An Effectual Machine Learning Based Coronary Artery Disease Classification for Low Error Rates

Archita Bhatnagar[a], Prof. (Dr.) Manoj Kapil[b]

[a] Research Scholar (Computer Science Engineering), Subharti Institute of Technology & Engineering, Swami Vivekanand Subharti University, Meerut
[b] Professor, Subharti Institute of Technology & Engineering, Swami Vivekanand Subharti University, Meerut

Article History: Received: 10 November 2020; Revised 12 January 2021 Accepted: 27January 2021; Published online: 5 April2021

Abstract:
Heart syndrome is the common and significant reasonfordeath in the world nowadays. Estimation of cardiovascular infection is a critical experiment in the region of clinical data breakdown. Machine knowledge has been presented to be operative in supportof making conclusions and estimations from the large number of recordscreated by healthcare engineering. The prediction model is familiarized with diversegroupings of structures and several well-known classification practices. The proposed approach deals with the efficient machine learning model for the detection of the CAD which are having low validation and testing errors and achieves high true positive error rates.Our proposed model consists of hybridization of optimization processes using PSO and firefly nature-inspired and the classification is performed on the data using discriminant analysis.The proposed approach is achieving above 95% accuracy of the detection on different test samples to achieve high-performance classifications.

Keywords: Artificial Intelligence, Coronary artery diseases, Classification, instance selections

1. Introduction

It is very complex to recognize heart problems because of various risk aspects such as diabetes, high B.P, heavyweight, irregular pulse rate, and countless other factors. Differentpractices in data miningare used and various classifications have been engaged to find out the solutionsto heart disease in humans. The severity of the illness is categorized on various procedures such as decision trees and many other hierarchical structures [1][2]. The environment of heart disease is difficult and hence, the infection must be handled sensibly. Not undertaking much heart or cause impulsive deaths. The perception of medical knowledge and data mining are recycled for determining various categories of metabolic disorders. Data mining with a groupingapproach plays a significant part in the estimation of heart problems and data explorations. Most of the studies are done in predicting the accuracy of the specific functions related to heart syndrome. Various procedures have been recycled for understandingthe heart-related data methods of data mining to diagnose efficiently and effectually [3][5].

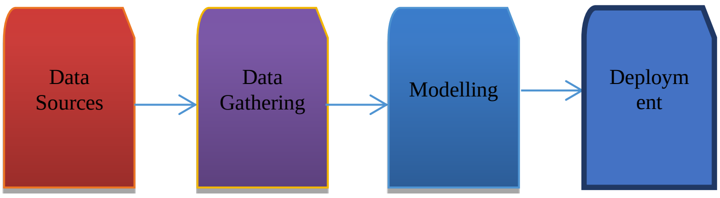


Fig 1: Data Mining Process [4]

The rest paper is divided into various phases. Section II shows the related work and review of the valid researches done. Section III shows the problem occurring in the current scenario. Section IV deals with the proposed workflow algorithm discussions. Section V shows the proposed result and discussion and the last section covers the conclusion and future scope of the proposed work.

2. Related Work

Heart problem these daysis a very serious problem which needs better for the proper diagnosis This section deals with the various research works which put light on the various previous researches done in the medical field.**Ilayaraja M, Meyyappan T [6**] worked on the predictive modeling to get the risk level heart disease patient. They have worked on the itemsets based on the support value. They have implemented their research in JAVA and achieved better accuracy.**KaanUyara et al. [7]** proposed an efficient genetic algorithm basedon the fuzzy logic for heart problem diagnosis where they have achieved 97.78% accuracy usingthe Cleveland dataset based on heart disease.**Rashmi G Sabojiet al. [8]** purposedthe diagnosis process of heart disease using Random Forest classification and they have achieved approx. 98% accuracy.**Mollet, Nico et al. [9]** proposed a HOG procedure thatis more efficient inappropriate for the exploration and ordering of heart diseases than the traditional Histogram of Oriented Gradient. The overall arrangement has been evaluated through intensive tests using various classifiers.

Kumar,Gandhi, et al. [10] presented a computer-aided system for the recognition of heart disease by usingIoT tools. They have used various IoT protocols for the real-time data and classify the diseases. Their analysis has worked on various significant parameters such as asymmetry, texture, analysis for regularization, and feature steps. The mined feature limitations are used to categorize the information.**Safdar, Saima, SaadZafar et al. [11]** proposed an evaluation for the arrangementto get the relationship of syndrome and non-disease through the classifiers. They collected information and perform data mining techniques for the handling of the data. They have performed a feature extraction process and classify the labels for the prediction using binarization.**Manogaran, R. Varatharajan et al. [12]** presented aninnovativeprocess for the recognition of heart infection. To sense the number of noises from records, the pre-processing phase is carried out by using filters. And consequently, the neuro-fuzzy methodology is executed to fill in the mysterious data. Fuzzy inference approaches are used to describe the information for the grouping of the infection. The process was estimated on a real-time dataset with efficientaccuracywhich is compared with the other same methods.**Ramalingam, Ayantan et al. [13]**performed analysis on several machines learning processes and classifiers. They have studied various current practices used for the categorizations and discussed the various difficulties used for the evaluations of their proposed method and implementation efficiently.**Kannan, Vasanthi, et al. [14]** presented their research on the predictive modeling using ROC curve using training and validation process. Their results are evaluated using true positive rates in terms of specificity and false negative which they have further comparedwith the previous studies.

3. Problem Statement

Classification and regularizations are certain tasks in the detection of coronary heart diseases. So it plays a significant role in the detection of disease data points in medical applications. Physical detection and examination of the diseased parts are a time-consuming and unpredictable task, and as the number of data points or values increases; the performance evaluation becomes very tough. However, it is comparatively time-consuming when the execution of the complex test cases becomes quite large and needs to be applied recurrently with different orientations. In many applied cases the simple thresholding is used but very less work is done in the optimization and classifications. The main focus of this research work is to enhance the accuracy and acceptance rate as well as to reduce the rejection rate of the detection. Lastly, the performance of the proposed system will be calculated to evaluate the performance and reliability of our proposed work.

4. Proposed Work

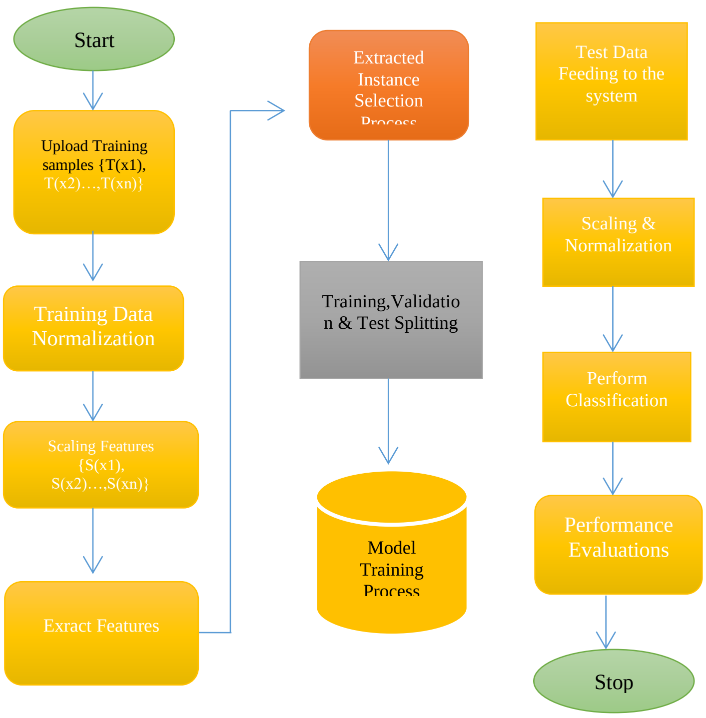


Fig 2: Proposed Flow Diagram

The proposed model is the following:-

1. Dataset

The proposed work uses the well-known dataset named as Z-AlizadehSanithrough which the performance and classification are performed. This dataset involves 303 patient proceedings and 56 features or columns. , 216 individuals with CAD, and 88 individuals with the normal position.The structures used in the information gatherings are divided into severalclusters i.e. demographic, indicator and investigation, ECG, and workshop and topographies. This dataset is generally used in the several latest types of researchhaving a large number of characteristics as compared to various other CAD datathrough which we came to know that this dataset is more informatics. The motive for consuming this dataset is that it contains features that put heavy weights and having high correlations within the attributes of the datasets. [19].

2. Data Normalization & Scaling

After uploading the data, normalization of the data is a crucial part. The values in the data are alphanumeric which needs to be normalized and scale for the less variance and standard deviations. The scaling intends to standardize the information. The strings are converted into the numeric form so that the processing will be performed in the vector form instead of the alphanumeric which is the significant part of the implementation. If this is not properly done then the training set will achieveoverfitting and underfitting of the model which is not good for our proposed work.

3. Feature Extraction

This performs the feature vector extraction process using PCA which is one of the significant processes to find the eigenvalues and eigenvectors by using the covariance process. In the proposed work the feature engineering is used so the relevant feature vector is attained to identify the patterns using correlations which is transformed into a new vector representation. Most of the feature algorithms are having one main problem which is thecomputation time and execution processing complexity which is overcome by PCA and it will also reduce the nonlinearity among the data in the N-dimensional space. So in the proposed approach, the linear kernel PCA is used which will reduce the non-linearity and reduces the variance among the data points to extract the highly correlated features which are then transformed to the feature vector and are the significant information for the feature vector.

4. Feature Optimization

This section deals with the hybrid optimization process of two nature-inspired algorithms that perform the instance selection process which is the significant part of the implementation to select the relevant features to form the extracted feature vector which is the eigenvector arranged in the multi-dimension vector. The explanation of the algorithm is given below.

a. Firefly-based modified instance selection-This is inspired by the blinking fireflies. Various norms are implemented for these algorithms.They help solve the nonlinear, dynamic, anomaly problems. The main steps covered in the proposed work for the instance selections

1) Particles are fascinatedby each other.

2) Magnetism is relative to the glow. Less optimistic is fascinatedby the brighter particle or instance.

3) If the intensityis the same for both instances, then there will be a random movement.

4) New best solutions are producedusing random walk.

There are various applications for Firefly algorithms

FA is applied in nonlinear problems, dynamic problems, feature selection, fault detection many more.

There are numerous advantages of FA over other optimization algorithms.

1) Automatic subdivision of the whole population into subgroups

2) Natural capability of dealing with multi-model optimization.

3) High randomness in the solutions.

Start

Determine the unbiased utility for the instance selection as an input.

such that: $\{xn\} = (xn\{1\}, xn\{2\}, \ldots, xn\{d\})\}$

Determine the instance as fireflies inhabitants such that $\{p(i)\} \rightarrow (i = 1,2,\ldots,n)$.

Estimate the intensity of the instance which is linked with f(xn) i.e. Int{val} = f(xn) where Int is the intensity value.

Define absorptioncapacity γ

While (t(pn) $< M$(gn) i.e. max iterations

forxp $= 1 :$ F(xn)

for: yp $= 1 :$ F(xn)

IF ({I{xp} $> I\{yp\}$})

Diverge A(x) i.e. fascination with dist € Distnace(xn{i})

Move F(xn) $from\ xp \rightarrow yp$;

EvaluateS(x{n}) and modifyInt{val}

end if

end for yp

end for xp

Perform best possible solutions and rank instances.

end While

Stop

b. Particle Swarm based modified instance selection -It is a metaheuristicprocess that cracks hard optimization complications based on complex computations. This procedure is encouraged by the communal flocking birds. It is a population-based searchingprocess. Each instance is called a swarm particle. Each instance in the swarm has a rate of movement. They act as optimal function to control structure. It is effectivelypractical to the heart disease classification and showing the efficientoutcome. Inmetaheuristicprocesses,the objective function is fed to minimize or maximize as per the requirementwhich is needed to be optimized. The proposed steps are given below.

For each occurrence as swarm inputx{p} $= 1, \ldots,$ S{p} do

Formulate the P{p} as position of the instanceby distribution trajectory: xp(i) $\rightarrow$ U(L{b}, U{b})

Arrange the x{p}instance known location to its initial location: s{i} $\leftarrow$ xp{i}

If func(p{i}) $< func(\{g\})$:

modify the swarm best knownloc: g{b} $\leftarrow$ s{i}

Fixed the instance rateinj a velocity vector: SV{i} ~ U(−|U{b} − L{b}|, |U{b} − L{b}|)

while a determinedquantity does not occur to do:

For each instancex{p} $= 1, \ldots,$ S{p} do

For each dimensiond{s} $= 1, \ldots,$ Nd{p} do

Check randomrecords: rnd{p}, r{g} ~ U{0,1}

Perform rate of speed updations of the each instance in the swarm:

SV(i, ds) $\leftarrow$ ω (SV{i}, ds) + φ{p} rnd{p} (xp(i, ds) − x(i, ds)) + φ(g{b}) r(g{b}) (g{b} − x(i, ds))

Modify the position of the instance in the vector: x{i} $\leftarrow$ x{i} + LR (SV{i})

if f(x{i}) < f(p{i}) then

Modify the instance best known locus: pi $\leftarrow$ xpi

iffunc(pi) $< func(g)$ then

Adjust the instance best-known pos: gb $\leftarrow$ pi

End For

End For

5. Discriminant Analysis and Classifying using k-fold cross-validation

In the proposed approach we have performed statistical analysis to evaluate the linear grouping of features that describes or splits two or more objects based on the classification process. It will act as a linear classifier which reduces the non-linearity in the training of the model and is used to perform high true positive based predictive modeling. It is also used to analyze the variance in the data because if the data perform high variance then it will increase the deviations in the data which will result in improper classifications. In our proposed approach discriminant analysis is performed on the independent variables in the continuous form and the categorical dependent variable as prediction labels.

5. Proposed Algorithm

Step 1: Input Records such that $t\{s\} = t\{s1\}, t\{s2\} \dots t\{n\}$as data and execute the framing of the data to process informationefficiently.

Step 2: Standardize&data scaling

For x=1 to $len(t\{s\})$

S_D = StdScaling { $t(s)$} to reduce the variances among data points.

EndFor

Where t is the total training samples

Step 3: Implement the extraction of the features & perform covariance data in the vector form

For p=1 to S_D

Cv(p) = COV(S_D)

EndFor

Step 4: Eigenvalues and vectors extractions for the transformation T{x}= $X \times W$ for the informationfor new space vector generation.

Where VE(v) where VE = {VE_1, VE_2… VE_N}is processed vector.

Step 5: Perform instance selections using hybridoptimization proceduresfor the selection of relevant instances usingthe feature reduction method

While ($t(pn) < M$(gn) i.e. max iterations

IF ({I{xp} $> I\{yp\}$})

Diverge A(x) i.e. fascination with dist € Distnace(xn{i})

Move F(xn) $from\ xp \rightarrow yp$;

Evaluate S(x{n}) and modify Int{val}

Perform best possible solutions and rank instances.

end While

For each occurrence as swarm input $x\{p\} = 1, \dots, S\{p\}$ do

Formulate the P{p} as position of the instance by distribution trajectory: $xp(i) \rightarrow U(L\{b\}, U\{b\})$

Arrange the x{p} instance known location to its initial location: $s\{i\} \leftarrow xp\{i\}$

If $func(p\{i\}) < func(\{g\})$:

while a determined quantity does not occur to do:

For each instance $x\{p\} = 1, \dots, S\{p\}$ do

For each dimension $d\{s\} = 1, \dots, Nd\{p\}$ do

Check random records: rnd{p}, r{g} ~ U{0,1}

Perform rate of speed updations of the each instance in the swarm:

$$SV(i, ds) \leftarrow \omega\ (SV\{i\}, ds) + \varphi\{p\}\ rnd\{p\}\ (xp(i, ds) - x(i, ds)) + \varphi(\ g\{b\})\ r(\ g\{b\})\ (\ g\{b\} - x(i, ds))$$

Modify the position of the instance in the vector: $x\{i\} \leftarrow x\{i\} + LR\ (SV\{i\})$

if f(x{i}) < f(p{i}) then

Modify the instance best known locus: pi ← xpi

iffunc(pi) $< func(g)$ then

Adjust the instance best-known pos: gb ← pi

Endif

endif

End For

EndFor

Step 6:Prepare data for training and testing phase

N_D= {Txs (N_D)}

70% training data and 30% for the test data.

Step 7:Perform Discriminant analysis classification

L{m} = {FitTransform[TR(N_D)]}

Where fit_transform generate the configuration of the training model and TR(x) is the training samples**Step 8:** Upload Test data ssuchthat TD_S = {TD_{S1}, TD_{S2}, TD_{S3}, TD_{S4}… TD_{SN}}

Step 9: Training model loading and Implement classification on T_{SN}.

Step 10: Evaluate Performance Evaluations and Repeat Step 5 to 9 until all configurations get completed.

6. Result And Discussions

This section covers the implementation part of the proposed model which is implemented in the MATLAB environment. No external library is used in the proposed system. A detailed explanation is given below.

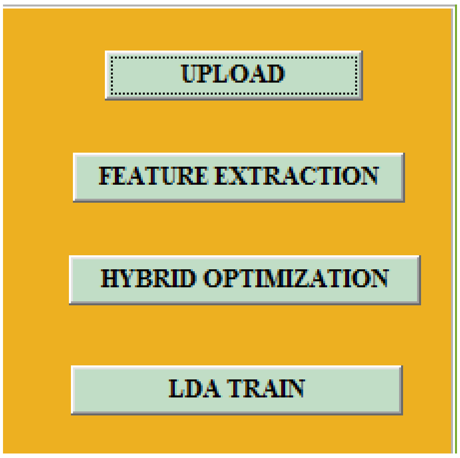
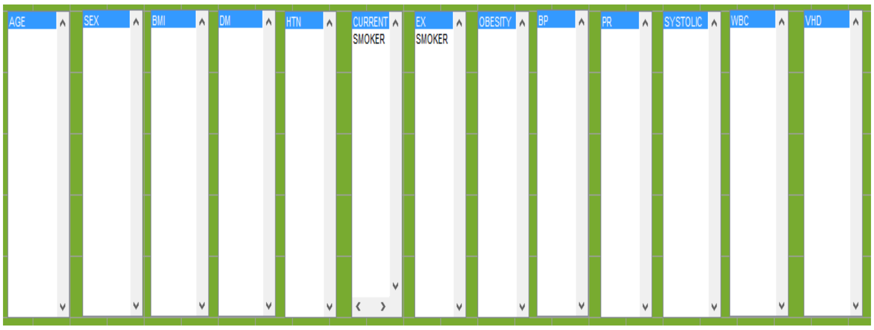


Fig 3: User Panel

Fig 3 shows the proposed panel shows the user interface consists of the graphical user interface tools which deal with the user interactions with the machines. The panel consists of the static texts, pushbuttons through which the user clicks on the pushbuttons and get the output which perform specific functions.



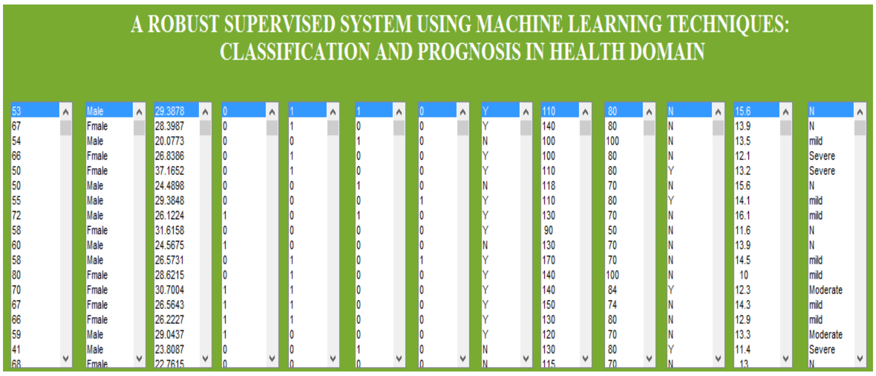


Fig 4: Training Panel

Fig 4 shows the training panel which consists of the list boxes in which the uploaded data is displayed from excel and the characteristics and nature of the data taken. It shows that the individual properties on which the processing will be performed. The data shown on the panel is limited but the back data which is processed is of a total of 56 attributes for the one individual whichwill give more insights into the data.

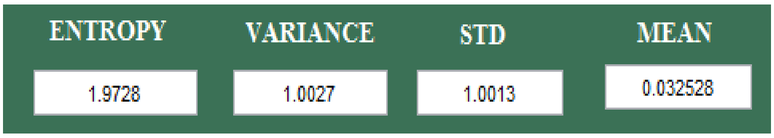


Fig 5: Extracted Features

Fig 5 shows the extracted features from the feature vector. From fig 6 the processing of the data can be controlled and it will give proper information of the feature values with the other data values. The entropy is considered as the significant parameter which shows the disorder among the data points. Variance, the standard deviation will tell us the measure of dispersion and spreading of the data, and mean shows us the tendency of data points to be centered for the groupings. It gives us important information about the total population concerning the observed values.

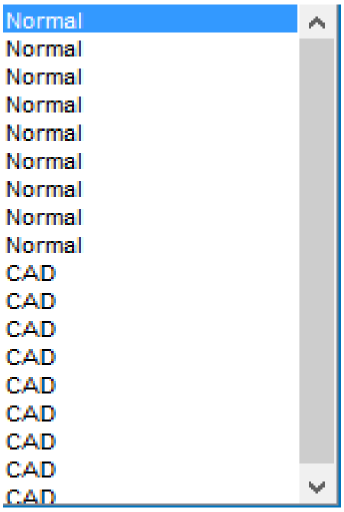


Fig 6: Classified Outcomes

Fig 6 shows the classified outcome as a result of which it can be distinguished that how many individuals are having CAD and how many are in a normal state. It is classified using discriminant analysis where the trained data model is loaded which is processed on the unknown data which is considered as the test data from our side. In the proposed model the 70% of the total training data and 30% of the test data are taken to check the validity and performance of the proposed system.

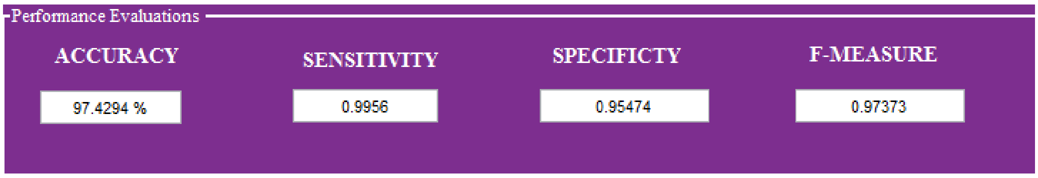


Fig 7: Performance Analysis

Fig 7 shows the performance evaluation of the proposed work. It can be noticed that the proposed supervised learning model is achieving high performance in terms of true positive and negative rates. It shows that the recognition accuracy is approx. 98% which is the desired outcome and also the sensitivity and specificity is increasing which increases the true positive and true negative rate of the proposed model. Also, the F-measure should be high which indicates that the proposed approach is retrieving efficient relevant information based on the training model which shows increases the precision and recall of the information retrieval process. The performance is evaluated using the following equations.

$$\mathrm{P(x) = tp(x) \div (tp(x) + fp(x))}$$

$$\mathrm{R(x) = tp(x) \div (tp(x) + fn(x))}$$

$$\mathrm{Sp(x) = tn(x) \div (tn(x) + fp(x))}$$

$$\mathrm{Sn(x) = tp(x) \div (tp(x) + fn(x))}$$

$$\mathrm{A(x) = tp(x + fp(x)) \div (tp(x) + fn(x) + fp(x) + tn(x))}$$

$$\mathrm{Fm(x) = 2 \times (\frac{[P(x) \times R(x)]}{[P(x) + R(x)]})}$$

Where $\mathrm{P(x)}$ and $\mathrm{R(x)}$ is the evaluated precision and recall of the proposed model. $\mathrm{Sp(x)}$ and $\mathrm{Sn(x)}$ is the sensitivity and specificity of the proposed model. $\mathrm{A(x)}$ and $\mathrm{Fm(x)}$ is the evaluated accuracy and F-measure of the proposed model.

Table 1: Accuracy Performance

Test No.	Accuracy (%)
1	97.429
2	97.319
3	98.185
4	98.401
5	96.297
6	96.310
7	96.739
8	97185
9	98.006
10	97.071

Table 2: Sensitivity Performance

Test No.	Sensitivity
1	0.972
2	0.979
3	0.985
4	0.981
5	0.962
6	0.968
7	0.969
8	0.974
9	0.986
10	0.978

Table 3: Specificity Performance

Test No.	Specificity
1	0.954
2	0.961

3	0.958
4	0.955
5	0.957
6	0.960
7	0.951
8	0.958
9	0.959
10	0.961

Table 4: F-Measure Performance

Test No.	F-Measure
1	0.973
2	0.956
3	0.978
4	0.981
5	0.976
6	0.976
7	0.962
8	0.978
9	0.974
10	0.968

Table 1, 2, 3, 4 shows the performance analysis on different test samples and can check the variation in different test samples. It can be seen that there are not that many variations after applying different samples of the test which shows that our proposed approach is having the highly precise evaluation of the system to achieve high accuracy in terms of true positive and negative rate and low classification error rates.

7. Conclusion & Future Scope

The health care data should be monitored timely and accurately which will help doctors to diagnosepatients efficiently. Still, the accuracies and precisions on the data are low and not upto the mark. So an efficient precise model is required which can give full and better insights among the data to gather information and characteristics about the patient's health. This paper put light on the robust predictive modeling using feature extraction and instance selection hybridization which gives betterment in the evaluation of the proposed model using discriminant analysis. It can be noticed from the proposed work performance that the developed predictive modeling is achieving satisfactory results with low false positive and negative rates which are the desired outputs and will help doctors with the effectual diagnosis of the CAD or normal state of the patient.

References

[1] Alizadehsani, R., Khosravi, A., Roshanzamir, M., Abdar, M., Sarrafzadegan, N., Shafie, D &Bishara, A. (2020). Coronary Artery Disease Detection Using Artificial Intelligence Techniques: A Survey of Trends, Geographical Differences, and Diagnostic Features 1991-2020. Computers in Biology and Medicine, 104095.

[2] Ghiasi, M. M., Zendehboudi, S., &Mohsenipour, A. A. (2020). Decision tree-based diagnosis of coronary artery disease: CART model. Computer methods and programs in biomedicine, 192, 105400.

[3] Setiawan, N. A., Venkatachalam, P. A., & Hani, A. F. M. (2020). Diagnosis of coronary artery disease using artificial intelligence based decision support system. arXiv preprint arXiv:2007.02854.

[4] Chen, M., Wang, X., Hao, G., Cheng, X., Ma, C., Guo, N., ...& Hu, C. (2020). Diagnostic performance of deep learning-based vascular extraction and stenosis detection technique for coronary artery disease.The British journal of radiology, 93(1113), 20191028.

[5] Orlenko, A., Kofink, D., Lyytikäinen, L. P., Nikus, K., Mishra, P., Kuukasjärvi, P., ...& Moore, J. H. (2020). Model selection for metabolomics: predicting diagnosis of coronary artery disease using automated machine learning. Bioinformatics, 36(6), 1772-1778.

[6] Ilayaraja M, Meyyappan T." Efficient Data Mining Method to Predict the Risk of Heart Diseases through Frequent Itemsets." 4th International Conference on Eco-friendly Computing and Communication Systems (2015) 586 – 592.

[7] KaanUyara, Ahmetİlhan. " Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks." 9th International Conference on Theory and Application of Soft Computing, Computing with Words and Perception, ICSCCW 2017, 22-23 August 2017, Budapest, Hungary.

[8] Rashmi G Saboji ,Prem Kumar Ramesh.” A Scalable Solution for Heart Disease Prediction using Classification Mining Technique.” International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS-2017).
[9] Mollet, Nico R., Steven Dymarkowski, WimVolders, JurgenWathiong, LievenHerbots, Frank E. Rademakers, and Jan Bogaert. "Visualization of ventricular thrombi with contrast-enhanced magnetic resonance imaging in patients with ischemic heart disease." Circulation 106, no. 23 (2002): 2873-2876.
[10] Kumar, PriyanMalarvizhi, and Usha Devi Gandhi. "A novel three-tier Internet of Things architecture with machine learning algorithm for early detection of heart diseases." Computers & Electrical Engineering 65 (2018): 222-235.
[11] Safdar, Saima, SaadZafar, NadeemZafar, and NaurinFarooq Khan. "Machine learning based decision support systems (DSS) for heart disease diagnosis: a review." Artificial Intelligence Review 50, no. 4 (2018): 597-623.
[12] Manogaran, Gunasekaran, R. Varatharajan, and M. K. Priyan. "Hybrid recommendation system for heart disease diagnosis based on multiple kernel learning with adaptive neuro-fuzzy inference system." Multimedia tools and applications 77, no. 4 (2018): 4379-4399.
[13] Ramalingam, V. V., AyantanDandapath, and M. Karthik Raja. "Heart disease prediction using machine learning techniques: a survey." International Journal of Engineering & Technology 7, no. 2.8 (2018): 684-687.
[14] Kannan, R., and V. Vasanthi. "Machine learning algorithms with ROC curve for predicting and diagnosing the heart disease." In Soft Computing and Medical Bioinformatics, pp. 63-72. Springer, Singapore, 2019