Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2199 – 2210
Received 05 March 2021; Accepted 01 April 2021.

Machine Learning Based Hybrid Model for Heart Disease Prediction

Jasjit Singh Samagh[1], Dilbag Singh[2]

[1](Research Scholar, Department of Computer Science and Engineering, Chaudhary Devi Lal University, Sirsa, Haryana, India, jasjitsamagh@gmail.com)

[2](Professor, Department of Computer Science and Engineering, Chaudhary Devi Lal University, Sirsa, Haryana, India, dscdlu7@gmail.com)

ABSTRACT

Coronary illness is one of the significant reason of death in the world today. An early detection and accurate prediction of the heart disease is needed so as to reduce the mortality rate. Prediction of Heart disease is very challenging, it is even difficult task for medical practitioners as it demands expertise and higher knowledge. Recent development in medical supportive technologies based on machine learning plays an important role in predicting cardiovascular diseases. Machine learning uses algorithms to analyse data, learn from that data and make well-informed learning based decisions. The present paper proposed a new hybrid model based on feature selection, feature optimization and ensemble technique. This exclusive combination will build an enhanced model that will have leverage over the existing models in predicting the heart disease more quickly and accurately and thus will assist the medical practitioners in taking the measures to control the mishap.

KeywordsHeart disease, Hybrid, Machine learning, Model, Prediction.

1. INTRODUCTION

Coronary disease remains the leading cause of death globally. Ischemic heart disease and strokes were estimated to account for approximately 15% to 20% of all deaths. For diagnosis of these diseases, investigations such as ECG, chest radiography, echo cardiograph are generally performed at the bed side, however more complex procedures such as cardiac catheterization, nuclear scanning, CT and MRI may also be performed. The data so get collected due to such investigations require lot of time for analysis and administering of medicines gets delayed which sometimes may have adverse effect on the patient. Machine learning may help doctors, pathologists to reduce the timing of such tests and results will be more accurate because amount of data is increased. With machine learning it is very easy to get knowledge from a large data which is not possible for man to analyse. The objective of this research work is not to replace the specialist physician, but to assist the doctor in obtaining an alternative opinion and its various feasibility in critical situations. [14]

1.1 Feature Selection

Feature selection is arguably the most critical pre-processing activity in any machine learning project. It intends to select a subset of system attributes or

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2199 – 2210
Received 05 March 2021; Accepted 01 April 2021.

features which makes a most meaningful contribution in a machine learning activity. [11]

Three approaches used are filter, wrapper and embedded [13]

1.1.1 Filter Approach

In this approach, feature subset is selected based on the statistical measures made to evaluate the qualities of the features from the data viewpoints. To evaluate the goodness of the feature selected, no learning algorithm is engaged. Some of the common statistical tests conducted on features as a part of filter approach are Pearson's Correlation, Chi Square etc. [11]

1.1.2 Wrapper Approach

In this approach, the best feature subset are identified using the induction algorithm as a black box. The feature selection algorithm searches for a good feature subset using the induction algorithm itself as a part of the evaluation function. Since for every candidate subset, the learning model is trained and the result is evaluated by running the learning algorithm, wrapper approach is computationally very expensive. However the performance is generally superior compared to filter approach. [13]

1.1.3 Embedded Approach

The approach is very similar to that of wrapper approach. This approach also uses the inductive algorithm in evaluating the general feature subset. The only difference it have as compare to that of wrapper approaches, it performs the feature selection and classification simultaneously. [13]

1.2 Optimization Algorithms

Optimization algorithms are used to find the best solution of the problem by minimization or maximization the objective function without violating resource constraints. Swarm optimization technique is new approach to problem solving that takes inspiration from social behaviours of insects and of other animals. These are used to solve the difficult problems that is computationally hard problem. [8]

1.2.1 Particle Swarm Optimization (PSO)

Swarm intelligence is a new approach to problem solving that takes inspiration from the social behaviours of insects and of other animals. It is disseminated answer for complex issues. Their aim is to interact with agent and their environment to solve the complicated problem. [9] In 1995, Russel Eberhart and James Kennedy proposed Particle Swarm Optimization. [10] This strategy depends on collaborating of particles between them, every particle moves and at every cycle, the one nearest to the ideal imparts its situation to the others so they can alter their position. Their thought was that a gathering of unintelligent people may have a complex worldwide association. [3]

Because of its ongoing nature, a lot of study is being done on P.S.O., yet the best so far is the extension to the framework of combinatorial optimization. [9]

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2199 – 2210
Received 05 March 2021; Accepted 01 April 2021.

In particle swarm optimization every individual of the population is called particle. In standard PSO, after the population has been initialized, at every iteration, particles update their velocity and position on the basis on their own experience that is their personal best (pbest) and with the best experience of all the surrounding particles that is global best (gbest). Using Equation. (1 and 2) given below, the performance of all particles will be evaluated by predefined cost functions at the end of each iterations. [9]

$v^{\mathrm{I}}[\mathrm{t}+1] = \mathrm{w}.\ v^{\mathrm{I}}[t] + c1r1\ (\mathrm{p}^{\mathrm{i,\ best}}[t] - p^{\mathrm{I}}[t] + c2r2\ (p^{\mathrm{g,best}}[t] - p^{\mathrm{I}}[t]$ (eq) ------- (1)

$p^{\mathrm{I}}[\mathrm{t}+1] = p^{\mathrm{I}}[t] + v^{\mathrm{I}}[t+1]$ (eq) ------- (2)

Where, $i = 1, 2\ldots N$, N represents the number of swarm population. $v^{\mathrm{I}}[t]$ represents the velocity vector in $[t]h$ iteration. $p^{\mathrm{I}}[t]$ is the current position of the ith particle. $p^{\mathrm{i,\ best}}[t]$ represents the previous best position of ith particle and $p^{\mathrm{g,best}}[t]$ represents the previous best position of a whole particle, ”w” has been used to control the pressure of local and global search. $c1$ and $c2$ represents the positive acceleration coefficients which is respectively the cognitive and social parameters lastly $r1$and $r2$ represents the random number between 0 and 1. [9]

1.2.2 Genetic Algorithm (GA)

Genetic Algorithm is an evolutionary computational method which is popular nowadays which was developed in early 1975 by Holland and later it was improvised by Goldberg. It is a search technique that solves a given problem by representing the natural process of evolution. Genetic Algorithm is an algorithm which is based on Darwin’s theory that utilizes the concept of survival of the fittest. Genetic Algorithm exploits new and higher optimal solutions with none presumption like continuity. Genetic Algorithm as a process has massive potential and because of this it has been used in many fields from gaming to financial analysis. Because of the fact that Genetic Algorithm can handle a large number of parameters it has become in high demand in many sectors and it comes with a solution which is optimal or close to the optimal which otherwise take lifespan to solve the problem.[15][16][17][18]

Genetic Algorithms consists of a set of solutions, chromosomes or individuals which are strings of binary values, “0”s and “1”s. Each value (“0” or “1”) defines the state of features in the chromosome. These set of chromosomes are referred as a population. Respective chromosome is then evaluated using a fitness function. After levelling the chromosomes according to their fitness values, they then undergo genetic operations such as crossover and mutation. For this, two chromosomes are selected on the basis of their positions on a roulette wheel which is prejudiced according to each chromosome’s fitness. The two chromosomes then passes through crossover and then mutation is applied to increase the local coverage of search space by the chromosomes, thereby decreasing the chances of being stuck at a local optimum. If this evolution process generates stronger descendant’s chromosomes than the previous ones, the algorithm replaces them. Hence the process repeats until it meets the end criteria.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2199 – 2210
Received 05 March 2021; Accepted 01 April 2021.

1.3 Ensemble Technique

Ensemble means collaboration. Ensemble methods are models that contain many weaker models that are autonomously trained and whose forecasts are combined approximately to create the overall prediction model. In this model, N different learning methods are applied on a single training data so that N different models can be achieved. On obtaining the different models, to make the final decision the system then combines their outputs. Thus the weighted average of all the individual model outputs is the resultant output. [6]

Ensemble Techniques are listed below:

- Bagging (Bootstrap Aggregation)
- Random Forest
- Boosting (AdaBoost)

1.3.1 Bootstrap Aggregation

Bootstrap Aggregation is also known as bagging. To generate multiple training datasets bootstrap sampling method is used. These training datasets are used to generate a set of models using the same learning algorithm. Then the outcomes of the models are combined by the majority voting that is classification or by regression that is average.Bagging is very simple ensemble technique which can perform really well for unstable learners like a decision tree, in which a slight change in data can impact the outcome of a model significantly. [13]

1.3.2 Random Forest

Random Forest is an ensemble classifier that is a combining classifier that uses and combines many decision tree classifiers. Ensembling is usually done using the concept of bagging with different features sets. The reason for using large number of trees in random forest is to train the trees enough such that contribution from each feature comes in a number of models. [13]

The random forest algorithm works as follows

- If there are N variables or features in the input dataset, select the subset of “m” (m<N) features at random out of N features. Also data instances should be randomly picked.
- Use the best split principle on the “m” features to calculate the number of nodes “d”.
- Keep splitting the nodes to the child nodes till the tree is grown to the maximum possible extent.
- Select a different subset of the training data ’with replacement’ to train another decision tree following steps (1) to (3). Repeat this to build and train “n” decision trees.
- Final class assignment is done on the basis of the majority votes from the “n” trees.

1.3.3 Boosting

Similar to that of bagging, boosting is another key ensemble based technique. Boosting is an iterative technique. It decreases the biasing error. If an observation was classified incorrectly, then it boosts the weight of that observation. In this type

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2199 – 2210
Received 05 March 2021; Accepted 01 April 2021.

of ensemble, weaker learning models are trained on resampled data and the outcomes are combined using a weighted voting approach based on the performance of different models. Adaptive boosting or AdaBoost is a special variant of a boosting algorithm. It is based on the idea of generating weak learners and learning slowly. [13]

2. RELATED WORK

Fatma Zahra Abdeldjouad et.al. (2020) proposed a new hybrid approach using different machine learning techniques to predict the heart disease. Classification algorithms like Logistic Regression, Adaptive Boosting and Multi Objective Evolutionary Fuzzy Classifier has been analysed using WEKA whereas Fuzzy Unordered Rule Induction, Genetic Fuzzy System-LogitBoost and Fuzzy Hybrid Genetic Based Machine Learning has been analysed with KEEL. Based on their performance best model is selected and hence compared with patient data to predict the heart disease. [1]

Youness Khourdifi et.al. (2019) compared the algorithms with different performance measures with the help of machine learning. For prediction algorithms worked differently.In this study the best results are shown by K-Nearest Neighbour, and Random Forest and Artificial Neural Network. Then he merged the algorithms and try to check the performance so as to check if it will be more efficient or not and later result is applied to heart disease data set and his proposed models resulted in better accuracy of 99.65%. [2]

Monther Tarawneh et.al. (2019) proposed the heart disease prediction system called “hybridization” in which all the techniques are combined into a single algorithm. This combined model of all the techniques confirmed and resulted in accurate diagnosis. The new algorithm can be used as expert system in hospitals to help doctors in diagnose heart disease quickly and save life. Also, can used for education purpose in medical schools. [7]

SenthilKumar Mohan. et.al. (2019) proposed a HRFLM approach which is the hybrid approach and made from combination of Random Forest and Linear Method. Stated that, for the long term in saving the human lives and for early recognition of symptoms in heart conditions, it is very helpful to identify the processing of raw heart healthcare data. Machine learning techniques plays an important role in processing the raw data and providing new and novel judgements towards heart disease. Thus HRFLM resulted in the prediction of heart disease quite accurately. [4]

Shadman Nashif et.al. (2018) developed system for transmission of the recorded data to a central server which are updated every 10 seconds with the help of which the doctors can see the patient’s real-time sensor data by using the application and start live video streaming if instant medication is required. Other important feature of the proposed system was that if any real-time parameter of the patient exceeds the threshold then the prescribed doctor will be notified at once through GSM technology. [5]

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2199 – 2210
Received 05 March 2021; Accepted 01 April 2021.

3.RESEARCH METHODOLOGY

A research method indicates the process in which the research is carried out. It specifies the research methodology used in the study. The Proposed research is based on exploratory and applied research design. It is exploratory because it consists of discovery of new ideas and possible insight in identifying further study. As the proposed work is for heart disease prediction, it is with the help of applied research we can easily use it in solving the problem.

4. PROPOSED MODEL

The Figure below represents the Proposed Machine Learning model for prediction of heart diseases. The proposed Model will not only assist the medical practitioners in effective predication of heart disease threat but will also help them in obtaining an alternative opinion and its various feasibility in critical situations.The below model is explained as follows:

- **Heart Disease Training Dataset:** Training data is the data to be used in machine learning model to predict the result that model is designed to predict. In simple words it is given as input to the model. In research the training data is to be collected from the UCI machine leaning repository.

- **Heart Disease Testing Dataset:**A testing dataset is a dataset that is not similar to that of the training dataset but it follows the same prospect as that of the training dataset.In simple words it is that Output on which the Input that is training dataset will get evaluated and obtain the result. In Proposed research work testing data is to be collected from the UCI machine leaning repository.

- **Preprocessing:** After loading the dataset it will passed through preprocessing stage. In this stage data cleaning will be done that is filtering of data set.

- **Feature selection:** It is a pretreatment step which reduces sizing, eliminates unresolved data, increases learning accuracy and improves understanding of results. The present research is performed with “Wrapper” technique to select features. In the proposed model the “wrapper” approach is being used in selecting the features.

- **Feature Optimization:**The proposed model is using the combination of Genetic Algorithm (GA) and Swarm optimization techniques such as PSO. These are novel techniques in evaluating the heart disease and will result in getting the high valued attributes from the set of attributes Optimization algorithms are used to find the best solution of the problem by minimization or maximization the objective function without violating resource constraints. Optimization algorithms can be applied to improve the performance of classifiers by tuning parameters so as to enhance the performance in heart disease prediction.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2199 – 2210
Received 05 March 2021; Accepted 01 April 2021.

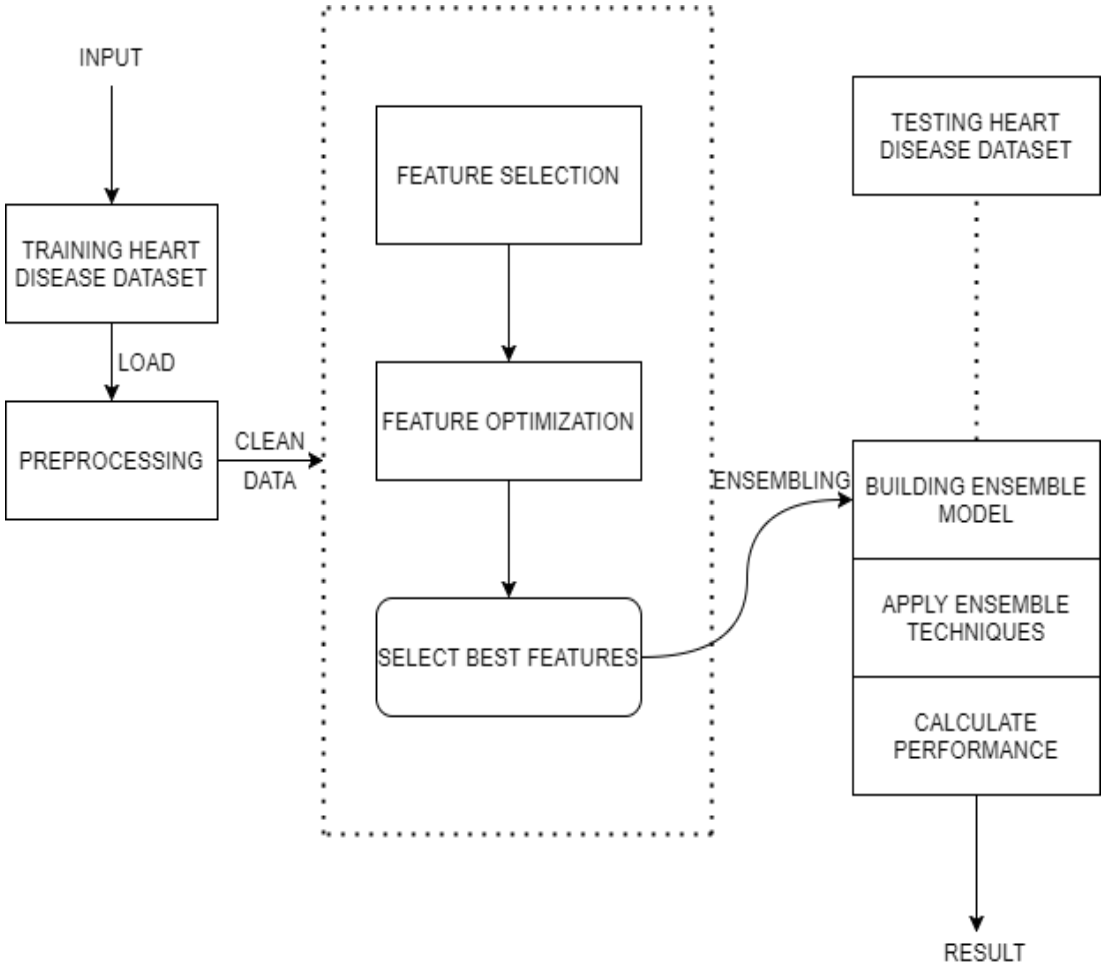

Figure1.Proposed Machine Learning Model for Heart Disease Prediction

- **Ensemble method:** When the result obtained after feature optimization that is "best features" obtained are applied and tested on the ensemble techniques. This technique will generate the best possible result as an ensemble method, it combines the weaker learners to create stronger ones which help in averaging out biases of different underlying techniques and also reducing the inconsistency in the prediction of heart disease. The result generated by ensemble techniques will be best and highly accurate

- **Result:** It is the last step of the proposed model which evaluates the performance and deliver the best output of the proposed model. Resulting model will assist the medical practitioners in early predicting the threats so as to take the measures to control the mishap.

5. FEATURE SELECTION AND FEATURE OPTIMIZATION

In heart disease dataset number of features can reach up to high extent, some of which are irrelevant and redundant. These irrelevant and redundant attributes

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2199 – 2210
Received 05 March 2021; Accepted 01 April 2021.

usually give incorrect results, high cost and consume more time. Lesser the attributes better would be the performance in predicting the heart disease. Feature selection and feature optimization plays an important role in reducing the features and keeping only the predominant features.

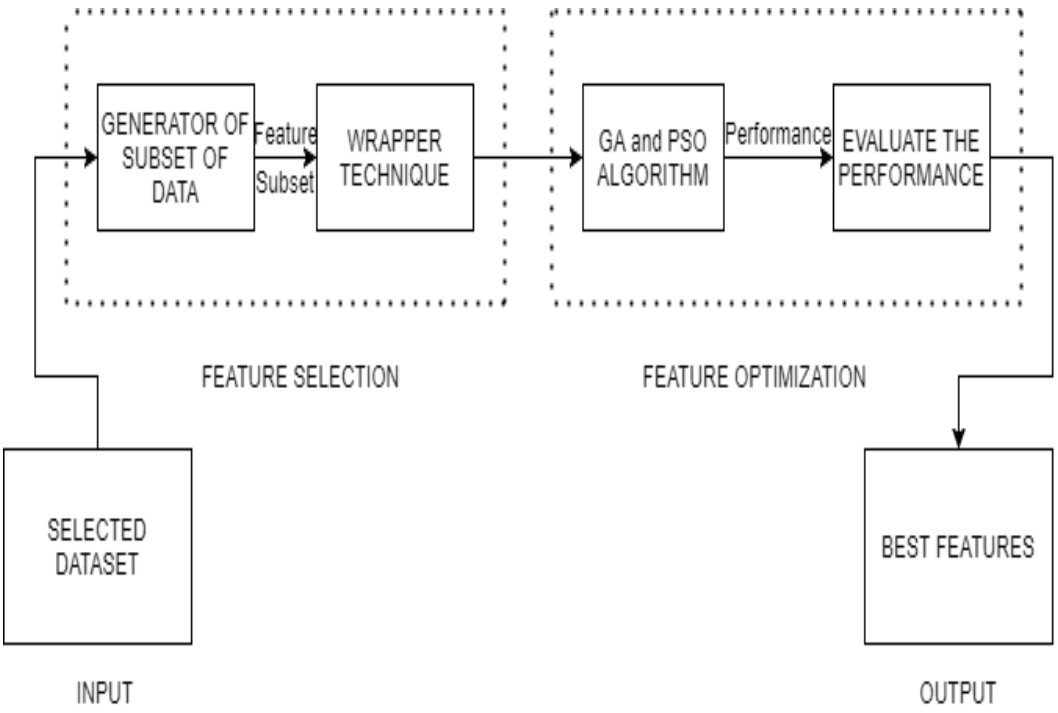

Figure 2. Mechanism of Feature Selection and Feature Optimization

The figure 2 represents the mechanism of feature selection and feature optimization. This mechanism starts with input of selected dataset and end with predominant dataset as an output. The dataset after filtering is sent for feature selection. This process can easily be understood by the figure shown. The dataset when passed through feature selection, it generates subset of features. These features subset, then applied to a feature selection technique. Irrelevant and redundant features are removed using this technique. Afterwards dataset passes through feature optimization where features are optimized using algorithms such as GA and PSO. At every iteration algorithms will evaluate the dataset by tuning parameters and finally resulting in Predominant features as the output.

6. PROCESS OF EXECUTION

The steps in process of execution for performance evaluation of the proposed hybrid model has been represented below

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2199 – 2210
Received 05 March 2021; Accepted 01 April 2021.

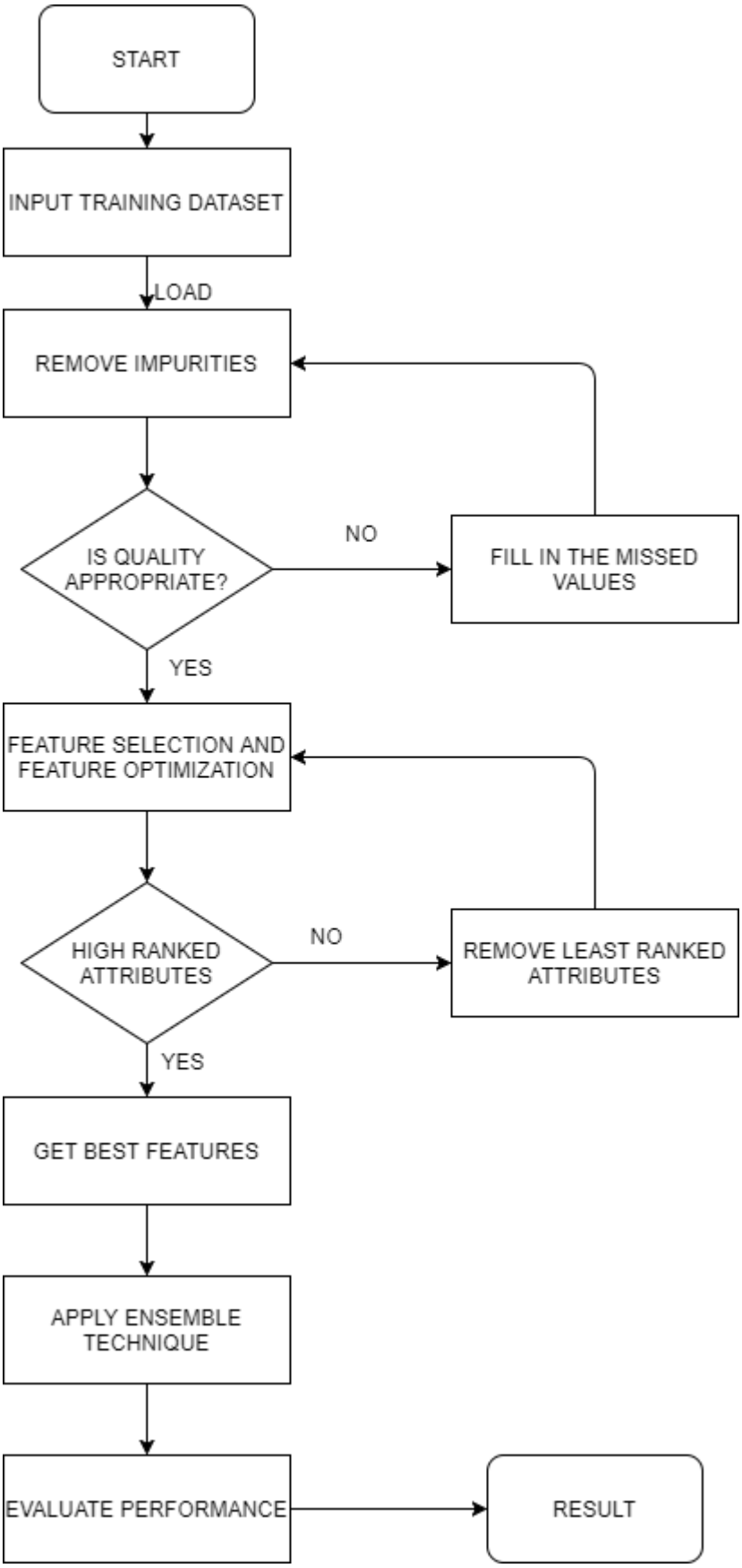

Figure 3. Process of execution of the Proposed Model

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2199 – 2210
Received 05 March 2021; Accepted 01 April 2021.

The Figure3 represents the Process of Execution of the proposed model. It is as follows, first of all Heart disease dataset is given as input, then dataset will be loaded and impurities be removed. After removing the impurities quality check will be assured to check the quality of the dataset, if the quality is good, then process proceeds to the next step else resend the dataset to remove impurities. After getting the appropriate quality dataset it is forwarded to feature selection and feature optimization. This is the important step as this step carries the removal of irrelevant and the redundant features from the dataset and further optimizes the dataset. The output of this step is the predominant features. After getting the predominant feature attribute check be done, whether the result obtained is the high level attributes or not, if it is no then it will resend it to feature selection and optimization phase. If the result found are high level attributes then, the resultants are the best features. After getting the best features dataset is tested on various ensemble techniques. This will evaluate the performance and generate the accuracy of the proposed model.

7. PROPOSED APPROACH

ALGORITHM

STEP1: Load the Heart Disease Dataset

STEP2: Remove the impurities and apply data cleaning by filling up the missing values

STEP3: Apply Feature Selection and Feature Optimization on the Heart Disease Dataset

Step4: Remove the irrelevant and redundant features and keep the Predominant features

Step5: Apply the Ensemble Technique on the Predominant Features

Step6: Calculate the performance of the proposed approach

The text box easily explains the functionality of model that is how the process executes from the beginning to the end. As shown in step1and2 dataset is loaded and performed data cleaning on it. In step3 feature selection and feature optimization is applied on dataset so as to remove irrelevant and redundant features and keep only the high level features which are termed as the best features which is shown in step4.In step5 the best features are then applied on the ensemble techniques in order to test and solve the problem. Finally the step6 will calculate the performance of the model.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2199 – 2210
Received 05 March 2021; Accepted 01 April 2021.

8. CONCLUSION

In this paper, a hybrid model for heart disease prediction using machine learning techniques has been proposed. An exploratory and applied research methods have been used to conduct present research. The proposed model is highly accurate and predicts the heart disease speedily. The proposed model is an integration of feature selection, feature optimization and ensemble techniques. Feature selection eliminates redundant data, increases learning accuracy and improves understanding of results whereas Feature Optimization is applied to improve the performance of classifiers by tuning parameters and resulted in delivering the high level attributes as a result. For this purpose, GA and PSO is used.GA resolves the given process by simulating the natural process of evaluation whereas PSO being a computational method optimizes the problem by iteratively trying to improve a candidate solution with regard to a given amount of features. Serial application of these optimizing techniques generates the predominant features. Ensemble technique, combination of two or more models, is used to test these features and it generates weighted average results of all the individual model outputs.This unique amalgamation resulted into building of an enhanced model leveraging over the existing models. The proposed model may assist the medical practitioners in early detection of heart disease to help in preventing mishaps.

REFERENCES

[1] Fatma Zahra Abdeldjouad, Menaouer Brahami, and Nada Matta, "A Hybrid Approach for Heart Disease Diagnosis and Prediction Using Machine Learning Techniques", ICOST 2020, LNCS 12157, pp. 299–306, **(2020)**.

[2] Youness Khourdifi and Mohamed Bahaji, "Heart Disease Prediction and Classification Using Machine Learning Algorithms Optimized by Particle Swarm Optimization and Ant Colony Optimization", International Journal of Intelligent Engineering and Systems, Vol. 12, no. 1, **(2019)**.

[3] Youness Khourdifi and Mohamed Bahaji, "The Hybrid Machine Learning Model Based on Random Forest Optimized by PSO and ACO for Predicting Heart", ICCWCS 2019, Kenitra, Morocco, **(2019)**,April 24-25.

[4] Senthilkumar Mohan, Chandrasegar Thirumalai, and Gautam Srivasatava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques", IEEE access vol. 7, **(2019)**.

[5] Shadman Nashif, Md. Rakib Raihan, Md. Rasedul Islam,and Mohammad Hasan Imam , ".Heart Disease Detection by Using Machine Learning Algorithms and a Real Time Cardiovascular Health Monitoring System", World Journal of Engineering and Technology, vol. 6, **(2018),** pp. 854-873.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2199 – 2210
Received 05 March 2021; Accepted 01 April 2021.

[6] Dilbag Singh and Jasjit Singh Samagh, "A Comprehensive Review of Heart Disease Prediction using Machine Learning", Journal of Critical Reviews, vol. 7, no. 12, **(2020),** pp.281-285.

[7] Monther Tarawneh and Ossama Embarak. "Hybrid Approach for Heart Disease Prediction Using Data Mining Techniques", Acta Scientific Nutritional Health, Vol. 3, no. 7, **(2019)**.

[8] N Satyanandam and Ch Satyanarayana. " Heart Disease Detection Using Predictive Optimization Techniques", I.J. Image, Graphics and Signal Processing, vol. 9, **(2019)**, pp.18-24.

[9] Indu Yekkala, Sunanda Dixit and M.A. Jabbar, "Prediction of heart disease using Ensemble Learning and Particle Swarm Optimization", International Conference on Smart Technology for Smart Nation, **(2017).**

[10] Russell Eberhart and James Kennedy," A New Optimizer Using Particle Swarm Theory", IEEE **(1995)**, pp.39-43

[11] L. Yu, and H. Liu, "Feature selection for high dimensional data: A fast correlation-based filter solution", in: Proc. of the 20th International Conference on Machine Learning, **(2003),** pp. 856-863.

[12] J. Kennedy, "Particle Swarm Optimization", in Encyclopedia of Machine Learning (Springer), C. Sammut and G. I. Webb, Eds. Boston, MA: Springer US, pp. 760–766, **(2010)**.

[13] Saikat Dutt, Subramanian Chandramouli and Amit Kumar Das, "Machine Learning", Pearson India Education Services Pvt. Ltd, India, **(2019)**.

[14] John A.A. Hunter, Editor, "Davidson's Principles and Practice of Medicine", Churchchill Livingstone An Imprint of Elsevier Limited, **(2002)**.

[15] H. Ceylan and M. G. H. Bell, Traffic signal timing optimisation based on genetic algorithm approach, including drivers' routing,Transport. Res. 38 (2004), 329–342

[16] B. Dengiz, F. Altiparmak and A. E. Smith, Local search genetic algorithm for optimal design of reliable networks, IEEE Trans. Evol. Comput. 1 (1997)

[17] G. R. Harik, F. G. Lobo and D. E. Goldberg, IEEE Trans. Evol. Comput. 3 (1999), 287–297

[18] C. Miles, S. J. Louis, N. Cole and J. McDonnell, Learning to play like a human: case injected genetic algorithms for strategic computer gaming, in: Proc. 2004 Congr. Evol. Comput. (IEEE Cat. No. 04TH8753), vol. 2, pp. 1441–1448, IEEE, Portland, OR, USA, 2004.