Turkish Journal of Computer and Mathematics Education Vol.12 No.6 (2021), 5351-5359

Research Article

An Effectual Supervised Learning Based Automatic Classification Of Coronary Heart Disease

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

[a] Assistant Professor, [b] Professor
[a, b] Subharti Institute of Technology & Engineering, Meerut

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

Abstract: Most of the deaths are in the world are due to heart diseases which should be controlledefficiently. Among all heart diseases, coronary artery disease is very common and dangerous worldwide. These diseases are not easily identifiable and need extra care and precision in the health monitoring of the patient. In this paper, an effectual classification and optimization process is developed which puts light on the machine learning approach to identify coronary artery disease. The simulation is taken place in the MATLAB environment on which the 303 patient's data is analyzed in terms of the characteristics which are helpful in the diagnosis of coronary heart disease. In the proposed approach the feature engineering is used in which the feature extraction and instance selection are evaluated and the system training is performed to achieve high accuracy and low error rates. Also, the over-fitting problem is resolved using the proposed approach which helps to achieve high sensitivity, specificity, and recognition rate with low false acceptance and rejection rates.

Keywords: Supervised Learning, CAD, and Sensitivity, True negative and positive rates

1. Introduction

Heart disease comprises the unnecessary growth of cancer in the heart or bloodvessels. As per W.H.O, heart infection is one of the importantreasons for world-wide death. Rendering to a fresh study, aprojected 18.7 million individualsdied as of CADs in 2018, demonstrating 31% of all deaths worldwide, and Ifexisting trends are permitted to stay, 24.6 million individualscan die from heart illness up to 2030 [1]. The core risk issues of Heart disease comprisediabetes, smoking, heaviness, high cholesterol, and many more. Robust automated heartinfection prediction arrangements can be favorable in the healthcare area for heartdisease estimation. This process automation will also diminish the number of assessments to beengaged by a patient. Therefore, it will save time and costboth forDoctors and patients.The identification of Heart disease in supreme cases depends on a difficultgrouping of clinical and uncontrolled data. Because of this difficulty,there happens a significant volume of interest between clinical specialists andacademicsabout the efficient and accurate estimation of Heart infection [2][3].

As per the statistical records from W.H.O, one-third of inhabitants worldwideexpired from Heart syndrome; Heart infection is established to be the important cause of deathworldwide.Computational ecology is often useful through the development of transformingbiological information into clinical training, as well as in the accepting ofbiological processes from the clinical records[4][5].

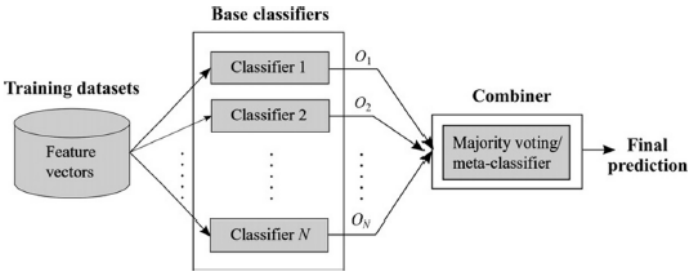

Fig 1: Schematic of Classification Process [16]

This development involves theimprovement of ananalytical model and the combination of distinctive varieties ofdata and understanding for problem-solving purposes. Moreover, this process involvesthe plan and arrangement of distinct procedures from statistical studyand data mining.Heart infection prediction arrangement can assist healthexperts in forecastingthe state of Sentiment, based on the scientific data of patients delivered into thesystem [13]. Surgeons may occasionally fail to take precise decisions while analyzingthe Heart infection of a patient, thus, Heart disease estimation systems, which recycled machine learning processes assist in such belongings to get accurate outcomes.Early recognition and correct judgment of heart syndrome are needed usingproper counseling and medications.Machine learning methodscanproduce a knowledge-richatmosphere which can benefit to suggestively help in Medical conclusions [14].Various managed machine learning and deep

learningmethodologies like SVM, CNN, and RNN can be recycled for estimation ofheart infection, and all need to be explored in terms of performance of heartdisease estimation [14].

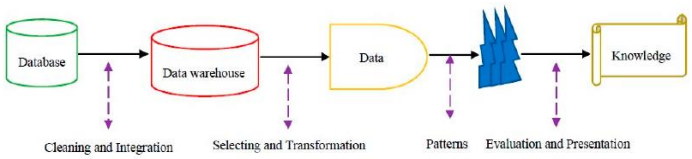

Fig 2: Knowledge Discovery Process [15]

The other section includes a literature survey (II) in brief which is discussed. The problem statement (III) covers the issues which still arise in the CAD. The proposed work (IV) shows the proposed working of the model and the result and discussion (V) covers the simulation implemented using the proposed work and the various tests implemented on the same. The last section (VI) shows the conclusion and future scope of the proposed system.

2. Related Work

Infection of heart diseaseis a seriousissuethatneeds to be controlledto reduce the deaths of the patients. This section shows good research works which show efficient advancements of the diagnosis of the patient in the medical domain. **Ilayaraja M, Meyyappan T [6**] presented the estimation model to acquire the risk of heart disease in the patient. They have controlledsomearrangementsgrounded on the support rate. They have executed their investigation in JAVA and accomplished better precision. **KaanUyara et al. [7]** projected an effectual genetic algorithm that is based on fuzzy logic for the analysis of the heartinfection where they have accomplished 97.78% precision. **Rashmi G Saboji et al. [8]** worked on the diagnosis development of heart disease utilizing Random Forest arrangement and they have accomplished approx. 98% recognition rate. **Mollet, Nico et al. [9]** worked on the HOG methodswhich are efficient for the heart disease diagnosis,and also they have compared their research with the traditional Histogram Gradient method and shows improvement. They have performed several tests to evaluate the performance of the system. **Kumar, Gandhi, et al. [10]**have used computer-aided arrangements for the heart disease recognition process using IoT processes. They have worked on various IoT proceduresto evaluate their system on real-time data. Their exploration has controlledvarious significant constraints such as irregularity, regularizations, and feature steps. The extracted feature boundaries are used to sort the information. **Safdar, Saima, SaadZafar et al. [11]**presented an arrangement to acquirea linear relationship of pattern for the heart infection through the classifiers. They composeddata and achieve data mining practices for the management of the information. They have achieved a feature extraction method and classify the categories of the diseases and perform estimation using the binarization process. **Manogaran, R. Varatharajan et al. [12]** presented thestate-of-the-art process for heart infection prediction and evaluate the performance in noisy environments. Also, their pre-processing segmentsare evaluated using different filters to reduce the noise levels. They have also worked on the neuro-fuzzy inference system on the real-time data through which they have evaluated the performance of the test cases on the trained model.

3. Problem Statement

Regression and identifications of the labels are certain processes in the classification of CAD (coronary heart diseases). So it deals with animportantpart in the disease detections in medical diagnosis. The medical diagnoses with physical machines aredifficult and less accurate tasks, and as the input increases; the evaluation in terms of the performance becomes a problematic situation that needs high accuracies and high precision of the information for the accurate diagnosis of the patient health. In many studies, the various threshold methods are used but they are less efficient in the classification scenarios. The main objective of this research is to increase the detection accuracy of the automatic classification model for the CAD and also increases the false measure by making use of the high precision and recall of the training model.

4. Proposed Work

The proposed work comprised of the normalization of the data, feature extraction and optimization, training of the model, Classification of the disease, and eventually the performance analysis of the system. The performance is evaluated using F-measure which deals with the high precision and recall of the system, sensitivity and specificity which deals with the true positive and negative rates and accuracy comprise of the low classification error rates on the unknown test data. Below the flow diagram is the proposed model which is explained in the parts.

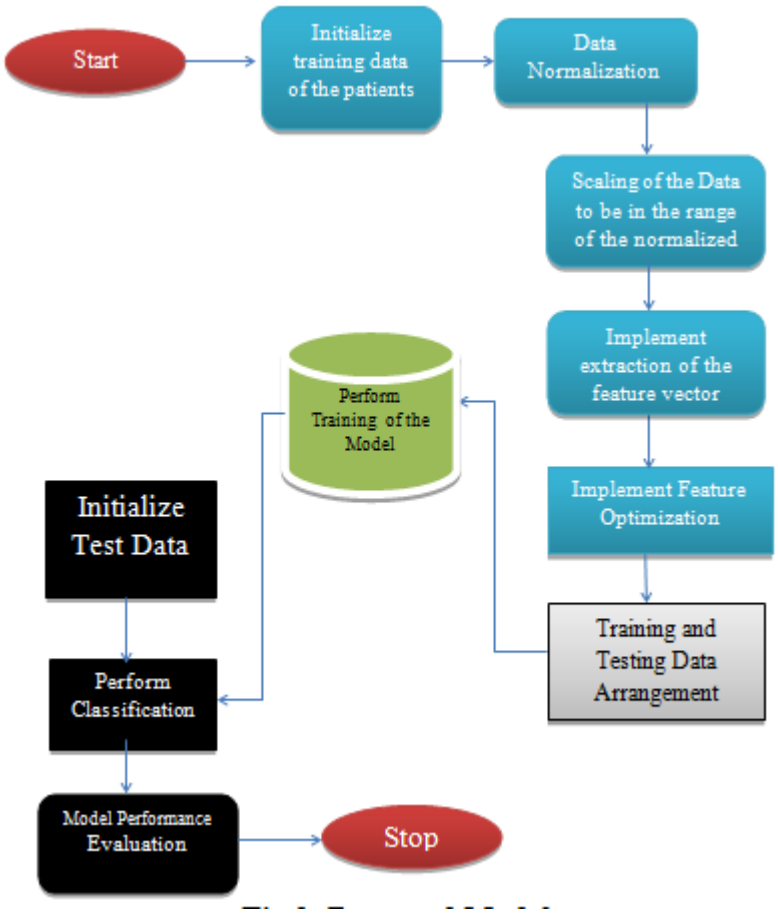

Fig 3: Proposed Model

Below are the proposed model evaluation steps for the classification of coronary artery diseases.

Input Dataset

The simulation works with the Z-AlizadehSani CAD datawhich is used as input for the model evaluations. This dataset covers 303 total patients and 56 characteristics of each patient. The dataset consists of various demographic characteristics which consider the various habits of the patients in arising the cancer of the coronary heart which is further divided into training and test samplesfor the classifications [19].

Normalization & Extraction of the features

This accomplishes the extraction of the features using PCA which makes use of the covariance evaluation among the data points of each character to achieve Eigenvectors. The feature engineering is recycled in the developed approach to recognize the formations using relationships of the data points. The feature engineering process is having the problem of time complexities in terms of execution which is overcome by the use of linear kernel PCA through which the non-linearity and the variance will be reduced the high correlations will be achieved among the data points.

Optimization of the features

This segmentuses the hybrid optimization approachusing two efficient instance selection procedures toaccomplish the instance selection method which is the important part of the execution to select the appropriatestructuresto form the meaningful feature vector.

- **Modified Firefly instance selection**-This is motivatedusing fireflies' blinking nature. Various processes are fulfilled for these systems. These are very helpful in solving nonlinear, problems. The leading steps covered for the implementation of the instance selection process is given below:

1) There are fascinating particles and are in contact with each other.

2) Main attraction is the glowing nature of the particles.

3) There will be random measures in terms of the movement in case of the same intensity.

4) Random walk process will be there for the selection of the best procedures.

The main concentration is given to solve the nonlinear problem of the selection of the instances. The positive point of selecting this algorithm because of the division of the whole particles into smaller groups and is having the ability to capture multi-objective problems having high randomness in the solutions.

- **Modified Particle Swarm Instance Optimization** - It usesa meta-heuristic process that is capable of solving complex procedures. This technique is stimulated by the sharednature of the chirping birds. The searching scenario takes place using the operations performed on the population. Each occurrence is a participating particle in the swarm. Each occurrence in the populationis having a calculated movement rate which is considered as an optimal task to control the nature of the flocking birds. It is wellhands-on to the heart disease arrangement and showing the resourcefuleffect. In meta-heuristic methods, the operations are performed to minimize the objective function to get the optimized instances or the relevant feature vector. The output of the firefly instances is the input as a population of the PSO to get the optimized instance selection which makes them hybrid optimization algorithms.

5. K-fold cross-validation based Discriminant Analysis

In the proposed approach statistical analysis for the classification of the CAD is performed through which the training of the model is done for achieving high true positive rates. It is based on the grouping of the extracted instances which generates the linear model and then the cross-validations are performed to reduce the overfitting of the training model for the proper classification labels which generates the high true positive and negative rates for the classification of the CAD. The proposed approach uses the discriminant analysis having independent variables and the definite dependent variable in terms of the categorical labels and is considered as prediction labels for the classifications.

5. Proposed Algorithm

Step 1: Input Samples in such a way that $\mathrm{tn\{s\}} = \mathrm{tn\{s1\}}, \mathrm{tn\{s2\}} \ldots \mathrm{tn\{sn\}}$as inputand fulfil the mounting of the records to process data.

Step 2: Perform standardization and scaling of the information

For xp=1 to $\mathrm{len(tn)}$

$\mathrm{STDN_D}$ = StdScaling { $\mathrm{tn(xp)}$} to diminish the inconsistenciesbetween data inputs.

EndFor

Where tnis the total input data for the training.

Step 3: Implement the characteristics extraction as feature vector &implement covariance of the data using Vectorization

For np=1 to $\mathrm{STDN_D}$

$\mathrm{C(p) = COV(STDN_D)}$

EndFor

Step 4: Implement the extractions of the featuresfor the conversion$\mathrm{Tp\{x\}} = \mathrm{X \times W}$ for the datato have new feature dimensional space VEC.

Where VEC(v) where VEC= $\{\mathrm{VEC_1}, \mathrm{VEC_2} \ldots \mathrm{VEC_N}\}$is the vector which is processes for the training.

Step 5: Perform selections of the instances using hybridoptimization measuresfor the collection of relevant featuresfor the feature reduction

While $\mathrm{(t(pn)} < MIter\mathrm{(gn)}$ i.e. max iterations

IF $(\{\mathrm{Inte\{xp\}} > Inte\{yp\}\})$

Separate A(x) i.e. attraction with D{i}. € Distance (X{current_particle})

Move $\mathrm{Fp(xn)}\ from\ xp \rightarrow yp$;

Evaluate S(x{n}) and modify Inte.{val}

Perform best conceivable solutions and rank occurrences.

End While

For each event as swarm input $\mathrm{SI\{p\}} = 1, \ldots, \mathrm{SI\{p\}}$ do

Frame the P{p} as place of the particle by distribution route: $\mathrm{xp(i)} \rightarrow \mathrm{Z(Lower\{b\}, Upper\{b\})}$

Arrange the SI{p} instance known place to its initial place: $\mathrm{s\{i\}} \leftarrow \mathrm{SI\{i\}}$

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

While a determined measure does not happen to do:

For each eventx{p} = 1, … , SI{p} do

For each dimension d{s} = 1, … , Nd{p} do

Check random records: Rand{p}, r{gl} ~ Z{0,1}

Perform velocity updations of the each occurrence in the group:

$$SVect(i, ds) \leftarrow \omega\,(SVect\{i\}, ds) + \varphi\{p\}\, rnd\{p\}\,(SI(i, ds) - SI(i, ds)) + \varphi(\, gl\{b\})\, r(\, gl\{b\})\,(\, gl\{b\} - x(i, ds))$$

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

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

Modify the occurrence best recognizedposition: pi ← xpi

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

Adjust the occurrencebest location pos: gb ← pi

End if

End if

End For

End For

Step 6: Make data for training and testing and divide it in the ratio in 70 (training) -30 (testing)
N_D= {Txs (N_D)}
Step 7:Implement a classification process using discriminant analysis.
LDA{m} = {FitTransform[TR(N_D)]}
Where FitTransform produces the arrangement of the training prototype and TR(x) is the total training input
Step 8: Upload Test records such that Tst_S = {Tst_{S1}, Tst_{S2}, Tst_{S3}, Tst_{S4}… Tst_{SN}}
Step 9: Load model training and perform classification on Tst.
Step 10: Estimatethe Performance of the classified model and Repeat Steps 5 to 9 until all arrangements get accomplished.

6. Result & Discussions

This section plays a significant role in the proposed model in which the whole simulation scenario is discussed and the performance is evaluated. The graphical user interface is used for the man-machine interaction which can help doctors to interact with the user interface and check the outputs by uploading the dataset of the patients directly. The simulated results are discussed below in detail.

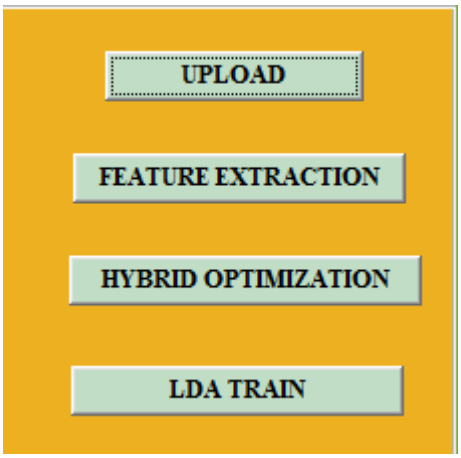

Fig 4: Training Panel

Fig 4 shows the training panel of the proposed work consists of the user interface in the MATLAB environment. The panel comprises the various user interface tools such as static texts, pushbuttons and helps users directly to check the specific operation of the button for the user from starting to the end operations for the classifications.

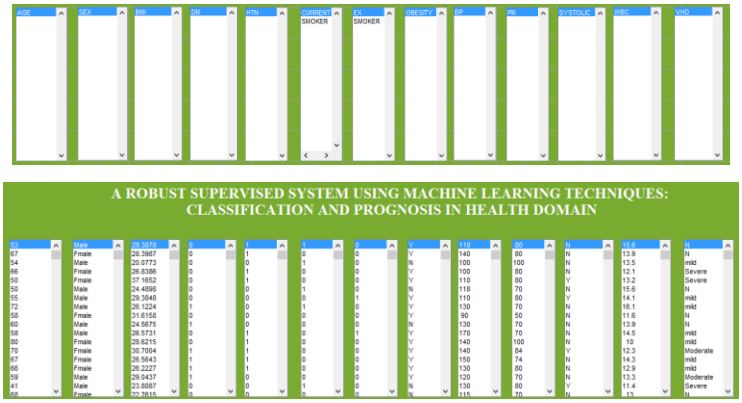

Fig 5: Input Panel

Fig 5 shows the input panel and shows the uploaded data which is displayed in the list boxes. The uploaded data is taken from excel which shows each health property on which the computations are performed such as the smoking nature, blood pressure, HB level, etc. So in the proposed approach,a total of 56 attributes are taken to have good insights into the data.

Fig 6: Extracted Features

Fig 6 shows the statistical analysis which is extracted from the feature vector. The statistical analysis shows the nature of the data in terms of the evaluations through which the data can be controlled and the proper information can be gathered. The entropy shows the rate of disorder. If the entropy is high then it can disturb the classifications and the model can't learn properly which can give rise to the overfitting and underfitting of the model. The other parameters like variance and STD (standard deviation) show the spreading and dispersion of the data points. Ifthe variance is high then there will be fewercorrelations among the data points which should not be there for the high performance of the training model in the supervised learning process. The center of the grouping of the data points is also necessary through which the variance and standard deviations can be analyzed which is done through the mean parameter. It gives us important statistics about the total population regarding the experimental values.

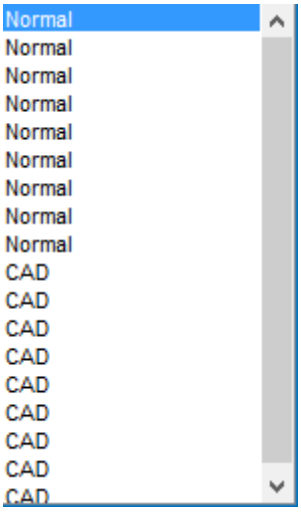

Fig 6: CAD Classification

Fig 5 shows the classification outcomes which are done through discriminant analysis. The classified result shows the total count of the normal and abnormal diseases. The classification is performed on the 30% untrained data. This step is performed in the testing phase where the training model is loaded and then the same process if the feature engineering is applied to the testing data and then the classifications are made on the unknown series of the data to check the high similarity measures of the trained model to validate the performance of the system classified labels.

Fig 7: Performance Analysis

Fig 7 shows the performance of the proposed work which is evaluated in terms of the sensitivity, specificity, F-measure, and accuracy rate. It can be seen from the figure that the proposed approach is achieving high precision and recall through which the F-measure performance can be evaluated, and also the high true positive and negative rates through which high sensitivity and specificity are achieved. The proposed model is also achieving good classification accuracy which means that the false acceptance rates are low for low false negatives and false-positive rates in terms of retrieving the information.

The below equations are used for evaluating the proposed performance.

$$\mathrm{PR(x)} = \mathrm{tposr(x)} \div (\mathrm{tposr(x)} + \mathrm{fposr(x)})$$

$$\mathrm{RE(x)} = \mathrm{tposr(x)} \div (\mathrm{tposr(x)} + \mathrm{fnegr(x)})$$

$$\mathrm{SPE(x)} = \mathrm{fnegr(x)} \div (\mathrm{tnegr(x)} + \mathrm{fposr(x)})$$

$$\mathrm{SEN(x)} = \mathrm{tposr(x)} \div (\mathrm{tposr(x)} + \mathrm{fnegr(x)})$$

$$\mathrm{ACR(x)} = (\mathrm{tposr(x)} + \mathrm{fposr(x)}) \div (\mathrm{tposr(x)} + \mathrm{fnegr(x)} + \mathrm{fposr(x)} + \mathrm{tnegr(x)})$$

$$\mathrm{Fmr(x)} = 2 \times \left(\frac{[\mathrm{Pre(x)} \times \mathrm{Re(x)}]}{[\mathrm{Pre(x)} + \mathrm{Re(x)}]}\right)$$

Where Pre(x) and Re(x) are the evaluated precision and recall of the proposed model. $\mathrm{SPE(x)}$ And $\mathrm{SEN(x)}$ is the sensitivity and specificity of the proposed model. $\mathrm{ACR(x)}$, and $\mathrm{Fmr(x)}$ is the evaluated accuracy and F-measure of the proposed model.

Table 1: Accuracy Performance

Test No.	**Accuracy (%)**
1	97.225
2	96.114
3	97.682
4	97.700
5	97.892
6	95.116
7	97.335
8	96.488
9	97.103
10	98.775

Table 1 shows the accuracy comparisons on different tests that were conducted on various cross-validations. The different test samples are uploaded which are not trained and the performance is evaluated on each test sample and it can be seen from Table 1 that the proposed approach is achieving high performance in terms of true detections as the accuracy of the model.

Table 2: Sensitivity Performance

Test No.	**Sensitivity**
1	0.954
2	0.966
3	0.987
4	0.974
5	0.979
6	0.984
7	0.987
8	0.988
9	0.975
10	0.977

Table 2 shows the sensitivity comparisons on different tests. It can be seen that the proposed approach is achieving high sensitivity i.e. the correct positives which must be high for the high classifications and correctly identified as infected. This statistical measure must be high for low false-positive rates.

Table 3: Specificity Performance

Test No.	**Specificity**
1	0.962
2	0.976
3	0.965
4	0.966
5	0.953
6	0.958
7	0.969
8	0.958
9	0.974
10	0.959

Table 3 shows the comparisons on different tests in terms of specificity.The specificity shows that the proposed model is correctly identified true negative proportions means the individuals which are not having CAD which is a desirable outcome in actuality also. If specificity is low then the accuracy will also reduce which results in high false detections.

Table 4: F-Measure Performance

Test No.	**F-Measure**
1	0.962
2	0.974
3	0.979
4	0.978
5	0.985
6	0.983
7	0.960
8	0.964
9	0.956
10	0.986

Table 4 shows the F-measure comparisons on different tests. The F-measureis measured using precision and recall of the test evaluations. It can be seen that the proposed model is achieving a high F-measure. So if the precision and recall of the model are high then the F-measure will also be high which a desirable outcome is. If the precision is low then the test classifications will have high deviations and if the recall is low then the information retrieval of the model is improper which can reduce the robustness of the model to perform accurate classifications.

7. Conclusion & Future Scope

Health care data plays a very significant role in the health monitoring systems. It should be managed properly from time to time. It will help the hospital staff to properly diagnose the patients.Still, the precisions and correctness of the data are small. So an effective precise prototype is essential which can offer full and improvedunderstandingsof the data to gather meaningful statistics and features about the health of the individual. This paper presented the robust modeling of the machineefficient feature engineering process which gives advancement in the estimation of the developed machine learning model using LDA classifications. It can be seen from the above comparisons and result evaluations that the proposed approach is achieving the appropriate outcome for the automatic classification of the true samples for the disease detections with low false acceptance and rejection rates..

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, AhmetİlhanI. “ 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.
15. HassannatajJoloudari, Javad, EdrisHassannatajJoloudari, Hamid Saadatfar, Mohammad GhasemiGol, Seyyed Mohammad Razavi, Amir Mosavi, NarjesNabipour, ShahaboddinShamshirband, and Laszlo Nadai. "Coronary Artery Disease Diagnosis; Ranking the Significant Features Using Random Trees Model." arXiv e-prints (2020): arXiv-2001.
16. Raza, Khalid. "Improving the prediction accuracy of heart disease with ensemble learning and majority voting rule." In U-Healthcare Monitoring Systems, pp. 179-196. Academic Press, 2019..