2020 3rd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)

Robust Heart Disease Prediction: A Novel Approach based on Significant Feature and Ensemble learning Model

1st Muhammad Affan Alim
College of Computing and Information Sciences
PAF-Karachi Institute of Economics and Technology
Karachi, Pakistan
affanalim@pafkiet.edu.pk

2nd Shamsheela Habib
College of Computing and Information Sciences
PAF-Karachi Institute of Economics and Technology
Karachi, Pakistan
shm.hbb@hotmail.com

3rd Yumna Farooq
College of Computing and Information Sciences
PAF-Karachi Institute of Economics and Technology
Karachi, Pakistan
yumnafarooq1995@gmail.com

4th Abdul Rafay
College of Computing and Information Sciences
PAF-Karachi Institute of Economics and Technology
Karachi, Pakistan
rafayjangsher2@gmail.com

***Abstract*—Among the different causes of human death, heart disease is one of the most common causes of non-communicable and silent death in the world. It is a challenge to early predict heart disease by using clinical data for better treatment. After evolving machine learning, its importance is incessantly being increased in every field of life. From the last couple of years, Machine learning is also the center of attention of researchers in field medical sciences. Researchers use different tools and techniques of machine learning for the early prediction of diseases. Essentially, heart disease prediction with available clinical data is one of the big challenges for researchers. State-of-theart results have been reported using different clinical data using different machine learning algorithms, nevertheless, there is some opportunity for improvement. In this paper, we propose to use a novel method that comprises machine learning algorithms for the early prediction of heart disease. Essentially, the aims of the paper are to find those features by correlation which can help robust prediction results. For this purpose UCI vascular heart disease dataset is used and compares our result with recently published article. Our proposed model achieved accuracy of 86.94% which outperforms compare with Hoeffding tree method reported accuracy of 85.43%**

***Index Terms*—Machine Learning, Stratified KFold, Random Forest and ROC**

I. Introduction

In today's chaotic world we all have very busy life, tough schedule and competitive activities for growing up and to achieve success in our life but we are neglecting our health issues because of this, we encounter many diseases that are threat to our lives. People don't pay attention to symptoms and this negligence cause death. Many diseases if they will not be treated properly they cause death. Heart disease is one of the chronic illnesses that produce different signals from early stage but we fail to recognize these signals which lead to long term illness or loss of life.

Heart disease mostly occurs in man then female according to the report of WHO world health organization's statistics 24% death in world that are not communicable happens because of heart illness [10]. One-third of death worldwide due to heart disease. Half of death occurs in America and half in other countries. 17 million people die due to this chronic illness [6]. A heart problem can occur in any age of life young, middle or old, environment where we live, it can also cause because of genetic and in heart disease gender plays an important role. Some causes of heart disease are not doing exercise or laziness, alcohol, smoking, tension, stress and consumption that exceed the body needs [1], [8]. The factors caused by healthy behavior factors due to diseases include high blood pressure hyperlipidemia, obesity, overweight and diabetes which the trend of risk factors for both overweight and obesity tears high blood and high cholesterol such factors are caused by improper self-care behaviors. For investing this with the help of a machine learning algorithm we have collected a dataset of vascular heart disease that is openly available at UCI [3]. This dataset has 14 attributes and 303 samples of heart disease named as age, sex, chest pain, resting blood pressure, cholesterol, fasting blood sugar, electrocardiographic results during rest, exercise, old peak, thalassemia, colored by fluoroscopy and heart disease presence. After collecting that dataset, we have applied different machine learning algorithms for predicting the output base on the existing data. Machine learning algorithms are giving beneficial results in medical science field for detecting many diseases, their treatments and

978-1-7281-4970-7/20/$31.00 ©2020 IEEE

so on. Many researchers have already used this dataset and produced the results. The reason for collecting and applying techniques on this paper is because we tried to improve the accuracy of model for this dataset. For this purpose we tried different algorithms and succeed in improving the accuracy of this dataset.

We have proposed to used Random Forest algorithm with correlation based selected features on a given data set. An improve results with highest accuracy compare with Hoeffding tree method [12]. We compare results with the following models Na¨ıve Byes, Logistic regression, Gradient Boost,Support Vector Classification and also compare the result with a paper named as A Classification for Patients with Heart Disease Based on Hoeffding Tree which was published in IEEE conference in October 2019 [12]. This paper consists of six sections, in section one introduction we defined overall research of the project. Section 2 consists of literature review in which heart disease, machine learning and model theoritical description included. In section 3 previous work of same dataset and machine learning is written. In section 4 Methodology is written in this dataset description, libraries and tool used, data preprocessing, correlatioon, scaling and model working with stratified KFold is described. in section Experiments and results of the model is defined and in last section conclusion and future work is described.

II. Previous Work

Many research papers have been published on predicting heart disease using a machine learning algorithm. Research with improved accuracy of heart disease prediction has been reported by using train test split validation followed by logistic regression was used for prediction which showed the improved results on UCI dataset set [5]. In [6] author has focused on ensemble classification techniques. In this regard, multiple classifiers are used followed by score level ensemble for improvement of prediction accuracy. A Hoeffding tree algorithm with a K-fold cross-validation method is used on UCI dataset and reported a high accuracy [12]. In [9] author has proposed to find a significant feature of clinical heart disease dataset by using hybrid machining algorithm. A combination of genetic algorithms and neural networks it was based on fuzzy logic for feature extraction exhibited an increase in accuracy of up to 99.97% [11]. A model was proposed in which coactive neuro-fuzzy inference system (CANFIS) used with neural networks, fuzzy logic, and genetic algorithms, and gave results improved accuracy for heart disease in which algorithm tuned automatically [7].

III. Proposed Methodology

This paper is based on classification technique Random Forest for identifying heart disease with accuracy and compare the results with other models of machine learning and a published paper [12].

TABLE I
Database Description

No.	Dataset Description	Ranges
1.	Age	29 to 79
2.	Sex Female 0, Male 1	0,1
3.	Chestpain (CP)	0,1,2,3
4.	Resting Blood pressure (Trestbps)	94 to 200 (mm Hg)
5.	Serum Cholestrol (chol)	126 to 564 (mg/dl)
6.	Fasting Blood pressure (FBS)	0,1 (mg/dl)
7.	Resting Electrocardiographicn Results (Restecg)	0,1
8.	Maximum heart rate achieved (Thalach)	71 to 202
9.	Exercise induced Angina (Exang)	0,1
10.	ST depression induced by exercise relative to rest (Oldpeak)	0 to 6.2
11.	Slope of the peak Exercise ST segment (Slope)	0,1,2
12.	Number of major vessels colored by fluoroscopy (ca)	0 to 3
13.	thalassemia (Thal)	1,2,3
14.	Target class	0,1

A. Dataset

We collected this dataset from UCI machine learning website [3]. Dataset has more than 300-tuple including 14 different attributes. A brief description of dataset attributes shown in Table I.

According to the table, we have 1st attribute of age which range starts from 29 and end on 79. as Gender is the important factor in heart disease that's why gender is the most important attribute, chest pain is the sign of heart disease so it is better to consult with doctor if you have this. Blood pressure plays an important factor in any disease, cholesterol also plays an important part, FBS, Restecg and other attributes also have correlation with each other which shown in Figure 1.

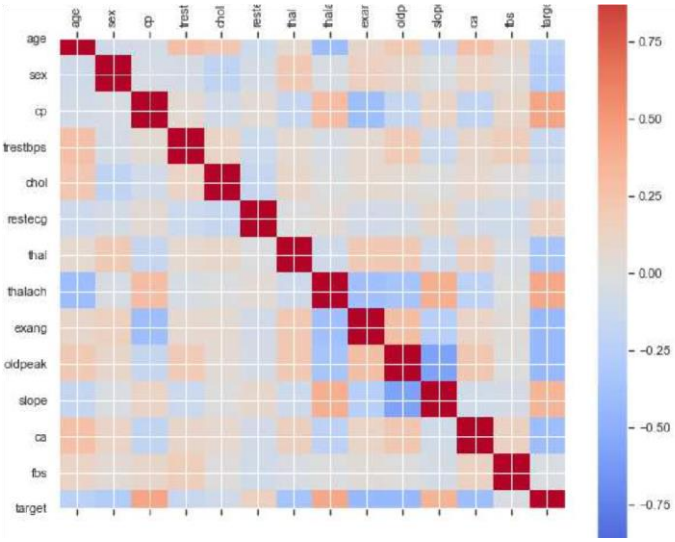
B. Libraries and tools

For applying this model, we have used python language and its libraries. For python, we have used a Spyder (Anaconda) tool. Libraries used in the model are numpy, pandas, sklearn, random forest classifier, cross-validation stratified KFold cross val score and matplotlib. Libraries for drawing graph of AUC curves.

C. Data Preprocessing

978-1-7281-4970-7/20/$31.00 ©2020 IEEE

Data Preprocessing is an important in machine Learning. without preprocessing it will give the biased or impure result. *1) Correlation:* First, we check the data for missing values and irregular information and we have found there is no missing values and outliers in the given dataset. Then we applied the correlation of the data attributes with each other and select only the significant features. The correlation coefficient of two



random variables X and Y can be calculated Fig. 1. Correlation between attributes

by

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_x \sigma_y} \tag{1}$$

Here ρ is correlation coefficient and Cov is covariance can be calculated as

$$Cov(X,Y) = E((X - \mu_x)(Y - \mu_y)) \tag{2}$$

D. Random Forest

Tin Kam Ho is considered to be a pioneer of random forest first algorithm [4]. Essentially, it works on the concept of tree. As defines in random forest algorithm, it makes number of trees from the given dataset and then combines the result this method is also called ensemble technique in which weak output produces strong output by combing all the results. In random forest we can change the inputs by tuning their parameters like criterion, depth of tree, maximum and minimum leaf, etc. Random forest is a supervised machine learning algorithm it can be used for both classification and regression. We used this algorithm for classification dataset. It does not suffer from overfitting. It is highly robust method and can replace missing values by own. Random forest algorithm is defined in algorithm section.

1) Normalization: Given dataset was not normalized, a min-max normalized technique is used which scaled the between 0 and 1. A normalized shown in Figure 2.

E. Model Working with Stratified KFold

for checking the accuracy and that model is completely fit the dataset. A Kfold is one type of cross-validation technique that split data into number of folds that are given by and use that data for training and testing iteratively for example if we give 10 Folds then it will split the data into 10 folds and use each fold for training and remaining data for testing. Stratified KFold validation is the type of KFold split validation. It is the variation in KFold in which splitting of data into folds may be governed by criteria such as ensuring that each fold has the same proportion of observations with a given categorical value, such as the class outcome value. This is called stratified cross-validation [2]. in short data is divided uniformly in each manner. Gradient boosting is also the technique of cross-validation Random Forest uses this technique and if we use Random Forest with KFold it will give the accurate, without overfitting and underfitting result. Random Forest makes number of decision trees bases on

After transformation, we have applied Random Forest with Stratified Kfold and tuned parameter. Cross-Validation prevents data from overfitting and underfitting in which we split the percentage of data in train and test apply model one it number of estimates we gave as parameter each tree produced similar or different result with each other testing data goes to majority voting decision tree sample either it is true or false. Random Forest uses bootstrapping technique and combined with stratified KFold it gives high accuracy. For dataset of 303 samples we used 5 splits of Stratified KFold cross-validation.

IV. Experiments and Results

A rigorous experiments were done on UCI heart disease dataset. The corpus of dataset has 303 instances including 14 multiple attributes attributes. Individual correlation with every attributes including the target were measured and found that attribute ”fbs” has less correlation with target class was excluded from experiment as shown in Figure 1. In first experiment we used all 303 tuples and apply the several machine learning models. The logistic regression achieved accuracy of 83.15%, Support vector machine obtained accuracy of 82.49%, the Naive Based attended the accuracy of 80.50%, and Gradient Boosting obtained accuracy of 81.50%. The proposed model Random Forest with Stratified KFold cross-validation method outperformed compare with other algorithms and attained accuracy of 86.12%. A detail compassion results graph is shown in Figure 3.

978-1-7281-4970-7/20/$31.00 ©2020 IEEE

For Second experiment 199 random samples were selected as proposed in [12]. According to our proposed method the less correlated attributes with target class were removed. The proposed method Random forest with Stratified KFold cross-validation method outperformed and obtained the highest accuracy of 86.94% compare with Hoeffding tree method

	0	1	2	3	4	...	8	9	10	11	12
0	0.708333	1.0	1.000000	0.481132	0.244292	...	0.0	0.370968	0.0	0.00	1.0
1	0.166667	1.0	0.666667	0.339623	0.283105	...	0.0	0.564516	0.0	0.00	0.0
2	0.250000	0.0	0.333333	0.339623	0.178082	...	0.0	0.225806	1.0	0.00	0.0
3	0.562500	1.0	0.333333	0.245283	0.251142	...	0.0	0.129032	1.0	0.00	0.0
4	0.583333	0.0	0.000000	0.245283	0.520548	...	1.0	0.096774	1.0	0.00	0.0
..	...	...	...	...	...	...	...	...	...	...	...
298	0.583333	0.0	0.000000	0.433962	0.262557	...	1.0	0.032258	0.5	0.00	0.0
299	0.333333	1.0	1.000000	0.150943	0.315068	...	0.0	0.193548	0.5	0.00	0.0
300	0.812500	1.0	0.000000	0.471698	0.152968	...	0.0	0.548387	0.5	0.50	1.0
301	0.583333	1.0	0.000000	0.339623	0.011416	...	1.0	0.193548	0.5	0.25	0.0
302	0.583333	0.0	0.333333	0.339623	0.251142	...	0.0	0.000000	0.5	0.25	0.0

Fig. 2. Normalized Dataset

attained accuracy of 85.43%. Our proposed method achieved the 86.92% and 100.00% precision and recall respectively which is higher with Hoeffding tree method with a margin 1.02% and 14.60% respectively. A detail results is shown in Table II. An AUC curve of proposed model is shown Figure 4.

From the given table we can see in any case our model producing best result with different number of records. Our model gives 86.94% accuracy, precision 86.42% and recall 100.00% with 199 sample and 86.12% accuracy, 83.94% precision and 92.72% recall with 303 samples.

Random Forest Algorithm

1 : **input:** $D = \{(x_1,y_1),\ldots, (x_n,y_n)\}$, be a training dataset
2 : **step 1:** Let $h = h_1(x),h_2(x),\ldots,h_k(x)$, are ensemble of weak classifiers
3 : **step 2:** If each h_k is a decision tree, the parameters of the tree are defined as $\phi = (\Theta_{k1}, \Theta_{k2},\ldots, \Theta_{kn})$
4 : **step 3:** Each decision tree k leads to a classifier $h_k(X) = h(X|\Theta_k)$
5 : **output:** Final classification $f(x)$ = Majority of $h_k(X)$

TABLE II
COMPARISON OF PROPOSED METHOD WITH HOEFFDING TREE

Score	RF with 303 samples	Hoeffding 199 with samples	RF with 199 samples
Accuracy	86.16	85.43	86.94
Precision	83.94	85.40	86.42
Recall	92.72	85.40	100.00
Error	14.50	13.87	13.05

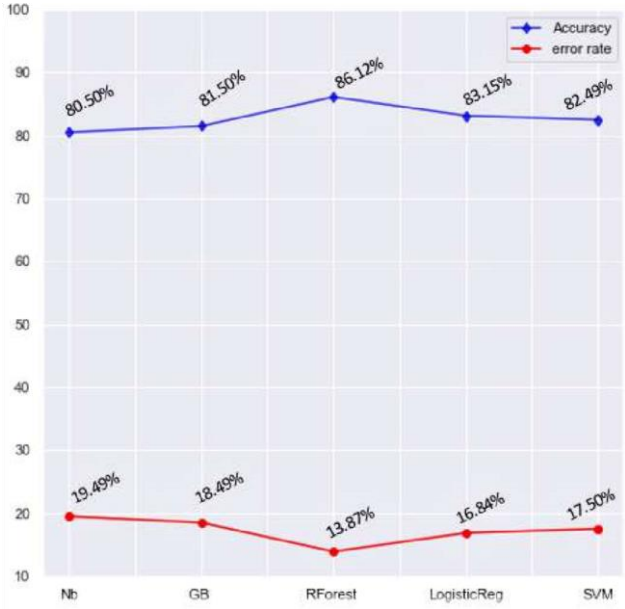
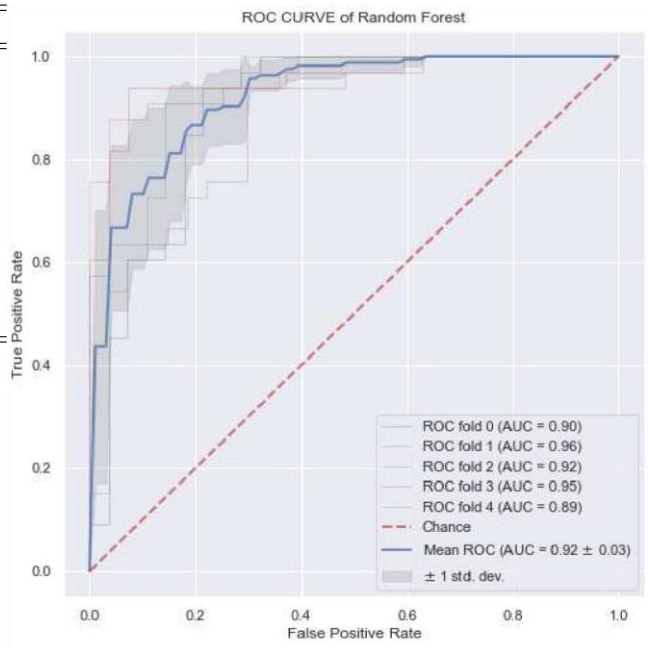


Fig. 3. Comparison of proposed method with other algorithms

Fig. 4. AUC graph of proposed model with 199 samples

A. ROC AUC Curve

For predicting the accuracy of the model ROC AUC curve is the best graphical representation of the model to spectate performance of the model we achieved 0.92 AUC area under the curve result which is nearest to 1. here is the graphical representation of ROC AUC curve of our model in which each fold graph and average graph has shown.



V. Conclusion and Future Work

Machine learning based diseases prediction helps the physician for treatment of patient. Vascular heart disease is a common issue Several articles have already been published for vascular heart disease. regarding the heart disease is very difficult to detect because diabetes, blood pressure, and other factors. Several techniques have already been reported in different research viz K-Nearest Neighbor Algorithm, Navie Based Algorithm, Genetic algorithm, and others. We introduced to used the correlation for selection of the significant attributes followed by Random Forest (RF) with stratified Kfold cross validation. Our proposed model attained a highest accuracy , precision and recall compared with recently published research. Some new machine learning algorithms has been proposed viz Naive Gradient Booting which is still untouched the in filed of vascular heart disease.

References

[1] C.-A. Cheng and H.-W. Chiu, "An artificial neural network model for the evaluation of carotid artery stenting prognosis using a national-wide database," in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, jul 2017. [2] H. Comin, R. Santos, D. Corradini, W. Morrison, C. Curme, D. L. Rosene, A. Gabrielli, L. da F. Costa, and H. E. Stanley, "Statistical physics approach to quantifying differences in myelinated nerve fibers," *Scientific Reports*, vol. 4, no. 1, mar 2014.

[3] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[4] T. K. Ho, "Random decision forests," in *Proceedings of 3rd international conference on document analysis and recognition*, vol. 1. IEEE, 1995, pp. 278–282.

[5] R. Kannan and V. Vasanthi, "Machine learning algorithms with ROC curve for predicting and diagnosing the heart disease," in *Soft Computing and Medical Bioinformatics*. Springer Singapore, jun 2018, pp. 63–72.

[6] C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," *Informatics in Medicine Unlocked*, vol. 16, p. 100203, 2019.

[7] S. R. LathaParthiban, "Intelligent heart disease prediction system using canfis and genetic algorithm," *Int. J. Biol. Biomed. Med. Sci*, 2008.

[8] T. Liu, "Notice of retraction: Cigarette smoking, glutathione stransferase p1 genetic variant, and cardiovascular fitness," in *2011 5th International Conference on Bioinformatics and Biomedical Engineering*. IEEE, may 2011.

[9] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019.

[10] W. H. Organization, P. H. A. of Canada, and C. P. H. A. of Canada, *Preventing chronic diseases: a vital investment*. World Health Organization, 2005.

[11] J. Singh and R. Kaur, "Cardio vascular disease classification ensemble optimization using genetic algorithm and neural network," *Indian Journal of Science and Technology*, vol. 9, no. S1, dec 2016.

[12] S. Thaiparnit, S. Kritsanasung, and N. Chumuang, "A classification for patients with heart disease based on hoeffding tree," in *2019 16th International Joint Conference on Computer Science and Software Engineering (JCSSE)*. IEEE, jul 2019.

978-1-7281-4970-7/20/$31.00 ©2020 IEEE