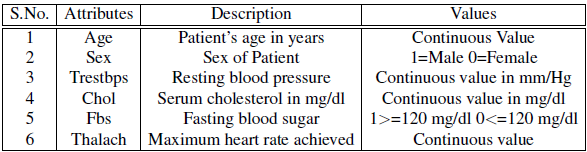
**Figure 1: Architecture of Heart Disease Prediction Systems**

The Heart disease system prediction system consists of flowing modules:

* Dataset Collection
* Feature Selection (Auto Correlation)
* Classification (SVM, Decision Tree and kNN)
* Voting Classification (Improve the Classification Accuracy)
* Final Prediction

## **Dataset Collection:** The heart disease dataset that is provided in the UCI Machine Learning repository is the source of the 282 observations and 75 characteristics that make up the input dataset.The dataset that was collected includes information such as the patient's gender, age, kind of chest discomfort, resting blood pressure, resting electrocardiogram findings, and the names of the variables that are associated with each individual patient.The dataset that was gathered is now accessible in a processed form that contains data that is both comprehensive and consistent. This sample dataset's contents are presented in Table 1.

**Table 1: Dataset attribute information**

****

## **Feature Selection:** The process of picking the subset of the features that make up the set of features that are the most relevant is referred to as feature selection. In this work, we have used auto correation as feature selection. Autocorrelation refers to the property of data where the correlation between values of the same variables is determined by their relationship to each other. This contradicts the notion of instance independence, which is a fundamental principle in most traditional models. It typically occurs in datasets where the data is sourced from the same origin, rather than being randomly picked.

## The original dataset is subjected to the Correlation based feature selection (CFS) process to identify class characteristics that are relevant.Collinear characteristics are characterised by a strong correlation between them. These factors in machine learning result in a reduction in the overall performance of the model on the test set due to increased variability and lower ability to comprehend the model. First-order serial correlation, which can be either positive or negative, is the most prevalent kind of autocorrelation. Positive serial correlation refers to a situation in which a positive deviation in one period persists as a positive deviation in the subsequent period.Negative serial correlation refers to the phenomenon where a negative deviation in one time period persists and leads to a negative deviation in the subsequent time period.

## **SVM Classifer:** The Support Vector Machine (SVM) is a highly popular method in Supervised Learning. It is commonly employed for both Classification and Regression tasks. However, its primary use is in solving classification problems within the field of Machine Learning. The objective of the Support Vector Machine (SVM) technique is to construct an optimal line or decision boundary that effectively separates an n-dimensional space into distinct classes. This enables us to accurately classify future data points into their appropriate categories. The decision boundary that optimally separates the data points is referred to as a hyperplane. SVM selects the most significant points/vectors that contribute to the formation of the hyperplane. The phrase "Support Vector Machine" refers to a method that involves extreme situations known as support vectors. Figure 2 illustrates the presence of two distinct categories that are differentiated by a decision boundary or ideal hyperplane.



**Figure 2: SVM and Hyperplane**

## **kNN Classifier:** The K nearest neighbour algorithm is a method that saves all existing instances and categorises new cases by using a similarity metric, such as distance functions.The K Nearest Neighbour (KNN) algorithm is a commonly employed classification method.There is either no clear training period or it is quite limited. Additionally, this implies that the training period is rather rapid. KNN's lack of generalisation is due to its retention of all the training data. Precisely, the testing phase requires the use of all (or the most of) the training data. The KNN algorithm relies on the similarity of features. The degree of similarity between out-of-sample attributes and our training set impacts the classification of a particular data point. Figure 3 illustrates the process of categorising a new data point using the KNN algorithm.



**Figure 3: kNN classifier separation**

## **Decision Tree Classifier:** Decision trees are a type of decision support tool that employ a tree-like graph or model of options and the potential outcomes of those decisions, which may include the outcomes of random events, the costs of resources, and the utility of those resources. This is one method for displaying an algorithm that only includes statements that control the conditions. In a structure that is similar to a flowchart, a decision tree is a structure in which each internal node represents a "test" on an attribute, each branch indicates the conclusion of the test, and each leaf node represents a distinct class label. Classes are represented by the pathways that lead from the root to the leaf. The construction of the Decision Tree is depicted and shown in Figure 4.



**Figure 4: Decision Tree classifier separation**

**Voting Classifier:** The techniques of voting are based on a weighted procedure that combines the predictions that are offered by the classification models that have been separately calibrated on a number of different analytical sources that are available. The class that has the highest number of predictions across all of the classifiers is the one that is selected to be the ensemble's prediction through the procedure known as majority voting. The method of voting from the majority is depicted in Figure 5.



**Figure 5: Ensemble voting classifier**

# EXPERIMENTAL RESULTS AND DISCUSSIONS

An autocorrelation analysis yielded a total of nine characteristics as a result. Cholesterol, the number of years spent smoking (years), resting electrocardiogram (restecg), the duration of the exercise test in minutes (thaldur), the maximum heart rate achieved (thalach), the number of major vessels coloured by flouroscopy (Ca), rest wall motion abnormality (restwm), along with thal, thalpul, and laddist are the ten features that were extracted. A representation of the code for auto correlation may be found in Figure 6.



**Figure 6: Feature Selection using Auto Correlations**

The KNN classifier searches for the classes of the K closest neighbours of a particular data point, and then assigns a class to the data point based on the class that is the majority of the neighbours. As shown in Figure 7, an accuracy of 83.55% was achieved by the utilisation of KNN data mining classification using 10 features.



**Figure 7: Heart Disease Prediction using k-NN classifiers**

The Decision Tree classifier is responsible for the creation of a decision tree, which serves as the basis for the assignment of class values to each different data point. As shown in Figure 8, with 10 characteristics the accuracy that can be achieved via the utilisation of decision tree data mining categorization accounts for 78.66%.



**Figure 8: Heart Disease Prediction using Decision Tree classifiers**

Through the process of adjusting the distance between the data points and the hyperplane, the support vector machine (SVM) classifier strives to create a hyperplane that is capable of separating the classes to the greatest extent feasible. The hyperplane is determined by a number of different kernels, of which there are numerous. For the purpose of determining the hyperplane, this project makes use of the linear kernel. As shown in Figure 9, 85.33% is the accuracy that can be achieved with the utilisation of SVM data mining classification using 10 features.

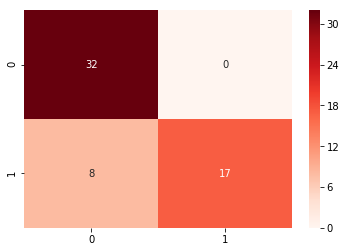
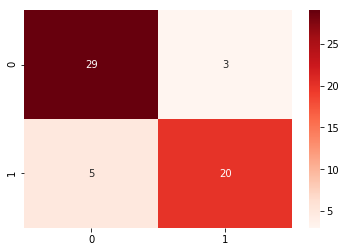


**Figure 9: Heart Disease Prediction using SVM classifiers**

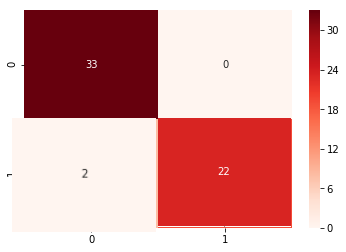
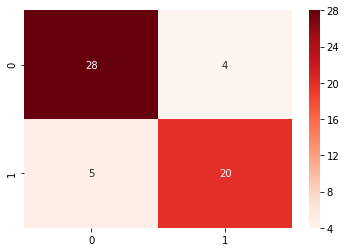


**Figure 10: Heart Disease Prediction using Ensemble Voting Classifiers**

In order to the Heart disease predctions, we have combined the three classifiers namely support vector machne, decision tree and k-Nearest Neighbors and calculate the voting ensemble stategies. As Shown in Figure 10, the voing classifiers have superier result rather than the other three classifiers. The consuion matrix of three traditional classifier as well as ensmble voring classifier are shown in Figure 11.

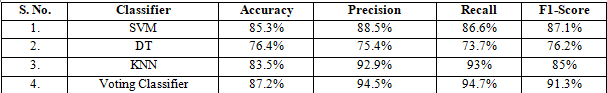


|  |  |
| --- | --- |
|  |  |



**Figure 11: Confuson matrix of all the classifiers**

**Table 2: Comparative Analysis of Traditional and Ensemble Classifier**



# CONCLUSIONS

The accuracy of heart disease prediction using an ensemble of classifiers is examined in this study. For training and testing, we used the Cleveland heart dataset that is part of the UCI machine learning library. In order to classify databases pertaining to cardiac diseases, we have investigated many classification techniques. Additionally, we have examined several categorization methods and the results they provide. Depending on the execution tools, this analysis reveals distinct technologies utilised in different publications with varied counts of characteristics and varying degrees of accuracy. Developing new permutations of data mining techniques can further improve the structure's accuracy. An accuracy of 83.5% was achieved when KNN was employed. A 78.6% success rate was achieved when a decision tree was employed.

With SVM, the accuracy reached 85.3%. When looking at the results side by side, it was clear that majority voting significantly improves accuracy. Using feature selection techniques substantially improved the performance. The accuracy of the ensemble algorithms was enhanced by the feature selection strategies. We discovered and put into action the Heart Disease Prediction System that makes use of a machine learning algorithm. The proposed system makes use of a multiple feature assessment technique and majority voting procedure for class label prediction, which are both efficient and accurate, due to the current technological developments that have led to a great deal of evolution in machine learning algorithms. Based on user input, the system also provides nearby trustworthy results. A decrease in the mortality rate from cardiovascular illnesses is possible if more people use the system and are therefore aware of their present heart health.

##### REFERENCES

1. Y. Indu, D. Sunanda and M. Jabbar, Prediction of heart disease using ensemble learning and Particle Swarm Optimization, p. 691-698, (2017).
2. L. Ladha and T. Deepa, Feature Selection Methods and Algorithms, International Journal of CSE, p.1787-1797, (2011).
3. J. Emakhu, S. Shrestha and S. Arslanturk, Prediction System for Heart Disease Based on Ensemble Classifiers, Proceedings of the 5thInternational Conference on Industrial Engineering and Operations Management Detroit, Michigan, USA, August p. 10 - 14, (2020).
4. H. Kathleen, H. Julia, and J. George, Diagnosing Coronary Heart Disease Using Ensemble Machine Learning Diagnosing Coronary Heart Disease Using Ensemble Machine Learning, International Journal of Advanced Computer Science and Applications, p. 1-12, (2016).
5. K. Mohamad Almustafa, Prediction of heart disease and classifiers sensitivity analysis, BMC Bioinformatics, p. 1-5, (2020).
6. X. Gao, A. Amin Ali, H. Shaban Hassan and E.M. Anwar, Improving the Accuracy for Analyzing Heart Diseases Prediction Based on the Ensemble Method, p. 1-12, (2021).
7. D. Shah, S. Patel, and S.K. Bharti, Heart Disease Prediction using Machine Learning Techniques, Soft Neuro Computing, p. 15-21, (2020).
8. I.D. Mienye, Y. Sun, Z. Wang, An improved ensemble learning approach for the prediction of heart disease risk, Informatics in Medicine Unlocked, p. 312-322, (2020).
9. T.K. Ho, The random subspace method for constructing decision forests, IEEE Transaction on Pattern Analysis Mach Intelligence, p. 832-844, (1998).
10. L. Breiman, Bagging predictors, Mach Learning, p. 123-140, (1996).
11. R.E. Schapire and Y. Singer, Improved boosting algorithms using confidence-rated predictions, Mach Learn, p. 297-336, (2011).
12. F. Leon, S.A. Floria, C. Badica, Evaluating the effect of voting methods on ensemble-based classification, IEEE International Conference on Innovations in intelligent Systems and Applications (INISTA), p. 1-6, (2017).
13. R.E. Banfield, L.O. Hall, K.W. Bowyer and W.P. Kegelmeyer, A comparison of decision tree ensemble creation techniques, IEEE Trans Pattern Anal Mach Intell, p. 173-180, (2017).
14. D. Ruta, B. Gabrys, C. Lemke, A generic multilevel architecture for time series prediction, IEEE Transaction on Knowledge Data Engineering, p. 350-359, (2011).
15. B. Zhang, Ensemble learners of multiple deep CNNs for pulmonary nodules classification using CT images, IEEE Access, p. 110358-110371, (2019).
16. L. Han, S. Luo, J. Yu, L. Pan, S. Chen, Rule extraction from support vector machines using ensemble learning approach: an application for diagnosis of diabetes, IEEE Journal of Biomedical and Health Informatics, p. 728-734, (2015).
17. J. Emakhu, S. Shrestha and S. Arslanturk, Prediction System for Heart Disease Based on Ensemble Classifiers, Proceedings of the 5thInternational Conference on Industrial Engineering and Operations Management Detroit, Michigan, USA, August p. 10 - 14, (2020).
18. H. Kathleen, H. Julia, and J. George, Diagnosing Coronary Heart Disease Using Ensemble Machine Learning Diagnosing Coronary Heart Disease Using Ensemble Machine Learning, International Journal of Advanced Computer Science and Applications, p. 1-12, (2016).
19. K. Mohamad Almustafa, Prediction of heart disease and classifiers sensitivity analysis, BMC Bioinformatics, p. 1-5, (2020).
20. X. Gao, A. Amin Ali, H. Shaban Hassan and E.M. Anwar, Improving the Accuracy for Analyzing Heart Diseases Prediction Based on the Ensemble Method, p. 1-12, (2021).
21. Pati, A., Parhi, M., & Pattanayak, B. K. (2022). IHDPM: An integrated heart disease prediction model for heart disease prediction. International Journal of Medical Engineering and Informatics, 14(6), PP. 1-12.
22. Shuvo, S. (2023). An early detection of heart disease using machine Learning (recurrent neural network). <https://doi.org/10.14293/s2199-1006.1.sor-.ppgv4id.v1>
23. VARKALA, K. (2022). Heart disease prediction system using Convolutional neural networks. <https://doi.org/10.21203/rs.3.rs-2009078/v2>