Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2020)
IEEE Xplore Part Number:CFP20K74-ART; ISBN: 978-1-7281-4876-2

Prediction of Cardiac Disease using Supervised Machine Learning Algorithms

R.Jane Preetha Princy
Karunya Institute of Technology and Sciences, Coimbatore
janer@karunya.edu.in

Saravanan Parthasarathy
B.S.Abdur Rahman Crescent Institute of Science and Technology, Chennai
saravanan_cse_2019@crescent.education

P. Subha Hency Jose
Karunya Institute of Technology and Sciences, Coimbatore
hency20002000@karunya.edu

Arun Raj Lakshminarayanan
B.S.Abdur Rahman Crescent Institute of Science and Technology, Chennai
arunraj@crescent.education

Selvaprabu Jeganathan
B.S.Abdur Rahman Crescent Institute of Science and Technology, Chennai
selva_cse_phd_17@crescent.education

Abstract — The healthcare industry is dealing with billions of patients all over the world and producing massive data. The machine learning-based models are dissecting the multidimensional medical datasets and generating better insights. In this study, a cardiovascular dataset is classified by using several state-of-the-art Supervised Machine Learning algorithms that are precisely used for disease prediction. The results indicate that the Decision Tree classification model predicted the cardiovascular diseases better than Naive Bayes, Logistic Regression, Random Forest, SVM and KNN based approaches. The Decision Tree bequeathed the best result with the accuracy of 73%. This approach could be helpful for doctors to predict the occurrence of heart diseases in advance and provide appropriate treatment.

Keywords — Cardiovascular Disease, Naive Bayes, Decision Tree, Logistic Regression, Random Forest, SVM, KNN, Risk prediction

I. INTRODUCTION

According to WHO, Cardiovascular Disease (CVD) is the number one cause of death worldwide. Every year almost 17.9 million deaths occur due to CVD, which is estimated to be 31% of deaths globally [1]. CVD is defined as any abnormal condition of the blood vessels of the heart, which gets build up by a substance called plague that narrows up the arteries and the veins that carry blood to and fro from the heart. This makes harder for the blood to flow, and may even cause blocks, which results in a heart attack or stroke. The risk factors which cause CVD include high blood pressure, poor nutrition, physical inactivity, high blood cholesterol levels, alcohol intake, tobacco usage, obesity and genetic mutations. The deaths from these factors can be avoided by early prognostication. On the other hand, the data collection mechanisms are getting upgraded day by day through the implementation of the Internet of Things. These advancements are yielding terabytes of data every day from healthcare organizations. It is not possible for a human being to consolidate millions of data and conclude the particular patient's illness. However, Machine learning could be exerted as a predictive mechanism to find the patterns inside the data.

By using Machine Learning, the factors are studied and predict people who are in the likeness of developing heart diseases. Machine learning methodologies can review large volumes of data, identify trends which might not be evident to human beings. It typically enhances efficiency and accuracy to the ever-increasing amounts of data that are being processed. It also allows instantaneous adaptation, without the demand for human intervention. Supervised Machine Learning is the process of accomplishing a task by providing labeled data and output patterns to the system. During the training, the algorithm searches for patterns in the data that correlates with the desired output. After the training, the supervised learning model can predict the correct label for newly presented input data. The objective of this study is to identify a well-performing approach by measuring the classification accuracy of various supervised machine learning algorithms.

II. LITERATURE REVIEW

Krumholz et al. utilized Machine Learning techniques to predict the mortality and hospitalization in heart failure subjects [2]. They have used five methods, LR with forwarding selection variable and LASSO regularization variable selection, Random Forest, Gradient Descent Boosting and SVM. Three years follow up was done and validated using 5-fold cross-validation. Random Forest gave the best results with 0.72 mean C-statistic value for predicting mortality and 0.76 for predicting hospitalization. The inclusion of time-to-event analysis could ameliorate the outcome of the proposed model.

Vilasi et al. used Machine Learning to gauge CVD in patients undergoing Dialysis [3]. Various Machine Learning algorithms were tested against Italian and American datasets.

The Support Vector Machine (SVM) with RBF Kernel algorithm gave the best results with 95.25% accuracy in the Italian Dataset and 92.15% in the American dataset. However, the bias in Italian dataset could impact the accuracy of predictions.

Shashikant et al. proposed an approach for the premonition of Cardiac arrest in smokers using the concept of Heart Rate Variability (HRV). HRV [4] is a non-invasive technique to assess the regulation of heartbeat. The optimum time and environment are required to obtain correct data points. The outcome of Decision Tree, Logistic Regression and Random Forest were compared. The 10-fold validation method is used to measure the performance of the entire classification techniques. The results showed an accuracy of Logistic Regression is 89.7%, Decision Tree is 92.59% and Random Forest is 93.61%. Random Forest was found to be the best among these methodologies.

An approach named 'Autoprognosis' was ideated by Ahmed et al. [5]. This system automatically selects and tunes Machine Learning models. The constructed model was tested using data of 423,604 participants. The outcomes were compared with well-established risk prediction algorithm 'Framingham Score'. The results illustrate that Autoprognosis model delivered better prediction than the Framingham Score, with an accuracy rate of 95%. The incorporation of other attributes such as triglycerides, markers of inflammation, natriuretic peptides were not considered during the prognostication process.

Zhou et al. propounded a learning process to overcome the missing value in the medical dataset, and to predict the CVD correctly [6]. Naive Bayes, SVM, Decision Tree, Logistic Regression, RBF and Random Forest algorithms were employed to predict the CVD. The results show that despite missing values RF was superior to other methods, having 88% sensitivity, 87.6% specificity, 88% precision.

A hybrid intelligent system to predict heart disease was designed by Amin et al. [7]. In that system, Logistic Regression, ANN, SVM, KNN, Decision tree, Naive Bayes and random forest were used for classification. To improve the efficiency of prediction, three feature selection algorithms were used, Relief, mRMR and LASSO which selects highly correlated features which greatly influences the target variable. Results show Logistic Regression with 10-fold validation gave 89% accuracy with Relief feature selection algorithm. The utilization of neural network-based optimization techniques could conceivably improve the outcome of this model.

Panagiotakos et al. compared Machine Learning methods with CVD established risk tool Hellenic Score [8]. A dataset named ATTICA was selected for this study. According to the type of classifier and training dataset, Hellenic Score showed 85% accuracy, 20% specificity, 97% sensitivity, 87% PPV and 58% NPV. Whereas the Machine Learning algorithms showed 65-84% accuracy, 46-56% specificity, 67-89% sensitivity, 89-91% PPV and 24-45% NPV. Random Forest provided the best outcome. The study has lacked an analysis of the relationship between lifestyle characteristics and occurrences of CVD.

Kang et al. applied Machine Learning to assess three risks - Hypertension, Hyperglycemia and Hyperlipidemia [9]. These are the main factors which lead to CVD. Two models were used to predict the risks; they are HCRT- Logistic model and Logistic CART model. Both the models were verified by 10-fold cross-validation. BMI, waist circumference, hip circumference, waist to hip ratio, waist to height ratio, disease history etc., were taken into consideration. The accuracy of the proposed model fluctuates based on gender type. The waist circumference was identified as the most valuable predictor in both genders.

Stephen et al. availed Machine Learning practices to envisage CVD using routine clinical data [10]. The data was obtained from 378,256 patients residing in the UK. Random Forest, Logistic Regression, Neural Network and Gradient Boosting techniques were exerted for this perusal. Neural Network clinched with better exactitude. Conversely, the logic nexus behind Neural Network process is perplexing to interpret.

Ashok has anticipated the occurrence of heart disorder using ANN, KNN, SVM, Logistic Regression, Classification Tree and Naive Bayes [11]. Moreover, these methods were also assessed by ROC curve. Here Logistic Regression gave the highest accuracy of 85%, with 89% sensitivity and 81% specificity. However, the model has to be tested with massive datasets to ensure reliability.

The technique of multi-level risk assessment was postulated by Aljaaf et al [12]. Three risk factors labeled as smoking, lack of physical activity and obesity were introduced along with existing attributes. Decision Tree method postulated the risks of heart failures with 86.53% accuracy. The implementation of advanced feature selection methods could maximize the denouement outcome of this model.

III. MATERIALS

Description of the dataset:

A cardiovascular disease dataset in Kaggle [11] has been used for this study. It contains twelve attributes which include one target variable. The depiction of the same is presented in Table 1. Persons from the age of 29 to 64 have been considered for the analysis. Their height and weight are also documented. The gender value 1 and 0 were assigned to male and female patients respectively. The systolic and diastolic blood pressures are deliberated to measure the influence. The Cholesterol, Glucose readings of the patients marked with categorical values as normal, above normal, well above normal.

The heart diseases are much inclined with one's smoking and drinking habits. These two variables are marked in binary values. The value of '1' represents that the patient is a 'smoker/drinker'; '0' denotes him or her as 'non-smoker/non-alcoholic'. The patients with regular physical activity marked with '1' and '0' for others. The presence or absence of cardiovascular disease is the target attribute. It comprises binary values. The '0' represents normal and '1' signifies the people confirmed with heart disease.

TABLE 1

FEATURE INFORMATION OF THE DATASET

S.No	Attribute Name	Description	Range of Values
1	age	Age	int (years)
2	height	Height	int (cm)
3	weight	Weight	float (kg)
4	gender	Gender	categorical code
5	ap_hi	Systolic blood pressure	int
6	ap_lo	Diastolic blood pressure	int
7	cholesterol	Cholesterol	1: normal, 2: above normal, 3: well above normal
8	gluc	Glucose	1: normal, 2: above normal, 3: well above normal
9	smoke	Smoking	binary
10	intake alco	Alcoholic	binary
11	active	Physical activity	binary
12	cardio	Presence or absence of cardiovascular disease	binary

IV. EXPERIMENTS AND RESULTS

The functional flow of this assessment is exemplified in Fig.1. The correlation between features in the dataset is portrayed in Fig. 2. The dark brown colour represents a high positive correlation and the dark blue colour indicates a negative correlation.

This research work focused on implementing a few classification algorithms and compares the outcomes. The dataset was divided into training and testing portions in the ratio of 70/30. Naive-Bayes, Decision Tree, Logistic-Regression, Random Forest, SVM and KNN classification models were used to predict CVD.

The confusion matrix is used for identifying the mislabeling or error in prediction. It matches the actual and predicted values with four elements (True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN)). Type-I and Type-II errors are seeded by False Positive and False Negative values. The confusion matrix is very expedient to calculate Precision, Recall, F1-score and Accuracy.

TABLE 2

CONFUSION MATRIX

		Actual values	
		Positive	**Negative**
Predicted Values	**Positive**	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

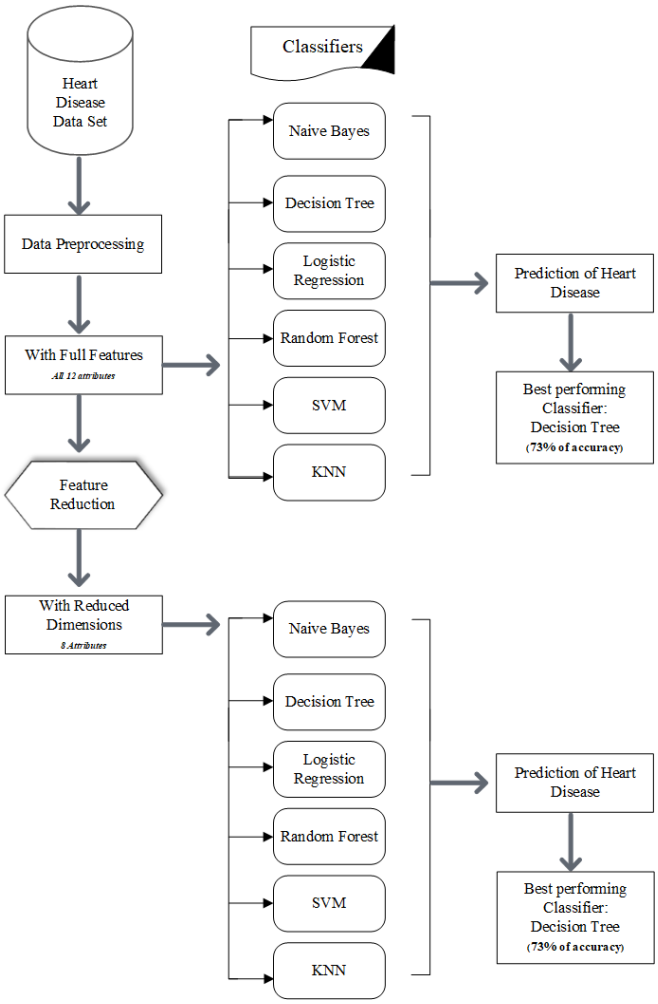


Fig. 1: Predicting CVD using supervised learning algorithms

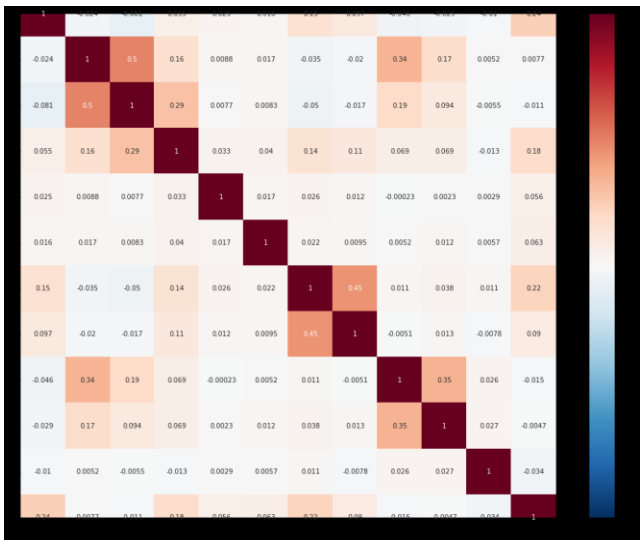


Fig. 2: Correlation between all available features

A. *Prediction Accuracy*

The accuracy denotes the properly predicted values. Fig. 3 represents the accuracy of each algorithm tested.

Accuracy = (True Positive+ True Negative)/ Total

The decision tree algorithm outperformed others by producing 73% accuracy. The logistic regression and SVM delivered 72% and Random forest made it with 71%. The KNN and Naive-Bayes algorithms delivered 66% and 60% of accuracy respectively.

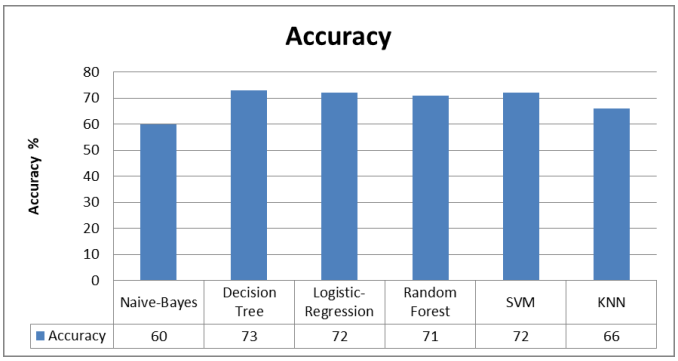


Fig. 3: Accuracy of various learning techniques

B. *Precision*

It represents the real positive cases from all the positive predictions. Fig. 4 indicates the precisions across different algorithms.

Precision = True Positive/ (True Positive + False Positive)

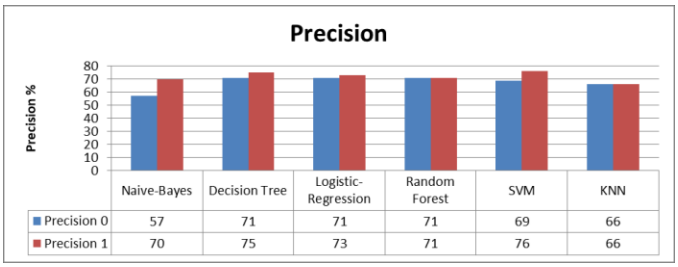


Fig. 4: Precision of learning techniques

C. *Recall*

It characterizes the correctly predicted values out of all the positive classes. Fig. 5 signifies the values of recall across tested algorithms.

Recall = True Positive/ (True Positive + False Negative)

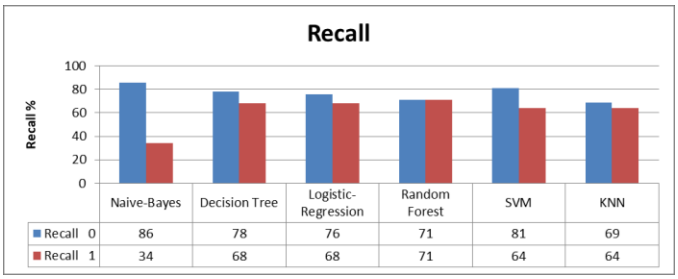


Fig. 5: Recall of learning techniques

D. *F1 score*

It appraises Recall and Precision and utilizes Harmonic Mean to calculate the test accuracy. Fig. 6 connotes the F1 score across various algorithms.

F1 score = (2 * Precision * Recall))/ (Precision + Recall)

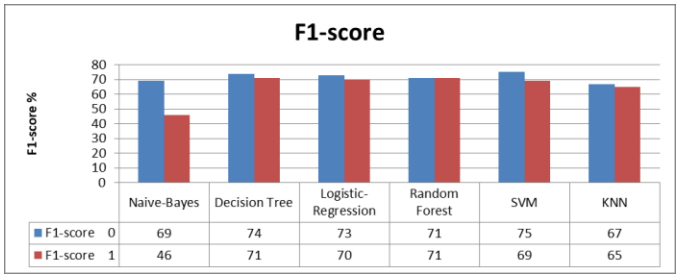


Fig. 6: F1 score of learning techniques

E. *Dimensionality Reduction*

From Fig. 2, the entities like height, smoking, drinking and physical activity are negatively correlated. These entities were removed from the dataset and tested with above-listed models. The algorithms predicted the CVD with reduced dimensions.

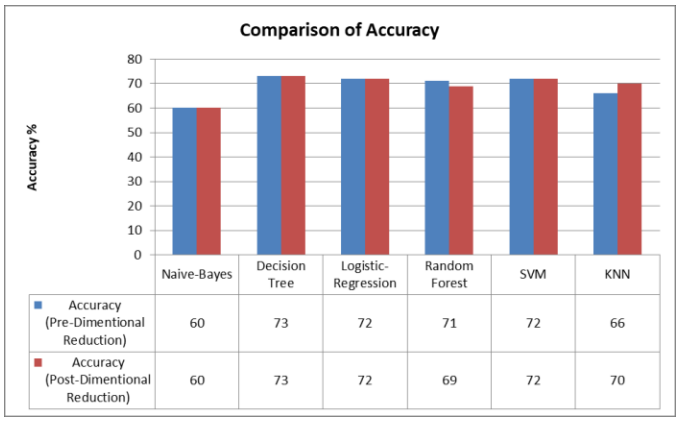
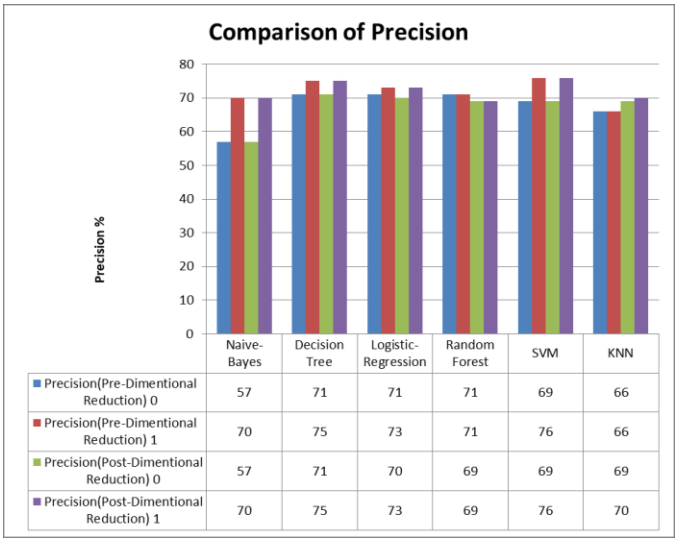


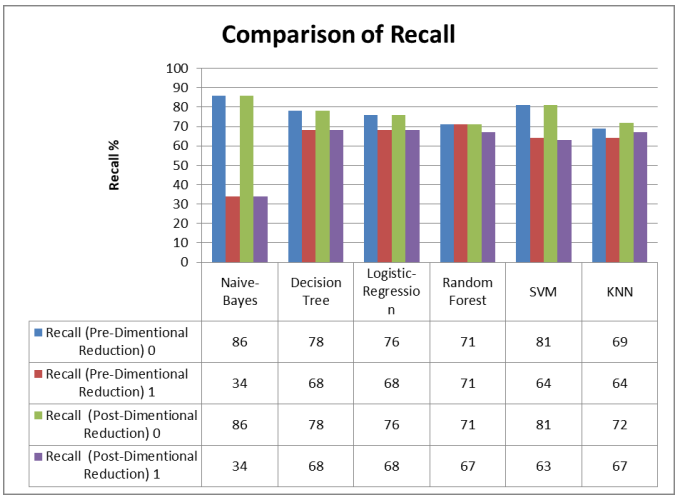
Fig. 7: Comparison of Accuracy

The pre and post dimensionality reduction accuracy of algorithms is compared in Fig. 7. After the attributes were reduced to eight, the outcome of measures also got affected. The accuracy of the Random Forest algorithm got deduced to 69% from 71%. However, the accuracy of the KNN algorithm got elevated from 66% to 70%. Fig. 8 signifies the precision values of algorithms before and after the dimensionality reduction. There is a minor change in the precision value of Logistic-Regression and considerable alterations in the precision of Random Forest and KNN. Fig. 9 and 10 are manifesting substantial changes in recall and F1 score of Random Forest and KNN.



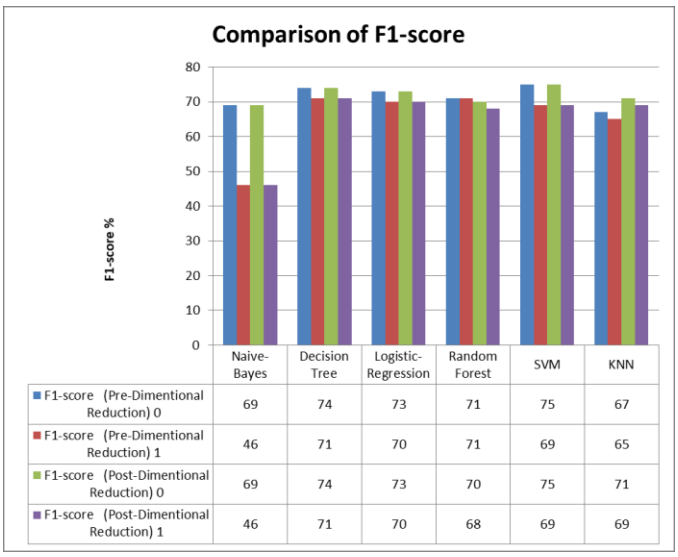
	Naive-Bayes	Decision Tree	Logistic-Regression	Random Forest	SVM	KNN
Precision(Pre-Dimentional Reduction) 0	57	71	71	71	69	66
Precision(Pre-Dimentional Reduction) 1	70	75	73	71	76	66
Precision(Post-Dimentional Reduction) 0	57	71	70	69	69	69
Precision(Post-Dimentional Reduction) 1	70	75	73	69	76	70

Fig. 8: Comparison of Precision



	Naive-Bayes	Decision Tree	Logistic-Regression	Random Forest	SVM	KNN
Recall (Pre-Dimentional Reduction) 0	86	78	76	71	81	69
Recall (Pre-Dimentional Reduction) 1	34	68	68	71	64	64
Recall (Post-Dimentional Reduction) 0	86	78	76	71	81	72
Recall (Post-Dimentional Reduction) 1	34	68	68	67	63	67

Fig. 9: Comparison of Recall



	Naive-Bayes	Decision Tree	Logistic-Regression	Random Forest	SVM	KNN
F1-score (Pre-Dimentional Reduction) 0	69	74	73	71	75	67
F1-score (Pre-Dimentional Reduction) 1	46	71	70	71	69	65
F1-score (Post-Dimentional Reduction) 0	69	74	73	70	75	71
F1-score (Post-Dimentional Reduction) 1	46	71	70	68	69	69

Fig. 10: Comparison of F1 score

V. Conclusion and Future Work

This study has been conducted on cardiovascular dataset by applying classification techniques. The Decision Tree algorithm delivered better prediction by providing 73% of accuracy. Since the dimension of dataset plays a major role in the performance of algorithms, the reduction of dimension affects the capability of Random Forest and KNN algorithms. The outcomes indicate that the dimension of dataset impacts the algorithms either positive or negative. In the next level, the High Correlation Filter and Principal Component Analysis to be applied for dimensionality reduction. The ensemble machine learning algorithms using CVD dataset are planned to assess and design a better disease prediction model.

VI. References

[1] WHO (World Health Organization): Cardiovascular Diseases - https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1

[2] Angraal S, Mortazavi BJ, Gupta A, Khera R, Ahmad T, Desai NR, Jacoby DL, Masoudi FA, Spertus JA, Krumholz HM , "Machine Learning Prediction of Mortality and Hospitalization in Heart Failure With Preserved Ejection Fraction", JACC : Heart Failure, vol. 8, Issue 1, January 2020.

[3] Sabrina Mezzatesta , Claudia Torino, Pasquale De Meo, Giacomo Fiumara , Antonio Vilasi, "A machine learning-based approach for predicting the outbreak of cardiovascular diseases in patients on dialysis" Computer Methods and Programs in Biomedicine, Elsevier, vol. 177, pp. 9-15, August 2019

[4] Shashikant R,Chetankumar P, "Predictive model of cardiac arrest in smokers using machine learning technique based on Heart Rate Variability parameter", Applied Computing and Informatics, June 2019

[5] Ahmed M. AlaaI, Thomas Bolton, Emanuele Di Angelantonio, James H. F. Rudd, Mihaela van der Schaar, "Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants", PloS One 14 (5): e0213653, May 2019.

[6] Runchuan Li, Shengya Shen, Xingjin Zhang, Runzhi Li, Shuhong Wang, Bing Zhou and Zongmin Wang, "Cardiovascular Disease Risk Prediction Based on Random Forest", Proceedings of the 2nd International Conference on Healthcare Science and Engineering, vol. 536, pp. 31-43, May 2019.

[7] Amin Ul Haq , Jian Ping Li , Muhammad Hammad Memon , Shah Nazir and Ruinan Sun, "A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms", Hindawi , Mobile Information Systems, vol. 2018, pp. 1-15, December 2018

[8] Alexandros C. Dimopoulos, Mara Nikolaidou, Francisco Félix Caballero, Worrawat Engchuan, Albert Sanchez-Niubo, Holger Arndt, José Luis Ayuso-Mateos, Josep Maria Haro, Somnath Chatterji, Ekavi N. Georgousopoulou, Christos Pitsavos and Demosthenes B. Panagiotakos, "Machine learning methodologies versus cardiovascular risk scores, in predicting disease risk" ,BMC Medical Research Methodology, vol. 18, Article number: 179, December 2018.

[9] Guixia Kang, Bo Yang, Dongli Wei, and Ling Li , "The Application of Machine Learning Algorithm Applied to 3Hs Risk Assessment", Big Data – BigData 2018, pp.169-181, June 2018

[10] Stephen F. Weng, Jenna Reps, Joe Kai, Jonathan M. Garibaldi, Nadeem Qureshi, "Can machine-learning improve cardiovascular risk prediction using routine clinical data?", PLoS One 12(4): e0174944, April, 2017.

[11] Ashok Kumar Dwivedi, "Performance evaluation of different machine learning techniques for prediction of heart disease", Neural Computing and Applications, vol. 29, pp. 685–693, September 2016

[12] A. J. Aljaaf, D. Al-Jumeily, A. J. Hussain,T. Dawson, P Fergus and M. Al-Jumaily, "Predicting the Likelihood of Heart Failure with a Multi Level Risk Assessment Using Decision Tree", 2015 Third International

Conference on Technological Advances in Electrical, Electronics and Computer Engineering, IEEE xplore, pp. 101 - 106, June 2015.

[13] https://www.kaggle.com/sulianova/cardiovascular-disease-dataset