The Expert System for Heart Disease Prediction using ADA Boosting and Gradient Boosting Algorithm

Pinjala Supriya, Preethi M.A.B , Sivaneshwaran R.N
Computer Science and Engineering Department , Velalar College of Engineering and Technology(Affliated by Anna University),
Erode, Tamilnadu,India.

Tamilselvi. V, M.E
Computer Science and Engineering Department , Velalar College of Engineering and Technology(Affliated by Anna University),
Erode, Tamilnadu,India.

Abstract- Healthcare domain has huge amount of data and to process this data we need to use advanced technologies which will help in delivering effective results and in making efficient decisions on data and getting appropriate results. Heart disease is the biggest problem and one of the main causes of poor health dead in the world. A machine learning algorithm using ADA boosting and gradient boosting algorithms implements an efficient heart disease prediction framework. In the first step, we upload the dataset file and run the algorithm to run on the selected dataset. The ECG signals are then extracted from the data set. Then plot the precision and generate a modality for the one with the highest frequency by training the corresponding dataset. In the next step, an input is given for each heart parameter and the risk stage of the heart is predicted based on the generated modality. Then we take preventive measures according to the patient's condition.Our strategy is effective in delivering the victim's heart disease. The heart disease prediction framework rendered by this aspect is one of the unique methods available in the heart disease category.

Keywords—Machine Learning, Data Mining, ADA Boosting algorithm, Gradient Boosting Algorithm, notch filter.

I. INTRODUCTION

Among several deadly diseases in the world today, heart disease is one of the most complex and deadly diseases affecting middle-aged and elderly people in the world. It's a tall order that could provide a computerized prediction of a patient's heart condition in hopes of justifying further treatment. This heart disease is usually based on its indications, presentation, and physical assessment of the patient. Factors such as cholesterol level in the body, high blood pressure, smoking habits, obesity, family history of coronary heart disease and lack of physical activity can increase the risk of contracting the disease. Because of these limitations, researchers have turned to current methods such as data mining and machine learning to predict disease.It is estimated that approximately 45,000,000 people are affected by heart disease. A growing number of young Indians are affected by heart diseases. The offspring looked unattractive. The prevalence of hypertension seems to have increased steadily over the last 5 years, more in urban than in pastoral areas. 25% to 30% in urban areas and 10% to 50% in rural areas.An inactive lifestyle is one of the leading causes of death, illness and disability which increases the risk of heart disease. There are a number of sophisticated investigative techniques to predict heart disease resulting in a multi-faceted nature which is an important cause affecting our daily life. Therefore, the treatment of heart disease is too complex, especially in developing countries, due to the abnormal accessibility of mechanical components. Considering the impact of cardiovascular diseases on the global population, machine learning models for early detection become very useful.

Efforts are underway to solve this huge and growing problem with the help of various technological advancements. According to (WHO), 12 million people worldwide die of heart disease every year. There are many causes of death from heart disease, including coronary heart disease (CAD), cardiomyopathy, and cardiovascular disease (CVD), which depend on blood flow throughout the body. Cardiovascular disease is a cause of serious illness, disability and death. Data mining is steadily expanding into a wide range of health science applications.

Great achievements in accuracy and low-cost social service management can be achieved using data mining classification methods and heart disease prediction frameworks. The large amount of information provided by health service companies with some hidden data helps to make wise choices to provide accurate results to make wise decisions on this information. These data mining methods are used to enhance the experience and conclusions given. Data mining applications are developed to aid in the estimation of effective medical care. Data mining can

provide potentially effective actions by comparing causes, symptoms and treatments.Data mining applications used in real life are fascinating because data miners face a different set of problems. One such practical problem concerns databases of cardiac patients.

In this study, recent studies on heart disease have been greatly reduced. We offer an ADA Boosting algorithm and a gradient boosting algorithm for prediction with better accuracy, reliability and scalability.

OBJECTIVE

1. Demonstrate ADA Boosting Algorithm, Gradient Boosting Algorithm and other algorithms.
2. Demonstrate the predictive improvement of each machine learning technique.
3. Provides the best prediction rate and prediction time to predict heart disease risk.

II. LITERATURE SURVEY

Kantar et al. (2014) introduced a decision support algorithm to automatically detect normal sinus rhythm or other conditions. The improved algorithm will support educational apps for doctors. In the main code, thirteen functions are used to mechanically diagnose eight multiple pathologies ECG. Of the techniques established for current research purposes, the greatest success in predicting future power is expected.

Miao etc. (2014) aimed at developing a heart failure risk prediction system with medical risk questions and testbed variables. Discrimination performance in several models was internally validated with bootstrap procedures, and to understand risk issues, additional experimental guidelines included possible disorders including blood urea nitrogen (BUN) and partial thromboplastin time activated (APTT) as HF morbidity. Alqaraawi et al. (2016) proposed a time-adaptive concept for estimating heart rate variability (HRV) from photoplethysmographic (PPG) signals recorded by wearable devices using linear predictive coding (LPC) and transform into methods. Wavelets. The projection algorithm performs well on both related algorithms, especially for low PPG signal-to-noise ratios. Evaluating this projection algorithm as a ground certainty calculated simultaneously from the ECG, the average temporal resolution was 8.7 ms, the sensitivity was 82.9% and the positive predictive value was 82.7%.

The heart disease prediction framework is implemented using computational techniques with machine learning models developed by Megha Shahi and Kaur.R. Sharmila and S. Chellammal presented a technical model to predict heart disease using data and engineering. Here are some reference articles published by various researchers, students and others, each with unique methods and procedures for predicting heart disease and giving advice on what is needed.

III. DATASET DESCRIPTION

In this project, the parameters we are taking are age, gender, chol, obesity, cp, trestbps, FBS, restecg, thalach, exang, old peak, slope, ca, thal. We have considered the data from the standard dataset that contains 304 records from UCI to predict heart illness.Heart disease risk predictor will utilize the mining information to give a user-arranged way to deal with new and hidden designs in the information. The information which is executed can be utilized by healthcare specialists to show signs of better improvement of service and to lessen the degree of medical impact.

The algorithms we have considered are gradient boosting and ADA boosting algorithms. Select the check box of algorithms which you want to use on the taken dataset. After selection, the accuracy of each algorithm we have selected gets predicted. A modal is generated for the algorithm having the highest accuracy. Of all, gradient boosting and ADA boasting algorithms is having the best accuracy result after pre-processing our datasets.

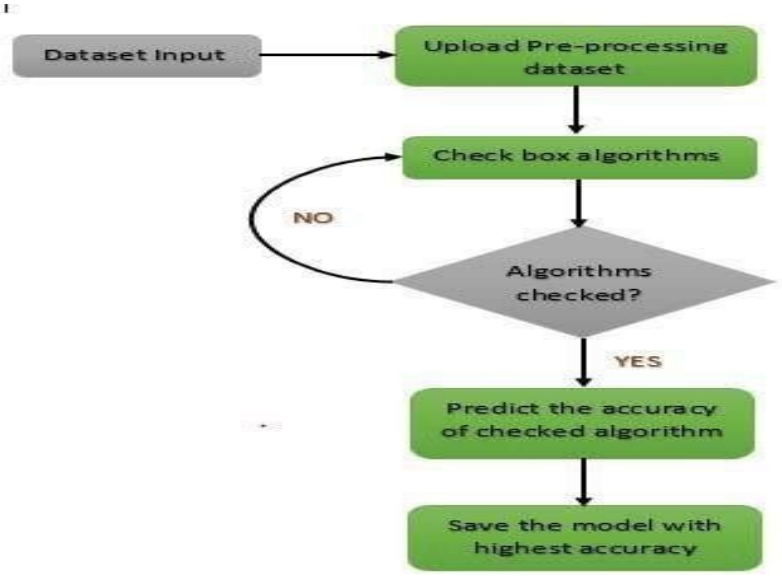

Fig. 3.1: Flowchart.

IV. EXISTING METHODOLOGIES

A. Classification Models

1) Naïve Bayes

In Naive Bayes algorithm, real data, features are often correlated. This assumption can lead to inaccurate forecasts. Naive Bayesian algorithms tend to perform well only when the data is small or the data has bounded values. As the data becomes more complex, the performance of the algorithms tends to degrade. The Naive Bayes algorithm is sensitive to outliers or outliers in the data. These outliers can significantly affect the accuracy of the model, leading to erroneous conclusions.

2) Support Vector Machine:

The SVM algorithm is sensitive to outliers which can affect the accuracy of the prediction. SVM algorithms require large amounts of data for training and testing, and obtaining such data is difficult. The SVM algorithm must iterate over a large number of parameters to determine the best hyperplane, which can be time consuming. Choosing the best kernel, the best regularization parameters, and the best complexity parameters can be difficult and requires expertise and experience. The SVM algorithm can overfit the training data, producing a model that performs well on the training dataset but poorly on the test dataset.SVM models are often difficult to interpret, especially nonlinear SVMs, which can obscure the underlying decision process.

3) Logistic Model Trees

The biggest challenge with the logistic model tree algorithm is the tendency to over fit the data. Overfitting means that the model is too complex and too close to the training data. This leads to poor generalization to new data and inaccurate predictions. Logic model trees are generally only suitable for problems with binary outcomes. This means that they may not be useful in predicting outcomes with more than two possible values.38 Logistic model trees can become very complex, with a large number of decision nodes and potential interactions between predictor variables. This can make them difficult to interpret and explain to non-expert users. Due to its complexity, the logic model tree can become a "black box", making it difficult to understand how the model arrived at a particular prediction or decision. Logistic model trees are highly dependent on the quality of the data used to train them. Any errors, missing values or outliers in the data can significantly affect the accuracy and reliability of the model.

4) Random Forest Algorithm

Random forest models can easily overfit training data, resulting in poor generalization to test data. Due to the aggregate nature of the model, it is difficult to interpret the results and assign the importance of the attributes. Random forests comes with several hyperparameters that need to be tuned, such as the number of trees, the number of features per split, and the maximum depth of each tree, which makes it difficult to configure the model efficiently. Random forest models do not perform well on unbalanced datasets because they tend to be biased towards the majority class. Training a random forest model is computationally intensive due to the creation of multiple decision trees.Although random forest models are effective in many applications, they are not as accurate as some other machine learning algorithms such as neural networks.

5) K- Nearest Neighbour Algorithm

The k-nearest neighbor algorithm is sensitive to noisy data because it directly uses information from the training set to make predictions. If the training set contains noisy or irrelevant data, the algorithm may give inaccurate results. Choosing the correct value of k is essential for obtaining accurate prediction results. However, choosing the best value of k can be difficult because it depends on the data used and the domain of the problem. The algorithm must compare the new data point 39 to all data points in the training set to find the k nearest neighbors, which makes the prediction process computationally expensive for large data sets.The algorithm relies heavily on how to calculate the distance between data points. If the distance metric is chosen incorrectly, it can lead to inaccurate predictions. The k-Nearest Neighbors algorithm does not perform well in high-dimensional data because it becomes difficult to accurately calculate distances between points. The curse of dimensionality refers to the problem where a dataset with too many dimensions can cause problems for algorithms. KNN can be used for both regression and classification prediction problems.However, when dealing with industrial issues, it is mainly used for classification because all parameters are evaluated to determine the usefulness of a technology.

V.PROPOSED METHODOLOGIES

A heart attack prediction is nothing but a description of the state of the heart. It is one of the biggest problems in our world. Therefore, his prediction became a subject of intelligence analysis. When the amount of data is large, the medical industry is difficult.Data mining and machine learning take large amounts of data and turn it into information that helps make predictions and make corresponding decisions. We are developing a heart disease prediction framework using the easy-to-understand ADA boost and gradient boost algorithms. Our proposed method is effective in predicting heart disease in patients. The heart disease prediction framework developed in the approach of coronary artery disease. The prediction accuracy is close to 97%.

1) ECG Signals filtering using notch filter:

Noise suppression in electrocardiogram (ECG) signals is very influential in distinguishing the essential features of the signal masked by noise. Power line interference (PLI) is the primary source of noise in most biopotential signals. Digital notch filters were used to remove PLI in ECG signals. However, there are issues such as transient interference and ringing effects, especially when PLI scanning does not meet full-time sampling conditions. This will be shown in Figure 5.1.

2) Filtered ECG Signals after wavelet and smooth filter:

This distorts the ECG signal during patient measurements. Error-free diagnosis of the ECG requires filtering out unwanted signals. Wavelet transform are used to remove this noise from the ECG. The wave creates an oscillating wave that rises from zero, rises to a maximum value, and then falls back to zero. These waves are scaled and translated copies (called "sub-wavelets") of finite length or rapidly decaying oscillatory waveforms (called "mother waves").Each analysis wave is different in duration, resolution and frequency band. Therefore, the duty cycle after the waveform transformation corresponds to the measured value of the electrocardiographic element in the frequency band and the time band. The wavelet threshold facilitates further signal processing. There are four types of thresholds in the wavelet transform. Rig sure and Sqtwolog for one-dimensional data. Hoursure thresholding is used for low SNR signals.

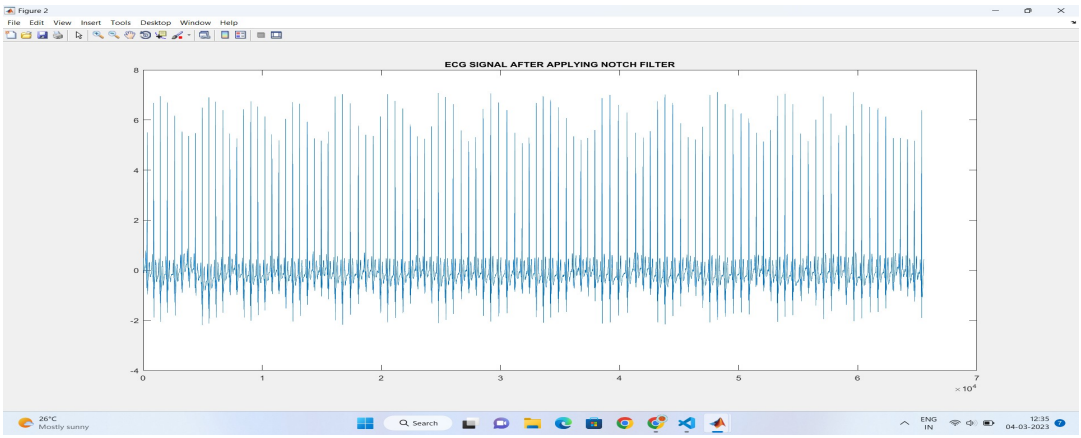

Fig. 5.1 : After applying Notch Filter.

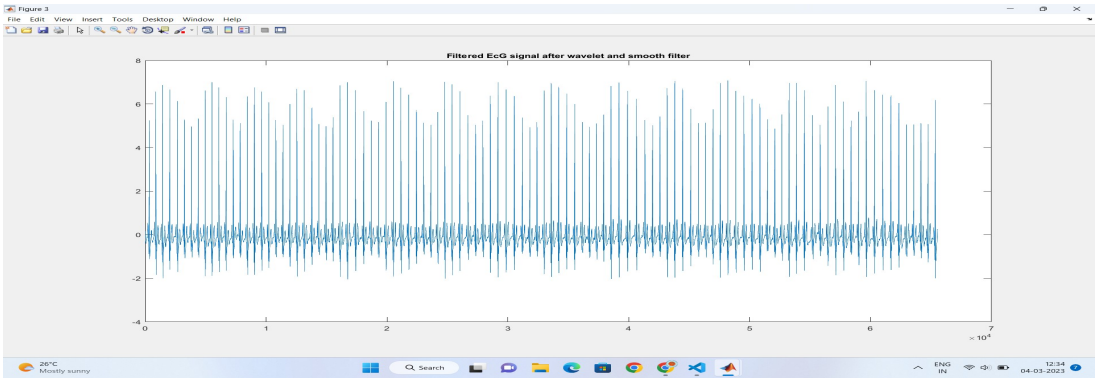

Fig. 5.2 : After applying Wavelet filter

A. Classification Models

In machine learning, there are many classification models that refer to predictive modeling by taking given information as input. Extraction is performed on the model of the data class. A classifier or classification model predicts a categorical class. In machine learning, we mainly perform two types of operations; one is prediction and the other is decision making. We used classification models from machine learning. The models we consider are ADA boosting and gradient boosting algorithms.

1) ADA Boosting algorithm

The ADA Boost algorithm is an acronym for Adaptive Boosting, which is a boosting technique used as an ensemble method in machine learning. This is called adaptive amplification because weights are reassigned to each instance, with higher weights assigned to misclassified instances. Reinforcement is used to reduce bias as well as variance in supervised learning. It works because learners grow sequentially. Except for the first, each subsequent learner grows from the previously developed learners.

2) Gradient Boosting algorithm

Gradient boosting is a machine learning technique used for regression and classification tasks, among others. This gives a predictive model as a set of weak predictive models, usually decision trees. When a decision tree is a poor learner, the resulting algorithm is called a boosted gradient tree; it usually outperforms random forests. Gradient-boosted tree models are built in stages like other boosting methods, but they generalize other methods by allowing optimization of arbitrary differentiable loss functions. If the training set is fitted too tightly, the generalizability of the model will decrease.Various so-called regularization techniques reduce this overfitting effect by limiting the fitting process. Another regularization parameter is the depth of the tree. The higher the value, the more likely the model is to overfit the training data.

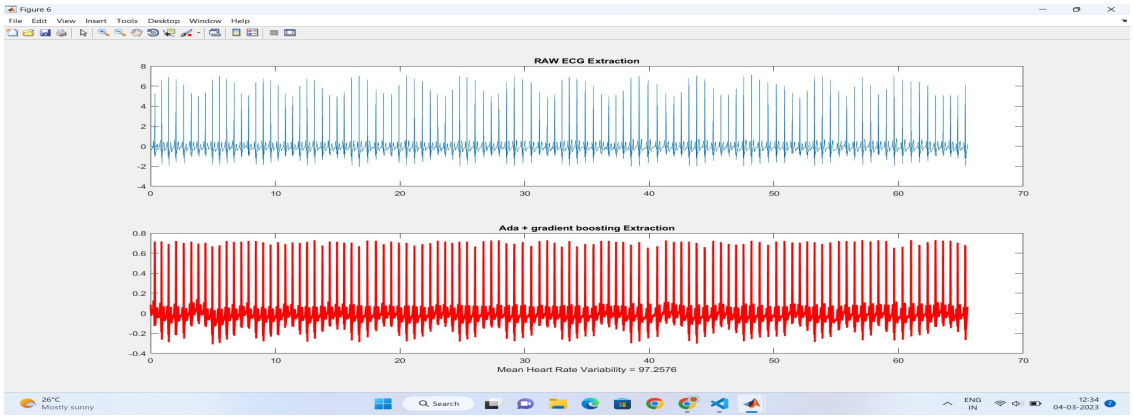

Fig. 5.3 : ADA Boosting and Gradient Boosting

B. Block Diagram

The complete block diagram of the project is displayed in the Fig. 5.4

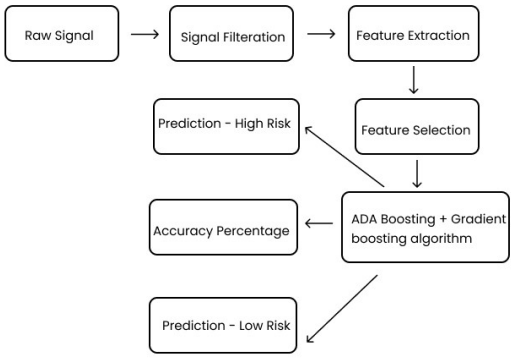

Fig.5.4 : Block Diagram

B. Prediction Result

Fig 5.5 displays the prediction result as “Low Risk” “High Risk” and “Normal” with accuracy percentage.

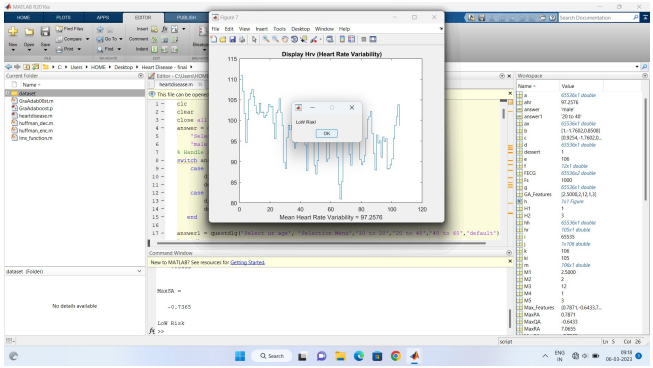

Fig. 5.5 : Predicting the result.

VI .EXPERIMENTAL RESULTS

After performing all the classification techniques, the gradient boosting and ADA boosting algorithms have an accuracy of 96.23%, which is good and superior compared to other models. The accuracies obtained by all the algorithmic models are mentioned in Table 6.1 below.

S No	Classification Algorithms	Accuracy
1	Gaussian Naïve Bayes	82.25%
2	Support Vector Machine	81.97%
3	Random Forest	86.32%
4	K- Nearest Neighbour	67.21%
5	Xg-Boost	78.69%
6	Gradient+ADA Boosting algorithm.	96.23%

Table 6.1. Algorithm accuracy comparison

After running all the classification techniques, the gradient boosting and ADA boosting algorithms achieved an accuracy of 96.23%, which is much higher than other models. The accuracies obtained by all the algorithmic models are presented in the table below. This is the accuracy obtained after pre-processing our dataset. Here is the accuracy map we obtained after pre-processing the classification model we chose for the dataset. The precision plot will be shown in Figure 6.1.

Displaying Accuracy

Fig 6.1 displays the accuracy bar graph compared with KNN algorithm, Random Forest and Proposed Algorithm.

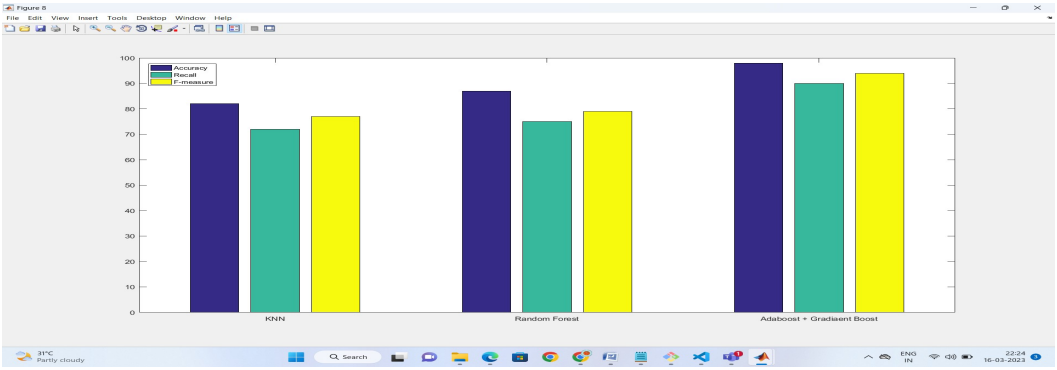

Fig. 6.1 : Accuracy chart

Algorithmic models and accuracy of each algorithmic model after the training dataset. The modalities are generated using the most accurate algorithm among the gradient boosting and ADA boosting algorithms mentioned in the figure.

VII. CONCLUSION

From our research, we understood how to achieve our exploration objectives. To predict heart disease, we used different machine learning algorithms. The cardiac dataset containing its factors was extracted from the UCI machine learning repository and another classification model was applied to the corresponding dataset. Our proposed method uses five machine learning models, namely Support Vector Machine, Random Forest, KNN, Gaussian Naive Bayes, Xg-Boost Algorithm, Gradient Boost and ADA Boost Algorithm. , to predict heart disease in a short time to get the result and human expenditure. We use these algorithms to improve normalization.In this project, a new method was developed to classify cardiac signals as normal or abnormal and then segment cardiac lesions from abnormal signals. Features were extracted from the co-occurrence matrix of high-frequency profile coefficients using notch filters and classified using a hybrid algorithm of ADA amplification and gradient amplification. The result is more accurate and time efficient, which we compare to the random forest algorithm. In conclusion, prognostication of heart disease using ECG signals has been successfully full-fledged using ADA boosting and gradient boosting algorithms. Notch filters, wavelet transforms and smoothing filters immensely improves the accuracy and efficiency of the anticipating process.With the increasing amount of data collected from ECG recordings, these techniques provide an effective method for the early detection and diagnosis of heart disease. Further research can be done to optimize parameters and investigate the effectiveness of other machine learning algorithms. Overall, this study demonstrates the potential of data-driven approaches to improve healthcare quality and improve individual well-being. In this study, we compared the performance of two popular machine learning algorithms (ADA augmentation and gradient augmentation) in predicting heart disease using ECG signals. We found that both algorithms were effective in identifying potential cases of heart disease from ECG signals. However, the gradient boosting algorithm outperformed the ADA scaling algorithm with a higher accuracy rate of 96% and an F1 score on the test set. Improved computational accuracy and performance of adopted and presented algorithms in desired areas. Improve the flexibility of the proposed framework by making any necessary improvements.

REFERENCES

[1] Han, J & Kamber, M 2012, “Data Mining Concepts and Techniques”, Morgan Kaufman Publishers, USA.

[2] Sathyadevi, G 2011, “Application of CART algorithm in Hepatitis disease diagnosis”, IEEE International Conference on Recent Trends in Information Technology, India, pp.1283-1288.

[3] Krishnaiah, V, Narsimha, G & Chandra, NS 2013, “Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques”, International Journal of Computer Science and Information Technologies, Vol. 4, No. 1, pp. 39-45.

[4] Nasser, HS, Tharwat, AA & Moniem, NKA 2010, “Support vector machine for diagnosis cancer disease: A comparative study”, Egyptian Informatics Journal, Vol.11, No. 2, pp. 81-92.

[5] Navid, KG & Saheb, A 2015, “Combination of PSO Algorithm and Naive Bayesian Classification for Parkinson Disease Diagnosis”, Advances in Computer Science: an International Journal, Vol. 4, No. 4, pp. 119-125.

[6] Chaves, R, Ramirz, J, Gorriz, JM & Puntonet, CG 2012, “Association rule-based feature selection method for Alzheimers disease diagnosis”, Vol. 39, No. 14, pp. 11766-11774.

[7] Venkatalakshmi, B & Shivsankar, MV 2014, “Heart disease diagnosis using predictive data mining”, 2014 IEEE International Conference on Innovations in Engineering and Technology (ICIET’14), Tamil Nadu, India, Vol. 3, No. 3, pp. 1873-1877.

[8] [8] Cherian, V & Bindu, MS 2017,“Heart Disease Prediction Using Naïve Bayes Algorithm and Laplace Smoothing Technique”, Vol. 5, No. 2, pp. 68-73.

[9] Shinde, SB & AmritPriyadarshi 2015, “Diagnosis of heart disease using data mining technique”, Vol. 4, No. 2, pp. 2301-2303.

[10] Palaniappan, S & Awang, R 2008,“Intelligent heart disease prediction system using data mining techniques”, Computer Systems and Applications, 2008. AICCSA 2008. IEEE/ACS International Conference on IEEE, pp. 108-115.

[11] Shouman, M, Turner, T & Stocker, R 2012,“Using Data Mining techniques in heart disease diagnosis and treatment”, Electronics Communications and Computers (JECECC), 2012 Japan-Egypt Conference on IEEE, pp. 173-177. 110

[12] Deekshatulu, BL & Chandra, P 2013,“Classification of heart disease using artificial neural network and feature subset selection”, Global Journal of Computer Science and Technology, Vol. 13, No. 3, pp. 4-

[13] Sonawane, JS & Patil, DR 2014, “Prediction of heart disease using learning vector quantization algorithm”, IT in Business, Industry and Government (CSIBIG), Conference on IEEE, pp. 1-5.

[14] Tugce Kantar, Ovul Koseoglu, Aykut Erdamar, 2014,“Analysis of Heart Diseases from ECG signal”, National Biomedical Engineering Meeting (BIYOMUT), 2014 18th National IEEE, pp. 1-4.

[15] Miao, F, Cai, YP & Zhang, YT 2014, “Risk prediction for heart failure incidence within 1-year using clinical and laboratory factors”, In Engineering in Medicine and Biology Society (EMBC), 36th Annual International Conference of the IEEE, pp. 1790-1793.

[16] Farooq, K, Karasek, J, Atassi, H, Hussain, A, Yang, P, MacRae, C & Slack, W 2014, “A novel cardiovascular decision support framework for effective clinical risk assessment”, IEEE Symposium on Computational Intelligence in Healthcare and e-health (CICARE), pp. 117-124.

[17] Mohawish, A, Rathi, R, Abhishek, V, Lauritzen, T & Padman, R 2015,“Predicting Coronary Heart Disease risk using health risk assessment data”, E-health Networking, Application & Services (Health Com), 17th International Conference on IEEE, pp. 91-96.

[18] Lafta, R, Zhang, J, Tao, X, Li, Y & Tseng, VS 2015, “An intelligent recommender system based on short-term risk prediction for heart disease patients”, Web Intelligence and Intelligent Agent Technology (WI-IAT), IEEE/ WIC/ ACM International Conference on IEEE, Vol. 3, pp. 102-105.

[19] Francis Guillemin, Alain Leplege, Serge Briancon, Elisabeth Spitz & Joel Coste 2017, “Perceived health and adaptation in chronic disease”, Taylor and Francis, First Edition.

[20] Bharti, S & Singh, SN 2015, "Analytical study of heart disease prediction comparing with different algorithms", Computing, Communication & Automation (ICCCA), International Conference on IEEE, pp. 78-82.
[21] Raihan, M, Mondal, S, More, A, Sagor, M. O. F, Sikder, G, Majumder, MA & Ghosh, K 2016,"Smartphone based ischemic heart disease (heart attack) risk prediction using 111 clinical data and data mining approaches, a prototype design", Computer and Information Technology (ICCIT), 19th International Conference on IEEE, pp. 299-303.
[22] Alqaraawi, A, Alwosheel, A & Alasaad, A 2016, "Towards efficient heart rate variability estimation in artifact-induced Photo plethysmography signals", Electrical and Computer Engineering (CCECE), IEEE Canadian Conference on IEEE, pp. 1-6.
[23] C.Nagarajan and M.Madheswaran - 'Experimental verification and stability state space analysis of CLL-T Series Parallel Resonant Converter' - Journal of ELECTRICAL ENGINEERING, Vol.63 (6), pp.365-372, Dec.2012.
[24] C.Nagarajan and M.Madheswaran - 'Performance Analysis of LCL-T Resonant Converter with Fuzzy/PID Using State Space Analysis'- Springer, Electrical Engineering, Vol.93 (3), pp.167-178, September 2011.
[25] C.Nagarajan and M.Madheswaran - 'Stability Analysis of Series Parallel Resonant Converter with Fuzzy Logic Controller Using State Space Techniques'- Taylor & Francis, Electric Power Components and Systems, Vol.39 (8), pp.780-793, May 2011.
[26] Nagarajan and M.Madheswaran - 'Experimental Study and steady state stability analysis of CLL-T Series Parallel Resonant Converter with Fuzzy controller using State Space Analysis'- Iranian Journal of Electrical & Electronic Engineering, Vol.8 (3), pp.259-267, September 2012.