Biosc.Biotech.Res.Comm. Special Issue Vol 13 No 11 (2020) Pp-24-27

BBRC
Bioscience Biotechnology Research Communications

Heart Disease Prediction Using Machine Learning Algorithms

Malavika G[1], Rajathi N[2], Vanitha V[3] and Parameswari P[4]

[1]*PG Scholar, Department of Information Technology, Kumaraguru College of Technology, Coimbatore, India.*
[2,3]*Professor, Department of Information Technology, Kumaraguru College of Technology, Coimbatore, India.*
[4]*Assistant Professor (SRG) Department of MCA, Kumaraguru College of Technology, Coimbatore, India.*

ABSTRACT

The rapidly growing field of data analysis plays a significant role in healthcare. The healthcare industry has become big business. The healthcare sector produces enormous amounts of data every day. This data helps to extract the hidden information, which is useful to predict disease at the earlier. In medical field, predicting heart disease is treated as one of the intricate tasks. Therefore, there is a necessity to develop a decision support system to forecast the cardio vascular disease in a patient. Machine learning plays a vital part in disease prediction. In this paper, various machine learning methods were used to predict the heart disease and their performances were compared. The results obtained show the superiority of the Random forest algorithm.

KEY WORDS: CLASSIFICATION ACCURACY, HEART DISEASE, MACHINE LEARNING.

INTRODUCTION

Data mining is used to examines and unearths important information from a massive collection of data. This can be further helpful in exploratory and illustration out patterns for making intelligent business-related decisions. One of the most threatening in medical domain is heart disease, which occurs instantly when its limitations are reached. Machine learning plays a vital role in disease prediction Rajathi N et al,. Cardiovascular disease generally refers to narrowed or blocked blood vessels, which can also lead to heart attack, chest pain or stroke. In general, blood pressure, cholesterol and pulse rate are the main reasons for a heart attack. Heart attack is the main heart disease.

ARTICLE INFORMATION

*Corresponding Author: rajathi.in.it@kct.ac.in
Received 9th Oct 2020 Accepted after revision 7th Dec 2020
Print ISSN: 0974-6455 Online ISSN: 2321-4007 CODEN: BBRCBA

Thomson Reuters ISI Web of Science Clarivate Analytics USA and Crossref Indexed Journal

Clarivate Analytics

crossref

NAAS Journal Score 2020 (4.31) SJIF: 2020 (7.728)

Online Contents Available at: http//www.bbrc.in/
Doi: http://dx.doi.org/10.21786/bbrc/13.11/6

Cardiovascular diseases (CVDs) are the most common explanation for global death. It is estimated that 17.9 million die annually. Heart attacks- once characterized as a part of "the old man's disease" but in this era it can causes for more young people. The heart attack occurs when the coronary arteries become blocked. It causes a serious attack when one or more coronary arteries become blocked.

Bad clinical results would be the doorway in the death of a patient. A computer-based support system can be developed to make a good decision in order to achieve correct and cost-effective treatment. Most of the hospitals maintains their patient data in the form of images, texts and numbers using database systems. This data contains much of the hidden information that has not yet explored, which are useful to make right decisions. Therefore, there is a need to develop an excellent system to help the practitioners to predict heart disease before it occurs. This paper mainly concentrates on the prediction of heart disease considering the past heart disease database records.

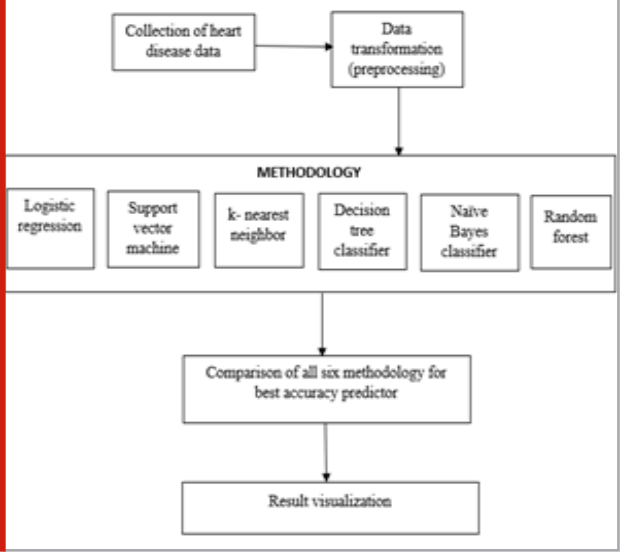
MATERIAL AND METHODS

Various studies with respect to diagnosis of heart disease are discussed in this section. Feixiang Huang et. al., used a data mining process to foresee hypertension from patient medical histories and concluded that J-48 classifier produces better results. M. Amiri et. al., developed diagnosis systems heart sounds. They used 116 heart sound signals to classification and regression trees. M.A. Nishara Banu et. al., used clustering and classification algorithm to forecast the hazard level of the patients. The authors Theresa Princy et. al.., discussed about classification methods including Naïve Bayes, neural network, KNN, decision tree for predicting the risk level of a patient they consider age, gender, pulse rate, blood pressure, cholesterol of each patient.

The various the machine learning algorithms are used by Min Chen et. al., for effective prediction of chronic disease. A multimodal disease risk prediction method was adopted for structured and unstructured data. The prediction accuracy the algorithm is better than other with a convergence speed. Tikotikar A et.al., data mining technique are used in the medical field for clinical diagnosis. It is inferred that an exhaustive survey of medical data help to make well informed diagnosis and decisions.

Cincy Raju et. al.., proved that the SVM technique is an efficient method for predicting heart disease. Praveen Kumar Reddy. M, et. al., used decision tree algorithm to prove the better prediction by comparing its performance with SVM. The authors Akash et. al., applied structured data and the text data of the patient to the k-mean algorithm and archived better accuracy. Reddy et. al., employed machine learning methods for heart disease prediction. All these created an interest to employ machine learning to prediction of heart disease.

Figure 1: Proposed System



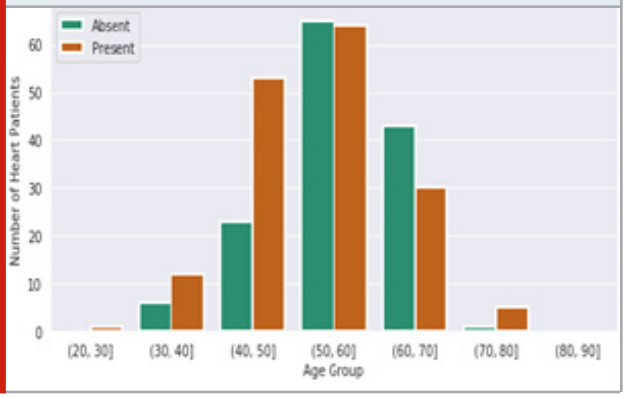
Proposed Methodology: In this paper, various machine learning methods including Naive Bayes classifier, logistic regression, random forest, support vector machine, decision tree classifier and KNN are employed to forecast heart disease. The Python language is used for implementation. The working of the model proposed is pictorially depicted in Figure 1. The dataset is pre-processed in order to remove irrelevant data which helps to achieve better accuracy.

Dataset: The heart disease dataset available in UCI repository taken for this study. The dataset consisting of the parameters including age, sex, chest pain type, serum cholesterol, resting blood pressure etc. After pre-processing the dataset was separated into training (70%) and testing (30%). The models used logistic regression, K-Nearest Neighbor, Support Vector Machine, Naïve Bayes, Decision tree and Random forest are trained using the training data and finally tested with the testing set.

RESULTS AND DISCUSSION

The overall objective of this paper is to forecast more accurately the occurrence of heart disease. Simulation based experiments were conducted using six methodologies named Naive Bayes Classifier, Logistic Regression, Random Forest, SVM, Decision Tree Classifier and KNN. From the result it's been seen that the random forest gives more accuracy as compared as other five techniques. The data set used is decomposed into a training set and testing set. Here, 70% of the dataset is taken for training and the remaining is considered for testing. From the dataset, it is identified that there are more people suffering from heart disease in the 50-60 age group. This is pictorially represented in figure 2.

Figure 2: Number of heart patients in different age group



From the dataset, it is inferred clearly that a greater number of men are suffering from heart diseases as compared to women. While the range of men suffering from heart disease lies between 80-100, the number of women suffering from heart disease lies between 60-80. This is shown in figure 3. The performance of the classification models on the test data was represented using confusion matrix, per class accuracy and classification accuracy and is given in table 1.

Figure 3: Presence of heart disease based on Gender

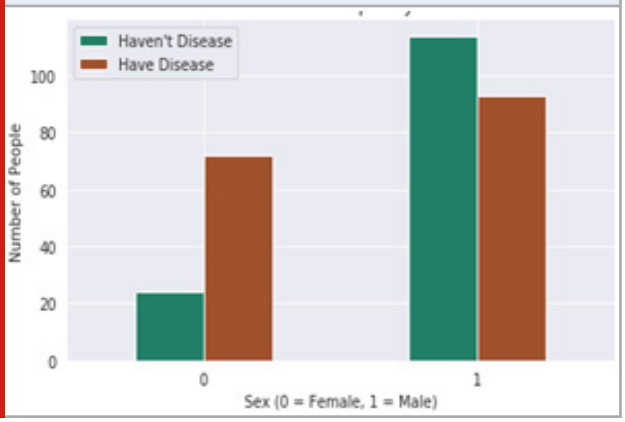


Table 1. Classification Performance of Various Algorithms

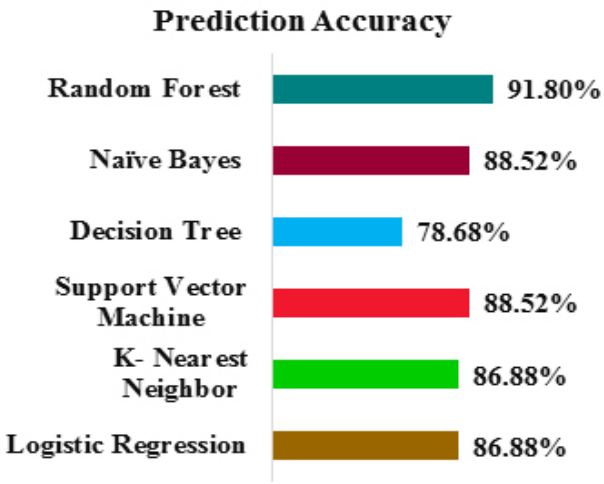
Methodology	Confusion Matrix		Per Class Accuracy
	0 (Female)	1(Male)	
Logistic Regression	23	4	85.18%
	4	30	88.23%
K- Nearest Neighbor	23	4	85.18%
	4	30	88.23%
Support Vector Machine	23	4	85.18%
	3	31	91.17%
Decision Tree	21	6	77.77%
	7	27	79.41%
Naïve Bayes	24	3	88.88%
	4	30	88.23%
Random Forest	25	2	92.59%
	3	31	91.17%

The classification accuracy of various algorithms is graphically represented in figure 4 and the results are presented in table 2. From the results achieved it is inferred that random forest algorithm gives best prediction accuracy than other algorithms.

Table 2. Classification Accuracy of Classifiers

Methodology	Prediction Accuracy
Logistic Regression	86.88%
K- Nearest Neighbor	86.88%
Support Vector Machine	88.52%
Decision Tree	78.68%
Naïve Bayes	88.52%
Random Forest	91.80%

Figure 4: Performance of Classifiers



CONCLUSION

In the field of disease prediction, machine learning plays a significant role. In this paper, various machine learning approaches are used for heart disease forecast. The experimental results illustrate that the Random Forest algorithm achieves the highest accuracy of 91.8% and thus successfully achieving the objective of improving the prediction accuracy. The future work is towards more investigation on evolutionary computation techniques for the problem undertaken and study their performances.

REFERENCES

Amiri, A.M. and Armano, G., 2013, August. Early diagnosis of heart disease using classification and regression trees. In The 2013 International Joint Conference on Neural Networks (IJCNN) (pp. 1-4).

Banu, M.N. and Gomathy, B., 2014, March. Disease forecasting system using data mining methods. In 2014 International conference on intelligent computing applications (pp. 130-133). IEEE.

Chen, M., Hao, Y., Hwang, K., Wang, L. and Wang, L., 2017. Disease prediction by machine learning over big data from healthcare communities. IEEE Access, 5, pp.8869-8879.

Huang, F., Wang, S. and Chan, C.C., 2012, August. Predicting disease by using data mining based on healthcare information system. In 2012 IEEE International Conference on granular computing (pp. 191-194).

Jamgade, A.C. and Zade, S.D., 2019. Disease prediction using machine learning. International Research Journal of Engineering and Technology, 6(5), pp.6937-6938.

Prasad, R., Anjali, P., Adil, S. and Deepa, N., 2019. Heart disease prediction using logistic regression algorithm using machine learning. International journal of Engineering and Advanced Technology, 8, pp.659-662.

Praveen Kumar Reddy, M., Sunil Kumar Reddy, T.,

Balakrishnan, S., Syed Muzamil Basha, & Ravi Kumar Poluru., 2019. Heart Disease Prediction Using Machine Learning Algorithm. International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-10.

Rajathi, N., Kanagaraj, S., Brahmanambika, R. and Manjubarkavi, K., 2018. Early detection of dengue using machine learning algorithms. International Journal of Pure and Applied Mathematics, 118(18), pp.3881-3887.

Raju, C., Philipsy, E., Chacko, S., Suresh, L.P. and Rajan, S.D., 2018, March. A Survey on Predicting Heart Disease using Data Mining Techniques. In 2018 Conference on Emerging Devices and Smart Systems (ICEDSS) (pp. 253-255). IEEE.

Thomas, J. and Princy, R.T., 2016, March. Human heart disease prediction system using data mining techniques. In 2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT) (pp. 1-5). IEEE.

Tikotikar, A., & Kodabagi, M., 2017. A survey on technique for prediction of disease in medical data. In 2017 International Conference on Smart Technologies for Smart Nation (Smart Tech Con) (pp. 550-555). IEEE

Enhanced Accuracy of Heart Disease Prediction using Machine Learning and Recurrent Neural Networks Ensemble Majority Voting Method

Irfan Javid[1], Ahmed Khalaf Zager Alsaedi[2], Rozaida Ghazali[3]
Faculty of Science Computer & Information Technology, Universiti Tun Hussein Onn, Malaysia[1, 3]
Department of Physics, College of Science, University of Misan, Maysan, Iraq[2]
Department of Computer Science & IT, University of Poonch,Rawalakot, AJK, Pakistan[1]

***Abstract*—To solve many problems in data science, Machine Learning (ML) techniques implicates artificial intelligence which are commonly used. The major utilization of ML is to predict the conclusion established on the extant data. Using an established dataset machine determine emulate and spread them to an unfamiliar data sets to anticipate the conclusion. A few classification algorithm's accuracy prediction is satisfactory, although other perform limited accuracy. Different ML and Deep Learning (DL) networks established on ANN have been extensively recommended for the disclosure of heart disease in antecedent researches. In this paper, we used *UCI* Heart Disease dataset to test ML techniques along with conventional methods (i.e. random forest, support vector machine, K-nearest neighbor), as well as deep learning models (i.e. long short-term-memory and gated-recurrent unit neural networks). To improve the accuracy of weak algorithms we explore voting based model by combining multiple classifiers. A provisional cogent approach was used to regulate how the ensemble technique can be enforced to improve an accuracy in the heart disease prediction. The strength of the proposed ensemble approach such as voting based model is compelling in improving the prognosis accuracy of anemic classifiers and established adequate achievement in analyze risk of heart disease. A superlative increase of 2.1% accuracy for anemic classifiers was attained with the help of an ensemble voting based model.**

Keywords*—Deep learning; machine learning; heart disease; majority voting ensemble; University of California; Irvine (UCI) dataset

I. Introduction

Heart disease is particular reason of millions of worldwide death per year confer to the World heart federation Report of 2018. Stroke or CVDs are medically familiar as Heart disease (HD) along with blood pressure (BP), artery disease (AD) and debilitated heart disease by cause of diminish, blockade or reinforce capillaries that hamper the required amount of blood circulation to brain, heart, lungs and other body parts. Congestive heart failure is the most trivial kind of heart disease in all other categories of cardiovascular disease. In human body, work of blood vessels is to provide blood to the heart. Alternate, there are some other reasons of heart disease as well alike valves in the heart not supply properly and may be the reason of heart failure. Chest pain, anesthesia, jaw pain, neck ache, throat burn and back agony, cramp in upper abdomen are the most prevailing syndromes of heart disease. Withal to curtail imperil of heart disease, there are a few predominant aspects such as inhibited blood pressure, under control cholesterol and legitimate exercise. Particularly, heart disease is diagnose after angina, dilated cardiomyopathy, stroke or congestive heart failure. Thus, it is significant to pay attention to CVDs parameter and turn to doctors.

Moreover, confer to the WHO, people expire around 17.9 million every year due to CVDs which coincide to 31% of all deaths globally [1]. This provoke a demand of acquiring an economical arrangement especially capable to provide preamble appraisal of patient established on comparatively elementary medical tests that are economical to everybody. Machine learning (ML) [2] methods have drawn maximum amount of understanding in research society. As illustrate in diverse ongoing studies ML techniques have eventual offering maximum accuracy in classification as associated to alternative procedures for testimony classification. Carry out spectacular accuracy in prediction is crucial as it can edge to pertinent stability. Different machine learning techniques may varies in prediction accurateness. Therefore, it is demanding to perceive gimmick efficient of generating maximum accuracy in heart disease (HD) prediction. Prediction accuracy adept in the take up work is coordinated with earlier research studies. The uttermost practical appraisal formation approach is ML classification for the here and now along with experimental position. Three machine learning (ML) techniques have been practiced consist of random forest (RF), Support Vector Machine (SVM), k-nearest neighbor (KNN). In biomedical field like in diabetes prediction [3] [4], accomplice of diabetes and CVDs [5], reasoning of diabetes proteins [6], machine learning (ML) has already been practiced. There are the divergent conventional approach to use these fettle data to grab the latent material, but the accuracy of the conventional approach is very low, along with prolonged. So, we require contemporary technology which can backing this complex data to be appraised and grab conducive information. Deep learning (DL) algorithms have the ability to learn features from the provided training data, which outrun extracted features used in traditional machine learning algorithms. There are modernity architectures like recurrent neural network (RNN), convolutional neural network (CNN), Long-short-Term memory (LSTM) and gated recurrent unit (GRU). The extant networks confide on disease definitive approach. For classification of cardiac disease in patient modernity

architectures like LSTM and GRU is applied on the extant dataset to evaluate the performance.

The Cleveland dataset from familiar *UCI* database was used to train and testing ML and DL models. It is substantiate dataset and it is extensively used for testing and training in deep learning (DL) and machine learning (ML) models [7]. The dataset consist of 303 patient records and 14 attribute features that are placed on acclaimed aspects and these features are consider to tie with risk of CVDs. We proposed a new hard voting ensemble method in this paper in which various deep learning and machine learning models are mixed and majority vote method is used to predict the result. By using this technique we can improved the overall accuracy in prediction result while aggregation of models produces collective comprehensive model.

The rest of the paper is formulated as follows. Section II, we have reviewed the earlier relevant work to the heart disease prediction and then in Section III we proposed the convoluted particulars of dataset, DL and ML techniques used and data preprocessing. Section IV shows the results produced by each model as well as the accuracy of the prediction proposed by hard voting model. Conclusion and future enhancement is outlined in Section V.

II. Review of Relevant Works

Deep learning and machine learning is advantageous for a divergent set of complications. One of the major application of these techniques is to predict the vulnerable variable from the values of autonomous variables. Even in the advanced countries one of the major reason of deaths is CVD [8]. In medical field artificial neural network (ANN) has been popularized to produce maximum accuracy [9]. The research conferred in [10] used the similar heart disease data as this study but divergent ML algorithms were enforced. Four discrete classification techniques were used which comprised Decision Tree, Naïve Bayes, Multi-layer perception and C4.5. Each of these models predict heart disease with maximum accuracy of 85.12% in the MLP classifier. Tree algorithms like J48 and Logistic model were implemented to predict CVDs also used the Cleveland HD dataset [11]. An observation of these approaches was conducted and maximum accuracy 84% was achieved with J48 algorithm.

With web base interface an application named "Intelligent Heart Disease Prediction System" was developed based on three classifier: DT, NB and ANN [12]. Several surveys conducted related to the ML utilization in Healthcare applications, especially in heart disease prognosis. The survey [13] conclude that Bayesian classification and DT surpass the others techniques like k-nearest neighbor, artificial neural network and clustering-based classification. Confer to the new study [14] by Kadi et al. has completed a pragmatic research after hands-on 149 papers proclaimed during the period from 2000-2015 for the prognosis of CVDs, DT, SVM and ANN were established to be the most periodically used ML techniques. An extreme machine learning (EML) were also implemented to predict heart disease (HD) by using UCI datasets repository and achieved highest accuracy of 80% [15]. GA and fuzzy logic (Hybrid genetic Fuzzy) approach trained and certified over similar UCI repository dataset with maximum accuracy of 86% [16].

According to [7] Raihan et al. developed an android based application to recommend a mock-up for data compilation for IHD. By practicing the P-value strategy and mobile interface they possessed 787 attributes and establish interrelationship amidst symptoms and Ischemic Heart Disease. They established a compelling correlation amidst features with P-value=0.0001 and Ischemic Heart Disease. Likewise, for scoring the symptoms statistical test chi-square, Fisher's exact test and risk score tree are used. BP algorithm is used to extract attributes and syndromes in recent past 2018 [17]-[21].

In RNN section, LSTM consider as the determination with four important factors (forget gate (f_g), input gate (I_g), output gate (O_g) and cell state) have an ample usage for the image analysis along with text and audio signal analysis but is extensively usage in time series analysis, transcribed analysis, voice recognition and health testimony [22]. The major detriment of the RNN model was vanishing gradient problem, LSTM increased the input and output capability of RNN to solve these issues and it uses logical memory to learn sequence vector. To deal with CVDs data temporal features could be learn by Intelligent Healthcare Platform (IHP) established on attention module based LSTM framework [23]. Moreover, to predict CVDs 4. distinct repositories in conjunction with Cleveland dataset is used [24]. Decision Tree (DT) algorithm is the only algorithm comprises of C4.5 and Fast Tree Decision. Formerly, trained technique is established on every attributes of dataset. Later the best sample from datasets are preferred and used to train the model. This approach enhanced the prediction accuracy of the technique from 76.3% to 77.5% adopting C4.5 (average accuracy from datasets) along with enhancement in average accuracy of Fast Tree Decision from 75.48% to 78.06%.

Furthermore, to achieve highest accuracy in the prediction of CVDs distinct methods were used in contemporary research, a few classification algorithms determine CVDs with low accuracy. In contrast with traditional algorithm, hybrid method (include classification algorithms) have produce high accuracy. Our research work proposed a technique to enhance the accuracy of weak classification algorithms by linking them with rest of the classification algorithms. Thus, this technique enhanced the competence of such algorithms along with prediction accuracy for CVDs. The proposed study using ensemble majority voting techniques is done and the results are figure out. The results are compute to illustrate that aforementioned models can have adequate significant usage in medical field.

III. Experimental Result Analysis

In this paper, the main objective is to demonstrate CVDs prediction system using prior dataset. The purpose of this research is to use dataset which reflect real life data and grant the prediction system to conclude to any advanced data.

A. Dataset Features Information

For the experiment *UCI* Cleveland heart dataset repository has been used. The most effective 14 attributes were found amongst the 76 based on the comprehensive experiment. The

Cleveland dataset consist of most dominant 14 attributes and 303 samples. Along with 8 absolute features and 6 numeric features. Table I depicts the description of dataset.

In this dataset selected patients had age from 29 to 77. The value 0 is used to depict the female patients and value 1 is used to depict the male patients. There are 3-types of chest pain might be an indicators of heart disease. Typical angina type-1 is because of the blocked heart arteries due to decreased blood discharge to the heart muscles. The basic reason of type 1 angina is mental or emotional stress. And, second type occurs due to numerous reasons but sometime it may not be the reason of actual HD are known as Non-angina chest pain. The next feature is trestbps depicts the readings of resting blood pressure. Cholesterol level is depicted by Chol. Fasting blood sugar level is represented by Fbs. If Fbs is above 120 mg/dl then the value 0 is assigned and value 1 depicts if the Fbs is below the 120 mg/dl. Resting electrocardiography result is represented as Restecg. Maximum heart rate is represented by thalach, exercise cajoled by angina reported as 0 depicts no pain and 1 depicts pain is represented by exang, ST depression is cajoled by exercise is represented as oldpeak, Peak exercise slope ST segment is depicts by slope, number of major vessels colored by fluoroscopy is represented by ca, exercise test duration is represented by thal and the last one target is as class attribute. Class attribute value is used to distinguish the patient with heart disease and patient with no heart disease. Value 1 depicts patients with heart disease and value 0 depicts normal.

A correlation value was determined among every attributes of dataset and the target diagnosis in order to evaluate the data. Oldpeak, Exang, cp and thalach features have the highest correlated value with target feature. Table II depicts the correlated value with target attribute. This is very helpful in making an analysis against the data that is being handle with.

Furthermore, a heat map is also used to show the clear analysis of the correlation among all the attributes in Fig. 1.

Along with, a bar chart depicts in Fig. 2 gender dissemination of samples in *UCI* Cleveland dataset. The male percentage is almost 68.3% and percentage of females is 31.7% in dataset.

Moreover, histograms are devise for discrete features data visualization to depict the marginal features distribution compared for disease and not disease as represented in Fig. 3 to Fig. 8. It is observed that all the discrete features acquire normal distribution. Age vs. Thalach is shown in Fig. 9.

For the age distribution attribute, Fig. 3 represents the people with CVDs and people with no CVDs commonly. It can be viewed that maximum measurements exist between 40-52 years old. It is also realized that if age has a relation to having CVDs, then people in age range from 50-52 and 40-41 had a dominant consolidation with heart diseases.

Furthermore, to depict the possibility of any relation, Fig. 4 represents the maximal correlated discrete feature (thalach) is devise adjacent to age. It is observed that heart rate is commonly higher for the people with heart disease as compared to the people with no heart disease. Moreover, maximal heart rate decreased noted to a -ve correlated value of -0.3 as age increased. It is represented previously in Fig. 1.

TABLE. I. FEATURES DESCRIPTION OF THE CLEVELAND HEART DISEASE DATASET

S#	Features	Features Description	Data Type
1	Age	Patient age (completed in years)	Numeric
2	Sex	Gender of the patient [0= Female, 1= Male]	Binary
3	Cp	Type of chest pain classified in to four values [1- Typical angina Type, 2- Atypical angina Type, 3- Non-angina pain]	Nominal
4	Trestbps	Level of the patient blood pressure at resting mode in mm/Hg	Continuous
5	Chol	Cholesterol serum, mg/dl	Numeric
6	Fbs	Level of blood sugar on fasting (>120 mg/dl): 1 depict in case of true & 0 depict in case of false.	Binary
7	Restecg	At resting result of ECG is depict in three different values: 0 represented Normal state, 1 represented abnormality in ST-T wave, 2 having LV hypertrophy defined	Nominal
8	Thalach	Maximum rate of Heart recorded	Continuous
9	Exang	Angina-induced by exercise (1 represent ‘yes’ and 0 represent ‘no’	Binary
10	Old peak ST	Exercise tempted ST depression comparative to rest state	Continuous
11	Slope	During peak exercise measured the ST segment in terms of slope represent in 3 values: [1. Up-sloping, 2. Flat, 3. Down-sloping]	Continuous
12	Ca	Ranges from 0-3 represent the number of vessels colored by fluoroscopy	Nominal
13	Thal	Status of the heart: [3. Normal, 6. Fixed defect, 7. Reversible defect]	Discrete
14	Target	Diagnosis represent in two categories: [0= Well, 1= possibility HD]	Binary

TABLE. II. CORRELATED VALUES WITH TARGET ATTRIBUTE ANALYSIS

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
age	1.000000	-0.098447	-0.068653	0.279351	0.213678	0.121308	-0.116211	-0.398522	0.096801	0.210013	-0.168814	0.276326	0.068001	-0.225439
sex	-0.098447	1.000000	-0.049353	-0.056769	-0.197912	0.045032	-0.058196	-0.044020	0.141664	0.096093	-0.030711	0.118261	0.210041	-0.280937
Cp	-0.068653	-0.049353	1.000000	0.047608	-0.076904	0.094444	0.044421	0.295762	-0.394280	-0.149230	0.119717	-0.181053	-0.161736	0.433798
Trestbps	0.279351	-0.056769	0.047608	1.000000	0.123174	0.177531	-0.114103	-0.046698	0.067616	0.193216	-0.121475	0.101389	0.062210	-0.144931
chol	0.213678	-0.197912	-0.076904	0.123174	1.000000	0.013294	-0.151040	-0.009940	0.067023	0.053952	-0.004038	0.070511	0.098803	-0.085239
fbs	0.121308	0.045032	0.094444	0.177531	0.013294	1.000000	-0.084189	-0.008567	0.025665	0.005747	-0.059894	0.137979	-0.032019	-0.028046
Restecg	-0.116211	-0.058196	0.044421	-0.114103	-0.151040	-0.084189	1.000000	0.044123	-0.070733	-0.058770	0.093045	-0.072042	-0.011981	0.137230
thalach	-0.398522	-0.044020	0.295762	-0.046698	-0.009940	-0.008567	0.044123	1.000000	-0.378812	-0.344187	0.386784	-0.213177	-0.096439	0.421741
Exang	0.096801	0.141664	-0.394280	0.067616	0.067023	0.025665	-0.070733	-0.378812	1.000000	0.288223	-0.257748	0.115739	0.206754	-0.436757
Oldpeak	0.210013	0.096093	-0.149230	0.193216	0.053952	0.005747	-0.058770	-0.344187	0.288223	1.000000	-0.577537	0.222682	0.210244	-0.430696
Slope	-0.168814	-0.030711	0.119717	-0.121475	-0.004038	-0.059894	0.093045	0.386784	-0.257748	-0.577537	1.000000	-0.080155	-0.104764	0.345877
Ca	0.276326	0.118261	-0.181053	0.101389	0.070511	0.137979	-0.072042	-0.213177	0.115739	0.222682	-0.080155	1.000000	0.151832	-0.391724
thal	0.068001	0.210041	-0.161736	0.062210	0.098803	-0.032019	-0.011981	-0.096439	0.206754	0.210244	-0.104764	0.151832	1.000000	-0.344029
Target	-0.225439	-0.280937	0.433798	-0.144931	-0.085239	-0.028046	0.137230	0.421741	-0.436757	-0.430696	0.345877	-0.391724	-0.344029	1.000000

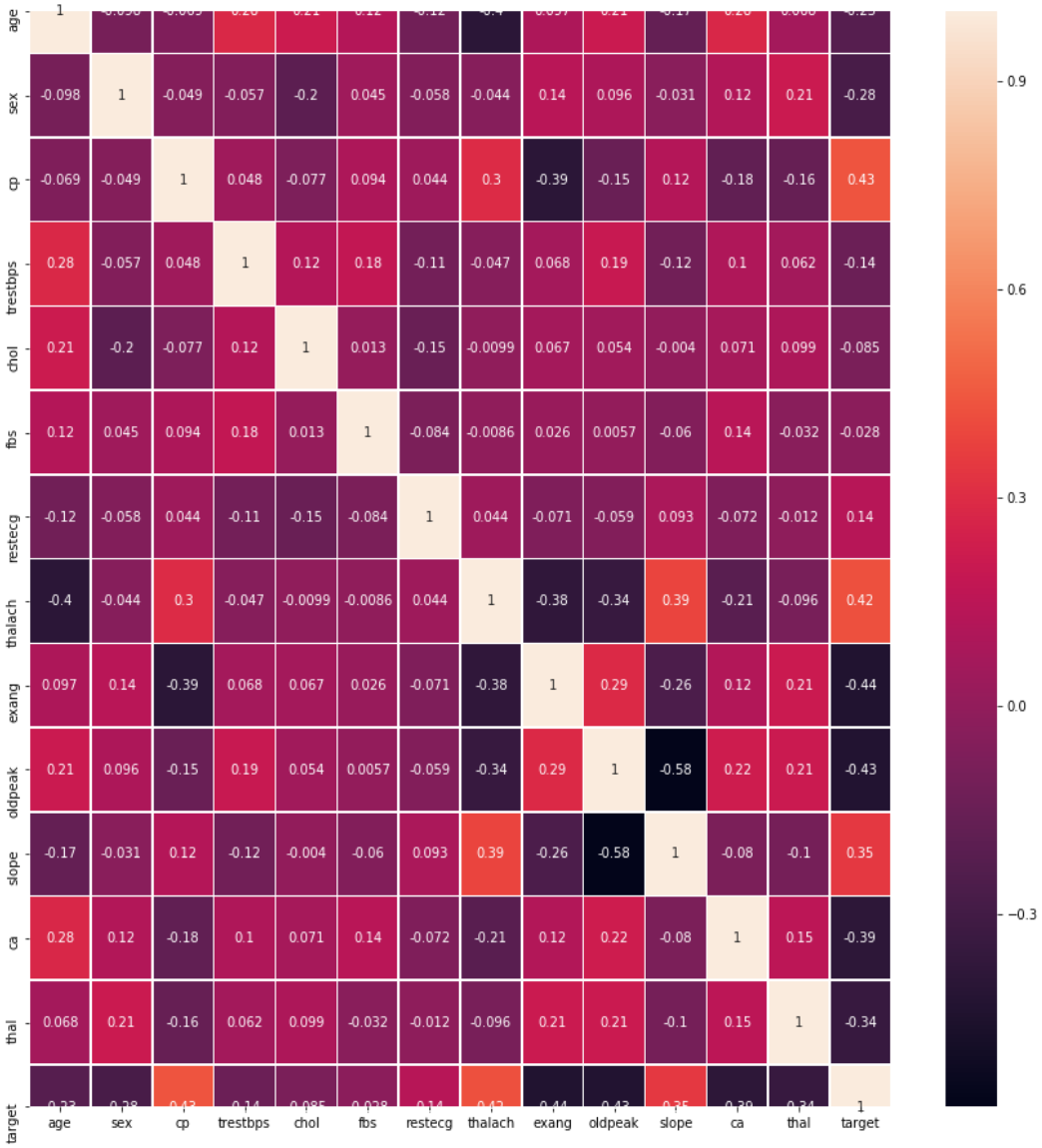


Fig. 1. Correlation-Cross Values Heat Map.

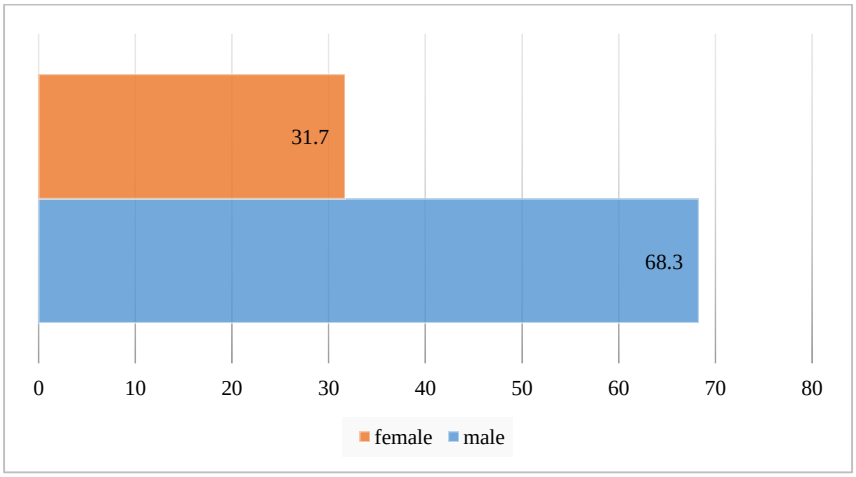


Fig. 2. Gender Dissemination

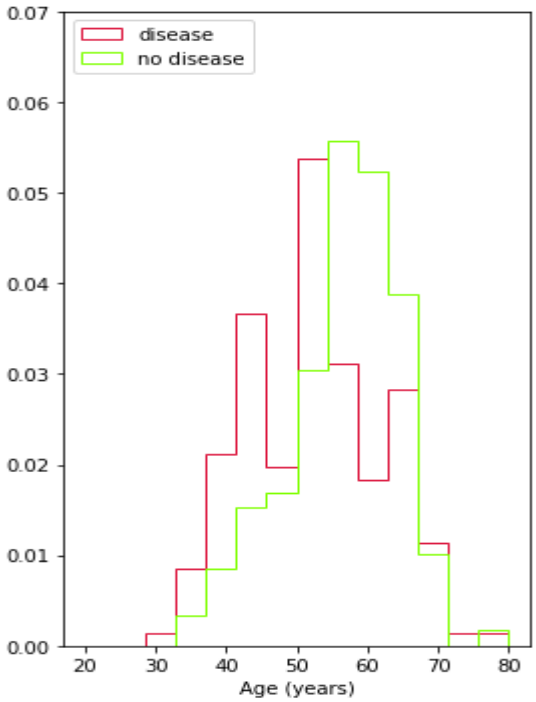


Fig. 3. Age Distribution.

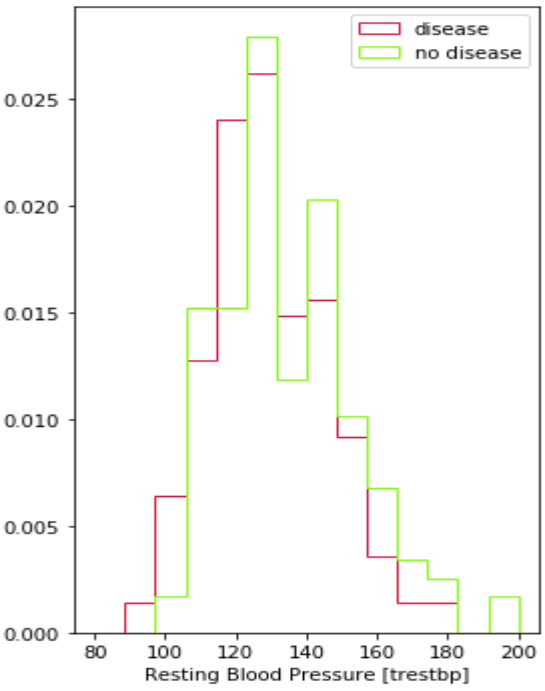


Fig. 4. Blood Pressure Distribution (at Rest).

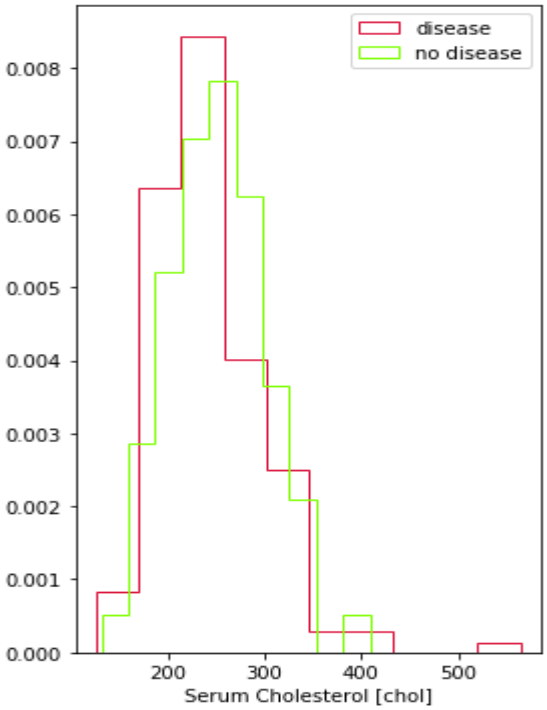


Fig. 5. Serum Cholesterol Distribution.

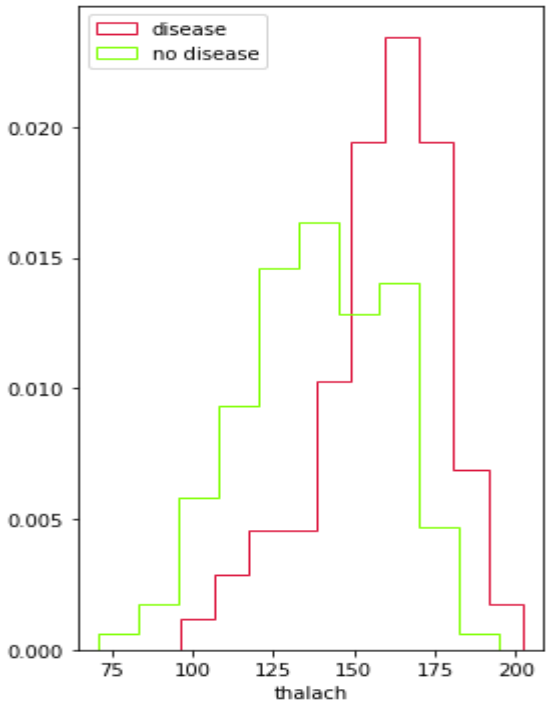


Fig. 6. Maximal Heart Rate Acquire.

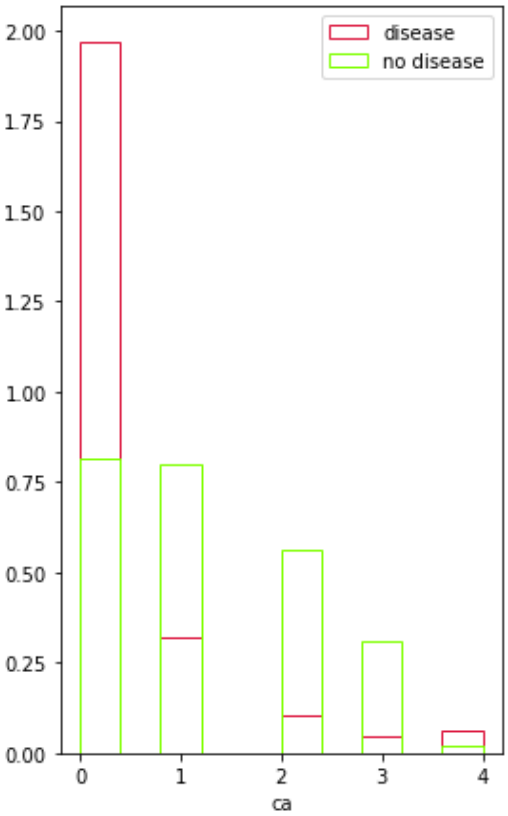


Fig. 7. Calcium Distribution.

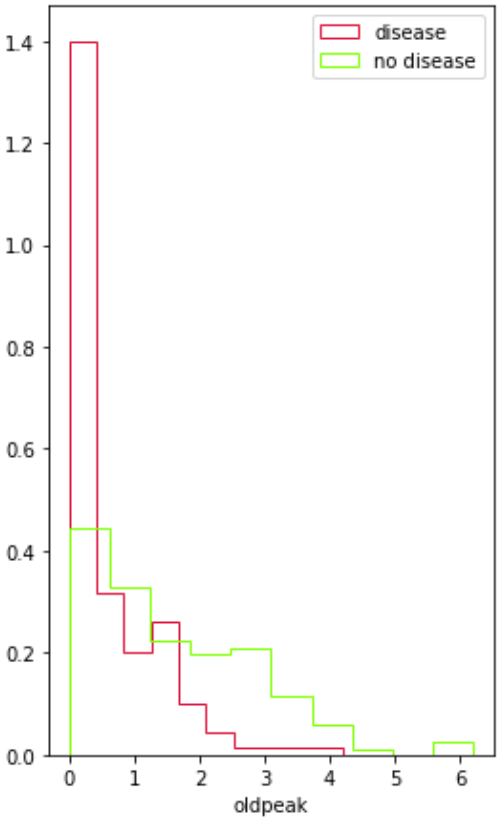


Fig. 8. Old Peak Distribution.

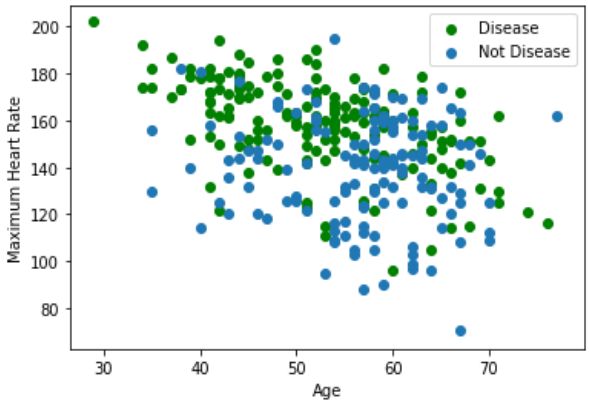


Fig. 9. Age vs. Thalach.

B. Attribute Preprocessing

In order to scale the maximum discrete values by using the Minimum and Maximal normalization approach, the attributes in Cleveland dataset acquire distinct proportions. As shown in eq (1), by using mentioned strategy data is transformed linearly by deducting the smallest and divide over the data range. So, the sample is categorized between zero and one which stimulate learning models to normalize the impact of distinct parameters and form a fair direction between data.

$$Z = \frac{\text{N}-\text{minimum}}{\text{maximum}-\text{minimum}} \quad (1)$$

IV. Machine Learning vs. Deep Learning Models

The Cleveland heart disease dataset has been split into a testing set and train set in the scale of 80% of training set and 20% of testing data and training data set is used to train particular models. Test data is used to check the ability of a models. The working of the particular models are described in the later part.

A. Random Forest Classifier

It is also known as tree based classifier algorithm. Basically, name of the classifier is the indication that the algorithm build a woodland surrounded by huge number of trees. In order to get a maximum accuracy and substantial prediction, RF is an ensemble algorithm comprises on constructing numerous trees and integrate them together. This model used random samples from the training set to build set of decision trees. RFC rerun with numerous samples and compose an eventual decision established on majority voting. To handle missing information RFC is very effective but it is prone to over fitting.

B. Support Vector Machine

SVM was first suggested by Vladimir N. V and Alexey Ya in his study related to theory of statistical learning [25, 26]. For classification and regression purposes a supervised learning machine approach known as support vector machine (SVM) is used. In SVM a technique named trick kernel is used to revamp the information and then it identify most appropriate solution based on these alteration. At present, patient with heart disease and patient with no heart disease are classified by SVM on the basis of binary classification for k_i = +1, -1 additionally. This approach can be protected for

classification in multiple classes by formulating two-multiclass classifiers [25]. A support vector machine classifier is a best approach to get reprieve hyper-plane which lie between two classes [27]. This reprieve clear hyperactive plane has numerous adequate statistical aspects. Finally, slack fickle is very informative to provide adversities of noisy data.

C. K-NN Classifier

The third classifier that was presented is the K-NN algorithm. The main purpose of this algorithm is to find the distance between the current sample along with all the trained samples, K depicts the predefined figures of adjacent points which are used for voting to the current test data's class. Certainly, classification follow established on the more classes of the K data points elected. On the bases of Grid-Search-CV more accurate results are produced and the predefined number for K in this study was selected to be 7.

D. Long-Short term Memory (LSTM)

LSTM was first proposed by Hochreiter al. is a special kind of Recurrent Neural Network (RNN) [28]. LSTM have two distinct states passed between the neurons – the cell state and the hidden layer. Cell state act as short term memory while hidden layer carry the long-term memory, commonly. There was a vanishing gradient problem with original RNN model. Therefore, RNNs are not suitable for long-term dependency data calculations. The vectors in the LSTM are added to the current node on the support of standard RNN model, which helps to solve the problems of RNN with long-term data calculations. Furthermore, LSTM model has been extensively used. LSTM layers consist of three vectors i.e., a forget vector, an input vector, and an output vector. With the passage of time many researchers proposed trivial changes to the standard LSTM model. One of the most attractive LSTM variant "peephole-connections" was introduced by Gers et al. [29]. There are numerous adaptations with small changes regarding the gated structure in the LSTM units. Here we will consider the one proposed by Graves et al. [32].

$$i_t = \sigma\,(W_{xi}\,x_t + W_{hi}\,h_{t-1} + W_{ci}\,c_{t-1} + b_i) \quad (2)$$

$$r_t = \sigma\,(W_{xr}\,x_t + W_{hr}\,h_{t-1} + W_{cr}\,c_{t-1} + b_r) \quad (3)$$

$$c_t = r_t \,\iota\, c_{t-1} + i_t \,\iota\, \tanh(W_{xr}\,x_t + W_{xc}\,x_t + W_{hc}\,h_{t-1} + b_c) \quad (4)$$

$$o_t = \sigma\,(W_{xo}\,x_t + W_{ho}\,h_{t-1} + W_{co}\,c_{t-1} + b_o) \quad (5)$$

$$h_t = o_t \,\iota\, \tanh(c_t) \quad (6)$$

where ι represent element wise product and r, o, i are the forget vector, output vector and input vector respectively. It is observed that the gating structure regulates how the new input and previous hidden state value must be unite to produce the new hidden state value.

The most attractive variant of LSTM is gated-recurrent unit (GRU) was introduced by Chung et al. [30]. The idea was to combine forget vector and input vector as single update vector. In GRU, cell state and hidden state are also merges and make some numerous changes as well. The GRU support the long term sequences and also carry the long-term memories. Therefore, proposed GRU architecture is simpler and most attractive than the original LSTM model.

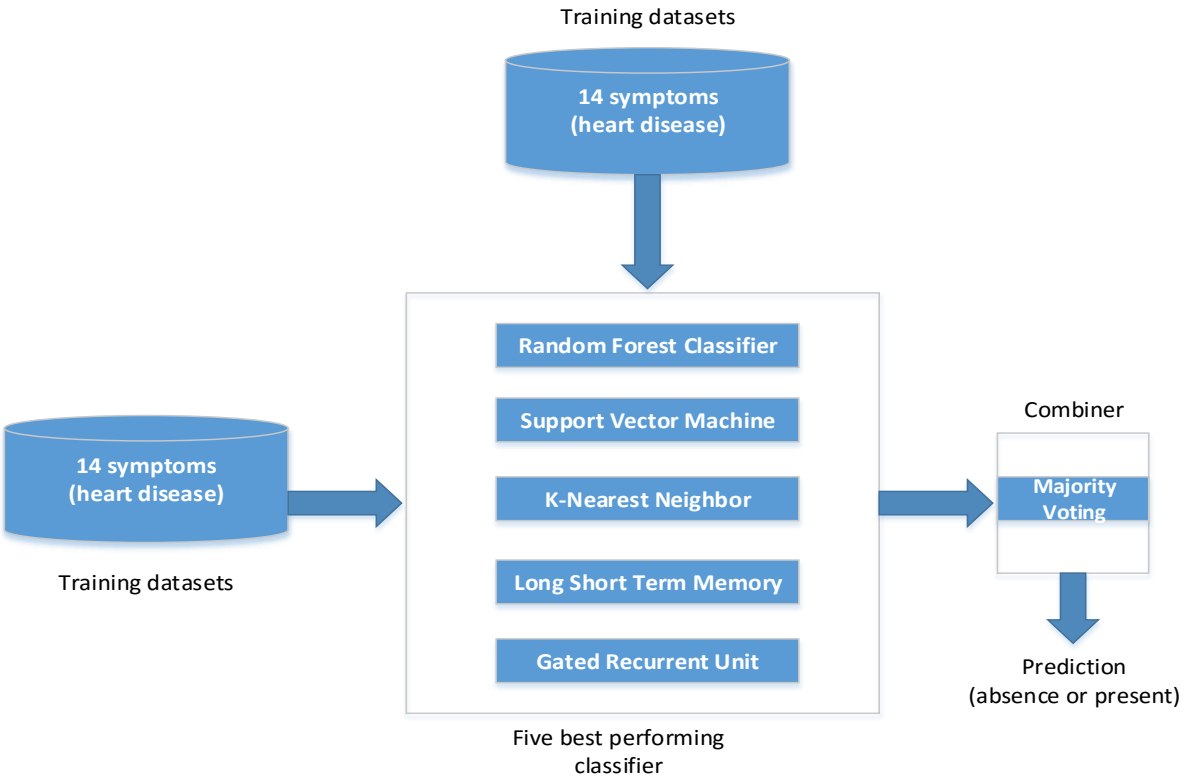


Fig. 10. Workflow of the Proposed Ensemble Vote based Model.

E. Gated Recurrent Unit (GRU)

Cho et al. [30] proposed another gating structure known as GRU (gated recurrent unit) with the purpose to carry long-term dependencies from the calculations within the GRU neuron to produce the hidden state. GRU have only one hidden state conveyed between time steps. Following are the equations determined by Chung et al. [31].

$$r_t = \sigma (W_r x_t + U_r h_{t-1}) \quad (7)$$

$$\tilde{h}_t = \tanh (W x_t + U(r_t \square h_{t-1})) \quad (8)$$

$$z_t = \sigma (W_z x_t + U_z h_{t-1}) \quad (9)$$

$$h_t = (1- z_t) h_{t-1} + z_t \tilde{h}_t \quad (10)$$

Where r and z are commonly the reset and update gates. It can be observed that, GRU is most simple than LSTM, and performance is far better in different experiments. In [31], Chung et al. provide a comparison related to the performance of original RNN, LSTM and GRU, using numerous datasets. It was observed that gated recurrent unit surpass the other techniques in different situations.

F. Ensemble Classifier

At the end, five models aforementioned are unite in an ensemble method where hard voting (majority vote of the models) technique is used for classification. The voting is based on the prediction of each model about each sample and final prediction is based on the majority votes, one that obtains more than 50% of the votes.

The independent classifiers output is united and plays an important role in the final output prediction of an ensemble system. As shown in the Fig.10. Therefore, one of the interesting research study is combination of classifiers in ensemble system. Majority voting approach is extensively used method for labeling the output [33]. In case of discrete outputs, like linear combination, a maximum, minimum, average or any other alternate like derriere possibilities may be used. Many times a classifier may be used as a meta-classifier for uniting outputs of ensemble-members. Due to better performance of majority voting approach over other linear and meta-classifiers has been applied in this work. Therefore, majority voting rule lies in 3 categories: (1) Unanimous-Voting method, here every models must acknowledge the prediction, (2) simple majority method, here prediction required to be partially higher than 50% of classifiers, and (3) majority voting method, here maximum figures of votes is required for the ensemble-decision. If the output of the individual classifier is independent than the majority voting rule combiner constantly enhance the prediction accuracy [34]. Suppose that a class define outputs of classifier O_i are shown as d-dimensional binary vectors:

$$[O_{i,1} \dots O_{i,d}] \in \{0,1\}^d, \; i=1,\dots\dots,N \quad (11)$$

Where $O_{i,I}$=1, if classifier O_i label y in w_j , and 0 differently. The majority voting method would provide an ensemble decision for class w_k , if the below equation is satisfied:

$$\sum_{i=1}^{N} O_{i,k} = \max^{c} j=1 \sum_{i=1}^{N} O_{i,j} \quad (12)$$

If we have 2 classes (c=2), the majority voting method correspond with simple majority approach (50% of vote +1). According to the equation (4) majority voting approach would predict an accurate class define at least [N/2+1] classifiers correctly predict the define class [35]. In our proposed research work, N = 5, it observes that our proposed approach would be able to predict correctly if more than half (at least 3 classifiers) predict the define class correctly.

V. Experimental Results and Discussion

The first classifier Random forest, to study unseen data prediction was run on the test dataset so that the approach has never overcome. Default parameters of the approach are used to run the early test and composed an accuracy of 83.6%. Along with, attributes importance was calculated in this approach and most important attributes were (ca, thalach, oldpeak). Confusion matrix obtained from this approach is shown in Fig. 11.

Also, the second classifier was Support Vector Machine (SVM) algorithm. To run the unseen test dataset the approach was developed with the default parameters. The prediction accuracy of this model was 81.31%. In Fig. 12, confusion matrix obtained from this classifier is depicted.

The third approach, known as K-Nearest Neighbor model. To run the unseen test dataset using default parameters we developed the model. The prediction accuracy get out to be 82.8%. Fig. 13 depicts confusion matrix obtained from this algorithm.

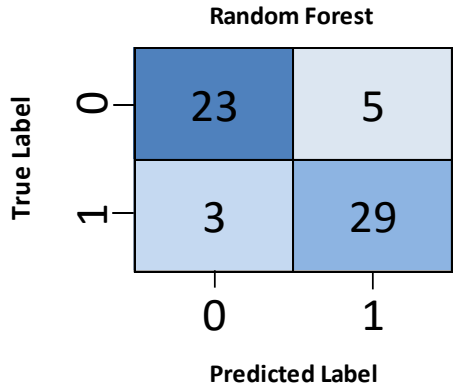


Fig. 11. Random Forest Model Confusion-Matrix.

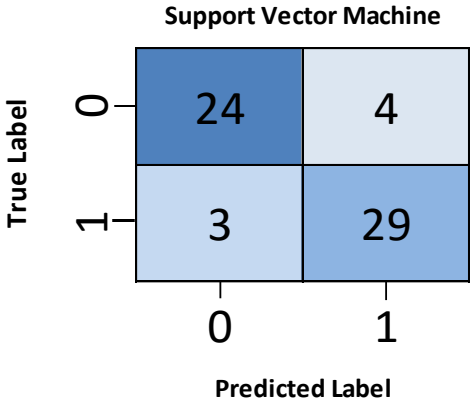


Fig. 12. SVM Model Confusion-Matrix.

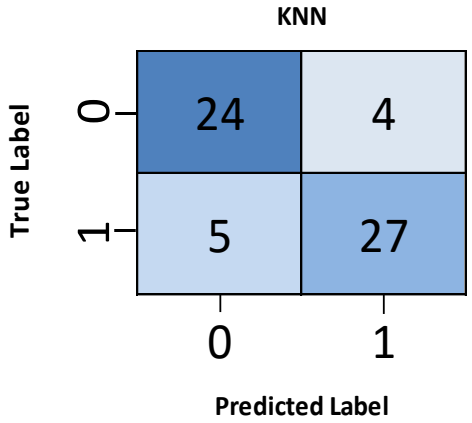


Fig. 13. K-Nearest Neighbor Model Confusion-Matrix.

Additionally, fourth approach that was developed known as LSTM model. Using the default parameters this approach was developed and classification established based on the hidden data test set. The prediction accuracy get out to be 81.31%. In Fig. 14, Confusion matrix obtained from this model is depicted.

Finally, the fifth approach that was developed was the GRU model. Using the default parameters this approach was developed and classification established based on the hidden data test set. The prediction accuracy get out to be 81.46%. Fig. 15 depicts the confusion matrix obtained from this model.

We have noticed that Random Forest and K-NN are constantly provide better prediction accuracy as compared to other classification models. The performance of the each model in accuracy prediction of Heart-Disease as shown in the Fig. 16.

Certainly, the overall prediction accuracy of this study after organizing the Hard Voting ensemble-method get out to be 85.71% which is treated a fairly required accuracy that can be further developed upon in future.

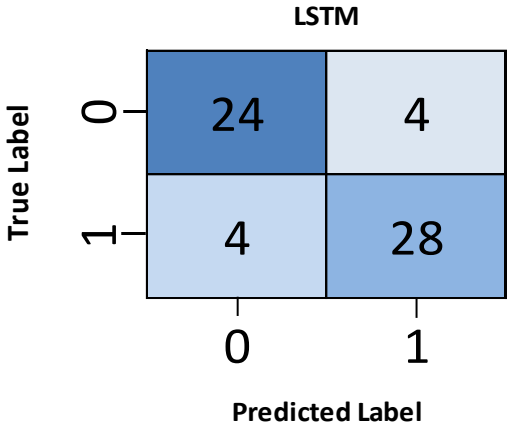


Fig. 14. LSTM Model Confusion-Matrix.

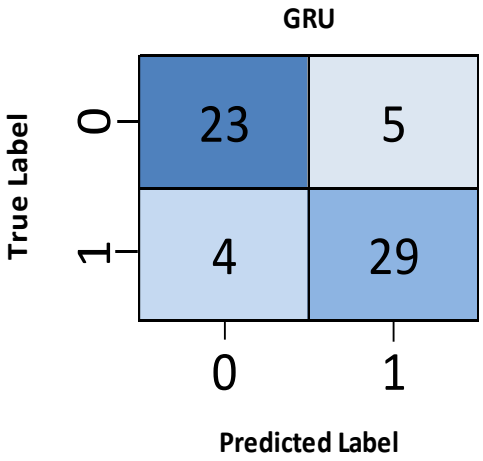


Fig. 15. GRU Model Confusion-Matrix.

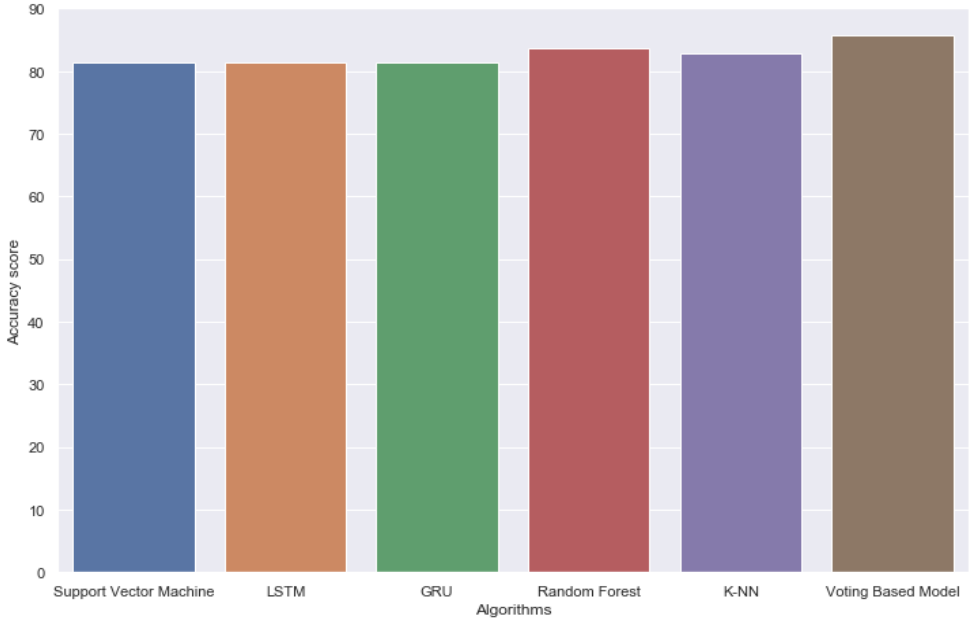


Fig. 16. The Performance Study of different ML and DL Models.

VI. Conclusion

To save the life of the human beings, early prediction of heart disease plays significant role. Here, in this paper we presented a ML and DL ensemble models that united multiple ML and DL models in order to give a maximum accuracy and vigorous model for the prediction of any possibility of having heart disease. Table III depicts the prediction accuracy comparison of Machine learning techniques (i.e. RF, SVM and KNN), deep learning models (i.e. LSTM and GRU) and proposed methodology. This Ensemble approach retained 85.71% accuracy, which surpass the prediction accuracy of every particular model. This approach may be very useful to assist the doctors to investigate the patient cases in order to legitimize their prescription. The future work of this study can be performed with different mixtures of ML and DL models to better prediction.

TABLE. III. Prediction Accuracy Comparison of the Models

Model Name	**Accuracy**
Random Forest	83.6%
Support Vector Machine	81.3%
K-NN	82.8%
LSTM	81.31%
GRU	81.46%
Hard Voting Ensemble Model	85.71%

References

[1] “Cardiovascular diseases (CVDs),” World Health Organization, 26-Sep2018.[Online]. Available: https://www.who.int/cardiovascular_ diseases /en/. [Accessed: 27-Apr-2019].

[2] Bache K, Lichman M (2013) UCI machine learning repository [http://archive.ics.uci.edu/ml]. University of California, School of Information and Computer Science. Irvine, CA.

[3] Bansal A, Agarwal R, Sharma R (2015) Determining diabetes using iris recognition system. Int J Diabetes Dev Ctries 35(4):432–438.

[4] Kalaiselvi C, Nasira GM (2015) Classification and prediction ofheart disease from diabetes patients using hybrid particle swarm optimization and library support vector machine algorithm. Int J Comput Algorithm 4(1):2278–2397.

[5] Bhramaramba R, Allam AR, Kumar VV, Sridhar G (2011) Application of data mining techniques on diabetes related proteins. Int J Diabetes Dev Ctries 31(1):22–25.

[6] King RD (1992) Statlog databases. Department of Statistics and Modelling Science, University of Strathclyde, Glasgow.

[7] Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[8] Vanisree K, JyothiSingaraju. Decision support system for congenital heart disease diagnosis based on signs and symptoms using neural networks. Int J Comput Appl April 2011;19(6). (0975 8887).

[9] L. Baccour, ``Amended fused TOPSIS-VIKOR for classication (ATOVIC) applied to some UCI data sets," Expert Syst. Appl., vol. 99, pp. 115125, Jun. 2018. doi: 10.1016/j.eswa.2018.01.025.

[10] D. Chaki, A. Das, and M. Zaber, “A comparison of three discrete methods for classification of heart disease data,” Bangladesh Journal of Scientific and Industrial Research, vol. 50, no. 4, pp. 293–296, 2015.

[11] R. G. Saboji, “A scalable solution for heart disease prediction using classification mining technique,” 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), 2017.

[12] S. Palaniappan, R. Awang, in: Intelligent heart disease prediction system using data mining techniques, 2008 IEEE/ACS International Conference on Computer Systems and Applications, 2008, pp. 108–115.

[13] J. Soni, U. Ansari, D. Sharma, S. Soni, Predictive data mining for medical diagnosis: an overview of heart disease prediction, Int. J. Comput. Appl. 17 (8) (2011) 43–48.

[14] I. Kadi, A. Idri, J.L. Fernandez-Aleman, Knowledge discovery in cardiology: a systematic literature review, Int. J. Med. Inform. 97 (2017) 12–32.

[15] S. Ismaeel, A. Miri, D. Chourishi, in: Using the Extreme Learning Machine (ELM) technique for heart disease diagnosis, IEEE Canada International Humanitarian Technology Conference, 2015, pp. 1–3.

[16] M. Raihan, S. Mondal, A. More, and M. Sagor et al., "Smartphonebased ischemic heart disease (heart attack) risk prediction using clinical data and data mining approaches, a prototype design," in International Conference on Computer and Information Technology (ICCIT), pp. 299-303, 2016.

[17] G. Shanmugasundaram, V. M. Selvam, R. Saravanan, and S. Balaji, "An Investigation of Heart Disease Prediction Techniques," in IEEE International Conference on System, Computation, Automation and Networking (ICSCA), pp. 1-6, 2018.

[18] P. Umasankar and V. Thiagarasu, "Decision Support System for Heart Disease Diagnosis Using Interval Vague Set and Fuzzy Association Rule Mining," in International Conference on Devices, Circuits and Systems (ICDCS), pp. 223-227, 2018.

[19] K. G. Dinesh, K. Arumugaraj, K. D. Santhosh, and V. Mareeswari, "Prediction of Cardiovascular Disease Using Machine Learning Algorithms," in International Conference on Current Trends towards Converging Technologies (ICCTCT), pp. 1-7, 2018.

[20] I. K. A. Enriko, M. Suryanegara, and D. Gunawan, "Heart Disease Diagnosis System with k-Nearest Neighbors Method Using Real Clinical Medical Records," in International Conference on Frontiers of Educational Technologies, pp. 127-131, 2018.

[21] H. Kahtan, K. Z. Zamli, W. N. A. W. A. Fatthi, and A. Abdullah et al., "Heart Disease Diagnosis System Using Fuzzy Logic," in International Conference on Software and Computer Applications, pp. 297-301, 2018.

[22] T. Ergen and S. S. Kozat, "Online Training of LSTM Networks in Distributed Systems for Variable Length Data Sequences," IEEE Transactions on Neural Networks and Learning Systems, vol. 29, no. 10, pp. 5159-5165, 2018.

[23] C. Lin, Y. Zhangy, J. Ivy, and M. Capan et al., "Early Diagnosis and Prediction of Sepsis Shock by Combining Static and Dynamic Information Using Convolutional-LSTM," in IEEE International Conference on Healthcare Informatics (ICHI), pp. 219-228, 2018.

[24] El-Bialy, R., Salamay, M., Karam, O. and Khalifa, M. (2015). Feature Analysis of Coronary Artery Heart Disease Data Sets. Procedia Computer Science, 65, pp.459-468.

[25] Vapnik VN, Vapnik V (1998) Statistical learning theory, vol 2.Wiley, New York.

[26] Vapnik V (2000) The nature of statistical learning theory.Springer, Berlin

[27] Burges CJ (1998) A tutorial on support vector machines for pattern recognition. Data Min Knowl Disc 2(2):121–167.

[28] S. Hochreiter and J. Schmidhuber, ``Long short-term memory," Neural Comput., vol. 9, no. 8, pp. 17351780, 1997.

[29] F. A. Gers and J. Schmidhuber, ``Recurrent nets that time and count," in Proc. IEEE-INNS-ENNS Int. Joint Conf. Neural Netw. (IJCNN), vol. 3. Jul. 2000, pp. 189194.

[30] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. (2014). “Empirical evalua-tion of gated recurrent neural networks on sequence modeling." [Online]. Available: https://arxiv.org/abs/1412.3555.

[31] Chung J, Gulcehre C, Cho K, Bengio Y. arXiv preprint arXiv:1412.3555.

[32] Graves A, Mohamed A, Hinton G. Speech recognition with deep recurrent neural networks. In ICASSP2013, pages 66456649. IEEE, 2013

[33] J. Kittler, M. Hatef, R.P. Duin, J. Matas, On combining classifiers, IEEE Trans. Pattern Anal. Mach. Intell. 20 (3) (1998) 226–239.

[34] J. Kittler, M. Hatef, R.P. Duin, J. Matas, On combining classifiers, IEEE Trans. Pattern Anal. Mach. Intell. 20 (3) (1998) 226–239.

[35] L.I. Kuncheva, Combining Pattern Classifiers: Methods and Algorithms, John Wiley & Sons, New York, 2004.

Prediction of Heart Disease Using Machine Learning Algorithms

Rachit Misra[1], Pulkit Gupta[2], Prashuk Jain[3]
[1,2,3]*Meerut Institute of Engineering and Technology, India*

Abstract **- heart disease prediction is one among the foremost complicated tasks in medical field. In the era, approximately one person dies per minute thanks to heart condition. Data science plays an important role in processing huge amount of knowledge within the field of healthcare. As heart condition prediction may be a complex task, there is a requirement to automate the prediction process to avoid risks related to it and alert the patient well beforehand. This paper makes use of heart condition dataset available in UCI machine learning repository. The proposed work predicts the probabilities of heart condition and classifies patient's risk level by implementing different data processing techniques like Naive Bayes, Decision Tree, Logistic Regression and Random Forest. Thus, this paper presents a comparative study by analysing the performance of various machine learning algorithms. The trial result verifies that Random Forest algorithm has achieved the highest accuracy of 90.16% compared to other ML algorithms implemented.**

Index Terms **- Decision Tree, Naive Bayes, Logistic Regression, Random Forest, heart condition Prediction.**

1.INTRODUCTION

The work proposed during this paper focus mainly on various data processing practices that are employed in heart condition prediction. Human heart is that the principal a part of the physical body. Basically, it regulates blood flow throughout our body. Any irregularity in the heart can cause distress in other parts of body. Any kind of disturbance to normal functioning of the guts are often classified as a heart condition. In todays times, heart condition is one among the first reasons for occurrence of most deaths. Heart disease may occur thanks to unhealthy lifestyle, smoking, alcohol and high intake of fat which can cause hypertension [2]. According to the planet Health Organization quite 10 million die thanks to heart diseases every single year round the world. A healthy lifestyle and earliest detection are only ways to stop the guts related diseases.

The main challenge in today's healthcare is provision of highest quality services and effective accurate diagnosis [1]. Even if heart diseases are found because the prime source of death within the world in recent years, they are also those which will be controlled and managed effectively. The whole accuracy in management of a disease lies on the right time of detection of that disease. The proposed work makes an effort to detect these heart diseases at early stage to avoid disastrous consequences.

Records of huge set of medical data created by doctors are available for analysing and extracting valuable knowledge from it. Data mining techniques are the means of extracting valuable and hidden information from the massive amount of knowledge available. Mostly the medical database consists of discrete information. Hence, deciding using discrete data becomes complex and hard task. Machine Learning (ML) which is subfield of knowledge mining handles large scale well-formatted dataset efficiently. In the medical field, machine learning are often used for diagnosis, detection and prediction of varied diseases. The main goal of this paper is to provide a tool for the doctors to detect the heart disease as early stage [5]. This successively will help to supply effective treatment to patients and avoid severe consequences. ML plays a really important role to detect the hidden discrete patterns and thereby analyse the given data. After analysis of knowledge ML techniques help in heart condition prediction and early diagnosis. This paper presents performance analysis of varied ML techniques like Naive Bayes, Decision Tree, Logistic Regression and Random Forest for predicting heart condition at an early stage [3].

2.RELATED WORK

A quiet Significant amount of work related to the diagnosis of Cardiovascular Heart disease using Machine

Learning algorithms has motivated this work. This paper contains a brief literature survey. An efficient Cardiovascular disease prediction has been made by using various algorithms some of them include Logistic Regression, KNN, Random Forest Classifier Etc. It can be seen in Results that each algorithm has its strength to register the defined objectives [7].

The model incorporating IHDPS had the ability to calculate the decision boundary using the previous and new model of machine learning and deep learning. It facilitated the important and the most basic factors/knowledge such as family history connected with any heart disease. But the accuracy that was obtained in such IHDPS model was far more less than the new upcoming model such as detecting coronary heart diseases using the artificial neural networks and other algorithms of machine and deep learning. The risk factors of coronary heart disease or atherosclerosis is identified by McPherson et al.,[8] using the inbuilt implementation algorithm using uses some techniques of Neural Network and were just accurately able to predict whether the test patient is suffering from the given disease or not.

Diagnosis and prediction of heart disease and Blood Pressure along with other attributes using the aid of neural networks was introduced by R. Subramanian. A deep Neural Network was Built incorporating the given attributes related to the disease which were able to produce a output which was carried out by the output perceptron and almost included 120 hidden layers which is the basic and most relevant technique of ensuring a accurate result of having heart disease if we use the model for Test Dataset. The supervised network has been advised for diagnosis of heart diseases. When the testing of the model was done by a doctor using an unfamiliar data, the model used and trained from the previous learned data and predicted the result thereby calculating the accuracy of the given model.

3. DATA SOURCE

An Organized Dataset of individuals had been selected Keeping in mind their history of heart problems and in accordance with other medical conditions [2]. Heart disease are the diverse conditions by which the heart is affected. According to World Health Organization (WHO), the greatest number of deaths in middle aged people are due to Cardiovascular diseases. We take a data source which is comprised of medical history of 304 different patient of different age groups. This dataset gives us the much-needed information i.e. the medical attributes such as age, resting blood pressure, fasting sugar level etc. of the patient that helps us in detecting the patient that is diagnosed with any heart disease or not. This dataset contains 13 medical attributes of 304 patients that helps us detecting if the patient is at risk of getting a heart disease or not and it helps us classify patients that are at risk of having a heart disease and that who are not at risk. This Heart Disease dataset is taken from the UCI repository. According to this dataset, the pattern which leads to the detection of patient prone to getting a heart disease is extracted. These records are split into two parts: Training and Testing. This dataset contains 303 rows and 14 columns, where each row corresponds to a single record. All attributes are listed in 'Table 1'

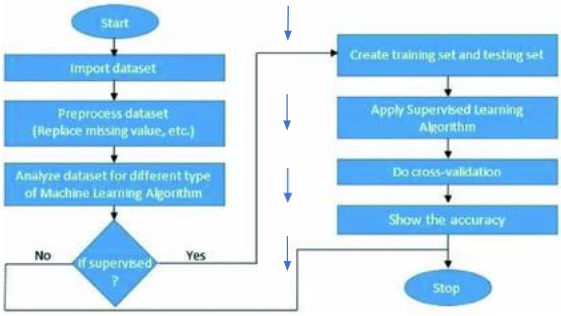
Table 1. Various Attributes used are listed

S. No	Observation	Description	Values
1.	Age	Age in Years	Continuous
2.	Sex	Sex of Subject	Male/Female
3.	CP	Chest Pain	Four Types
4.	Trestbps	Resting Blood Pressure	Continuous
5.	Chol	Serum Cholesterol	Continuous
6.	FBS	Fasting Blood Sugar	< ,or > 120 mg/dl
7.	Restecg	Resting Electrocardiograph	Five Values
8.	Thalach	Maximum Heart Rate Achieved	Continuous
9.	Exang	Exercise Induced Angina	Yes/No
10.	Oldpeak	ST Depression when Workout compared to the Amount of Rest Taken	Continuous
11.	Slope	Slope of Peak Exercise ST segment	up/ Flat /Down
12.	Ca	Gives the number of Major Vessels Coloured by Fluoroscopy	0-3
13.	Thal	Defect Type	Reversible/Fixed/Normal
14.	Num(Disorder)	Heart Disease	Not Present /Present in the Four Major types.

4.METHODLOGY

This paper shows the analysis of various machine learning algorithms, the algorithms that are used in this paper are K nearest neighbors (KNN), Logistic Regression and Random Forest Classifiers which can be helpful for practitioners or medical analysts for accurately diagnose Heart Disease. This paperwork includes examining the journals, published paper and the data of cardiovascular disease of the recent times. Methodology gives a framework for the proposed model [13]. The methodology is a process which includes steps that transform given data into recognized data patterns for the knowledge of the users. The proposed methodology (Figure 1.) includes

steps, where first step is referred as the collection of the data than in second stage it extracts significant values than the 3rd is the preprocessing stages where we can explore the data. Data preprocessing deals with the missing values, cleaning of data and normalization depending on algorithms used [15]. After pre-processing of data, classifier is used to classify the pre-processed data the classifier used in the proposed model are KNN, Logistic Regression, Random Forest Classifier. Finally, the proposed model is undertaken, where we evaluated our model on the basis of accuracy and performance using various performance metrics. Here in this model, an effective
Heart Disease Prediction System
(EHDPS) has been developed using different classifiers. This model uses 13 medical parameters such as chest pain, fasting sugar, blood pressure, cholesterol, age, sex etc. for prediction [17].



5.RESULTS & DISCUSSIONS

From these results we can see that although most of the researchers are using different algorithms such as SVC, Decision tree for the detection of patients diagnosed with Heart disease, KNN, Random Forest Classifier and Logistic regression yield a better result to out rule them [23]. The algorithms that we used are more accurate, saves a lot of money i.e. it is cost efficient and faster than the algorithms that the previous researchers used. Moreover, the maximum accuracy obtained by KNN and Logistic Regression are equal to 88.5% which is greater or almost equal to accuracies obtained from previous researches. So, we summarize that our accuracy may be improved due to the increased medical attributes that we used from the dataset we took. Our project also tells us that Logistic Regression and KNN outperforms Random Forest Classifier in the prediction of the patient diagnosed with a heart disease. This proves that KNN and Logistic Regression are better in diagnosis of a heart disease. The following 'figure 2', 'figure 3', 'figure 4', 'figure 5' shows a plot of the number of patients that are been segregated and predicted by the classifier depending upon the age group, Resting Blood Pressure, Sex, Chest Pain:

- Risk of Heart Attack
- No Risk of Heart Attack

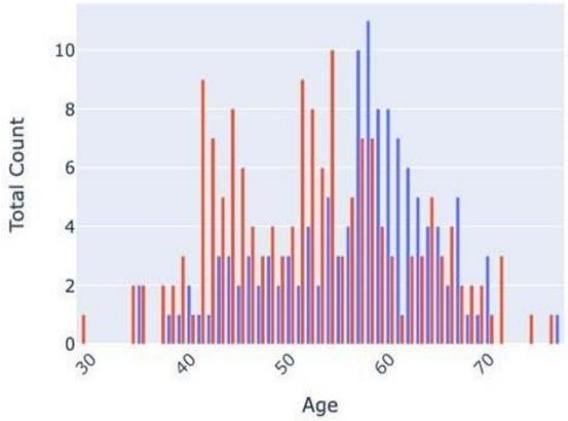


Figure 2. Shows the Risk of Heart Attack on the basis of their age.

TABLE 1. Values Obtained for Confusion Matrix Using Different Algorithm

Algorithm	True Positive	False Positive	False Negative	True Negative
Logistic Regression	44	10	8	30
Naïve Bayes	42	12	6	62
Random Forest	44	10	12	56
Decision Tree	50	4	8	60

TABLE 2. Analysis of Machine Learning Algorithm

Algorithm	Precision	Recall
Decision Tree	0.845	0.823
Logistic Regression	0.857	0.882
Random Forest	0.937	0.882
Naïve Bayes	0.837	0.911

6. CONCLUSION

With the increasing number of deaths thanks to heart diseases, it's become mandatory to develop a system to predict heart diseases effectively and accurately.

The motivation for the study was to seek out the foremost efficient ML algorithm for detection of heart diseases. This study compares the accuracy score of Decision Tree, Logistic Regression, Random Forest and Naive Bayes algorithms for predicting heart condition using UCI machine learning repository dataset. The result of this study indicates that the Random Forest algorithm is the most efficient algorithm with accuracy score of 90.16% for prediction of heart disease. In future the work are often enhanced by developing an internet application supported the Random Forest algorithm also as employing a larger dataset as compared to the one utilized in this analysis which can help to supply better results and help health professionals in predicting the guts disease effectively and efficiently.

REFERENCES

[1] Sonam, A.M. “Predictions of Heart Condition Using Machine Learning Algorithms” in International Journal of Advanced Engineering, Management and Science (IJAEMS) June2016 vol-2

[2] Kelley “Heart Disease: Causes, Prevention, and Current Research” in JCCC Journal

[3] Costas Sideris, Mohammad, Haik K, "Remote Health Monitoring Outcome Success Prediction using Baseline and First Month Intervention Data" in IEEE Journal of Biomedical and Health

[4] Po Athi, Brad Jenkins, Marcia Johansson, Miguel Labrador "A Mobile Health Intervention to Improve Self-Care in Patients Having Heart Failure: Pilot Randomized Control Trial" in JMIR Cardio 2017, vol. 1, issue 2, pgno:1

[5] Dh, J K. Al, Mohamed Ibrahim, Mohammad. Naeem "The Utilization of Machine Learning Approach for Medical Data Classification" in Annual Conference on New Trends in Information & Communication Technology Applications - march2017

[6] Applying k-Nearest Neighbour in Diagnosing Heart Disease Patients Mai Shou, Tim Turner, and Rob Stocker International Journal of Information and Education Technology, Vol. 2, No. 3, June 2012

[7] Amu, J., Pad, S., Nandhini, R., Kavi, G., D, P., Venkata, V.S.K., "Recursive ant colony optimization routing in wireless mesh network", (2016) Advances in Intelligent Systems and Computing, 381, pp. 341-351.

[8] Ala, B.P., Kavitha, A., Amu, J., "A novel encryption algorithm for end-to-end secured fib optic communication", (2017) International Journal of Pure and Applied Mathematics, 117 (19 Special Issue), pp. 269-275.

[9] Amu, J., In, P., B, B., Ananda, B., Ven, T., Prem, K., "An effective analysis on harmony search optimization approaches", (2015) International Journal of Applied Engineering Research, 10 (3), pp.2035-2038.

[10] Amu, J., Kath, P., Reddy, L.S.S., Aa, A., "Assessment on authentication mechanisms in distributed system: A case study", (2017) Journal of Advanced Research in Dynamical and Control Systems, 9 (Special Issue 12), pp. 1437-1448.

[11] Amu, J., Kode, C., Prem, K., Jai, S., Raja, D., Ven, T., Hari, R., "Comprehensive analysis on information dissemination protocols in vehicular ad hoc networks", 6 (2015) International Journal of Applied Engineering Research, 10 (3), pp. 2058-2061.

[12] Amu, P., Reddy, L.S.S., Satyanarayana, K.V.V., "Effects, challenges.

[13] Amu, J., Ila, R., Mo, N., Ravishankar, V., Baskaran, R., Prem, K., "Performance analysis in cloud auditing.

Global Transitions Proceedings 2 (2021) 233–237

Contents lists available at ScienceDirect

Global Transitions Proceedings

journal homepage: http://www.keaipublishing.com/en/journals/global-transitions-proceedings/

Analysis of performance metrics of heart failured patients using Python and machine learning algorithms

Rachana R Sanni[a,*], H.S. Guruprasad[a]

[a] Department of ISE, BMSCE, Bangalore- 560019, India, 7829315318

ARTICLE INFO

Keywords:
Analyzing
Disease
Heart Failure
Machine Learning Algorithms
Patients
Python Programming
Python Scripting

ABSTRACT

Heart is one of the most important organs in Human's body. In life, some changes may happen that may bring various diseases like, blood pressure, sugar, etc. Similarly, heart failure is also a dreadful disease. Heart failure is a serious condition and there is no cure for this disease. It is a situation in which the patient's heart is not pumping the blood well as the normal heart pumps. Heart Failure prediction is a complex task in the medical field. The rates of heart failure have been increasing day by day as the rate of population is also increasing day by day. This paper aims at analyzing the machine learning algorithms based on the percentage of various performance metrics (such as, Accuracy, Precision and Recall). The machine learning methodology is proposed. The most suitable algorithm for each metrics is predicted. It is analyzed using the specific variables in the dataset by using the python programming as well as different supervised machine learning algorithms which include, Decision Tree, Logistic Regression, KNN and Random Forest. Anaconda jupyter notebook is used for implementing python scripting.

1. Introduction

We know that heart is the most important organ in our body; it plays a very vital role in functioning of our body. The heart failure conditions describe abnormal health which affects the heart and all its parts [1]. The rate of the patients who are suffering from this disease is increasing day by day. Due to consumption of high fatty foods such as cheese, cholesterol, etc, and fast foods like, pizza, burger, etc, the rate of getting this disease is high. Hence, this disease influences the heart.

To control the death rates, the preventive measures are suggested by World Health Organization (WHO) after the prediction of rates of heart failured patients. The rates are increasing yearly and hence the preventive measures need to be taken.

To find the death rates, there are different techniques used by World Health Organization (WHO). In this paper, there are different types of machine learning algorithms analyzed to predict the performance metrics.

2. Related Works

The heart failure disease has emerged and has been emerging as one of the most deathful diseases from decades around the world. The death rates are increasing from years. China had the largest death rates last year (2020) followed by Russia, India, America and Indonesia. Some others such as, Japan, France, and Peru had low death rates. Among men, the death rate was underlying 9.6 million and in female, the death rate was underlying 8.9 million. Among the people of age between 30 and 70 years, the death rates were about 6 million. Different unique datasets have been used to predict the heart failure. The supervised and unsupervised innovations have raised the accuracy of diagnostics of models [2]. Due to increase in heart death rates, 17.9 million of people are dying. Hence, to predict the causes of the heart failure, the machine learning models are built on related parameters. To predict the death rate of heart failured patients, the datamining techniques are used [3]. The crucial heart attack disease is a threatening and most common disease. There is a need to predict the occurrence of this disease based on combination of risk factors. Hence, different techniques needs to implemented and compared based on standard metrics [4].

In Medical Field, the heart disease prediction at an early stage is very necessary to save the patient's life. Hence, there should be a tool available to predict the heart disease. This tool is provided to doctors to detect the presence of disease. Hence, performance analysis is done to provide this tool [5].

There are various different machine learning algorithms, among those, one algorithm, KNN, is used to predict the heart disease at an early stage [6].

The heart disease predicts the occurrence of heart failure using Random Forest Algorithm using a dataset and Anaconda Jupiter [7]. The heart failure can be analyzed using various algorithms in machine learning. The algorithms can be compared and best out of those is

* Corresponding author.
E-mail addresses: rachanar.scn19@bmsce.ac.in (R.R. Sanni), guru.ise@bmsce.ac.in (H.S. Guruprasad).

https://doi.org/10.1016/j.gltp.2021.08.028
Received 15 June 2021; Accepted 5 July 2021
Available online 12 August 2021

analyzed [8]. The decision tree algorithm is used to predict the heart disease in a patient. The occurrences of heart disease will be reduced in the patients [9]. The heart failure is predicted through the model with the accuracy through the hybrid random forest with linear model [10].

Heart failure prediction accuracy has been improved using UCI dataset. Many techniques of machine learning have been used to predict the chances of heart failure [11]. To predict the heart disease, the accuracy of different machine learning algorithms is calculated [12]. The WHO (World Health Organization) collects the data from various health centers to predict the count of heart failured patients. It uses many required techniques to count the death rates and brings on some preventive measures to be taken to prevent the cause of heart failure.

The EHMS (Electronic – Hospital Management System) which is also called as HIS (Hospital Information System) was designed to manage the entire hospital administration. HMS/HIS are purely responsible for managing the details regarding the patients [13].

The medical records of the patients are stored in database and risk of databases is secured using the electronic database [14–16].

3. Methods

This paper uses the "heart_failure_clinical_records_dataset.csv" dataset that has reviews on heart failured patients. Dataset consists of 377,650 views [17–19]. The records have the data format for data in the dataset as shown in the below Table 1.

The Proposed system has the process diagram as shown in the below Fig. 1. The following are the steps carried out.

Table 1
Dataset Set Description

Column_ID	Description
Age	Age
Anaemia	Decrease of red blood cells or haemoglobin (Boolean)
creatinine_phosphokinase	Level of the CPK enzyme in the blood (mcg/L)
Diabetes	If the patient has diabetes (Boolean)
ejection_fraction	Percentage of blood leaving the heart at each contraction
high_blood_pressure	If the patient has hypertension (Boolean)
Platelets	Platelets in the blood (kiloplatelets/mL)
serum_creatinine	Level of serum creatinine in the blood (mg/dL)
serum_sodium	Level of serum sodium in the blood (mEq/L)
Sex	Woman or man (binary)
Smoking	If the patient smokes or not (Boolean)
Time	Follow-up period (days)
DEATH_EVENT	If the patient deceased during the follow-up period (Boolean)

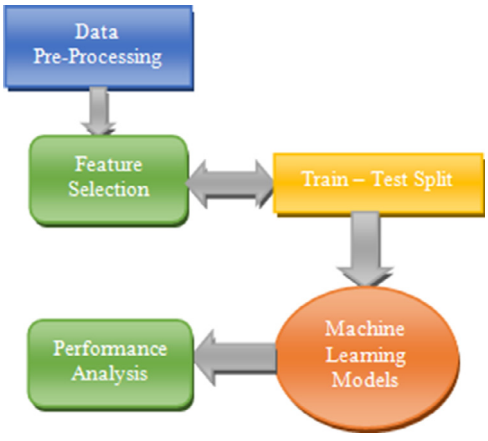


Fig. 1. Proposed system process diagram

3.1. Data Pre-Processing

It is a process of modifying the required columns in a dataset (if the dataset contains raw data) for analyzing the data. For example, the dataset used in this paper had few string formats which represent labels, genders, etc. Data Pre-processing was applied to retrieve numerical data, string data, categorical data appropriately. In this paper, the data preprocessing is carried out to remove the unnecessary values in the columns such as, missing values, false values and null values. Our dataset may be noisy sometimes, so to remove these unnecessary and missing columns, this step is carried out.

3.2. Feature Selection

Feature Selection is a process of selecting the necessary variables to increase the accuracy. The selection of variables during the feature selection can be manual or automatic. In this paper, the necessary features are selected to raise the percentage of accuracy. This step involves selecting the relevant features and discarding the irrelevant features. In this paper, the features regarding the heart failure are selected and access features are removed [20–22].

3.3. Train - Test Split

Train – Test Split is a technique to divide the given dataset into subsets and train further. This technique can be used to rate the performance of a machine learning algorithms. In this paper, the dataset is split to fit into any of the models. The data is splitted into train and test subsets by taking the size of test.

3.4. Machine Learning Models

These models take request in the form of input data, analyze, process and then serve the response. It is the one that take request in the form of input data and then serves the response. The models can be trained over a set of data and the algorithms reason out and learn from this data. In this paper, mathematical models (also called "Supervised Machine Learning Models") are given by; Logistic Regression, Decision Tree, KNN, and Random Forest are used for implementation.

3.5. LR (Logistic Regression)

It is a supervised algorithm used to predict the probability of the target variable. In this algorithm, there is a binary classification of two classes; they may be 0 or 1, true or false. This paper uses this algorithm to check the probability of heart failured. In this algorithm, the predicted value will be categorical. It is one of the mathematical algorithms used for classification.

3.6. DT (Decision Tree)

This is a supervised algorithm where data can be split into subsets based on certain parameter. The decision tree can be used for classification as well as regression. This mathematical model basically gives a graphical representation of all possible solutions to which the decision needs to be made. The graphical representation purely includes a tree with conditions. The decisions might depend on some conditions. These decisions made can be explained easily.

3.7. KNN

KNN is a simple, supervised machine learning algorithm which can be used for solving both the classification as well as regression problems. This algorithm predicts the output based on the similarity measures such as Euclidean distance, Manhatten distance and Q norm distance. It stores the available cases and classifies new cases based on similarity measures.

It is the nearest neighbour from which the votes can be taken. As the value of K varies, the results change.

3.8. RF (Random Forest)

It is a supervised machine learning algorithm used for classification and regression. The decision trees are created on data samples and prediction is got from each of those and then best solution is selected by voting means by using this algorithm. It has simplicity and diversity since it is most used algorithm. This algorithm builds multiple decision trees and merges together. It gives more accurate and stable prediction. It corrects the over fitting of random decision forests to their training set. It is the algorithm trained with "bagging" method.

3.9. Performance Analysis

Performance Analysis is the one used for predicting the algorithm based on various metrics such as accuracy, precision, recall/F1-Score, etc. Performance Analysis aims at comparing the accuracy and performance of machine learning models. The metrics is evaluated with four measures.

3.10. TP [True Positive]

If the positive input (from dataset) is given to the classifier, it gives positive output (predicted value). It predicts the total true positive cases identified correctly. True positive values will be 1(True) for heart failured patients. The predicted value will be value 1 if the true positive value is 1.

3.11. TN [True Negative]

If the negative input (from dataset) is given to the classifier, it gives negative output (predicted value). It predicts the total true negative cases identified correctly. The true negative values will be 0(False) for heart failured patients. The predicted value will be 1(True) if the true negative value is 0(false).

3.12. FP [False Positive]

If the negative input (from dataset) is given to the classifier, it gives positive output (predicted value). It predicts the total false positive cases identified incorrectly. The false positive values will be 0 for heart failured patients. The predicted value will be 1 if the false positive value is 0.

3.13. FN [False Negative]

If the positive input (from dataset) is given to the classifier, it gives negative output (predicted value). It predicts the total false negative cases identified incorrectly. The false negative values will be 0 for heart failured patients. The predicted value will be 0 if the values predicted are false.

The important metrics for performance analysis are accuracy, recall (F1-Score) and precision. These above measures are used to define the metric.

3.14. Accuracy (A)

Accuracy is a performance metric that has the correct predictions for the test data. It gives the percentage of correct predictions for testing the data. In machine learning, accuracy is calculated using the formula as shown in equation (1),

$$(TP + TN)/(TP + TN + FP + FN) \quad (1)$$

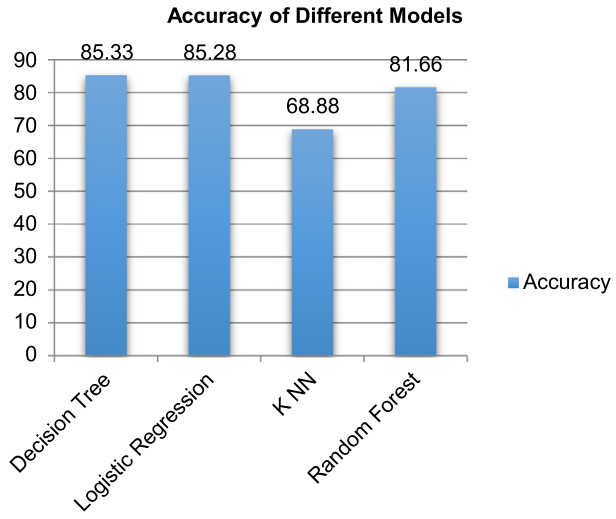


Fig. 2. Accuracy of Different Models

3.15. Recall (R) or F1-Score

Recall metric is used to predict the number of correct samples (all samples identified as positive). It is the fraction of values (results) returned to the total number of values that can be returned. It is the ratio between true positives and all the actual positives. It measures the correctly identified positive samples out of all the actual positive samples. The recall is calculated using the formula as shown in equation (2),

$$TP/(TP + FN) \quad (2)$$

3.16. Precision (P)

Precision which is also called as "positive predictive value". It gives the percentage of true positives. It is the ratio between true positives and all the predicted positives. It measures the correctly identified positive samples out of all positively predicted samples. The precision is calculated using the formula as shown in equation (3),

$$TP/(TP + FP) \quad (3)$$

4. Results

The results obtained by the four machine learning algorithms are analyzed here. The dataset undergoes appropriate pre-processing. The algorithm uses this dataset and performance metrics are shown below.

Comparison of previous paper performance metrics is shown in the Table 2.

Accuracy of different supervised machine learning algorithms is plotted in Fig. 2. The Decision Tree algorithm gives the highest accuracy of 85.33% and the KNN algorithm gives the lowest accuracy of 68.88%.

Precision of different supervised machine learning algorithms is plotted in Fig. 3. The Logistic Regression gives the highest precision of 80.95% and the KNN algorithm gives the lowest precision of 68.88%.

Recall of different supervised machine learning algorithms is plotted in Fig. 4. The KNN algorithm gives the highest recall of 100% and the Decision Tree algorithm gives the lowest recall of 68.88%.

5. Discussion

The prediction of heart failure is analyzed using different algorithms. The algorithms used here are compared and the algorithm with highest rates in the performance metrics is predicted. The algorithms used are supervised machine algorithms.

Table 2
Comparison of the results with other papers

Methods	Accuracy	Precision	Recall/F1-Score
Proposed Model(Decision Tree)	85.33	73	70
Apurb Rajdhan, Milan Sai, Avi Agarwal, Dundigalla Ravi, Dr. Poonam Ghuli (2020)	81.97	-	-
Prasanta Kumar Sahoo, Pravalika Jeripothula (2019)	71	71	70
Proposed Model(Logistic Regression)	85.28	80.95	75.55
Apurb Rajdhan, Milan Sai, Avi Agarwal, Dundigalla Ravi, Dr. Poonam Ghuli (2020)	85.25	-	-
Prasanta Kumar Sahoo, Pravalika Jeripothula (2019)	76	76	75
Ms. Preet Chandan Kaur (2020)	61.45	-	-
Proposed Model(KNN)	68.88	68.88	100
Prasanta Kumar Sahoo, Pravalika Jeripothula (2019)	-	-	81
Proposed Model(Random Forest)	81.66	78.26	97.29

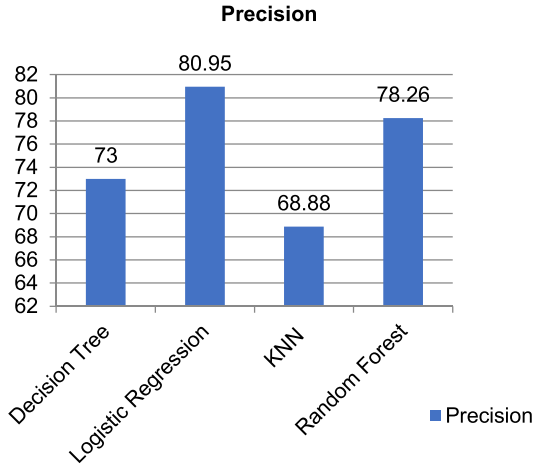


Fig. 3. Precision of Different Models

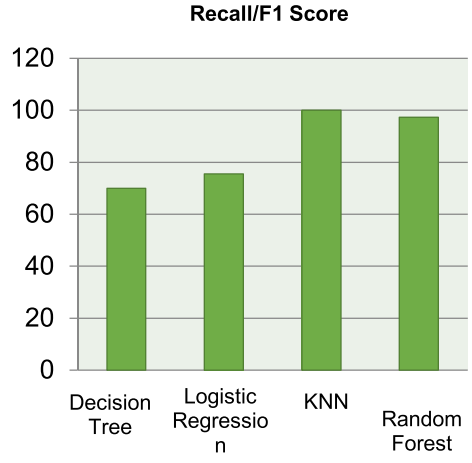


Fig. 4. Recall/F1 Score of Different Models

The expectation of the result is satisfied. A dataset is taken; it is split into train and test. The confusion matrix is obtained by fitting into the respective algorithms. By using the confusion matrix, the four values are obtained. By using those values, the performance metrics values are obtained. To get the result, the metrics of different algorithms are compared and the algorithm with highest rates is the one having the highest performance metrics.

6. Conclusion

The performance metrics, Accuracy, Precision and Recall, obtained by the four machine learning algorithms are compared. These mathematical models are being proposed to get the better analysis on the performance metrics. As per the comparisons shown in the above Table 2, the performance metrics of this paper are more accurate than the previous papers. Hence, the results of performance metrics meet the expectation of implementation. It is observed that the highest accuracy is obtained by decision tree, the highest precision is obtained by Logistic Regression and the highest recall is obtained by KNN. The random forest algorithm gives promising results across all the performance metrics.

References

[1] Nidhi Bhatla, Ludhiana GNDEC, GNDEC, An analysis of heart disease prediction using different data mining techniques, International Journal of Engineering Research & Technology (IJERT) 1 (8) (2012) ISSN: 2278-0181.
[2] U.S. The, National Heart, Lung, and Blood Institute offers a guide to a healthy heart, J. Am. Coll. Cardiol. (2020) news release.
[3] Vijeta Sharma, Shrinkhala Yadav, Manjari Gupta, Heart disease prediction using machine learning techniques, in: 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 2020.
[4] Akash Chandra Patel, Anash Shameem, Sunil Chaurasiya, Manish Mishra, Abhishek Saxena Prof., Prediction of heart disease using machine learning, International Journal of Scientific Development and Research (IJSDR) 4 (2019) www.ijsdr.org. ISSN: 2455-2631 ©IJSDR |.
[5] Apurb Rajdhan, Milan Sai, Avi Agarwal, Dundigalla Ravi, Dr.Poonam Ghuli, Heart disease prediction using machine learning, International Journal of Engineering Research & Technology (IJERT) 9 (2020) http://www.ijert.org. ISSN: 2278-0181 IJERTV9IS040614 (This work is licensed under a Creative Commons Attribution 4.0 International License.) Published by: www.ijert.org.
[6] A. Geetha Devi, Surya Prasada Rao Borra and K. Vidya Sagar, "A method of cardiovascular disease prediction using machine learning", International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 Published by, www.ijert.org ICRADL - 2021 Conference Proceedings.
[7] Madhumita Pal, Smita Parija, Prediction of heart diseases using random forest, J. Phys. Conf. Ser. 1817 (2021) 012009, doi:10.1088/1742-6596/1817/1/012009.
[8] Vivek Ruban, Krithi, Heart disease prediction using machine learning models, International Journal of Recent Technology and Engineering (IJRTE) 8 (2020) ISSN: 2277-3878Issue-5S.
[9] Praveen Kumar Reddy M, T Sunil Kumar Reddy, S. Balakrishnan, Syed Muzamil Basha, Ravi Kumar Poluru, Heart disease prediction using machine learning algorithm, International Journal of Innovative Technology and Exploring Engineering (IJITEE) 8 (2019) ISSN: 2278-3075.
[10] Galla Siva Sai Bindhika, Munaga Meghana, Manchuri Sathvika Reddy, Rajalakshmi, Heart disease prediction using machine learning techniques, International Research Journal of Engineering and Technology (IRJET) 07 (2020) | www.irjet.net .
[11] Fahd Saleh Alotaibi, Implementation of machine learning model to predict heart failure disease, (IJACSA) International Journal of Advanced Computer Science and Applications 10 (6) (2019).
[12] Archana Singh, Rakesh Kumar, Heart disease prediction using machine learning algorithms, in: International Conference on Electrical and Electronics Engineering, 2020 (ICE3-.
[13] Ashmita Gupta, Ashutosh Niranajan, Hospital management and control system, European Journal of Molecular & Clinical Medicine 07 (2020) ISSN 2515-8260.
[14] R.K. Dash, T.N. Nguyen, K. Cengiz, A. Sharma, Fine-tuned support vector regression model for stock predictions, in: Neural Computing and Applications, 2021, pp. 1–15.
[15] D.T. Do, T.T.T. Nguyen, T.N. Nguyen, X. Li, M. Voznak, Uplink and downlink NOMA transmission using full-duplex UAV, IEEE Access 8 (2020) 164347–164364.
[16] K.K. Bharath, Ankit Kumar, Adithya Varma, R. Rajyashree, Secured electronic hospital database management system, International Journal of Recent Technology and Engineering (IJRTE) 8 (2019) ISSN: 2277-3878.
[17] T.N. Nguyen, B.H. Liu, S.I. Chu, D.T. Do, T.D. Nguyen, WRSNs: toward an efficient scheduling for mobile chargers, IEEE Sens. J. 20 (12) (2020) 6753–6761.
[18] T. Kowsalya, R.G. Babu, B.D. Parameshachari, A. Nayyar, R.M. Mehmood, Low Area PRESENT Cryptography in FPGA Using TRNG-PRNG Key Generation, CMC-Computers Materials & Continua 68 (2021) 1447–1465.

[19] Davide Chicco, Giuseppe Jurman, Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone, BMC Med. Inf. Decis. Making 20 (2020) 16.
[20] M.K. Kumar, B.D. Parameshachari, S. Prabu, S. liberata Ullo, Comparative Analysis to Identify Efficient Technique for Interfacing BCI System, IOP Confer. Ser.: Mater. Sci. Eng. 925 (2020) 012062 IOP Publishing.
[21] S. Rajendrakumar, V.K. Parvati, Automation of irrigation system through embedded computing technology, Proceed. 3rd Int. Confer. Cryptogr., Secur. Privacy (2019) 289–293.
[22] F. Ding, G. Zhu, M. Alazab, X. Li, K. Yu, Deep-learning-empowered digital forensics for edge consumer electronics in 5G HetNets, IEEE Consum. Electron. Magazine (2020), doi:10.1109/MCE.2020.3047606.

Journal of Pharmacological and Toxicological Methods 112 (2021) 107127

Contents lists available at ScienceDirect

Journal of Pharmacological and Toxicological Methods

journal homepage: www.elsevier.com/locate/jpharmtox

Research Paper

Prediction of GABA receptor antagonist-induced convulsion in cynomolgus monkeys by combining machine learning and heart rate variability analysis

Shoya Nagata [a], Koichi Fujiwara [a,*], Kazuhiro Kuga [b], Harushige Ozaki [b]

[a] Department of Material Process Engineering, Nagoya University, Nagoya, Japan
[b] Drug Safety Research and Evaluation, Takeda Pharmaceutical Company Ltd., Kanagawa, Japan

ARTICLE INFO

Keywords:
Convulsion biomarker
Heart rate variability analysis
Machine learning
Anomaly detection
Telemetry
Cynomolgus monkeys

ABSTRACT

Drug-induced convulsion is a severe adverse event; however, no useful biomarkers for it have been discovered. We propose a new method for predicting drug-induced convulsions in monkeys based on heart rate variability (HRV) and a machine learning technique. Because autonomic nervous activities are altered around the time of a convulsion and such alterations affect HRV, they may be predicted by monitoring HRV. In the proposed method, anomalous changes in multiple HRV parameters are monitored by means of a convulsion prediction model, and convulsion alarms are issued when abnormal changes in HRV are detected. The convulsion prediction model is constructed based on multivariate statistical process control (MSPC), a well-known anomaly detection algorithm in machine learning. In this study, HRV data were collected from four cynomolgus monkeys administered with multiple doses of pentylenetetrazol (PTZ) and picrotoxin (PTX), which are GABA receptor antagonists, as convulsant agents. In addition, low doses of pilocarpine (PILO) were administered as a negative control. Twelve HRV parameters in three hours after drug administration were monitored by means of the prediction model. The number and duration of convulsion alarms from HRV increased at medium and high doses of PTZ and PTX (1/3 or 1/4 of convulsion dose). On the other hand, the frequency of convulsion alarms did not increase with PILO. Although vomiting was observed in all drugs, convulsion alarms were not associated with the vomiting. Thus, convulsion alarms can be used as a biomarker for convulsions induced by GABA receptor antagonists.

1. Introduction

Convulsions are one of the typical neurological symptoms observed as a crucial toxicological finding in nonclinical toxicology studies. When a drug candidate has a convulsion liability, its dosage in clinical studies must be conservative since there is no appropriate biomarker for convulsions. In other words, investigators are required to set low dose levels to secure a large safety margin. Appropriate biomarkers could contribute to drug development (Wong, Siah, & Lo, 2019) through proper assessments of convulsion risk of drug candidates, and they could lead to a reduction in the number of dropped candidates in the drug development process. There have been *in vitro* methods for predicting convulsion risks of drug candidates (Easter, Sharp, Valentin, & Pollard, 2007). *In vitro* methods have been developed for predicting convulsion risk of drug candidates; however, *in vivo* studies are still required for evaluating convulsion risk before the start of clinical studies. A new biomarker for demonstrating convulsion potential in animals is desirable to estimate the risk of convulsions in clinical studies.

Heart Rate Variability (HRV) is defined as fluctuations between an R wave and the next R wave on an electrocardiogram (ECG), which reflects activities of the autonomic nervous system (ANS) (Camm, Thomas Bigger, Cohen, & Fallen, 1996; Shaffer & Ginsberg, 2017). HRV has been widely used for various medical purposes in humans, such as stress estimation (Dishman et al., 2000; Kim, Cheon, Bai, Lee, & Koo, 2018), cardiovascular diseases (Colhoun, Francis, Rubens, Underwood, & Fuller, 2001; Tsuji et al., 1996), and sleep disorders (Iwasaki et al., 2021; Lado et al., 2011; Sumi et al., 2020). In addition, we previously developed an HRV-based epileptic seizure prediction algorithm in humans (Fujiwara et al., 2016). It was assumed that drug-induced convulsions can be predicted by utilizing HRV in the same manner as epileptic seizures.

We aimed to develop a new method based on HRV for predicting drug-induced convulsions. In this study, we attempted to predict convulsions in monkeys as a first step, because electrocardiograms for HRV analysis can be recorded in animals without restriction, and HRV has been adopted for animals as well as humans to monitor autonomic

* Corresponding author at: Department of Material Process Engineering, Nagoya University, Furo-cho, Chikusa-ku, Nagoya-shi, Aichi 464-8601, Japan.
E-mail address: fujiwara.koichi@hps.material.nagoya-u.ac.jp (K. Fujiwara).

https://doi.org/10.1016/j.vascn.2021.107127
Received 25 April 2021; Received in revised form 30 September 2021; Accepted 30 September 2021
Available online 6 October 2021

functions. The effect of alcohol was evaluated by means of HRV through experiments using cynomolgus monkeys (Shively et al., 2007). Kerem tried to predict generalized epileptic seizures of rats based on HRV (Kerem & Geva, 2005). An HRV-based ischemic stroke detection method was developed utilizing a rat middle cerebral artery occlusion model (Kodama et al., 2018). Iyer et al. suggested that cardiorespiratory parameters, including HRV, may be used as biomarkers for sudden unexpected death in epilepsy (SUDEP) through experiments with Kv1.1 knockout mice (Iyer et al., 2020).

When HRV data with and without the use of a convulsant are regarded as abnormal and normal data, respectively, collecting normal data is much easier than abnormal data. A convulsion prediction method is developed only from the HRV data without a convulsant. In the machine learning field, such a problem can be formulated as an anomaly detection problem whose aim is to detect low-frequency anomalous events from a large amount of normal data. In this study, we adopt multivariate statistical process control (MSPC) to construct a convulsion prediction model—a well-known anomaly detection algorithm used in various fields, including the biomedical field (Fujiwara et al., 2016; Fujiwara et al., 2019; Nakayama et al., 2019). The convulsion prediction model issues convulsion alarms when an anomalous alteration of HRV is detected.

We collected HRV data of cynomolgus monkeys treated with picrotoxin (PTX) and pentylenetetrazol (PTZ), which are $GABA_A$ receptor antagonists and are well-known convulsants. In addition, pilocarpine (PILO) was administered to animals at low doses as a negative control. We applied the proposed biomarker to the collected HRV data for predicting drug-induced convulsions to evaluate its validity.

2. Materials and methods

2.1. Animals

This study was conducted under the approval of the Institutional Animal Care and Use Committee (IACUC), Shonan Health Innovation Park and conforms to the Guide for the Care and Use of Laboratory Animals published by the National Institutes of Health. Four male cynomolgus monkeys were used because monkeys are one of the most frequently-used animals for ECG measurements in a safety pharmacology study, which is essential for drug development, and the number of animals used was determined based on the standard test protocol for safety pharmacology studies on the evaluation of the cardiovascular system in non-rodents.

The monkeys M1-M4 were aged 6–7 years old and weighing 5–7 kg. The monkeys had been implanted intraperitoneally with a transmitter (TL11M2-D70-PCT or TL11M3-D70-PCTP, Data Sciences International). The pressure sensor catheter of the transmitter was inserted into the right femoral artery and placed in the abdominal aorta. The negative electrocardiographic (ECG) electrode (solid tip) was inserted into the right jugular vein, and the tip of the electrode was placed in the superior vena cava. The positive ECG electrode was fixed to the diaphragm near the cardiac apex in the abdominal cavity. The animals were individually housed in metal cages (floor area: 0.433 m2, height: 0.765 m) set on racks in an animal room with manipulable toys removed from individual cages during the telemetry recording period. The animal room conditions were as follows: temperature control range: 22 °C to 27 °C, relative humidity control range: 40% to 70%, air exchange: 10 to 25 times/h, and 12-h light/dark cycle (lights on from 7:00 am to 7:00 pm, over 150 lx at 85 cm above floor level). Each animal was fed 100 g of a pelleted diet once daily. The food was supplied between 5:15 pm and 5:45 pm on the days before dosing and days of dosing. The remaining food was withdrawn the following morning. The animals were allowed free access to tap water.

Table 1
Dosage settings [mg/kg].

	PTZ	PTX	PILO
Baseline	0	0	0
Low	7	0.05	0.6
Middle	18	0.15	2
High	35	0.5	6

PTZ: pentylenetetrazol, PTX: picrotoxin, PILO: pilocarpine.

2.2. Test drugs

PTZ (Tront Research Chemicals Inc., USA) and PTX (Tokyo Chemical Industry Co., Ltd., Japan) were used as convulsants, and pilocarpine hydrochloride (PILO, Tokyo Chemical Industry Co., Ltd., Japan), a muscarinic receptor antagonist, was used as a non-convulsant. Range-finding studies were performed in female monkeys before this study. PTX and PTZ induced tonic convulsions at 1 mg/kg and 70 mg/kg, respectively. Hence, half of those doses were set as the high doses in this study. Dose levels of PTX were 0.05, 0.15, and 0.5 mg/kg, and those of PTZ were 7, 18, and 35 mg/kg. The dose-volume was set to 2 mL/kg for each subcutaneous dose. Although PILO is a convulsant, it did not induce convulsions up to 100 mg/kg in a dose-ranging experiment. We set PILO doses to 0.6, 2, and 6 mg/kg, which are much lower than the expected convulsive dose (> 100 mg/kg). The dose volume was set to 5 mL/kg for each oral dose. The dosage settings in this study are summarized in Table 1. Before the start of dosing with each drug, approximately 10 mL of saline subcutaneously, or 20 mL of tap water orally, was administered to the animals once daily for four days to habituate them to the dosing procedure and to acquire training data.

2.3. Data collection

Telemetry and video data were recorded during the four-day vehicle administration acclimatization period and the day before administering each drug. The vehicles used in this study were saline for PTZ and PTX and 0.5% methylcellulose solution for PILO. After that, PTZ and PTX were administrated subcutaneously, and PILO was administrated orally. We titrated doses at three- or four-day intervals in the experiment of each drug, and an interval of at least one month was interposed between experiments. We dosed diazepam to suppress convulsions shortly after occurrence.

Blood pressure (BP), ECG waveforms, body temperature, and activity counts were recorded continuously from approximately 24 h pre-dose to approximately 24 h post-dose using a telemetry system. BP, body temperature, and activity counts were not used in this study. Human access to the animal room was limited during the recordings. Concurrently, the animals' behavior was recorded with a CCD camera and digital recorder. All parameters were recorded by means of a transmitter, a receiver (RMC-1, Data Sciences International), and a telemetry data collection and analysis computer system (OpenART, ver. 4.3/Ponemah Physiology Platform ver. 5.0, Data Sciences International). All RR intervals obtained in this study (including the pre-dose period) were output to an excel file for each beat. However, data including arrhythmia or abnormal waveforms were excluded. The timings of vomiting/retching, convulsion, dose (test article or diazepam), feeding, and the technician's entry and exit were also recorded in the excel file.

In the dose-ranging experiments, most neurological symptoms occurred within three hours after administration, suggesting that drug effects might not persist for a long period. In addition, HRV is known to be affected by nap and sleep (Gong et al., 2016; Gosselin, Michaud, Carrier, Lavigne, & Montplaisir, 2002). Thus, we focused on data from the three-hour period from the time of drug administration (around 9 am) in this analysis, although 24-h telemetry data after drug administration were also collected.

The animal experiment was performed by Axcelead Drug Discovery

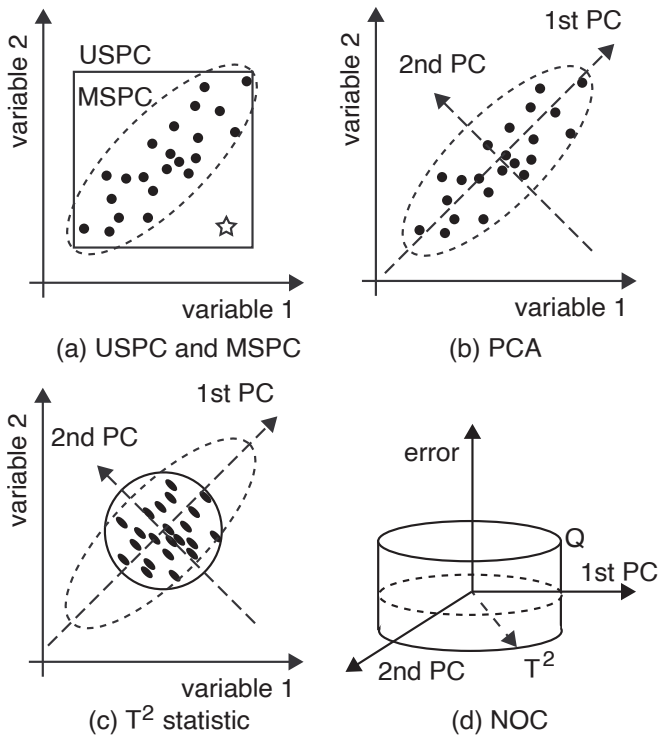


Fig. 1. Schematic diagram of MSPC: (a) the difference between USPC and MSPC in two-dimensional space. USPC cannot detect an anomaly (☆) that does not follow the correlation between variables 1 and 2 because it monitors each variable independently. On the other hand, MSPC monitors the correlation among variables by using an ellipsoid control limit. (b) The concept of PCA, which is a useful statistical tool for extracting a major trend in the training data. (c) The T^2 statistic is defined as the distance between the origin and a sample normalized by the standard deviation of each principal component. It is difficult to evaluate the correlation changes if the standard deviations of principal components are significantly different. (d) MSPC uses the T^2 and the Q statistics for monitoring changes in the correlation. The T^2 statistic monitors changes in the subspace spanned by the principal components, while the Q statistic monitors changes in residual subspace. Thus, the normal operating condition (NOC) is defined by control limits of the T^2 and the Q statistics, expressed as a cylinder-like volume.

Partners Inc.

2.4. Multivariate statistical process control (MSPC)

The proposed HRV-based method for predicting drug-induced convulsions detects abnormal changes in HRV by using a machine learning model. Such a model is trainable from only HRV data without the use of drugs, because it is difficult to collect a large amount of HRV data around the time of a drug-induced convulsion. In the machine learning field, such a problem is formulated as anomaly detection. Although any anomaly detection algorithm can be used in the proposed method, we adopted multivariate statistical process control (MSPC).

The simplest way of detecting anomalies is to monitor each variable with upper and lower constraints independently, which is a method called univariate statistical process control (USPC). However, changes in the correlation between variables cannot be observed with USPC. For example, when two variables have a positive correlation, as shown in Fig. 1(a), USPC cannot detect the anomaly (x), which does not follow a positive correlation because its constraints form a rectangular area. If an ellipsoid control limit shown by the dashed line is defined, the anomaly (x) can be detected since the correlation between variables is taken into account. MSPC is a correlation-based anomaly detection method widely used in artificial intelligence for medical purposes (Fujiwara et al., 2016; Kodama et al., 2018). Therefore, MSPC can detect a sample that does not follow the major trend in the modeling data.

In MSPC, the correlation between variables is modeled using principal component analysis (PCA), which finds linear combinations of variables referred to as principal components (PC). PCA can describe major trends in a dataset, as shown in Fig. 1(b). Usually in MSPC, the normal operating condition (NOC) is defined using two monitored indices, *i.e.*, the T^2 and Q statistics instead of an ellipsoid control limit.

The T^2 statistic is defined as the Mahalanobis distance between a sample and the origin in the subspace spanned by principal components, which expresses a circular control limit as shown in Fig. 1(c). When the T^2 statistic is small, the sample is close to the mean of the modeling data. Further, the Q statistic is a measure of dissimilarity between the sample and the modeling data from the viewpoint of the correlation between variables. MSPC detects an anomaly when either the T^2 or the Q statistic exceeds a predefined control limit. Fig. 1(d) shows a schematic diagram of the NOC of MSPC.

The control limit controls the sensitivity and the specificity, and the 99% or 95% confidence limit is usually adopted. The control limit is set so that 99% or 95% of the normal data are judged to be below normal. Since MSPC is based on PCA, the number of principal components R in PCA has to be carefully determined as a tuning parameter. The cumulative proportion of principal components can be used to assess R. For example, R can be determined so that the cumulative proportion reaches a predefined value, such as 80% or 90%.

In the proposed method for drug-induced convulsions, the HRV data without the use of drugs and the HRV data with the use of drugs are defined as normal data and anomalous data, respectively. MSPC requires only normal data, which is easy to collect compared to anomalous data, and it can be easily implemented because it is a linear method. The detailed mathematical description of MSPC is described in (Fujiwara et al., 2016).

2.5. Convulsion prediction

The proposed method consists of two phases: convulsion prediction model training and convulsion prediction. The procedure of the training phase is as follows:

1. Collect ECG data without using convulsion-inducing drugs and extract the RRI from the collected ECG data.
2. Calculate HRV data from the extracted RRI data.
3. Apply PCA to the extracted HRV data to build an MSPC model that expresses the relationship among HRV parameters *sans* drugs.
4. Determine the control limits of the T^2 and Q statistics for each patient.

The number of principal components R and the control limits of the T^2 and Q statistics have to be determined appropriately to achieve good prediction performance.

Convulsion prediction is performed by following the procedure below:

i) Measure ECG of an animal and extract RRI from the measured ECG.
ii) Derive the HRV data from the measured RRI data.
iii) Calculate the T^2 and Q statistics by using the convulsion prediction model.
iv) Issue a convulsion alarm when either the T^2 or Q statistic continuously exceeds the control limits for more than a predefined period τ.

In step iv), the wait period τ prior to issuing the convulsion alarm is set in order to suppress false positives, since the T^2 and Q statistics can fluctuate due to ECG artifacts.

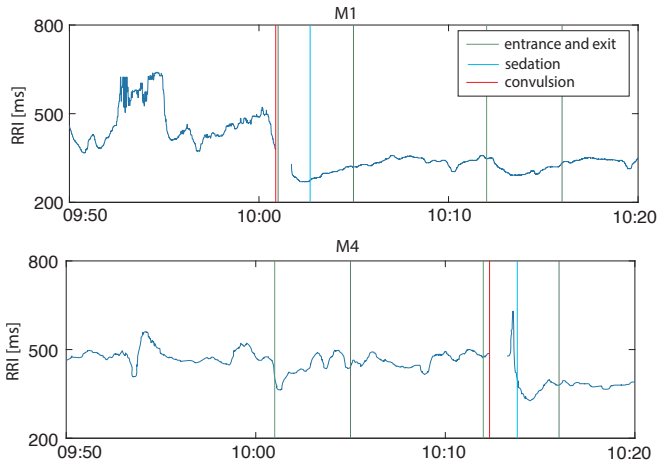


Fig. 2. Changes in the RRI data of M1 (top) and M4 (bottom) around the time of convulsion onset. The red, blue, and green vertical lines denote the times of convulsion onset, diazepam administration, and entrance/exit of persons into the animal room. The RRI data of M4 after the convulsion occurrence were lost due to significant body movement caused by the convulsion. In addition, the RRI of M4 changed around the time of animal room entrance/exit. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

2.6. Analysis procedure

The R waves in the collected ECG data of four cynomolgus monkeys were detected using a first derivative-based peak detection algorithm, and each RRI was calculated. We used twelve standard HRV parameters (six time-domain parameters— meanNN, SDNN, Total Power (TP), RMSSD, NN50, and pNN50—and six frequency-domain parameters— LF, HF, VLF, LF/HF, LF norm, and HF norm) for convulsion prediction. The detailed description of HRV parameters used in this study is provided in the Appendix.

The convulsion prediction model was trained using the HRV data of all animals during acclimatization. The control limits of the T^2 and Q statistics were determined for each subject so that they represented the 99% confidence limit. The number of principal components R was set to four so that the cumulative proportion reached more than 90%, which was a standard setting in PCA. A convulsion alarm was issued when either the T^2 or Q statistic continuously exceeded its control limit for more than $\tau = 30$ s because a tranquilizer is usually given within a couple of minutes after convulsion occurrence, and a 30-s delay is acceptable for treatment.

2.7. Statistical analysis

We used Levene's test and the Kruskal-Wallis test with a significance level $p < 0.05$ for comparing the alarms by means of MSPC among different dosages of drugs. All computations in this study were performed in MATLAB 2020a (Mathworks Inc.).

3. Results

In this study, convulsion occurred in two cynomolgus monkeys, M1 and M4, when PTX with 0.5 mg/kg was administrated. Fig. 2 illustrates the changes in their RRI data around convulsion onset and the changes in the corresponding HRV data are shown in Figs. 3 and 4. Table 2 shows the numbers of vomiting and retching events observed in video during a three-hour period after drug administration. According to Figs. 3 and 4,

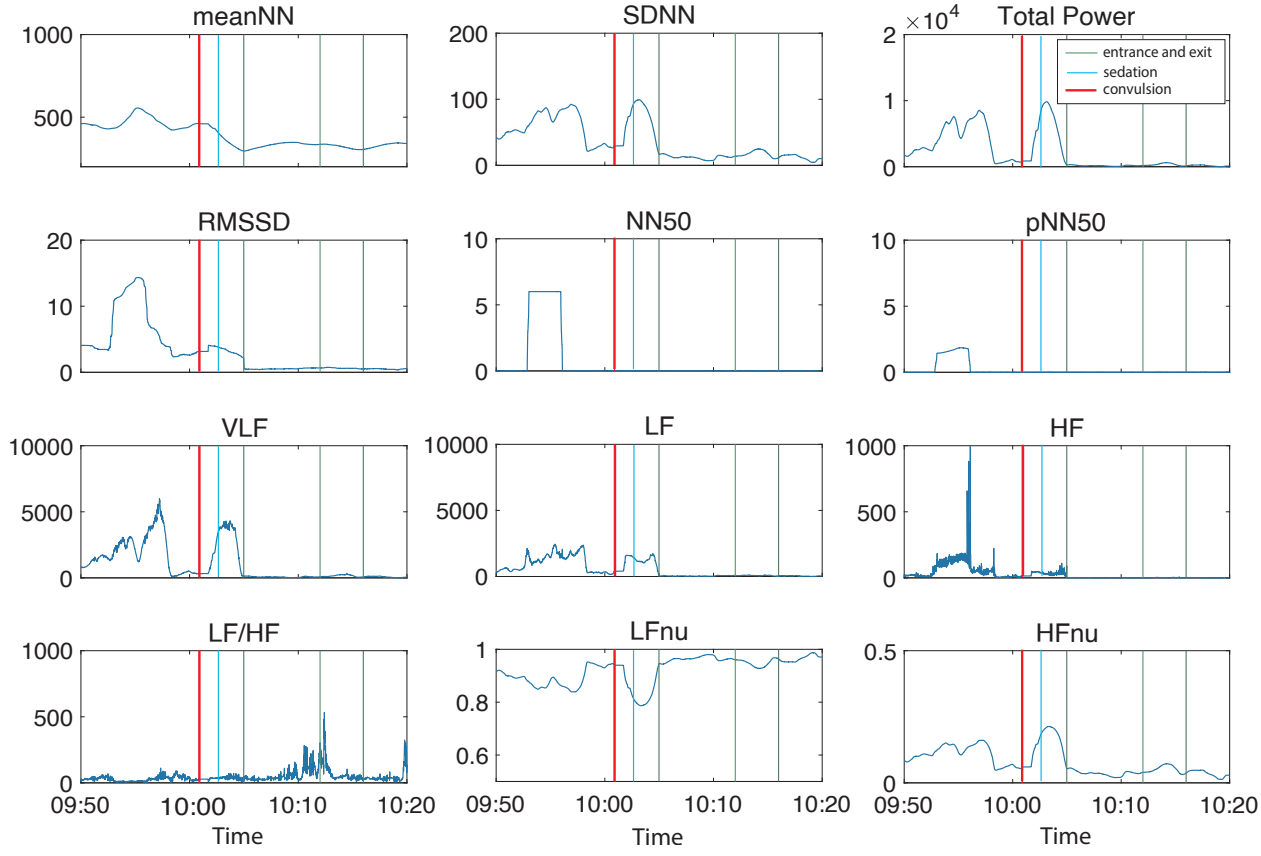


Fig. 3. Changes in HRV of M1: The red, blue, and green vertical lines denote the times of convulsion onset, diazepam administration, and entrance/exit of persons into the animal room, respectively. These HRV parameters fluctuated around the time of convulsion onset. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

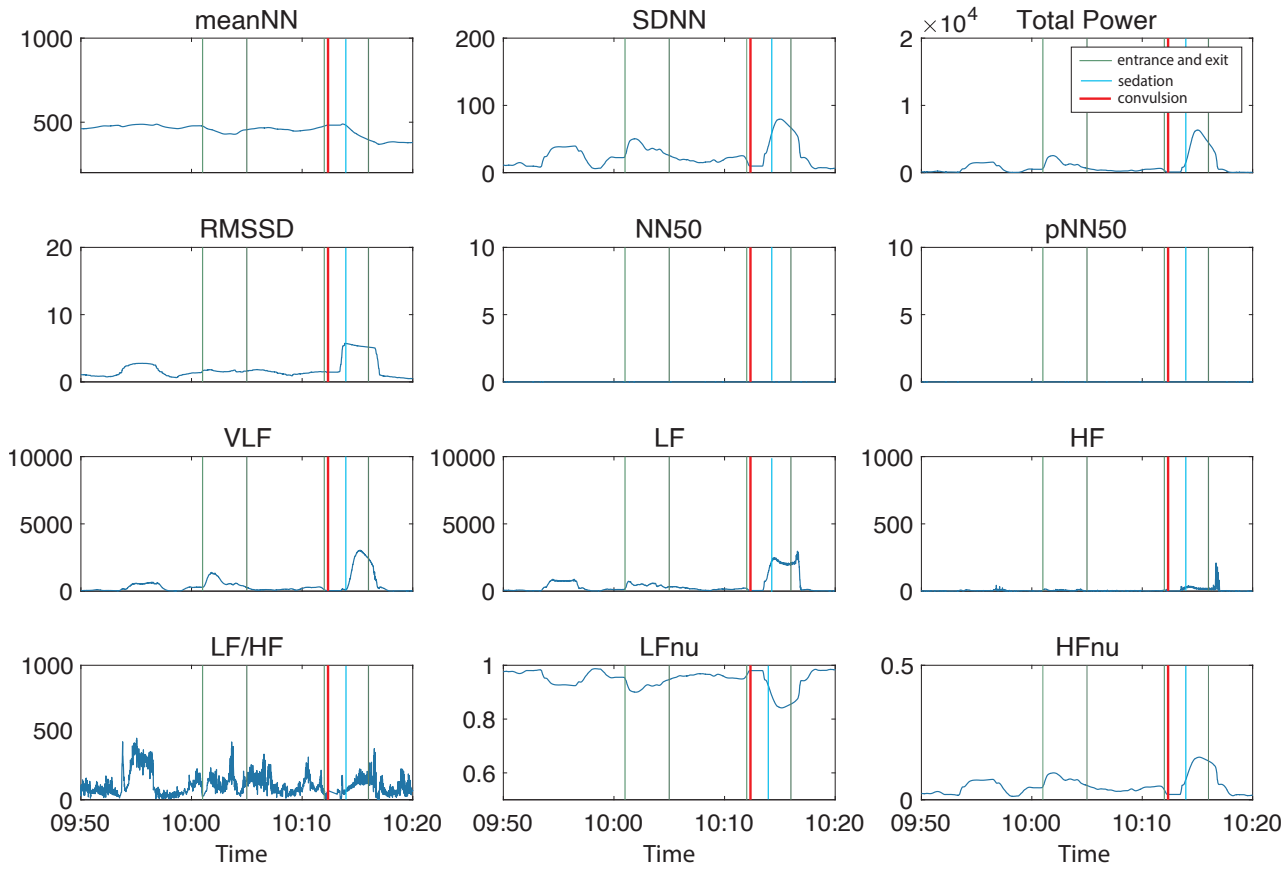


Fig. 4. Changes in HRV of M4: The red, blue, and green vertical lines denote the times of convulsion onset, diazepam administration, and entrance/exit of persons into the animal room, respectively. The HRV parameters during the time in which RRI data were lost due to the convulsion were calculated using spline interpolation. These HRV parameters fluctuated around the time of convulsion onset. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Number of vomiting and retching events in video observation [times/head].

	PTZ	PTX	PILO
Low	0	0	0
Middle	0	2.5	2.25
High	6.75	19.25	9.5

the HRV parameters changed around the time of convulsions in M1 and M4. On the other hand, vomiting did not greatly affect HRV, although it occurred with all drugs. The R waves in the collected ECG data of four cynomolgus monkeys were detected using a first derivative-based peak detection algorithm, and each RRI was calculated.

The convulsion prediction models for each cynomolgus monkey were trained from the HRV data collected from each cynomolgus monkey during acclimatization. We applied these trained models to the HRV data collected with various dosage conditions listed in Table 1 and calculated the T^2 and Q statistics in each dosage condition. Finally, we extracted the periods when either the T^2 or Q statistic exceeded its control limit as convulsion alarms. Figs. 5 and 6 show the number and the total duration of alarms issued by the convulsion prediction model in response to the drug dosages of PTZ, PTX, and PILO.

The number and the total duration of convulsion alarms increased with the medium and high doses of PTZ and PTX. On the other hand, a dose-dependent relationship of the number and the total duration of convulsion alarms was not observed with PILO. We used Levene's test, and the equality of the variance of the number and the total duration of convulsion alarms among different doses was not confirmed in any of the drugs ($p < 0.05$). According to the Kruskal-Wallis test, there were significant differences in the number and the total duration of

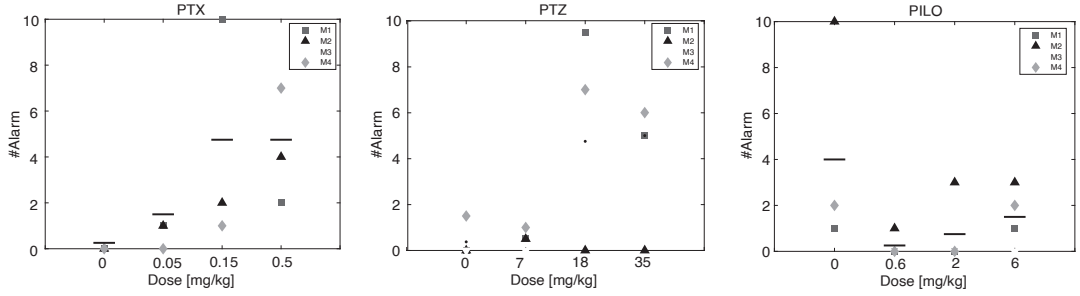


Fig. 5. The number of convulsion alarms of M1-M4 when PTX (left), PTZ (center), and PILO (right) were dosed. The black lines denote their means. The number of alarms increased with the medium and high doses of PTZ and PTX.

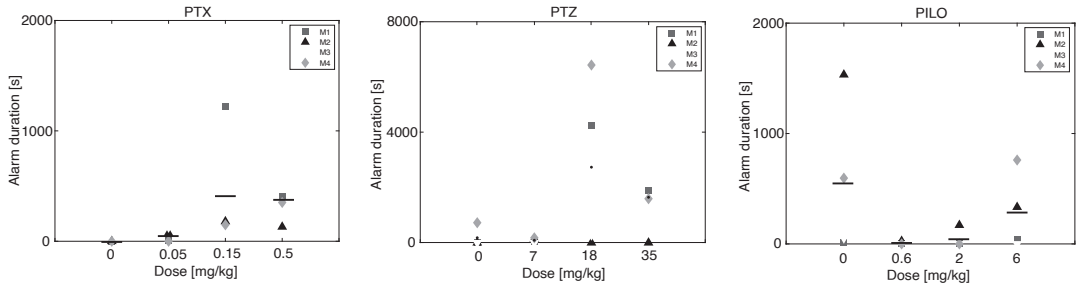


Fig. 6. The total duration of convulsion alarms of M1-M4 when PTX (left), PTZ (center), and PILO (right) were dosed. The black lines denote their means. The total duration of alarms increased with the medium and high doses of PTZ and PTX.

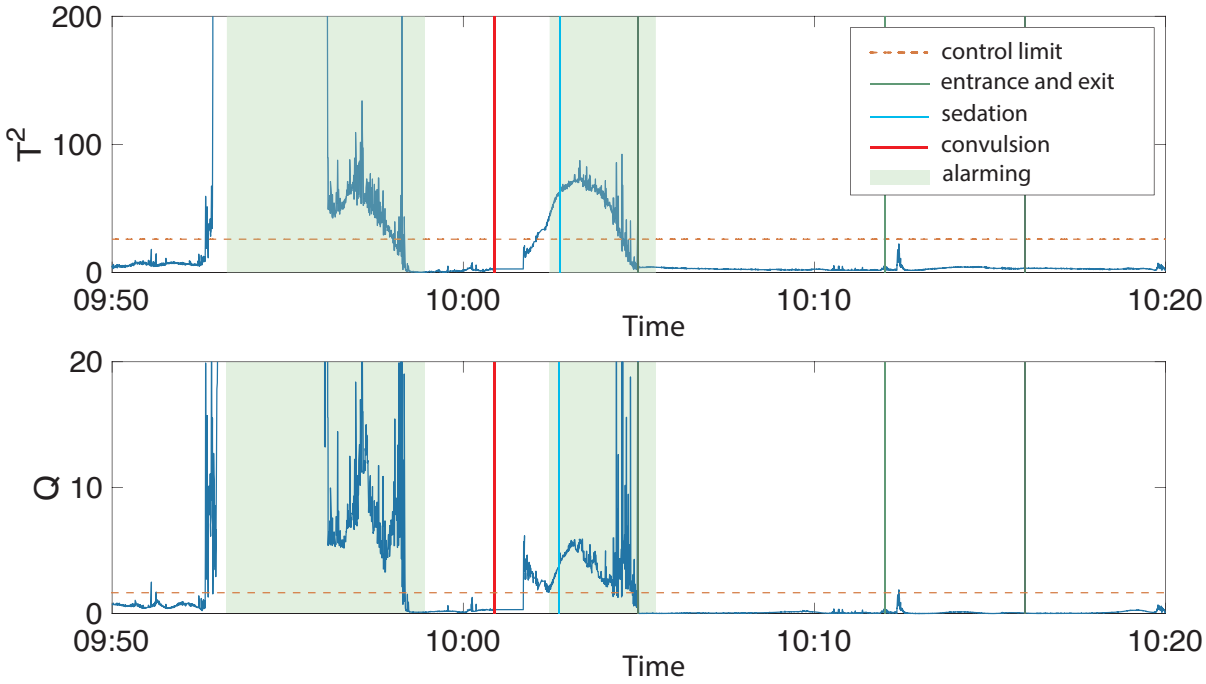


Fig. 7. The T^2 and Q statistics of M1 around the time of convulsion onset with 0.5 mg/kg-PTX. The red, blue, and green vertical lines denote the times of convulsion onset, diazepam administration, and entrance/exit of persons into the animal room, respectively. The horizontal dashed lines denote the control limits of the T^2 and Q statistics, and the colored bands show the periods during which the convulsion alarms were issued by the proposed method. Both the T^2 and Q statistics exceeded their control limits before and after the convulsion onset. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

convulsion alarms among different doses in PTX ($p < 0.05$). Thus, there was a dose-dependent relationship between the convulsion alarms and PTX.

The T^2 and Q statistics of M1 and M4 around the time of convulsion onset with 0.5 mg/kg-PTX are shown in Figs. 7 and 8, which demonstrates that both the T^2 and Q statistics exceeded their control limits before and after the convulsion onset. These convulsion prediction results clearly show that the proposed method succeeded in predicting the drug-induced convulsion before onset. Thus, the number and the duration of convulsion alarms calculated by the proposed method may be used as a biomarker for drug-induced convulsions.

4. Discussion

In this study, convulsion occurred in only two monkeys in PTX; however, it is important for a biomarker to be able to detect convulsion risk below the convulsion-inducing dosage. Thus, this experiment and data analysis meet the aim of the study, even though not all animals had convulsions.

We accurately issued convulsion alarms at the medium and high doses of PTZ and PTX. The medium doses were 1/3 and 1/4 of the convulsion doses suggesting that we could predict convulsion liability without inducing a convulsion using our prediction model.

In addition, the proposed method predicted a convulsion before its onset when 0.5 mg/kg of PTX was administrated. This result shows that the convulsion alarm by means of the proposed method can be used as a biomarker for drug-induced convulsions because the number and the total duration of convulsion alarms may indicate the possible occurrence of a convulsion in the near future.

A convulsion alarm may be issued due to ECG artifacts caused by measurement failure or arrhythmia. Although we visually checked the collected ECG signals around the time of convulsion alarms, no artifacts or arrhythmia had occurred, which indicates that the convulsion alarms might not have been affected by artifacts or arrhythmia in this study.

According to Fig. 5, many convulsion alarms occurred at 0 mg/kg-PILO administration in M2, which were not associated with the drug administration. ECG artifacts or arrhythmia were not observed around the time of convulsion alarm occurrences. That is, these alarms were false positives. Since changes in sleep condition significantly affect HRV (Gosselin et al., 2002) and may lead to false positives (Fujiwara et al.,

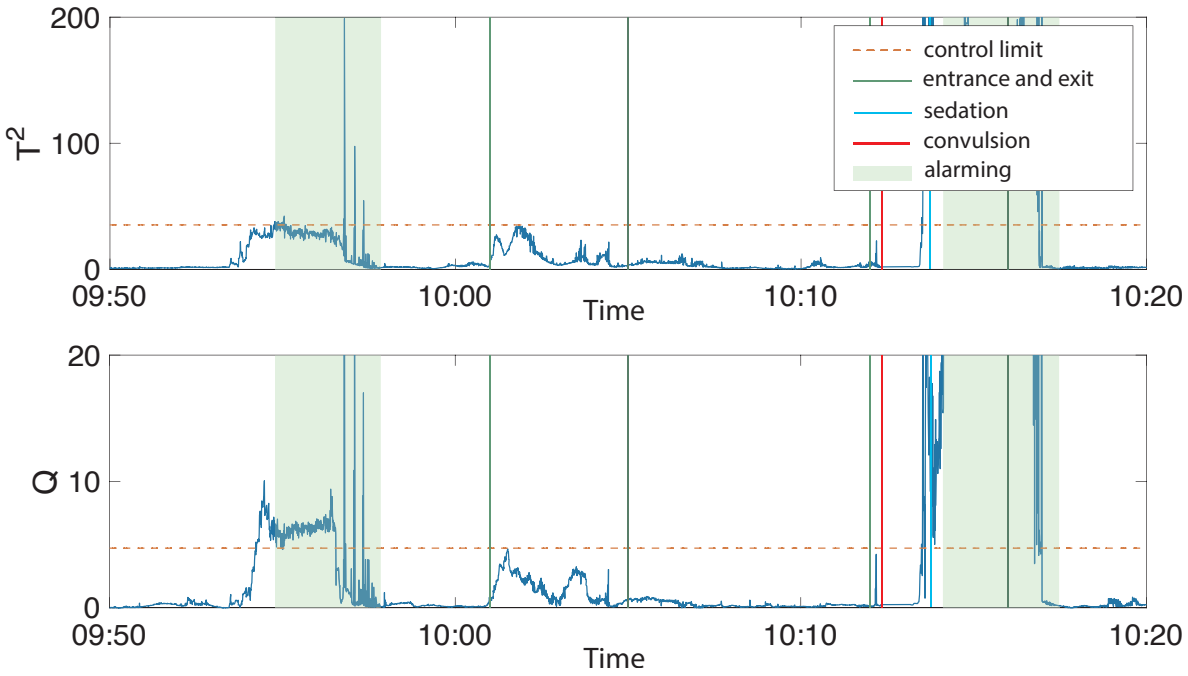


Fig. 8. The T^2 and Q statistics of M4 around the time of convulsion onset with 0.5 mg/kg-PTX. The red, blue, and green vertical lines denote the times of convulsion onset, diazepam administration, and entrance/exit of persons into the animal room, respectively. The horizontal dashed lines denote the control limits of the T^2 and Q statistics, and the colored bands show the periods during which the convulsion alarms were issued by the proposed method. Both the T^2 and Q statistics exceeded their control limits before and after the convulsion onset. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
Number of alarms in PTX with the convulsion monitoring model trained with 24-h HRV data.

	M1	M2	M3	M4
0 mg/kg	64	31	51	30
0.05 mg/kg	39	21	67	32
0.15 mg/kg	43	33	79	36
0.50 mg/kg	35	23	69	17

2016), there is the possibility that M2 was sleeping during this measurement. Although we could not determine if the animal slept, video observation and activity count demonstrated that the animal did not move during these alarms.

We trained the convulsion prediction models using the three-hour data with administration between the start of administration and noon, because such data might not contain sleep data in this study. On the other hand, we collected the telemetry data over 24 h in this experiment, and we tried to train the convulsion prediction models using the 24-h HRV data. Table 3M summarizes the number of convulsion alarms issued by the models of M1-M4 when PTX was dosed. This result shows that the number of convulsion alarms significantly increased, particularly with 0 mg/kg of PTX, whose number of alarms have been zero. Thus, the models trained with the 24-h HRV data did not function appropriately. The 24-h data contained significant changes in the activities of ANS due to feeding and sleeping; HRV data containing such changes were not appropriate for model training. Thus, the convulsion prediction model has to be trained from HRV data during resting conditions, in which ANS activities do not significantly fluctuate. This indicates that the modeling data have to be selected carefully, which is a common problem in machine learning.

The limitations of this study include the collected data, such as the fact that all of the animals used in this study were male cynomolgus monkeys of six to seven years of age; thus, we could not consider sex, age, or species differences, which may affect HRV. Accordingly, we need to collect data from various animals to verify the proposed method.

In this study, we proposed a method for predicting drug-induced convulsions for cynomolgus monkeys by combining HRV analysis and MSPC, which is an anomaly detection algorithm used in the machine learning field. Applying the proposed method to the HRV data with convulsion-inducing drugs showed a dose-dependent relationship between convulsion alarms and PTX, and that the convulsion alarms can predict convulsion onsets. This suggests that the convulsion alarms issued by the proposed method can be used as a biomarker for drug-induced convulsions.

In future works, we will evaluate the validity of the proposed method through animal experiments using other animals, such as dogs, and convulsant agents with different GABAA antagonistic mechanisms. Since other anomaly detection algorithms such as an autoencoder or a one-class support vector machine have been proposed in the machine learning field, we will validate whether they can be used for convulsion prediction.

Funding

The authors declare no fundings associated with this manuscript.

Declaration of Competing Interest

Koichi Fujiwara is with Quadlytics Inc. as well as Nagoya University. Other authors declare that the research was conducted in the absence of any commercial or financial relationship that could be construed as a potential conflict of interest.

Acknowledgment

We would like to thank Yoshiyuki Furukawa, Ryouta Hayashi, and Tomoki Shimada in Axcelead Drug Discovery Partners Inc. for the animal experiments.

Appendix

The ordinal HRV data consist of the time-domain parameters and the frequency domain parameters (Camm et al., 1996; Shaffer & Ginsberg, 2017). This Appendix describes the HRV parameters used in this study.

Time-domain parameters

The following six time-domain parameters were calculated directly from raw RRI data.

- meanNN: Mean of RRI.
- SDNN: Standard deviation of RRI.
- RMSSD: Root means square of the difference of adjacent RRI.
- Total Power (TP): Variance of RRI.
- NN50: The number of pairs of adjacent RRI whose difference is more than 50 ms within a given length of measurement time.
- pNN50: The value of NN50 divided by the total number of RRI.

Frequency-domain parameters

Frequency-domain analysis must be modified for cynomolgus monkeys since their heart rate (about 90–150 bpm) is much faster than that of humans (about 60–80 bpm). The frequency-domain parameters are defined as powers of a specific frequency range in a power spectrum density (PSD) of the RRI data. Although the powers in 0.04 Hz - 0.15 Hz, 0.15 Hz - 0.4 Hz, and $\leq$ 0.04 Hz are usually used as LF, HF VLF for humans (Camm et al., 1996), these frequency range definitions are modified for cynomolgus monkeys. The following six frequency-domain parameters are adopted in this study (Shively et al., 2007).

- LF: The power in the low-frequency range (0.01 Hz – 0.2 Hz) in the PSD. LF reflects sympathetic nervous activity and parasympathetic nervous activity.
- HF: The power in the high-frequency range (0.2 Hz – 0.8 Hz) in the PSD. HF reflects parasympathetic nervous activity.
- VLF: Power in the very-low-frequency range ($\leq$ 0.04 Hz).
- LF/HF: Ratio of LF to HF. LF/HF expresses the balance between sympathetic and parasympathetic nervous activities.
- HF norm: The ratio of HF to the entire frequency range.
- LF norm: The ratio of LF to the entire frequency range.

A rectangular moving window with a size of three minutes was used. For the frequency-domain parameter calculation, the RRI data were resampled so that the sampling points are arranged at equal intervals, which were interpolated by means of the third-order spline, and 4 Hz resampling was adopted. An autoregressive model of order 40 was used.

References

Camm, J. A., Thomas Bigger, J., Cohen, R. J., & Fallen, E. L. (1996). Heart rate variability: Standards of measurement, physiological interpretation and clinical use. Task force of the european society of cardiology and the north american society of pacing and electrophysiology. *Circulation., 93*(5), 1043–1065.

Colhoun, H. M., Francis, D. P., Rubens, M. B., Underwood, S. R., & Fuller, J. H. (2001). The association of heart-rate variability with cardiovascular risk factors and coronary artery calcification: A study in type 1 diabetic patients and the general population. *Diabetes Care, 24*(6), 1108–1114.

Dishman, R. K., Nakamura, Y., Garcia, M. E., Thompson, R. W., Dunn, A. L., & Blair, S. N. (2000). Heart rate variability, trait anxiety, and perceived stress among physically fit men and women. *International Journal of Psychophysiology, 37*(2), 121–133.

Easter, A., Sharp, T. H., Valentin, J. P., & Pollard, C. E. (2007). Pharmacological validation of a semi-automated in vitro hippocampal brain slice assay for assessment of seizure liability. *Journal of Pharmacological and Toxicological Methods, 56*(2), 223–233.

Fujiwara, K., Abe, E., Kamata, K., Nakayama, C., Suzuki, Y., Yamakawa, T., , … Masuda, F., et al. (2019). Heart rate variability-based driver drowsiness detection and its validation with eeg. *IEEE Transactions on Biomedical Engineering, 66*(6), 1769–1778.

Fujiwara, K., Miyajima, M., Yamakawa, T., Abe, E., Suzuki, Y., Sawada, Y., , … Sasai-Sakuma, T., et al. (2016). Epileptic seizure prediction based on multivariate statistical process control of heart rate variability features. *IEEE Transactions on Biomedical Engineering, 63*(6), 1321–1332.

Gong, X., Huang, L., Liu, X., Li, C., Mao, X., Liu, W., , … Wu, W., et al. (2016). Correlation analysis between polysomnography diagnostic indices and heart rate variability parameters among patients with obstructive sleep apnea hypopnea syndrome. *PLoS One, 11*(6), Article e0156628.

Gosselin, N., Michaud, M., Carrier, J., Lavigne, G., & Montplaisir, J. (2002). Age difference in heart rate changes associated with micro-arousals in humans. *Clinical Neurophysiology, 113*(9), 1517–1521.

Iwasaki, A., Nakayama, C., Fujiwara, K., Sumi, Y., Matsuo, M., Kano, M., & Kadotani, H. (2021). Screening of sleep apnea based on heart rate variability and long short-term memory. *Sleep & Breathing*. https://pubmed.ncbi.nlm.nih.gov/33423183/.

Iyer, S. H., Aggarwal, A., Warren, T. J., Hallgren, J., Abel, P. W., Simeone, T. A., & Simeone, K. A. (2020). Progressive cardiorespiratory dysfunction in kv1.1 knockout mice may provide temporal biomarkers of pending sudden unexpected death in epilepsy (sudep): The contribution of orexin. *Epilepsia., 61*(3), 572–588.

Kerem, D. H., & Geva, A. B. (2005). Forecasting epilepsy from the heart rate signal. *Medical & Biological Engineering & Computing, 43*(2), 230–239.

Kim, H. G., Cheon, E. J., Bai, D. S., Lee, Y. H., & Koo, B. H. (2018). Stress and heart rate variability: A meta-analysis and review of the literature. *Psychiatry Investigation, 15*(3), 235–245.

Kodama, T., Kamata, K., Fujiwara, K., Kano, M., Yamakawa, T., Yuki, I., & Murayama, Y. (2018). Ischemic stroke detection by analyzing heart rate variability in rat middle cerebral artery occlusion model. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 26*(6), 1152–1160.

Lado, M. J., Vila, X. A., Rodríguez-Liñares, L., Méndez, A. J., Olivieri, D. N., & Félix, P. (2011). Detecting sleep apnea by heart rate variability analysis: Assessing the validity of databases and algorithms. *Journal of Medical Systems, 35*(4), 473–481.

Nakayama, C., Fujiwara, K., Sumi, Y., Matsuo, M., Kano, M., & Kadotani, H. (2019). Obstructive sleep apnea screening by heart rate variability-based apnea/normal respiration discriminant model. *Physiological Measurement, 40*(12), 125001.

Shaffer, F., & Ginsberg, J. P. (2017). An overview of heart rate variability metrics and norms. *Frontiers in Public Health, 5*, 258.

Shively, C. A., Mietus, J. E., Grant, K. A., Goldberger, A. L., Bennett, A. J., & Willard, S. L. (2007). Effects of chronic moderate alcohol consumption and novel environment on heart rate variability in primates (macaca fascicularis). *Psychopharmacology, 192*(2), 183–191.

Sumi, Y., Nakayama, C., Kadotani, H., Matsuo, M., Ozeki, Y., Kinoshita, T., , … Hasegawa-Ohira, M., et al. (2020). Resting heart rate variability is associated with subsequent orthostatic hypotension: Comparison between healthy older people and patients with rapid eye movement sleep behavior disorder. *Frontiers in Neurology, 11*, 567984.

Tsuji, H., Larson, M. G., Venditti, F. J., Manders, E. S., Evans, J. C., Feldman, C. L., & Levy, D. (1996). Impact of reduced heart rate variability on risk for cardiac events. The Framingham heart study. *Circulation., 94*(11), 2850–2855.

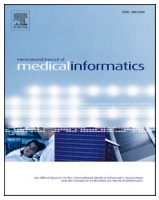
Wong, C. H., Siah, K. W., & Lo, A. W. (2019). Estimation of clinical trial success rates and related parameters. *Biostatistics., 20*(2), 273–286.

International Journal of Medical Informatics 163 (2022) 104786

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf



Automating and improving cardiovascular disease prediction using Machine learning and EMR data features from a regional healthcare system

Qi Li [a], Alina Campan [b], Ai Ren [a], Wael E. Eid [c,d,e,f,*]

[a] School of Business, State University of New York at New Paltz, New Paltz, NY, USA
[b] Department of Computer Science, Northern Kentucky University, Highland Heights, Kentucky, USA
[c] Department of Internal Medicine, Division of Endocrinology, St. Elizabeth Physicians Regional Diabetes Center, Covington, KY, USA
[d] Department of Internal Medicine, University of Kentucky College of Medicine, Lexington, KY, USA
[e] Department of Internal Medicine, Division of Endocrinology, University of South Dakota Sanford School of Medicine, Sioux Falls, SD, USA
[f] Department of Internal Medicine, Division of Endocrinology, Alexandria University, Alexandria, Egypt

ARTICLE INFO

Keywords:
Cardiovascular disease
Machine learning
Electronic health record
Risk
Mass screening

ABSTRACT

Background: The ACC/AHA Pooled Cohort Equations (PCE) Risk Calculator is widely used in the US for primary prevention of atherosclerotic cardiovascular disease (ASCVD), but may under- or over-estimate risk in some populations. We therefore designed an automated, population-specific ASCVD risk calculator using machine-learning (ML) methods and electronic medical record (EMR) data, and compared its predictive power with that of the PCE calculator.

Methods and Findings: We collected data from 101,110 unique EMRs of living patients from January 1, 2009 to April 30, 2020. ML techniques were applied to patient datasets that included either only cross-sectional (CS) features, or CS combined with longitudinal (LT) features derived from vital statistics and laboratory values. We compared the utility of the models using a proposed new cost measure (Screened Cases Percentage @ Sensitivity level).

All ML models tested achieved better predictive power than the PCE risk calculator. The random forest ML technique (RF) applied on the combination of CS and LT features (RF-LTC) produced the best area under curve (AUC) score of 0.902 (95% confidence interval (CI), 0.895–0.910). To detect 90% of all positive ASCVD cases, the best ML model required screening only 43% of patients, while the PCE risk calculator required screening 69% of patients.

Conclusions: Prediction models built using ML techniques improved ASCVD prediction and reduced the number of screenings required to predict ASCVD when compared with the PCE calculator, alone. Combining LT and CS features in the ML models significantly improved ASCVD prediction compared with using CS features, alone.

1. Introduction

Atherosclerotic cardiovascular disease (ASCVD) carries immense health and economic implications, globally. North America, the Middle East, and Central Asia carry the highest prevalence of CVD, while Eastern Europe and Central Asia have the highest mortality rates attributable to CVD [1]. Risk assessment is a critical step in primary prevention. In high-risk individuals, providing ASCVD risk scores may reduce risk by initiating preventive interventions and decreasing risk from diagnostic procedures [2–4]. In clinical practice, there is

Abbreviations: AUC, Area under curve; BMI, Body mass index; CVS, Cerebrovascular stroke; CHF, Congestive heart failure; CA, Coronary angiogram; CABG, Coronary artery bypass graft; CCTA, Coronary computed tomography angiogram; CS, Cross-sectional; LTC, Cross-sectional and longitudinal features; DM, Diabetes mellitus; EMR, electronic medical record,; HTN, Essential hypertension; FHX, Family history; HbA1c, Hemoglobin A1c; HDL-C, High-density lipoprotein cholesterol; HTN, Hypertension; CVS, Ischemic cerebrovascular stroke; LR, Logistic regression; LT, Longitudinal; LDL-C, Low-density lipoprotein cholesterol; MAX, Maximum; MIN, Minimum; ML, Machine learning; NB, Naïve Bayes; NN, Neural networks; non-HDL-C, Non high-density lipoprotein cholesterol; OB, Obesity; PCI, Percutaneous coronary intervention; PAD, Peripheral artery disease; PCE, Pooled cohort equation.

* Corresponding author at; Department of Endocrinology, St. Elizabeth Physicians Regional Diabetes Center, 1500 James Simpson Jr. Way, Suite 301, Covington, Kentucky, USA.

E-mail address: Wael.eid@usd.edu (W.E. Eid).

https://doi.org/10.1016/j.ijmedinf.2022.104786
Received 15 February 2022; Received in revised form 23 April 2022; Accepted 25 April 2022
Available online 29 April 2022

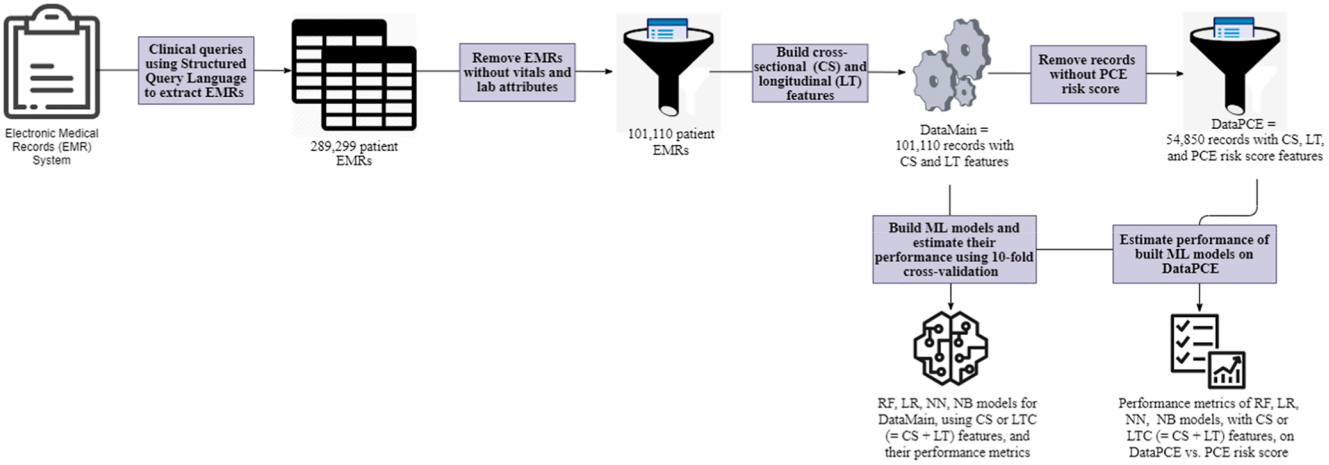


Fig. 1. Workflow for building and testing ML models. Data is extracted from EMRs and filtered. ML models are built on DataMain, compared with each other for performance, and compared against the PCE risk score for the DataPCE subgroup. Abbreviations: ASCVD, atherosclerotic cardiovascular disease; CS, cross-sectional; EMR, electronic medical record; LT, longitudinal; machine-learning models: RF, random forest; LR, logistic regression; NN, neural networks; NB, naïve Bayes.

heightened interest in optimizing ASCVD risk scores to improve cost effectiveness and to reduce risks associated with expensive or invasive testing [5,6]. Current guidelines employ pooled cohort equations (PCE) to calculate the 10-year risk for hard ASCVD [non-fatal myocardial infarction (MI), non-fatal stroke, or death from CHD or stroke] and to guide therapeutic decisions [4,7]. Scores from other studies are available for selected ASCVD risk outcomes and selected populations [8–11].

Various models currently are used in clinical practice to predict ASCVD; however, these models may misclassify risk either by underestimating or overestimating actual observed risk in populations with different comorbidities or with different demographic or socioeconomic determinants [2,5,12–14]. Inaccurate risk prediction precipitates potential over- or under-treatment and can be a contributing factor to clinical inertia among physicians and patients [2] Available risk calculators increasingly are embedded in EMR platforms that have decision support; [2] however, a need remains for tools that provide a more comprehensive estimate of ASCVD risk over time, and that are indicative of near-term risks not estimated by the current PCE calculators [2,5,8].

Machine learning (ML) leverages statistics, mathematics, and computer science to create data-driven predictive tools [15]. It is applied in various medical decision-making systems [16–19] and has shown equal or better performance compared with human-made risk prediction decisions in cardiology [5,12]. Longitudinal data from EMR systems can be used in ML to facilitate ASCVD clinical risk assessment [12]. Previous studies have focused on applying ML in ASCVD-related events detection by combining CS clinical variables with or without coronary artery calcium (CAC) scores [12,20]. Although CAC is a robust, cost-effective re-classifier for most ASCVD risk calculators (as a score and as a distribution), [4,5,12] barriers still impede its widespread use [21,22]. In this study, we developed a clinically-based ASCVD prediction model that can fill the gap in current methodologies.

2. Material and methods

We conducted a retrospective, records-based, longitudinal study using datasets from unique EMRs of living patients maintained by a US regional healthcare system. Using a dynamic EMR-based clinical decision-support tool, we used structured query language (SQL) to extract data from records of patients in the St. Elizabeth Health Care System (Kentucky, USA) who had a clinical encounter between January 1, 2009 and April 30, 2020 that involved checking low-density lipoprotein cholesterol (LDL-C). The query identified every record of living patients who had a documented LDL-C level throughout the identified date range. Since statin treatment lower LDL-C values with expected percentages, we used a validated formula (last recorded LDL-C multiplied by 1.43) to calculate the estimated pretreatment LDL-C for all individuals with an active statin prescription at the time of the last recorded LDL-C [4,23–29]. Data used in this research were anonymized according to US Health Insurance Portability and Accountability Act (HIPAA) regulations and are available upon reasonable request from the author with support from the St. Elizabeth Healthcare Clinical Research Institute. The study was approved by the St. Elizabeth Health Care Institutional Review Board and a waiver for informed consent was approved, allowing for retrospective data anonymization.

A total of 289,299 inpatient and outpatient records were screened. Records containing pertinent laboratory and vital sign data for the study

Table 1
Diagnostic Criteria for Comorbidities in the Study Population.

Diagnosis	Diagnostic criteria	Reference
Atherosclerotic cardiovascular disease (ASCVD)	Having either coronary artery disease (CAD), cerebrovascular stroke (CVS), or peripheral artery disease (PAD)	
Atherosclerotic coronary artery disease (CAD)	Active CAD diagnosis or ICD-10: I20, I21, I22, I23, I24, or I25 on the EMR problem list or having at least 3 instances of CAD appearing as an encounter diagnosis in the last 2 years or at least 3 CAD claim diagnoses in the last 2 years	[32]
Premature coronary artery disease (Premature CAD)	CAD occurring before age 55 years in males or 60 years in females	[33]
Ischemic cerebrovascular stroke (CVS)	Active CVS diagnosis or ICD-10: I63, I74, or I75 on the EMR problem list	[32]
Peripheral artery disease (PAD)	Active PAD diagnosis or ICD-10: I63, I74, or I75 on the EMR problem list	[32]
Diabetes mellitus (DM)	Active DM diagnosis on the EMR problem list or HbA1c $\geq$ 6.5% more than once or random peripheral blood glucose $>$ 200 mg/dl plus HbA1c $\geq$ 6.5% and no gestational diabetes	[34]
type 1 or type 2		
Obesity (OB)	Active obesity diagnosis on the EMR problem list or most recent BMI $\geq$ 30 kg/m^2	[35]
Essential hypertension (HTN)	Active essential HTN diagnosis on the EMR problem list	[36]
Congestive heart failure (CHF)	Active CHF diagnosis on the EMR problem list	[37]

Table 2
Demographic, Clinical Characteristics, and Diagnostic Interventions for the DataMain Group and for the DataPCE Subgroup.

Features	DataPCE (n = 54 850)		DataMain (n = 101 110)	
	ASCVD (n = 6339, 11.56%)	No ASCVD (n = 48511, 88.44%)	ASCVD (n = 17578, 17.39%)	No ASCVD (n = 83532, 82.61%)
Age (yrs): Mean (Standard Deviation)	65.76 (±8.74)	59.13 (±10.06)	69.32 (±11.63)	56.22 (±15.71)
Latest available insurance carrier:				
Carrier_General Managed Care	Total number: 5516		Total number: 9275	
Yes	385	5131	888	8387
No	5954	43,380	16,690	75,145
Carrier_Humana	Total number: 2165		Total number: 3785	
Yes	145	2020	322	3463
No	6194	46,491	17,256	80,069
Carrier_Medicaid	Total number: 229		Total number: 746	
Yes	34	195	120	644
No	6305	48,316	17,458	828,888
Carrier_Medicaid Managed Care	Total number: 3260		Total number: 8022	
Yes	358	2902	940	7082
No	5981	45,609	16,638	76,450
Carrier_Medicare	Total number: 8153		Total number: 17,447	
Yes	1580	6573	5505	11,942
No	4759	41,938	12,073	71,590
Carrier_Medicare Managed Care	Total number: 8673		Total number: 16,302	
Yes	1711	6962	5069	11,233
No	4628	41,549	12,509	72,299
Carrier_SelfPay	Total number:68		Total number: 158	
Yes	10	58	22	136
No	6329	48,453	17,556	83,396
Carrier_United Healthcare	Total number: 6358		Total number: 10,413	
Yes	379	5979	774	9639
No	5960	42,532	16,804	73,893
Carrier_Worker's Comp	3854		6594	
Yes	446	3408	961	5633
No	5893	45,103	16,617	77,899
Gender:				
Female = 1	2784 (43.92%)	27,389 (56.46%)	7173 (40.81%)	46,903 (56.15%)
Male = 0	3555 (56.08%)	21,122 (43.54%)	10,405 (59.19%)	36,629 (43.85%)
Hypertension	45,344 (84.30%)	35,051 (72.25%)	13,991 (79.59%)	52,843 (63.26%)
Diabetes T1 and T2	3071 (48.45%)	17,728 (36.54%)	9806 (55.79%)	32,091 (38.42%)
Obesity	3835 (60.50%)	33,889 (69.86%)	8787 (49.99%)	54,826 (65.63%)
PCE risk score				
Null = unavailable	0	0	11,239	35,021
Mean (Standard Deviation)	21.62 (14.78)	12.16 (11.70)		
PCE [0, 5)	633 (633/6339 = 10%)	16,456 (16456/48511 = 33.92%)	633	16,456
PCE [5, 18.6)	2544 (2544/6339 = 40%)	21,196 (21196/48511 = 43.69%)	2544	21,196
PCE [18.6,)	3162 (3162/6339 = 50%)	10,859 (10859/48511 = 22.38%)	3162	10,859
Diagnostic tests & interventions				
CAC	135 (2.13%)	703 (1.45%)	275 (1.56%)	955(1.14%)
Not done				
Done				
Range:				
0	7	375	13	485
1–100	48	247	77	341
101–300	36	48	67	72
>300	44	33	118	57
CCTA				
Not done	5921 (92.98%)	46,968 (96.72%)	16,875 (95.64%)	81,436 (97.40%)
Done	445 (7.02%)	1589 (3.28%)	767 (4.36%)	2171 (2.60%)
Range:				
0	42	1043	71	1382
1–2	179	436	277	607
3	86	50	161	81
4–5	111	14	194	26
Coronary angiogram				
Not done	2863 (45.16%)	46,454 (95.76%)	8314 (47.30%)	80,378 (96.22%)
with subsequent diagnosis/procedure:				
PCI				
	1484 (2.71%)	0	4455 (4.41%)	0
CABG	962 (1.75%)	0	2769 (2.74%	0
PCI done	1484 (2.71%)	0	4455 (4.41%)	0
CABG done	962 (1.75%)	0	2769 (2.74%)	0
PCI and CABG done	220 (0.40%)	0	692 (0.68%)	0
PCI or CABG done	2226 (4.06%)	0	6532 (6.46%)	0

timeframe were selected to build the ML models (the DataMain group, n = 101,110). Records containing PCE risk scores were assigned to the DataPCE subgroup (n = 54,850). Our EMR system uses patients' most recent data records to compute the PCE score following current US clinical guidelines [4,30]. PCE score can be calculated for patients with unstable angina (ICD10 – I20.0) or with atherosclerotic heart disease of the native coronary artery (ICD10- I25.1), but is not calculated for ineligible patients [i.e., those with MI or cerebrovascular stroke (CVS)] or if the patient's record is missing components required for PCE calculation [7]. Our study design (Fig. 1) enabled us to compare the predictive power of the ML models (S1 Table) with the existing PCE score and also to explore the impact of longitudinal features (S2 Table) on the models' predictive power, since these features summarize 10 years of patient data. We also used the IJMEDI medical AI assessment checklist to cross reference our model design, result reports, and discussion [31].

Diagnostic criteria for comorbidities are defined in Table 1. ASCVD refers to patients diagnosed as having either coronary artery disease (CAD), cerebrovascular stroke (CVS), or peripheral artery disease (PAD).

When patients meet the CAD/CVS inclusion criteria, their record is 'time stamped' as having the disease, and the remains in that registry throughout the study. A patient might have more than one disease (e.g., CAD and CVS) and/or other comorbidities.

2.1. Study population

Table 2 shows detailed demographic, clinical characteristics, and diagnostic interventions for the DataMain group (101,110 records) and for the DataPCE subgroup (54,850 records).

DataMain includes 17% of patients (17,578) with ASCVD (mean age, 69.32 ± 11.63 years) and 83% of patients without ASCVD (mean age, 56.22 ± 15.71 years). DataPCE includes 12% of patients (6,339 records) with ASCVD (mean age, 65.76 ± 8.74 years), and 88% without ASCVD, (mean age, 59.13 ± 10.06 years).

2.2. Cross-sectional features and longitudinal features

We used CS and LT features, respectively, to build the ML models. The 31 CS features included demographics, aggregate risk scores, family history, clinical care group, laboratory values, vital signs, and comorbidities (Table 3a). To capture time-sequential LT features, including lipid profile, HbA1c, and blood pressure, we designed an additional 63 LT features (Table 3b) where for every patient record and time-sequential feature, we calculated the minimum, maximum, average, reading-time range, reading-value range, standard deviation of the readings, average reading days, and coefficient of variation for the readings. For patients without ASCVD diagnosis, we calculated these statistics based on all readings collected throughout the study timespan. For patients with an ASCVD diagnosis, we considered only readings registered in the EMR preceding the ASCVD diagnosis. Our data include two possible scenarios for which values might not be available and the process methods are described in S1 Text.

2.3. ML models

To predict the likelihood a patient would develop ASCVD, we built automated models for each of the selected four ML methods (logistic regression (LR); naïve Bayes (NB); neural networks (NN); and random forest (RF) (S1 Table). These models were built once using only CS features as predictors and once with a combination of CS and LT features (LTC) as predictors (S2 Table). The overall experiments are listed in Table 4. Given the two-by-two contingency table (S3 Table), we used several measures to evaluate the predictive performance of each ML model (S4 Table) based on 10-fold cross-validation on the evaluation datasets (DataMain and DataPCE, respectively).

2.4. Screened cases percentage @ Sensitivity level

Since a key goal for any prediction model is to achieve a high sensitivity while screening as few patients as possible, we created the Screened Cases Percentage@Sensitivity (SCP@Sensitivity) to measure

Table 3a
Cross-sectional (CS) features used in the ML models.

Feature	Description(s)
Demographics	• Age • Gender • Age categories: <30, [30,40), [40,50), [50,55), [55,60), [60,65), [65,70), [70,75), [75,80), >=80
Aggregate risk scores	• ASCVD 10-year risk score (PCE) • ASCVD 10-year risk score (PCE) categorical, discretized to 3 categories: null value, <5, and ≥ 5 • Numerical score for the family history group of the Dutch Lipid Clinic Network (DLCN) (0,1) • Hierarchical Condition Category Risk Score (Risk Score) • Numeric score for the LDL-C group of the DLCN (0,1,3,5,8)
Family history (FHX)	• Family history of any coronary artery disease (FHX-++) • Family history of premature coronary artery disease (FHX Premature) • Family history of non-premature coronary artery disease (FHX Non-premature)
Clinical care group	• Current insurance carrier (Carrier) • Current primary care provider is an employee of the healthcare system where the study is conducted or not (SEP Affiliation) • Have seen endocrinologist in the past or not (Saw Endo) • Patient has account with the MyChart personal health record or not (MyChart)
Laboratory values	• Maximum LDL-C (whether EHR-documented or last estimated pretreatment) ≥ 190 mg/dL at least twice (LDL-C > 190 x2) • Maximum LDL-C (whether EHR-documented or last estimated pretreatment) ≥ 190 mg/dL at least once (LDL-C > 190) • The last LDL-C reading before a CAD diagnosis, or the last LDL-C reading in absence of CAD LDL (Num Before CAD Avg) • The last Non-HDL-C reading before a CAD diagnosis, or the last Non-HDL-C reading in absence of CAD (Non-HDL-C Num Before CAD Avg) • The last VLDL-C reading before CAD diagnosis, or the last VLDL-C-reading in absence of CAD (VLDL-C-Num-Before-CAD-Avg) • Maximum Lp(a) (MAX LPA) • Maximum Lp(a) group (whether MAX LPA < 29 or > 50 or null value) (MAX LPA cat)
Vital signs	• Last mean arterial blood pressure (MAP) reading before a CAD diagnosis, or the last MAP reading in absence of CAD (MAP BEFORE CAD Avg) • Last systolic arterial blood pressure (SYS) reading before a CAD diagnosis, or the last SYS reading in absence of CAD (SYS BP BEFORE CAD Avg) • Last diastolic arterial blood pressure (DIA) reading before a CAD diagnosis, or the last DIA reading in absence of CAD (DIA BP BEFORE CAD Avg)
Comorbidities	• Diabetes (T1 or T2) (Yes or No, Number of months of having diabetes before being CAD diagnosis) • Hypertension (Yes or No, Number of months of having HTN before CAD diagnosis) • Obesity (Yes or No, Number of months of having OB before CAD diagnosis)

Table 3b
Longitudinal features (LT) for vital signs* and laboratory values.†.

Feature	Description
Minimum (MIN)	Lowest value of all patient recorded values for a feature
Maximum (MAX)	Highest value of all patient recorded values for a feature
Average (MEAN)	Average value for all recorded values of the feature
Readings number (COUNT)	Number of recorded readings for each measure
Reading-time range (TRANGE)‡	Time difference, in days, between the first and last recorded values for the feature
Reading-value range (VRANGE)§	Difference between the smallest and largest recorded value for the feature
Standard deviation (STDEV)	Amount of variation between the recorded values for the feature
Average reading days (Avg-Test-Day)||	Average time, in days, between consecutive recorded values for the feature
Coefficient of variation (CV)#	Standard measure of dispersion of a probability distribution or frequency distribution

*Vital signs: diastolic BP, systolic BP, mean arterial pressure (MAP).
† Laboratory values: LDL-C, total cholesterol, HDL-C, non-HDL-C, triglycerides, HbA1c.
‡ Reading-time range: to determine if the length of the patient's care (reflected by the history of vital and laboratory records) has had any impact on the risk for developing an ASCVD.
§ Reading-value range: might be significant for the cases with large reading differences.
|| Average reading days: to determine if the frequency of patient care, as reflected by patients' vital signs (e.g., BP), being checked in a professional environment and more frequent laboratory tests (e.g., lipid profile, LDL-C) have any impact on the risk for developing ASCVD.
Coefficient of variation (also known as the relative standard deviation): to study whether the fluctuation of laboratory values and vital sign readings contribute to a patient's risk for developing ASCVD.

Table 4
ML Models Evaluated Using Feature Sets.

	Feature Set*	
ML technique	CS	LTC
Naive Bayes	NB-CS	NB-LTC
Logistic regression	LR-CS	LR-LTC
Neural network	NN-CS	NN-LTC
Random forest	RF-CS	RF-LTC

* The labeling system for each experimental group included the ML acronym (NB, LR, NN, RF) combined with the feature set used (CS when using only cross-sectional features, LTC when using cross-sectional features *and* longitudinal features together).

the percentage of the population that must be screened, in order to achieve the desired sensitivity level:

$$\text{SCP @Sensitivity(S)} = \frac{|\text{Subpopulation of patients who must be screened to achieve the target sensitivity level of S}|}{|\text{Overall Patient Population}|}$$

where $|X|$ is a notation for the cardinality of a set X.

This measure also can be used clinically to assess the resources needed to screen patients so to achieve a desired sensitivity level in the overall population, which ideally could reduce unnecessary testing. Theoretically, the percentage of patients need to be screened in order to detect all ASCVD positive patients, i.e., SCP@Sensitivity(1), can be as low as 11.56% in the DataPCE cohort, which is total ASCVD patients (6339) out of DataPCE (54,850) cases.

3. Results

We compared all models and reported their AUC scores in Section 3.1 in order to find the best model and best feature sets. We analyzed the features used on the best model in Section 3.2 to discover the important features. We compared the models with or without PCE as a feature, in Section 3.3, to answer if the PCE as a feature contributes to our models. We further compared our model to PCE calculator, in Section 3.4, to answer the question if our model performance better than the current PCE calculator. We compared our model statistically in Section 3.5. We reported our novel measurement, screened cased percentages @ sensitivity level, in Section 3.6. In order to help finding the optimal threshold, the performance of our best model, according to AUC score, with its probability threshold is reported in Section 3.7.

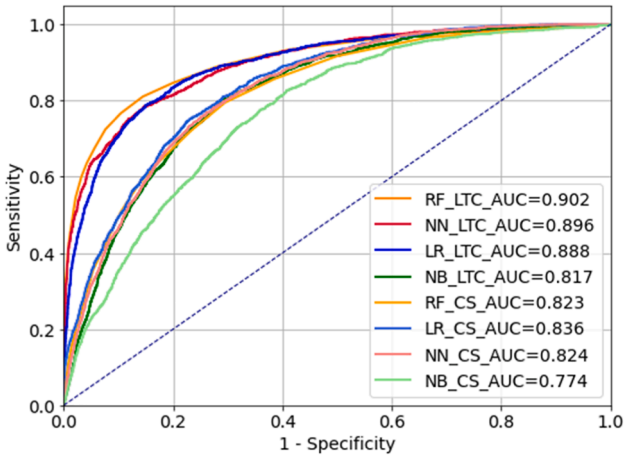


Fig. 2. Area under the curve (AUC) for DataMain. As a measure of individual model performance for predicting ASCVD in the DataMain cohort, RF-LTC produced the best AUC for ASCVD prediction.

3.1. Model performance

The RF-LTC model produced the best AUC (Fig. 2) and performance metrics (Table 5) for ASCVD prediction (AUC = 0.902; 95% CI, 0.895–0.910) compared with RF-CS (AUC = 0.82; 95% CI, 0.814–0.831), NN-LTC (AUC = 0.896; 95% CI, 0.889–0.904), LR-LTC (AUC = 0.888; 95% CI, 0.881–0.896), and NB-LTC (AUC = 0.817; 95% CI, 0.809–0.826). The AUC in Fig. 2 is an aggregate measure based on all probabilities that are associated with paired sensitivity and specificity on the ROC curve.

NN with backpropagation algorithm consists of three layers: input layer, hidden layer (150 nodes), and output layer. Other important parameters such as learning rate and momentum are 0.01, and 0.9, respectively. RF uses bootstrap method with the criterion of gini without limiting the max depth. LR is built with important parameters of C = 1.0, max iteration of 100, and l2 penalty method. And NB uses Gaussian naive Bayes without priors. We use Scikit-learn package to implement our algorithms [38].

3.2. Impact of features used to build the models

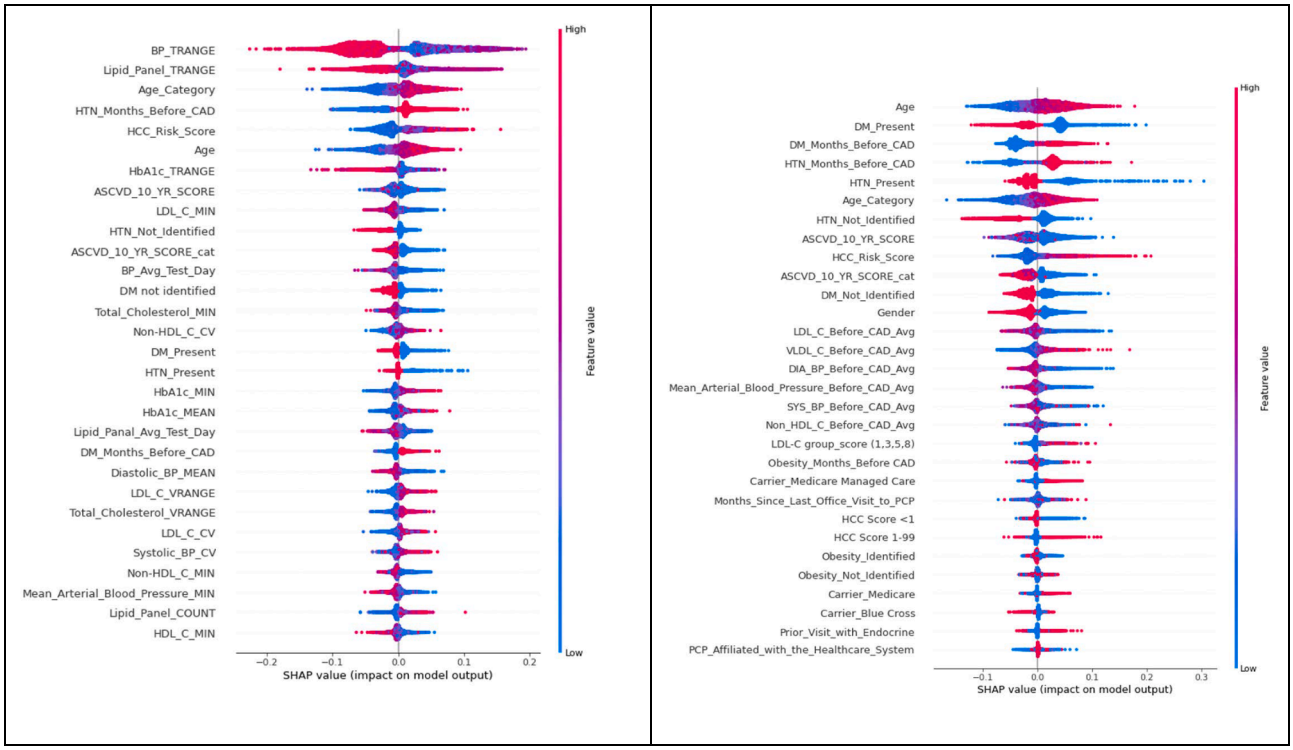
We used Shapley Additive Explanations (SHAP) [39] Diagram, in Fig. 3, to illustrate the relative importance of features for the machine learning models for the DataMain dataset. Age, comorbidities, and aggregate risk scores were the most predictive features in the RF-CS model, followed by LDL-C values. The most predictive features in the LTC model included BP, lipid levels, and HbA1c. In both models (RF-LTC and RF-CS), age was one of the most important features.

Moreover, we noted that PCE features were not listed in the top 5 most important features in SHAP diagrams. ASCVD_10_YR_SCORE_score was ranked at the 8th in both RF_LTC and RF_CS while ASCVD_10_YR_SCORE_cat was ranked at the10th and 11th in RF_LTC and RF_CS models respectively.

Table 5
RF-LTC Model Performance on DataMain for Various Probability Threshold Values*†.

Cut-off Probability	AUC	NPV	Specificity	F0	PPV	Sensitivity	F1	SCP@Sensitivity
0.05	0.673	0.982	0.378	0.546	0.247	0.967	0.393	68.2%
0.1	0.759	0.975	0.590	0.735	0.323	0.928	0.479	50.0%
0.15	0.803	0.968	0.718	0.824	0.398	0.888	0.550	38.8%
0.2	0.820	0.960	0.797	0.871	0.466	0.842	0.600	31.4%
0.25	0.826	0.953	0.853	0.900	0.533	0.799	0.640	26.1%
0.3	0.824	0.945	0.892	0.918	0.596	0.756	0.666	22.1%
0.35	0.816	0.938	0.920	0.929	0.651	0.712	0.680	19.0%
0.4	0.806	0.932	0.941	0.936	0.704	0.672	0.687	16.6%
0.45	0.793	0.925	0.956	0.940	0.751	0.629	0.685	14.6%
0.5	0.776	0.917	0.968	0.942	0.795	0.584	0.673	12.8%
0.55	0.759	0.910	0.977	0.942	0.832	0.541	0.655	11.3%
0.6	0.740	0.903	0.983	0.941	0.861	0.497	0.630	10.0%
0.65	0.720	0.895	0.988	0.940	0.892	0.451	0.599	8.8%
0.7	0.699	0.888	0.992	0.937	0.916	0.406	0.563	7.7%
0.75	0.676	0.880	0.995	0.934	0.934	0.358	0.517	6.7%
0.8	0.652	0.872	0.997	0.930	0.952	0.307	0.465	5.6%
0.85	0.626	0.864	0.998	0.926	0.964	0.254	0.402	4.6%
0.9	0.598	0.855	0.999	0.922	0.975	0.197	0.328	3.5%
0.95	0.568	0.846	1.000	0.916	0.987	0.136	0.239	2.4%

* Various methods may be used to calculate the probability threshold, *t*: i) assigned as 0.5 (halfway within the 0–1 range); ii) based on model performance [select the value from the set (0.05, 0.1, 0.15, …, 0.95) that achieves the best performance metric (AUC, PPV or sensitivity); or iii) based on the SCP, PPV, and sensitivity of the model for that threshold value (the higher *t*, the higher the PPV, but the lower the sensitivity and SCP of the model).
† Color variation indicates low (lightest) to high (darkest) AUC.



(a) RF-LTC **(b) RF-CS**

Fig. 3. Shapley Additive Explanations (SHAP)[39] **Diagram**. This illustrates the relative importance of features for: (a) longitudinal features plus cross-sectional features (LTC), and (b) cross-sectional features (CS) only on the RF models, since RF-LTC showed the best performance according to the AUC measure. The blue and red points in each row represent data cases having low to high values of the specific variable: blue for low and red for high. The X-axis represents the SHAP value, which quantifies the variable's impact on the model [i.e., tendency to drive the predictions toward an event (positive SHAP value, i.e., ASCVD) or non-event (negative SHAP value, i.e., non-ASCVD)]. The top 20 variables contributing most to the class separability of the model are shown in the figure. Age, comorbidities, and aggregate risk scores were the most predictive features in the RF-CS model, followed by LDL-C features. The BP TRANGE measure, lipid TRANGE measure, and HbA1c TRANGE measure were the most predictive LT features. In both models, age was one of the most important features. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

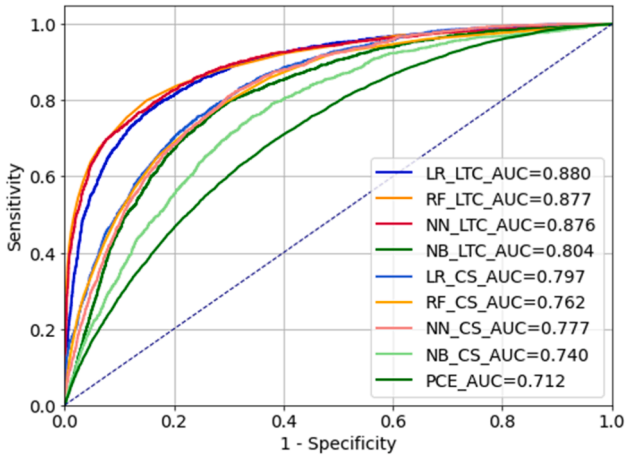


Fig. 4. Area under the curve (AUC) for DataPCE. As a measure of individual model performance for predicting ASCVD in the DataPCE group, LR-LTC produced the best AUC.

3.3. ML comparison with the PCE features and without the PCE features

We also evaluated the same models in the dataset, DataPCE, with PCE features (PCE scores and PCE categorical) and without PCE features in order to evaluate the impact of the PCE score on model performances. The results without PCE features are as shown in S2 Fig. The NN-LTC model produced the best AUC score of 0.896, which is the same as the one with PCE features. And the AUC score of RF_LTC model is 0.894, which drops 0.006.

3.4. ML comparison with the PCE calculator

We compared the automated ML methods with the PCE 10-year risk score currently used in clinical settings. Although the ML models were built using the DataMain dataset, comparison with the PCE score was performed using the DataPCE dataset, because PCE risk scores were available only within this dataset. Comparison of the AUC results from 10-fold cross-validation are shown in Fig. 4.

All the ML models produced a better ASCVD prediction than the PCE risk calculator (AUC = 0.712; 95% CI, 0.700–0.730). The LR-LTC model produced the best AUC in the DataPCE dataset (AUC = 0.880; 95% CI, 0.867–0.894).

3.5. Calibration curves and Net reclassification index

Model calibration was performed in order to assess the certainty of a given new observation from each new model belonging to each of the already established classes as shown in S1Fig [40]. The Brier score ranges from 0 to 1. The Brier score of 0 is the best achievable score, and 1 is the worst achievable score. The Brier scores of LR, RF, and NN models with LTC were 0.104, 0.079, and 0.083, respectively, which were all better than the Brier scores of models with CS feature (0.126, 0.113, 0.114, respectively). And the Brier Score of PCE was 0.262.

Net reclassification indexes (NRI) are presented in S6 Table [41]. Continuous NRI was performed to compare various machine learning models. NRIs were 0.3239, 0.3949, 0.8532, and 0.5462 when NB, LR, NN, and RF models with LTC features were compared to the models with CS features. All the improvements were significantly ($P < 0.001$), which implied the importance of the derived longitudinal features in building ASCVD prediction models (S6 Table a). We compared the models built on DataMain set, and found that RF-LTC was significantly better than NN-LTC model, NN-LTC model was significantly better than LR-LTC model, and LR-LTC model was significantly better than NB-LTC in the DataMain, as shown in S6 Table b. We also evaluated that NB, LR, NN, and RF models with LTC features were significantly better than PCE models (S6 Table c).

3.6. Screened cases percentage at sensitivities 50% and 90% [SCP@Sensitivity(0.5) and SCP@Sensitivity(0.9)]

Since risk-prediction methods must be accompanied by laboratory and other diagnostic tests that either confirm or refute the diagnosis, the method is more valuable if it correctly predicts a higher percentage of potential positive cases (e.g., ASCVD), while requiring further tests on a lower percentage of the total population (SCP@Sensitivity). If a method is adjusted to increase its sensitivity, it may incur increased costs, based the percentage of the population requiring testing.

At a chosen target sensitivity level (S), a clinically useful ML model will have a higher PPV and a lower SCP@Sensitivity(S). Since 5% and 20% are popular PCE values in clinical practice for discriminating between low- and high-risk patients [4,30], we compared SCP@Sensitivity of the different ML models at sensitivity levels 0.90 (corresponding to the PCE risk score of 5%) and 0.50 (corresponding to the PCE risk score of 18.5%).

In order to achieve high sensitivity (90%), theoretically, SCP@Sensitivity(0.9) can be as low as 10.40% in the DataPCE cohort. The PCE method requires testing all patients with a score of 5% or above which, in our dataset, was 68.8% of the total population [i.e., SCP@Sensitivity (0.90) = 0.688 for the PCE risk score.], as shown in S7 Table. However, the NN-LTC model required screening only 43.4% of the population [i.e., SCP@Sensitivity(0.90) is 43.3%] with a probability cutoff of 3.2% (S7 Table). All the ML models performed better than the PCE score relative to SCP@Sensitivity(0.90) (all ML models require screening a lower percentage of cases than the PCE model in order to achieve sensitivity of 90%).

Theoretically, in order to cover 50% of the true positive cases (i.e., sensitivity of 0.5), SCP@Sensitivity(0.5) can be as low as 5.78% in the DataPCE cohort. As shown in S8 Table, the PCE calculator requires screening 25.6% of the population [i.e., SCP@Sensitvity(0.50) is 25.6%] at a probability cutoff of 18.6% (i.e., PCE score = 18.6), while the RF-LTC model requires screening 7.1% of the population to achieve the same sensitivity [i.e., SCP@Sensitivity (0.50) is 7.1%] at a probability cutoff of 18.6% (S8 Table). All the ML models performed better than the PCE calculator relative to SCP@Sensitivity(0.50). That is, all ML models required screening a lower percentage of cases than the PCE to achieve a sensitivity of 50%.

3.7. Probability threshold for DataMain model performance

In clinical practice, the decision maker makes binary predictions [i.e., outcomes are either 1 (positive) or 0 (negative)], while ML models make a continuous prediction (i.e., a risk score in the 0–1 interval). This predicted risk score requires an interpretation based on a threshold value t: given a selected probability threshold value t, a predicted risk score $\geq t$ is interpreted as a positive prediction, and a predicted risk score $< t$ is interpreted as a negative prediction. Since the RF-LTC model has the best overall AUC values according to Fig. 2, we reported the RF-LTC model performance with their corresponding threshold probability values (Table 5) in order to help clinicians choosing the optimal threshold according to different needs. The AUC in Table 5 is the point-level AUC, which is associated with one paired sensitivity–specificity according to the optimal threshold. The other two best model performance with their corresponding threshold probability values were in S9 Table (NN_LTC) and S10 Table (LR_LTC.). The other tested models never outperformed these three models.

According to the AUC, the best performance for RF_LTC model was for the threshold of 0.25, with the AUC value 0.826. While according to the F1, the best performance was for the threshold of 0.40 with the F1 value 0.687 and other metrics as shown in the table, the cut-off probability could be chosen according to clinical need. For example, if we need 80% sensitivity, we can choose the cut-off probability of 0.25 and if

we need 80% of PPV, we can choose the cut-off probability of 0.5. This is helpful to minimize misclassifying high-risk patients as low risk and to maximize opportunities to initiate potentially life-saving managements [20].

4. Discussion

Previous studies of ML ASCVD [12,20] estimators have shown variable outcomes. Some studies have supported the use of ML to augment cardiac imaging results, [42] while others have supported ML use for predicting a spectrum of ASCVD outcomes [5,12,20,43–46]. Motwani et al. [43] (n = 10,030), van Rosendael et al. [45] (n = 8,844), and Nakanishi et al. [12] (n = 66,636 asymptomatic patients) evaluated 3- and 5-year all-cause mortality and 10-year risk for CHD- and ASCVD-related mortality respectively, based on baseline CS metrics, including cardiac imaging. Ward et al. [20] (n = 262,923) evaluated 5-year ASCVD risk prediction using structured CS EMR data [20]. We found a similar mean ASCVD score for DataMain (7.19 ± 11.313), but a higher score for DataPCE (13.26 ± 12.47) compared with that of the ML model study by Nakanishi et al. [12].

Although we did not evaluate ASCVD-related deaths (records for deceased patients were excluded), we did evaluate current comprehensive ASCVD risk, given the effect of both CS and LT metrics prior to ASCVD diagnosis. Prediction outcomes of the Van Rosendael et al. [45] and Al'Aref et al. [5] studies were non-fatal MI and presence of obstructive CAD on CCTA, respectively, using CS clinical features and CCTA or CAC. Most studies incorporating ML in prediction models showed improved prediction accuracy, compared with using clinical features, CAC, or CT, alone; [12] however, some studies [43,45] may have incurred selection bias, because patients were referred for cardiac imaging due to clinical suspicion of CAD [45].

Our study like that of Ward [20], have found the ML models to have better performance than PCE score in ASCVD risk predication and can be widely applicable to subjects for whom the pooled cohort equations are not applicable [20].

Our study included 94 clinical variables (31 CS and 63 LT). Although the number of clinical features included has varied among studies [12,20,43] and few ML studies have integrated clinical variables into ASCVD prediction, none have assessed the effect of LT features on clinical outcome. To our knowledge, this is the first ML-based modeling study to assess current comprehensive ASCVD risk (excluding death) using both CS and LT clinical and laboratory data abstracted from EMR records of a large regional healthcare system.

In our study population (n = 101,110), we have shown that ML models integrating both CS and LT features can achieve superior prognostic performance on ASCVD prediction (Figs. 2 and 3). This emphasizes the importance of the numerous variables and cumulative risk, as shown in the SHAP figure (Fig. 3), leading to ASCVD, [47] and supports the concept of annual averaged assessment of single risk factors (e.g., LDL-C, BP, HbA1c) as a measure of pharmacotherapy adherence and assessment of long-term benefit [47,48]. Similar to other studies, [12,20] we found the key predictors for ASCVD are age, PCE risk score, HTN, and DM (T1 and T2). Our results show that incorporating novel features as LT features (laboratory data and vital signs) can capture data patterns and consequently can improve the model's predictive power compared with similar studies using primarily structured EMR CS features [20].

Our most accurate model, RF-LTC (0.902), achieved the highest AUC of those reported in previous studies, despite the different ASCVD outcomes and different populations studied [5,12,20,43–46]. Compared with the PCE risk calculator, the ML models in our study showed better performance for AUC, SCP@Sensitivity(0.9), and SCP@Sensitivity(0.5), indicating these models would require screening a lower percentage of the population in clinical practice. We propose using SCP@Sensitivity (percentage of patients requiring further diagnostic tests) as a new performance measure for evaluating ASCVD prediction models at the targeted sensitivity level, to facilitate estimation of the volume and associated costs required for further testing [5].

4.1. Study limitations

Although some studies showed that CAC and CCTA scores are important for estimating ASCVD-related disease in individuals with available risk scores, [2,8,49,50] only 2% of our study cohort had CAC and CCTA data, and were not included in our feature set. Despite this, we obtained an AUC of 0.92. In addition, although patient- or physician-reported signs and symptoms contribute to ASCVD diagnosis, our study included only structured EMR data. In the future, we expect to include signs and symptoms extracted from unstructured data, thus further improving the prediction power of our ML models. We did not validate our models in an independent external cohort, and therefore cannot generalize conclusions beyond our population. Including a calculated PCE in our ML models might create some information bias, because knowledge of the PCE score might alter a patient's treatment course and affect the development of ASCVD. However, our ML models included many other risk factors that contribute to CVD (beyond those used to calculate the PCE) that may mitigate such bias. Besides running the models without the PCE feature did not significantly reduce the accuracy of the models.

4.2. Study strengths

To our knowledge, this is the first study to assess the use of ML models to predict population- specific total ASCVD (excluding death) independent of the CAC score, and using both CS and LT EMR data for an entire regional healthcare system (including symptomatic and asymptomatic patients). In clinical practice, the short-term risk assessment capacity provided by our ML models may facilitate provider-patient communication of risk and healthy behavioral changes for ASCVD prevention (including among asymptomatic individuals) [8]. In contrast to similar ML studies [5,43,45] that employed a well-defined, preselected cohort, our study utilized a real dataset (i.e., not collected specifically for the study) from an EMR containing a wealth of descriptive clinical features representative of our study population and free from referral bias. We also propose the new SCP@Sensitivity measure to further assist cost effective screening for ASCVD which can be compared, in the future, to the commonly used Net benefit tool [51].

This study demonstrates that an ML approach integrating clinical CS and LT features can provide improved risk assessment models for ASCVD than the traditional PCE risk calculator. Summarized LT features for laboratory values and vital signs significantly improved the models' accuracy for ASCVD prediction. Since CS clinical data, and LT laboratory and vital records are readily available in many EMRs, automated ASCVD prediction is likely to become a routine tool for ASCVD-related disease assessment. Leveraging EMR data in ML models not only may improve clinical assessment and enhance early intervention with preventive therapies, [4] but also potentially can reduce the need for further diagnostic testing (e.g., CAC, CCTA or stress testing). Our team has initiated additional studies utilizing this research to examine more specific ASCVD endpoints.

Authors statement

The study was approved by the St. Elizabeth Health Care Institutional Review Board and a waiver for informed consent was approved, allowing for retrospective data anonymization.

Summary Table.

What is already known on this topic?

Machine learning based models are used for the ASCVD prediction

What did this study add to our knowledge?

The comparison of machine learning based ASCVD prediction to the commonly used ACC/AHA (PCE) risk calculator is evaluated and provides healthcare practitioner better estimate of the ASCVD risk.

ASCVD risk prediction was significantly improved by using both longitudinal and cross-sectional features from electronic medical record data. We also explored the most important features for the ASCVD prediction.

Our proposed Screened Cases Percentage@Sensitivity level (SCP@Sensitivity) measure may help to better estimate the number of screenings required to predict ASCVD over the whole population, without compromising sensitivity.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Additional contributions: We thank Amy Neil McBride, MS, MAP, for editing assistance, and the Information Systems department at St. Elizabeth Physicians for data extraction assistance.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Codes

Machine learning programming codes are publicly available through GitHub (https://github.com/QiLi-NewPaltz/CardiologyModeling).

Disclosures

Dr. Eid is on the Speaker Bureau of Amgen and Esperion Pharmaceuticals.

References

[1] S.S. Virani, A. Alonso, H.J. Aparicio, E.J. Benjamin, M.S. Bittencourt, C. W. Callaway, A.P. Carson, A.M. Chamberlain, S. Cheng, F.N. Delling, M.S.V. Elkind, K.R. Evenson, J.F. Ferguson, D.K. Gupta, S.S. Khan, B.M. Kissela, K.L. Knutson, C. D. Lee, T.T. Lewis, J. Liu, M.S. Loop, P.L. Lutsey, J. Ma, J. Mackey, S.S. Martin, D. B. Matchar, M.E. Mussolino, S.D. Navaneethan, A.M. Perak, G.A. Roth, Z. Samad, G.M. Satou, E.B. Schroeder, S.H. Shah, C.M. Shay, A. Stokes, L.B. VanWagner, N.-Y. Wang, C.W. Tsao, Heart Disease and Stroke Statistics—2021 Update: A Report From the American Heart Association, Circulation 143 (8) (2021), https://doi.org/10.1161/CIR.0000000000000950.

[2] D.M. Lloyd-Jones, L.T. Braun, C.E. Ndumele, S.C. Smith, L.S. Sperling, S.S. Virani, R.S. Blumenthal, Use of Risk Assessment Tools to Guide Decision-Making in the Primary Prevention of Atherosclerotic Cardiovascular Disease: A Special Report From the American Heart Association and American College of Cardiology, Circulation 139 (25) (2019), https://doi.org/10.1161/CIR.0000000000000638.

[3] Karmali KN, Persell SD, Perel P, Lloyd-Jones DM, Berendsen MA, Huffman MD. Risk scoring for the primary prevention of cardiovascular disease. Cochrane Database Syst Rev. 2017;3:CD006887. Epub 2017/03/16. doi: 10.1002/14651858.CD006887.pub4. PubMed PMID: 28290160; PubMed Central PMCID: PMCPMC6464686.

[4] Grundy SM, Stone NJ, Bailey AL, Beam C, Birtcher KK, Blumenthal RS, et al. 2018 AHA/ACC/AACVPR/AAPA/ABC/ACPM/ADA/AGS/APhA/ASPC/NLA/PCNA Guideline on the Management of Blood Cholesterol: Executive Summary. Circulation. 2018:CIR0000000000000624. Epub 2018/12/20. doi: 10.1161/CIR.0000000000000624. PubMed PMID: 30565953.

[5] Al'Aref SJ, Maliakal G, Singh G, van Rosendael AR, Ma X, Xu Z, et al. Machine learning of clinical variables and coronary artery calcium scoring for the prediction of obstructive coronary artery disease on coronary computed tomography angiography: analysis from the CONFIRM registry. Eur Heart J. 2020;41(3):359-67. Epub 2019/09/13. doi: 10.1093/eurheartj/ehz565. PubMed PMID: 31513271; PubMed Central PMCID: PMCPMC7849944.

[6] K.M. Chinnaiyan, P. Peyser, T. Goraya, K. Ananthasubramaniam, M. Gallagher, A. DePetris, J.A. Boura, E. Kazerooni, C. Poopat, M. Al-Mallah, S. Saba, S. Patel, S. Girard, T. Song, D. Share, G. Raff, Impact of a Continuous Quality Improvement Initiative on Appropriate Use of Coronary Computed Tomography Angiography, J. Am. Coll. Cardiol. 60 (13) (2012) 1185–1191.

[7] Goff DC, Jr., Lloyd-Jones DM, Bennett G, Coady S, D'Agostino RB, Sr., Gibbons R, et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. J Am Coll Cardiol. 2014;63(25 Pt B):2935-59. Epub 2013/11/19. doi: 10.1016/j.jacc.2013.11.005. PubMed PMID: 24239921; PubMed Central PMCID: PMCPMC4700825.

[8] M.O. Gore, C.R. Ayers, A. Khera, C.R. deFilippi, T.J. Wang, S.L. Seliger, V. Nambi, E. Selvin, J.D. Berry, W.G. Hundley, M. Budoff, P. Greenland, M.H. Drazner, C. M. Ballantyne, B.D. Levine, J.A. de Lemos, Combining Biomarkers and Imaging for Short-Term Assessment of Cardiovascular Disease Risk in Apparently Healthy Adults, JAHA 9 (15) (2020), https://doi.org/10.1161/JAHA.119.015410.

[9] McClelland RL, Jorgensen NW, Budoff M, Blaha MJ, Post WS, Kronmal RA, et al. 10-Year Coronary Heart Disease Risk Prediction Using Coronary Artery Calcium and Traditional Risk Factors: Derivation in the MESA (Multi-Ethnic Study of Atherosclerosis) With Validation in the HNR (Heinz Nixdorf Recall) Study and the DHS (Dallas Heart Study). J Am Coll Cardiol. 2015;66(15):1643-53. Epub 2015/10/10. doi: 10.1016/j.jacc.2015.08.035. PubMed PMID: 26449133; PubMed Central PMCID: PMCPMC4603537.

[10] R.B. D'Agostino, R.S. Vasan, M.J. Pencina, P.A. Wolf, M. Cobain, J.M. Massaro, W. B. Kannel, General Cardiovascular Risk Profile for Use in Primary Care: The Framingham Heart Study, Circulation 117 (6) (2008) 743–753.

[11] P.M. Ridker, J.E. Buring, N. Rifai, N.R. Cook, Development and Validation of Improved Algorithms for the Assessment of Global Cardiovascular Risk in Women: The Reynolds Risk Score, JAMA 297 (6) (2007) 611, https://doi.org/10.1001/jama.297.6.611.

[12] R. Nakanishi, P.J. Slomka, R. Rios, J. Betancur, M.J. Blaha, K. Nasir, M. D. Miedema, J.A. Rumberger, H. Gransar, L.J. Shaw, A. Rozanski, M.J. Budoff, D. S. Berman, Machine Learning Adds to Clinical and CAC Assessments in Predicting 10-Year CHD and CVD Deaths, JACC Cardiovasc Imaging. 14 (3) (2021) 615–625, https://doi.org/10.1016/j.jcmg.2020.08.024.

[13] M. Kavousi, M.J.G. Leening, D. Nanchen, P. Greenland, I.M. Graham, E. W. Steyerberg, M.A. Ikram, B.H. Stricker, A. Hofman, O.H. Franco, Comparison of Application of the ACC/AHA Guidelines, Adult Treatment Panel III Guidelines, and European Society of Cardiology Guidelines for Cardiovascular Disease Prevention in a European Cohort, JAMA 311 (14) (2014) 1416, https://doi.org/10.1001/jama.2014.2632.

[14] Rana JS, Tabada GH, Solomon MD, Lo JC, Jaffe MG, Sung SH, et al. Accuracy of the Atherosclerotic Cardiovascular Risk Equation in a Large Contemporary, Multiethnic Population. J Am Coll Cardiol. 2016;67(18):2118-30. Epub 2016/05/07. doi: 10.1016/j.jacc.2016.02.055. PubMed PMID: 27151343; PubMed Central PMCID: PMCPMC5097466.

[15] J.A.M. Sidey-Gibbons, C.J. Sidey-Gibbons, Machine learning in medicine: a practical introduction, BMC Med. Res. Method. 19 (1) (2019), https://doi.org/10.1186/s12874-019-0681-4.

[16] Y. Ye, F. Tsui, M. Wagner, J.U. Espino, Q. Li, Influenza detection from emergency department reports using natural language processing and Bayesian network classifiers, J. Am. Med. Inform. Assoc. 21 (5) (2014) 815–823.

[17] H. Zhai, P. Brady, Q.i. Li, T. Lingren, Y. Ni, D.S. Wheeler, I. Solti, Developing and evaluating a machine learning based algorithm to predict the need of pediatric intensive care unit transfer for newly hospitalized children, Resuscitation 85 (8) (2014) 1065–1071.

[18] F. Doshi-Velez, R.H. Perlis, Evaluating Machine Learning Articles, JAMA 322 (18) (2019) 1777, https://doi.org/10.1001/jama.2019.17304.

[19] N. Hong, H. Park, Y. Rhee, Machine Learning Applications in Endocrinology and Metabolism Research: An Overview, Endocrinol Metab 35 (1) (2020) 71, https://doi.org/10.3803/EnM.2020.35.1.71.

[20] A. Ward, A. Sarraju, S. Chung, J. Li, R. Harrington, P. Heidenreich, L. Palaniappan, D. Scheinker, F. Rodriguez, Machine learning and atherosclerotic cardiovascular disease risk prediction in a multi-ethnic population, npj Digit. Med. 3 (1) (2020), https://doi.org/10.1038/s41746-020-00331-1.

[21] E.A. Gill, M.J. Blaha, J.R. Guyton, JCL roundtable: Coronary artery calcium scoring and other vascular imaging for risk assessment, Journal of Clinical Lipidology 13 (1) (2019) 4–14.

[22] C.E. Orringer, M.J. Blaha, R. Blankstein, M.J. Budoff, R.B. Goldberg, E.A. Gill, K. C. Maki, L. Mehta, T.A. Jacobson, The National Lipid Association scientific statement on coronary artery calcium scoring to guide preventive strategies for ASCVD risk reduction, J. Clinical Lipidology 15 (1) (2021) 33–60.

[23] S.D. de Ferranti, A.M. Rodday, M.M. Mendelson, J.B. Wong, L.K. Leslie, R. C. Sheldrick, Prevalence of Familial Hypercholesterolemia in the 1999 to 2012 United States National Health and Nutrition Examination Surveys (NHANES), Circulation 133 (11) (2016) 1067–1072, https://doi.org/10.1161/CIRCULATIONAHA.115.018791. PubMed PMID: 26976914.

[24] M. Benn, G.F. Watts, A. Tybjaerg-Hansen, B.G. Nordestgaard, Familial hypercholesterolemia in the danish general population: prevalence, coronary artery disease, and cholesterol-lowering medication, J. Clin. Endocrinol. Metab. 97 (11) (2012) 3956–3964, https://doi.org/10.1210/jc.2012-1563. PubMed PMID: 22893714.

[25] Myocardial Infarction Genetics Consortium I, Stitziel NO, Won HH, Morrison AC, Peloso GM, Do R, et al. Inactivating mutations in NPC1L1 and protection from coronary heart disease. N Engl J Med. 2014;371(22):2072-82. Epub 2014/11/13.

doi: 10.1056/NEJMoa1405386. PubMed PMID: 25390462; PubMed Central PMCID: PMCPMC4335708.
[26] M. Benn, G.F. Watts, A. Tybjærg-Hansen, B.G. Nordestgaard, Mutations causative of familial hypercholesterolaemia: screening of 98 098 individuals from the Copenhagen General Population Study estimated a prevalence of 1 in 217, Eur. Heart J. 37 (17) (2016) 1384–1394.
[27] Eid WE, Sapp EH, Flerlage E, Nolan JR. Lower-Intensity Statins Contributing to Gaps in Care for Patients With Primary Severe Hypercholesterolemia. J Am Heart Assoc. 2021;10(17):e020800. Epub 2021/09/02. doi: 10.1161/JAHA.121.020800. PubMed PMID: 34465130.
[28] W.E. Eid, E.H. Sapp, A. Wendt, A. Lumpp, C. Miller, Improving Familial Hypercholesterolemia Diagnosis Using an EMR-based Hybrid Diagnostic Model, J. Clin. Endocrinol. Metab. 107 (4) (2022) 1078–1090, https://doi.org/10.1210/clinem/dgab873.
[29] W.E. Eid, E.H. Sapp, T. McCreless, J.R. Nolan, E. Flerlage, Prevalence and Characteristics of Patients With Primary Severe Hypercholesterolemia in a Multidisciplinary Healthcare System, Am. J. Cardiol. 132 (2020) 59–65, https://doi.org/10.1016/j.amjcard.2020.07.008.
[30] D.C. Goff, D.M. Lloyd-Jones, G. Bennett, S. Coady, R.B. D'Agostino, R. Gibbons, P. Greenland, D.T. Lackland, D. Levy, C.J. O'Donnell, J.G. Robinson, J.S. Schwartz, S.T. Shero, S.C. Smith, P. Sorlie, N.J. Stone, P.W.F. Wilson, 2013 ACC/AHA Guideline on the Assessment of Cardiovascular Risk: A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines, Circulation 129 (25_suppl_2) (2014), https://doi.org/10.1161/01.cir.0000437741.48606.98.
[31] F. Cabitza, A. Campagner, The need to separate the wheat from the chaff in medical informatics, Int. J. Med. Inf. 153 (2021) 104510, https://doi.org/10.1016/j.ijmedinf.2021.104510.
[32] National Center for Health Statistics. Center For Disease Control and Prevntion. https://icd10cmtool.cdc.gov/?fy=FY2021. Accessed May 29, 2021. Available from: https://icd10cmtool.cdc.gov/?fy=FY2021.
[33] B.G. Nordestgaard, M.J. Chapman, S.E. Humphries, H.N. Ginsberg, L. Masana, O. S. Descamps, et al., Familial hypercholesterolaemia is underdiagnosed and undertreated in the general population: guidance for clinicians to prevent coronary heart disease: consensus statement of the European Atherosclerosis Society, Eur. Heart J. 34 (45) (2013) 3478–3490, https://doi.org/10.1093/eurheartj/eht273. PubMed PMID: 23956253; PubMed Central PMCID: PMCPMC3844152.
[34] American Diabetes A. 2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes-2020. Diabetes Care. 2020;43(Suppl 1):S14-S31. Epub 2019/12/22. doi: 10.2337/dc20-S002. PubMed PMID: 31862745.
[35] Defining Adult Overweight and Obesity. Center For Disease Control and Prevntion. https://www.cdc.gov/obesity/adult/defining.html. Accessed May 29, 2021. Available from: https://www.cdc.gov/obesity/adult/defining.html.
[36] J.J. Boisvenue, C.U. Oliva, D.P. Manca, J.A. Johnson, R.O. Yeung, Feasibility of identifying and describing the burden of early-onset metabolic syndrome in primary care electronic medical record data: a cross-sectional analysis, cmajo 8 (4) (2020) E779–E787.
[37] Y. Xu, S. Lee, E. Martin, A.G. D'souza, C.T.A. Doktorchik, J. Jiang, S. Lee, C. A. Eastwood, N. Fine, B. Hemmelgarn, K. Todd, H. Quan, Enhancing ICD-Code-Based Case Definition for Heart Failure Using Electronic Medical Record Data, J. Cardiac Fail. 26 (7) (2020) 610–617.
[38] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al., Scikit-learn: Machine Learning in Python, J. Mach. Learn Res. 12 (null) (2011) 2825–2830.
[39] Lundberg SM, Lee S-I, editors. A Unified Approach to Interpreting Model Predictions. NIPS; 2017.
[40] B. Van Calster, D.J. McLernon, M. van Smeden, L. Wynants, E.W. Steyerberg, Calibration: the Achilles heel of predictive analytics, BMC Med 17 (1) (2019), https://doi.org/10.1186/s12916-019-1466-7.
[41] K.F. Kerr, Z. Wang, H. Janes, R.L. McClelland, B.M. Psaty, M.S. Pepe, Net Reclassification Indices for Evaluating Risk Prediction Instruments: A Critical Review, Epidemiology 25 (1) (2014) 114–121.
[42] V. Brandt, T. Emrich, U.J. Schoepf, D.M. Dargis, R.R. Bayer, C.N. De Cecco, C. Tesche, Ischemia and outcome prediction by cardiac CT based machine learning, Int. J. Cardiovasc. Imaging 36 (12) (2020) 2429–2439.
[43] Motwani M, Dey D, Berman DS, Germano G, Achenbach S, Al-Mallah MH, et al. Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: a 5-year multicentre prospective registry analysis. Eur Heart J. 2017;38(7):500-7. Epub 2016/06/03. doi: 10.1093/eurheartj/ehw188. PubMed PMID: 27252451; PubMed Central PMCID: PMCPMC5897836.
[44] M. van Assen, A. Varga-Szemes, U.J. Schoepf, T.M. Duguay, H.T. Hudson, S. Egorova, K. Johnson, S. St. Pierre, B. Zaki, M. Oudkerk, R. Vliegenthart, A. J. Buckler, Automated plaque analysis for the prognostication of major adverse cardiac events, Eur. J. Radiol. 116 (2019) 76–83.
[45] A.R. van Rosendael, G. Maliakal, K.K. Kolli, A. Beecy, S.J. Al'Aref, A. Dwivedi, G. Singh, M. Panday, A. Kumar, X. Ma, S. Achenbach, M.H. Al-Mallah, D. Andreini, J.J. Bax, D.S. Berman, M.J. Budoff, F. Cademartiri, T.Q. Callister, H.-J. Chang, K. Chinnaiyan, B.J.W. Chow, R.C. Cury, A. DeLago, G. Feuchtner, M. Hadamitzky, J. Hausleiter, P.A. Kaufmann, Y.-J. Kim, J.A. Leipsic, E. Maffei, H. Marques, G. Pontone, G.L. Raff, R. Rubinshtein, L.J. Shaw, T.C. Villines, H. Gransar, Y. Lu, E. C. Jones, J.M. Peña, F.Y. Lin, J.K. Min, Maximization of the usage of coronary CTA derived plaque information using a machine learning based algorithm to improve risk stratification; insights from the CONFIRM registry, J. Cardiovasc. Comput. Tomogr. 12 (3) (2018) 204–209.
[46] K.M. Johnson, H.E. Johnson, Y. Zhao, D.A. Dowe, L.H. Staib, Scoring of Coronary Artery Disease Characteristics on Coronary CT Angiograms by Using Machine Learning, Radiology 292 (2) (2019) 354–362.
[47] J. Brandts, K.K. Ray, Low Density Lipoprotein Cholesterol–Lowering Strategies and Population Health: Time to Move to a Cumulative Exposure Model, Circulation 141 (11) (2020) 873–876.
[48] K. Khunti, M.D. Danese, L. Kutikova, D. Catterick, F. Sorio-Vilela, M. Gleeson, S. R. Kondapally Seshasai, J. Brownrigg, K.K. Ray, Association of a Combined Measure of Adherence and Treatment Intensity With Cardiovascular Outcomes in Patients With Atherosclerosis or Other Cardiovascular Risk Factors Treated With Statins and/or Ezetimibe, JAMA Netw Open 1 (8) (2018) e185554, https://doi.org/10.1001/jamanetworkopen.2018.5554.
[49] A. Khera, M.J. Budoff, C.J. O'Donnell, C.A. Ayers, J. Locke, J.A. de Lemos, J. M. Massaro, R.L. McClelland, A. Taylor, B.D. Levine, Astronaut Cardiovascular Health and Risk Modification (Astro-CHARM) Coronary Calcium Atherosclerotic Cardiovascular Disease Risk Calculator, Circulation 138 (17) (2018) 1819–1827.
[50] B. Ó Hartaigh, H. Gransar, T. Callister, L.J. Shaw, J. Schulman-Marcus, W. J. Stuijfzand, V. Valenti, I. Cho, J. Szymonifka, F.Y. Lin, D.S. Berman, H.-J. Chang, J.K. Min, Development and Validation of a Simple-to-Use Nomogram for Predicting 5-, 10-, and 15-Year Survival in Asymptomatic Adults Undergoing Coronary Artery Calcium Scoring, JACC: Cardiovascular Imaging 11 (3) (2018) 450–458.
[51] A.J. Vickers, B. van Calster, E.W. Steyerberg, A simple, step-by-step guide to interpreting decision curve analysis, Diagn Progn Res 3 (1) (2019), https://doi.org/10.1186/s41512-019-0064-7.

KICS The Korean Institute of Communications and Information Sciences

Available online at www.sciencedirect.com

ScienceDirect

ICT Express 8 (2022) 109–116

www.elsevier.com/locate/icte

A comparative study of classification and prediction of Cardio-Vascular Diseases (CVD) using Machine Learning and Deep Learning techniques

M. Swathy*, K. Saruladha

Department of CSE, Puducherry Technological University, Puducherry, India

Received 8 December 2020; received in revised form 3 May 2021; accepted 18 August 2021
Available online 3 September 2021

Abstract

Cardio-Vascular Diseases (CVD) are found to be rampant in the populace leading to fatal death. The statistics of a recent survey reports that the mortality rate is expanding due to obesity, cholesterol, high blood pressure and usage of tobacco among the people. The severity of the disease is piling up due to the above factors. Studying about the variations of these factors and their impact on CVD is the demand of the hour. This necessitates the usage of modern techniques to identify the disease at its outset and to aid a markdown in the mortality rate. Artificial Intelligence and Data Mining domains have a research scope with their enormous techniques that would assist in the prediction of the CVD priory and identify their behavioural patterns in the large volume of data. The results of these predictions will help the clinicians in decision making and early diagnosis, which would reduce the risk of patients becoming fatal. This paper compares and reports the various Classification, Data Mining, Machine Learning, Deep Learning models that are used for prediction of the Cardio-Vascular diseases. The survey is organized as threefold: Classification and Data Mining Techniques for CVD, Machine Learning Models for CVD and Deep Learning Models for CVD prediction. The performance metrics used for reporting the accuracy, the dataset used for prediction and classification, and the tools used for each category of these techniques are also compiled and reported in this survey.

Keywords: Cardio-Vascular Diseases; Classification; Regression; SVM; Deep Learning; Data Mining; Machine Learning; ANN; Artificial Intelligence

1. Introduction

Cardio-Vascular Disease (CVD) is an overall term referring to the conditions that affect the heart and blood vessels of a human body. This can also include the damage of the arteries in the organs such as the kidneys, heart, eyes and brain [1]. CVD is one of the main causes of death in many developed and developing countries all over the world even with young people. But, the fact is that it can be extensively prevented by leading a healthy lifestyle.

There are four main categories of the Cardio-Vascular Diseases. First and foremost is the Coronary Heart Disease which occurs due to the blockage of blood to heart muscle. This causes an increased strain on the heart and leads to angina, heart attacks and heart failure. The second type is Strokes and Transient Ischaemic Attack (TIA) which occurs due to blockage of blood to the brain and temporary disruption in the blood flow. The third type is Peripheral Arterial Disease which occurs due to blockage of blood to the limbs. This causes worst leg pains, hair loss on legs and feet, weakness in legs and persistent ulcers. The last type is the Aortic Disease which affects the largest blood vessel — Aorta. This has no symptoms, but causes a life-threatening bleeding when there is a chance of burst.

Cardiovascular disease comprises of the coronary artery diseases (CAD) like angina and myocardial infarction (commonly known as a heart attack) and coronary heart diseases (CHD), in which a waxy substance called plaque is developed inside the coronary arteries. Without a fast initiate to recovery, a heart attack can lead to serious health problems and even death, as this stays as a common cause of death worldwide.

Though, the exact cause for the CVD is not clearly found yet, there are lot of possibilities of one getting it. There are several risk factors involved with the chances of getting a CVD. Some of the most prominent factors are High Blood Pressure, Smoking, High Cholesterol, Diabetes, Obesity, Family History, Age, etc.,

* Corresponding author.
E-mail addresses: mswat97@pec.edu (M. Swathy), charuladha@pec.edu (K. Saruladha).

Peer review under responsibility of The Korean Institute of Communications and Information Sciences (KICS).

https://doi.org/10.1016/j.icte.2021.08.021

Identification of people at risk due to CVD is a cornerstone. Due to many constraints in the manual identification of the heart diseases, scientists have moved towards modern approaches like Data Mining, Machine and Deep Learning methodologies for predicting the disease. These proved [2] to be effective to assist in decisions-making and predictions from enormous amount of data produced by the health care industry.

CVD can be diagnosed using an array of lab tests and imaging studies. However, the primary part of diagnosis is medical and family history of the patient, risk factors and physical examination. Through the statistical data, we can coordinate the findings and predict the presence of disease from results and procedures. Automation with Machine and Deep Learning can enable doctors to make informed decisions.

Automation in disease predictions can create a single platform from which structured data can be retrieved and efficient care can be given to the patients. Thus, it redefines the level of personalized health-care. Through artificial intelligence and machine learning, computers are taught to recognize patterns in which the disease occurs and convert them as structural data for predicting the same.

Innovations are made in Electronic Health Records (EHR), revenue cycle and operations through AI. In future, it will be integrated with the clinical work-flow with the existing tools empowering the practitioners with real-time data at the point of care.

2. Survey Organization

This survey gives a comparison of various classification and predictions for CVD. It follows a threefold organization with Data Mining Techniques for CVD, Machine Learning Models for CVD and Deep Learning Models for CVD prediction.

This study gives a brief description about the methods and algorithms which are used for predictions, the classification techniques, performance metrics and tools used for evaluation of their model.

2.1. Classification and Data-Mining techniques for CVD

Heart is the vital organ of a human body. It pumps blood to all other organs through the body. If there is any distortion in the circulation of blood (or) insufficiency in blood, it can lead to serious effects such as brain fever or even death that occurs within minutes [3]. A Heart disease is a collective term that refers to the diseases in par with the heart and its associated blood vessel system.

Data Mining (DM) is a field of Computer Science that is used for extraction of useful data from huge data sets which can be used in predictions or the data can be described using techniques such as classification, clustering, association, etc., [4,5]. Data Mining combines machine learning, mathematical analysis and information technology to evaluate large pre-existing databases to extract the hidden patterns among the data. In the healthcare domain, DMT can be used to predict the presence of a disease by evaluating the giant databases and exploring the relationships between the data using the useful trained patterns. This automation of this can be extremely advantageous in the prediction systems. Research based on DMT has already been applied to diabetes, asthma, CVD and AIDS. DMT such as Naïve Bayesian (NB) classification, Artificial Neural Networks (ANN), Support Vector

Machines (SVM), Decision Trees (DT), Logistic Regression (LR), etc. are used in the medicinal research. Decision tree gives a procedural approach for classification of categorical data based on the features or attributes. Hence, it stands the most widely used method for classification in DMT for processing large amount of data.

DMT basically applies to Predictive Analytics. This refers to extracting uncovered patterns from vast datasets and uses them to build an appropriate network for prediction. The general workflow of prediction systems is given below (Fig. 1).

This research work used the Artificial Neural Network (ANN) [6] which is an effective way to design the heart disease prediction system. Similarly [7] utilizes DMT for the cerebral-vascular disease prediction system. It uses Decision Tree and Bayesian classifiers for classification and a Multilayer perceptron with Back Propagation (BP) for training the model. The result of this work indicates that, DMT can make the prediction of CVD more efficiently. Advanced DMT can even help in identifying the hidden patterns and relationships which often goes unexploited.

The healthcare and medicinal industry deal with large volume of data. Many of them are not structured and viable to identify the existing patterns. Making the process of discovering the relationships can make an effective decision-making system [8]. Choosing the appropriate data needed for implementing the above procedure from the huge loads of data is the most needed work. DMT would be the best approach to derive the useful details in depth on all the different perspectives [9,10].

The prediction of CVD is the hot topic of research in the milestone of medical industry. DMT based prediction systems can help in determining the disease at the starting stage itself which can minimize the risk associated. The research work [11], proposed a slight modification in the Weighted Associated Classifier (WAC) which showed an accuracy of 81.51%. Similarly [12], proposed a rule based discovery model with Associative Classification Mining to construct the classification systems. This uses an algorithm called MAFIA (Maximal Frequent Item-set Algorithm) to determine the frequent patterns. Medical Data Mining is a part of healthcare with lot of imprecision and uncertainty. This work [13] proves the working efficiencies of ID3 and CART algorithms on decision trees and concludes stating CART is more efficient than ID3 both theoretically and practically. The decision making capability [14] can be enhanced using the K-Means Clustering.

According to WHO estimation till 2030, very nearly 23.6 million individuals will pass because of Heart Malady [15]. CVD though is the major cause, can be controlled and prevented. To minimize the seriousness caused, analysis is very important. The most complex part is to choose the right ailment [16]. Gives the summary of recent works done in the field of data mining related to CVD. It gives a beautiful conclusion

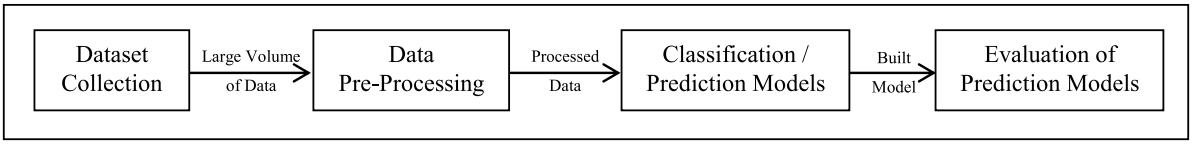


Fig. 1. General workflow of prediction systems.

stating instead of relying upon a very specific DMT, hybrid or combination techniques can produce more efficient and accurate results.

But the tough procedure here is to find the appropriate combination of the algorithms to form the accurate model. The frequently used list of datasets with DMT and ML models are listed in the table (see Table 1).

From the results of [17] it can be concluded that the accuracy of prediction depends on the DMT used, Data Set handled by the model and the number of attributes. According to [18], if 102 cases are analysed, SVM has a highest accuracy of 90.5% and Logistic Regression has the lowest of 73.9%. Survey of 1000 patients showed SVM with 92.1%, ANN with 91% and Decision Tree with 89.6% accuracy. It can also be inferred that results with large sensitivity rate and specificity rate but are with lower accuracy will be abstained from the results, which makes the model a highly efficient.

2.1.1. Observation and Inferences: Data-Mining & Classification Techniques

Data Mining is a generalized term that includes many techniques to extract meaningful information without having pre-conceived notations about what will be discovered.

(1) In general the most widely used method for classification is Decision Tree and Naïve Bayes Classifiers. Observations show that, Neural Network based classification has more performance than the above two methods.

(2) On the other hand, other data mining methods such as Clustering, Association Rule based, Time Series based can also be analysed for the usage of predictions.

(3) Almost all the models discussed above uses categorical data for their classification and prediction. In real-time, the usage of continuous data can be more advantageous for analysis.

(4) Further, we should also think upon extending the models to ensemble based algorithms.

2.2. Machine Learning models for CVD

Machine Learning is the basic practice of using algorithms to make predictions by parsing data and learning from it. These models have the capability of learning by itself from prior experience or from historical data. These algorithms can figure out extract the important tasks to be performed by generalizing from examples provided to them as training sets.

Different types of ML algorithms have evolved. These are grouped by either learning style (i.e. supervised learning, unsupervised learning, and semi-supervised learning) or by similarity or by their functioning (i.e. classification, regression, decision tree, clustering, deep learning, etc.). All machine learning algorithms comprise of three different components namely:

1. Representation: Set of classifiers in Computer understandable form
2. Evaluation: The objective defined for the classifier model (algorithm) - Scores
3. Optimization: Search method for the most scored classifier.

The primary goal of ML algorithms are to generalize beyond the training samples provided to them, which involves the successful interpretation of data that it has never noticed before.

The main difference that holds between ML and AI is that ML works for increasing accuracy and AI works for the increasing chance of success.

The basic three types of ML Techniques are: Supervised Learning, Unsupervised Learning and Reinforcement Learning. The selection of the algorithm and the learning type can be made by different approaches like depending on the task accomplished, (or) the amount of data involved (or) the different types of data that are available. This exhibits a dynamic role that plays out in applications of medical diagnostics as it involves creating self-learning algorithms. CVD prediction involves supervised learning technique as labelled data is required for training the model.

In case of prediction of CVD, regular diagnosis is very important in the initial stages of treatment which in turn reduces the risks associated with it. The most common and vital diagnostic tests include echocardiography (echo), cardiac magnetic resonance imaging (MRI), and computed tomography (CT) [19]. High quality cardiac images are produced by MRI and CT scans, which are not preferred for predictions as they have prolonged acquisition time, limited availability and involve the use of radiations.

Electrocardiogram (ECG) is a graphical representation that is produced by repolarization and depolarization of the ventricles and atria. Though there are several advancements made in the field of prediction and diagnostics, the accurate way of avoiding heart attack is not known as there are no proper symptoms associated with it. To discover a disease that forms the main causes of death such as HIV, Cancer, CVD, machine learning [20] can be used. It is a great consequence to research.

CVD are caused due to the deposits of fat (cholesterol) in the inner walls of arteries which narrows down or blocks coronary arteries. An efficient heart disease prediction system can be a beneficial way to exactly predict the diseases and save the patient's life. The system model presented in [21] can interpret human patterns and accurately determine trends in the patients' records.

The efficient functioning of heart is very vital for human life. In the prediction of heart-based diseases, automation

Table 1
Comparison of datasets used with the prediction models.

S. No	Data-set	No. of I/P attributes	Data mining/Machine learning techniques[a]							
			DT	NB	RF/J48	CL	RB	SVM	NN	GA
1		13	✓	✓					✓	
2	Cleveland database	14		✓	✓	✓	✓			
3		15	✓	✓					✓	
4	UCI repository	13	✓			✓		✓	✓	
5	Publicly available heart	13	✓	✓					✓	✓
6	Disease dataset	6	✓	✓		✓				
7	Kaggle							✓		
8	Deep Learning techniques									
	Echo-cardiography datasets (Frames of complete cardiac cycles)	29–55 Frames with an average of 45 frames	DenseNet, ResNet, LSTM, GRU, CNN and RNN							

[a]DT – Decision Tree, NB – Naïve Bayes, SVM – Support Vector Machines, NN- Neural Network, A-NN – Artificial Neural Network, CL – Clustering, RB- Rule Based, RF – Random Forest, GA – Genetic Algorithm.

Table 2
Comparison of evaluation metrics with deployment models.

S. No	Evaluation metric	Data mining/Machine learning models[a]											
		DT	NB	SVM	NN	BP	MLP- NN	GA	FL	CL	PCA-Knn	RB	RF
1	Sensitivity	✓	✓	✓		✓	✓				✓	✓	
2	Specificity	✓		✓		✓	✓				✓	✓	
3	Accuracy	✓	✓	✓	✓	✓			✓	✓	✓	✓	✓
4	Precision	✓	✓			✓	✓					✓	
5	Confusion matrix	✓	✓			✓	✓						
6	Efficiency	✓						✓					
7	Lift chart	✓	✓		✓								
8	Deep Learning models												
	R^2 – Regression score	ResNet, DenseNet, LSTM, GRU, CNN, RNN											

[a]DT – Decision Tree, NB – Naïve Bayes, SVM – Support Vector Machines, NN- Neural Network,
BP – Back Propagation, MLP-NN – Multi Layer Perceptron with NN, GA – Genetic Algorithm, FL – Fuzzy Logics, CL – Clustering, PCA- kNN – Principal Component Analysis with k-Nearest Neighbour, RB- Rule Based,
RF – Random Forest.

plays an important role with a swift examination of accurate result. This research work [22] incorporates the classes of Heart Disease through Support Vector Machine (SVM). They initially analyse the historical data of the patient and fetch the real-time ECG values. Prediction is then made with these as input through SVM.

In the prevailing lifestyle, everyone is more tensed and is tested to have high blood pressure and sugar levels at a very young age. The less attention shown towards the quality of food taken and the own medications they tend to possess can lead to major threat of heart diseases [23]. This system extracts hidden knowledge associated with CVD diseases. The two important notices are, continuous data is preferred instead of categorical data and the integration of data mining and text mining through the machine learning model. The evaluation metrics commonly used to test the prediction model with DMT and ML Techniques is discussed (see Table 2).

The most popular machine learning algorithm called Naive-Bayes (NB) forms the basis of several other algorithms and data processing ways.

This algorithm uses the Bayesian Rule that calculated the predictive capabilities through probabilistic approach [24]. This helps in exploring new ways of knowledge-oriented training, classification and prediction. Data Mining combined with Naive Bayes can give efficient results of prediction. In this [25] research work, DSHDPS is deployed as web-based questionnaire application. Based on the user responses, the model can discover and extract the hidden relationships associated with the heart disease. This can be the most significant way of prediction of heart diseases.

2.2.1. Observation and inferences: Machine Learning techniques

Different types of Machine Learning oriented research papers have been analysed. The general conclusion drawn from the implemented models are listed below:

(1) More featured medical attributes can be used for providing better model with more accuracy and performance.

(2) Integrate data-mining and text-mining with the existing models for constructing efficient prediction systems.

(3) Continuous information can be used in place of categorical information to build a heart disease system with early detection.

(4) Execution can be improved by using Genetic Algorithms and Swarm Intelligence Techniques to provide more concentration on the feature selection and input parameters.

2.3. Deep Learning models for CVD prediction

Deep Learning is referred to as the subset of ML and AI that has a network which capable of learning unsupervised forms of data. In Machine Learning, when the learning algorithm is not working properly, to make it accurate we feed more amounts of data for training the model. This may lead to issues with scalability and the learning time of the model increases exponentially. To overcome the data handling issues, we can switch over to Deep Learning techniques (DLT) which is capable of learning better representation of unstructured (or) unlabelled data with multiple levels of abstraction.

DLT utilizes hierarchical level of artificial neural networks (built like a human brain) to carry out the exact process of ML. This is one of the biggest advantages of DLT as it processes the data in a non-linear fashion while others process in a linear way. This makes the system quickly adapts the healthcare domain as it offers the ability to analyse data with a greater speed and precision. It also has an added benefit of being able to take decisions with a significantly less involvement of human trainers. Deep Learning requires less pre-processing of data when compared to ML and DMT. The DLT network itself is capable of filtering and normalization tasks, which is done by human programmers in other MLT. According to [26], clinical decisions are made based on the heuristic's experiences and on the doctors' intuition.

The knowledge based on the hidden data can be used to efficiently diagnose a heart disease thus reducing medical errors and it also decreases the diagnostic time. Through regular practices, it can also enhance the patient's satisfaction and safety. Diagnosing a disease is the most crucial part because it is proven to be based on the doctor's knowledge and experiences. But, training a machine to act like a human and making the machine learn the algorithms to carry over can in turn make it more [27] accurate and time efficient. The accuracy of the prediction is the prime concern in the predictive methodology and models. DLT offers a wide range of applications to represent trust-worthy text analytics in spite of the biased and skewed data. So many types of DLT are offering roles in decision making and predictive analytics because it combines the more advanced methods in order to increase the power and creates new method to benefit for prediction and prevention. Some of their applications are Personalized Treatment for Diabetes(Type II) and Cancer Patients, Using radiology for Image Classification for Cancer, Tumour and Lung treatment, Drug Discovery and Data Augmentation using GAN, Treatment Identification for Cancer and HIV, etc., In today's world, many people are found to be living with heart diseases [28] without any awareness of themselves. Once they are predicted in advance, an accurate treatment can be provided to reduce the consequence. DLT can be used for Chat Bots and Medical Imaging Solutions which can identify the patterns and symptoms associated with specific kinds of diseases.

Neural Network which is a DLT follows a graph topology. It is a parallel, distributed information processing structure consisting of multiple nodes. Each node corresponds to the neurons and the weight associated with them corresponds to the edge. It has many hierarchical layers which are finite in order to decrease the time of problem solving. Every node has a single output connection that branches into many connections. The layers of NN typically include one input and output with multiple hidden layers. Neuro-Fuzzy is a combination of Neural Networks and Fuzzy Logics which can be used to solve wide range of real-time problems with an ease. This can well suit the limitations pertaining to other models of the automated techniques.

Artificial Neural Networks (ANN) is used in many areas of medicine [29], such as cardiology, Electroencephalography, Pulmonology, Genetics, Clinical chemistry, Pathology, Ophthalmology, Obstetrics and Gynecology. Considerable researches are also being taken over the heart diagnosis. Prediction of CVD from various factors and symptoms is a multi-layered issue [30] which may lead to false presumptions and unpredictable effects.

DLT is so adept for Image Processing that many of the AI researchers are using Neural Networks to create medical images — to read them, analyse them and use them for the prediction of diseases. Convolution Neural Networks (CNN) is a type which is particularly suited to analyse the MRI Scans and X-rays. These can operate more effectively on larger images also thus surpassing the diagnosticians' accuracy on imaging studies. In prediction of CVD, there is a demand to accurately detect the ED and ES frames using an automated image-driven method. CNN and RNN (Recurrent Neural Networks) has gained an enormous growth in the medicinal applications such as CVD Prediction Systems, Tumour Detection, Cancer Detection, Gene Classification, Neural Cells Classification, etc.,. The main advantages of this is compared to other techniques, this can automatically detect the important features for prediction networks without human intervention [19]. CNN can be used for Image Feature Extractions while RNN can be used for learning about the temporal dependencies between them. These are also computationally efficient when compared with other techniques for prediction. These models can also suit the time-series kind of data [31,32] like the patient's diagnostic history, EHR, E-Prescriptions etc., for predictions as the information is remembered throughout the network. The final output from these networks can be given to Regression Modules for prediction.

ResNet and DenseNet are a form of NN. The former uses skip-connections to forward the features from one layer to another whereas the latter uses the entire information by concatenating them from the preceding layers and pass them as input features to the next layer. Effective and accurate diagnosis can only lead to appropriate treatment for the patient. This can be completely done on deep study of CV analysis of the patient. The [33] table represents the performance metrics of various algorithms.

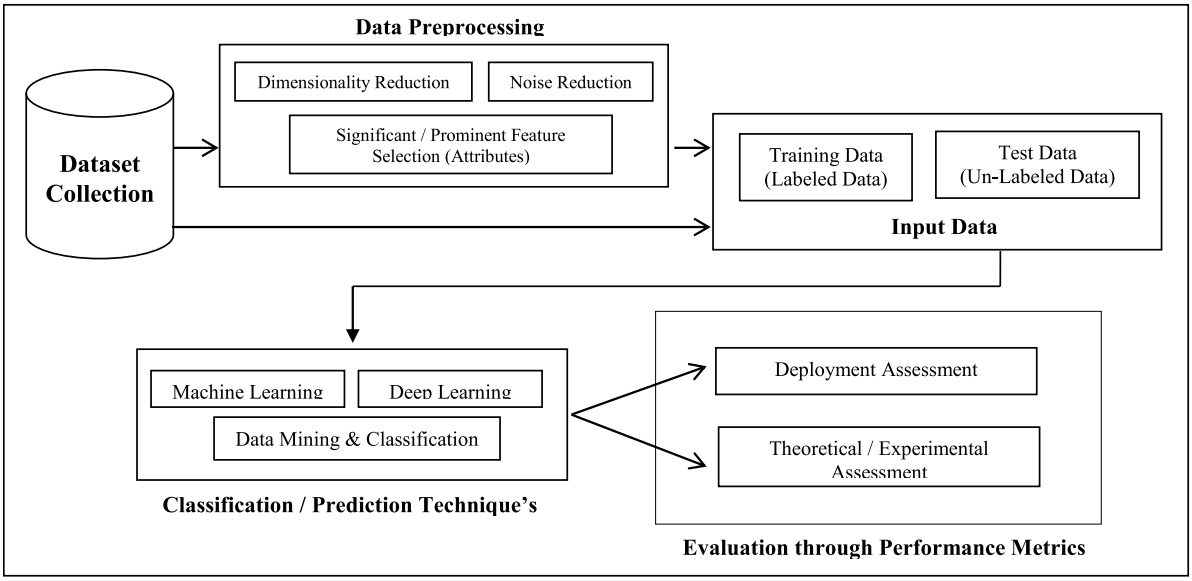


Fig. 2. Prediction/classification workflow.

Table 3
Comparison table – Prediction models with deployment tools.

Models[a]	Tools used		
	WEKA	TANGARA	MATLAB
DT	✓	✓	✓
NB	✓	✓	
NN/A-NN	✓	✓	✓
FL		✓	✓
GA	✓		✓
SVM		✓	✓
PCA			✓
CL	✓	✓	
J48	✓		
RF	✓		

[a]DT – Decision Tree, NB – Naïve Bayes, SVM – Support Vector Machines, NN- Neural Network, FL- Fuzzy Logics, GA – Genetic Algorithm, CL –Clustering, RF – Random Forest, PCA- Principal Component Analysis.

2.3.1. Observation and inferences: Deep Learning techniques

Present NN based models suit only for specific or minimal kind of heart diseases. Hence, NN based systems should be expanded to suit wide range of heart based diseases.

(1) With respect to ANN, we can make changes in the architecture and train algorithms for achieving more accurate results.

(2) Generally, for 15 attributes, the Multi-Layer Perceptron Neural Networks with Back Propagation provides better results than other models.

(3) Mostly, in all DL models, accuracy is considered to be the performance metric. But, we can also try to consider other metrics (such as sensitivity, specificity, efficiency, etc.,) based on the demands of the diagnosis.

(4) In future, Fuzzy Logics can be incorporated with NN to include more discrete valued attributes for prediction of CVD.

Table 4
Pros and cons of WEKA.

Advantages	Disadvantages
Freely available and portable	Algorithm does not cover sequential modelling
Easy to use GUI and command line	It cannot suit multi-relational data mining
Provides access to SQL – Databases through JDBC	It is memory bound

3. Tools for prediction/classification models

There are several tools available for evaluation of the proposed prediction models. Some of the widely used tools are listed below. The tools suitable for different kind of prediction models is also discussed (see Table 3).

3.1. Waikato Environment for Knowledge Analysis (WEKA)

It is a Data-Mining/ Machine-Learning tool that can be used to apply the algorithm directly on the dataset or through JAVA Code. It contains tools for data pre-processing, classification, regression, clustering, association rules and visualization. It is open-source software in JAVA issued under GNU. (General Public Licensed) (see Table 4).

3.2. TANGARA

It is free Data Mining software for academic and research purposes. It contains tools for exploratory data analysis, statistical learning, machine learning, and databases. It allows analysing both real and synthetic data and also allows adding our own data mining methods, to compare their performances. It is open source software in C++ issued under GNU.

Table 5
Pros and cons of MATLAB.

Advantages	Disadvantages
Platform independent & Easy to use	It can be slow as MATLAB is an interpreted language
Predefined functions and Toolboxes	It is not a free - software
Device independent plotting	Consumes large amount of memory
Good visualization of results	Consumes large amount of time – making real time applications very complicated.

3.3. Matrix Laboratory (MATLAB)

It provides user accurate solutions for problems with flexible graphics. It is highly interactive and programmable environment. It supports all mathematical computations, visualizations and programming. It is composed of High-Level programming language which is similar to C (see Table 5).

4. Future research perspectives

This survey paper is a consolidation of works done in the field of Prediction of Cardio-Vascular Diseases using Machine and Deep Learning Techniques. The lifestyle has changed over the past few years leading to a lot of health complications which goes unnoticed in major of the population. Taking right measures at the right time can lead to save an individual's life.

Heart Diseases form a vital part of mortality rate. These have no specific symptoms for their occurrences. This can be prevented by making custom lifestyle changes such as stop smoking, having a controlled BP, checking cholesterol rate, keeping diabetes under control, diet patterns and exercises and maintaining pressures and stress. The existing texhniques and the workflow prediction is depicted as a overview in the above figure (Fig. 2).

In order to diagnose a heart disease, several kinds of lab tests and imaging studies are required. The latest research includes examination of risks of heart attacks and its possible recovery of open heart surgeries with angioplasty and stenting in patients with diabetes and blockages in more than one coronary artery. Researchers are exploring the use of diagnostic technologies in detection of heart diseases.

Automation applied to the field of prediction can lead to a high benefit of asset in the medicinal industry. AI applied in CVD — includes precisional medicine, clinical prediction, cardiac imaging and analysis and intelligent Robots. Development of Sensor Technology has furthered the application of AI. Machine Learning helps to assess the risk of patients suffering from CVD.

The above discussed three-fold approach confines to the specific kinds of automation used along with their algorithms and methods to effectively predict CVD. Each method has its own kind of pros and cons. According to our research perspective, we can either relate them and choose a specific method or use a combination of methods to achieve an accurate model.

In conclusion, the prediction methods listed are clearly in their basic level and further efforts and resources are needed to gather more data with length follow-ups to derive population-specific methods that addresses all the concerns of existing prediction models. This can also end up in producing personalized risk assessments in the future. The future still is in the hands of medical professionals who are now being supported by the technology to understand their needs and reduce the stress they experience upon.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by AICTE Research funding under the Research Promotion scheme-2019- VIDE NO:F.No.8-35/RIFD/RPS. The authors acknowledge the support of Dr.V.Akila of Department of Computer Science and Engineering, Puducherry Technological University, for her assistance in improving the quality of the manuscript.

References

[1] www.nhs.uk/conditions/cardiovascular-disease/, 0000.
[2] towardsdatascience.com/heart-disease-prediction-73468d630cfc, 0000.
[3] S.H.Ms. Ishtake, Intelligent heart disease prediction system using data mining techniques, Int. J. Healthc. Biomed. Res. 1 (3) (2013) 94–101.
[4] T. Mythili, Dev Mukherji, Nikita Padalia, Abhiram Naidu, A heart disease prediction model using SVM-decision trees-logistic regression (SDL), Int. J. Comput, Appl. (0975–8887) 68 (16) (2013).
[5] Jyoti Soni, Ujma Ansari, Dipesh Sharma, Predictive data mining for medical diagnosis: An overview of heart disease prediction, Int. J. Comput. Appl. (0975–8887) 17 (8) (2011).
[6] Gadoya Komal, D.R. Vipul Vekariya, Novel approach for heart disease prediction using decision tree algorithm, Int. J. Innov. Res. Comput. Commun. Eng. 3 (11) (2015) (An ISO 3297: 2007 Certified Organization).
[7] A. Sudha, P. Gayathri, N. Jaisankar, Effective analysis and predictive model of stroke disease using classification methods, Int. J. Comput. Appl. (0975–8887) 43 (14) (2012).
[8] T. Chandrasekhar, S.S.V Harsha Pavan, B. Vinay, Detection of heart diseases using data mining techniques, Int. J. Eng. Sci. Res. Technol. IJESRT (ISSN: 2277-9655) 8 (3) (2019).
[9] Sonali S. Jagtap, Prediction and analysis of heart disease, Int. J. Innov. Res. Comput. Commun. Eng. 5 (2) (2017).
[10] T.K. Keerthana, Heart disease prediction system using data mining method, Int. J. Eng. Trends Technol. (IJETT) 47 (6) (2017).
[11] Chaitrali S. Dangare, Sulabha S. Apte, Improved study of heart disease prediction system using data mining classification techniques, Int. J. Comput. Appl. (0975–888) 47 (10) (2012).
[12] Aditya Methaila, Prince Kansa, Himanshu Arya, Pankaj Kumar, Early heart disease prediction using data mining techniques, in: CCSEIT, DMDB, ICBB, MoWiN, AIAP – 2014, 2014, pp. 53–59, http://dx.doi.org/10.5121/csit.2014.4807.
[13] T. Chandrasekhar, S.S.V Harsha Pavan, B. Vinay, Detection Of Heart Diseases Using Data Mining Techniques, Int. J. Eng. Sci. Res. Technol. - ISSN: 2277-9655, Impact Factor: 5.164, IC^{TM} Value: 3.00 CODEN: IJESS7, 0000.
[14] Rucha Shinde, Sandhya Arjun, Priyanka Patil, Prof. Jaishree Waghmare, An intelligent heart disease prediction system using K-means clustering and Naïve Bayes algorithm, (IJCSIT) Int. J. Comput. Sci. Inf. Technol. 6 (1) (2015) 637–639, IISN: 0975-9646.

[15] Shweta Gupta, Prof. Aditi Nema, Prof. Kiran Agrawal, Heart Disease Prediction using PCA-KNN in Data Mining, in: IDES Joint International Conferences on IPC and ARTEE – 2017, 0000.

[16] Animesh Hazra, Subrata Kumar Mandal, Amit Gupta, Arkomita Mukherjee, Asmita Mukherjee, Heart disease diagnosis and prediction using machine learning and data mining techniques: A review, Adv. Comput. Sci. Technol. (ISSN: 0973-6107) 10 (7) (2017) 2137–2159.

[17] Mudasir M. Kirmani, Mudasir m kirmani cardiovascular disease prediction using data mining techniques: A review, Orient. J. Comput. Sci. Technol. (ISSN: 0974-6471) 10 (2) (2017) 520–528.

[18] T. Mythili, Dev Mukherji, Nikita Padalia, Abhiram Naidu, A heart disease prediction model using SVM-decision trees-logistic regression (SDL), Int. J. Comput. Appl. (0975–8887) 68 (16) (2013).

[19] Fatemeh Taheri Dezaki, Zhibin Liao, Christina Luong, Hany Girgis, Neeraj Dhungel, Amir H. Abdi, Delaram Behnami, Ken Gin, Robert Rohling, Purang Abolmaesumi, Teresa Tsang, Cardiac phase detection in echocardiograms with densely gated recurrent neural networks and global extrema loss, IEEE Trans. Med. Imaging 38 (8) (2019).

[20] S. Vinothini, Ishaan Singh, Sujaya Pradhan, Vipul Sharma, Heart disease prediction, Int. J. Eng. Technol. 7 (3.12) (2018) 750–753.

[21] Era Singh Kajal, Prediction of Heart Disease using Data Mining Techniques, Int. J. Adv. Res. Ideas Innov. Technol. 2 (3), 0000, ISSN: 2454-132X.

[22] Bhavana Baad, Neha Bhatkhande, Kanchan Kalokhe, Prof. Jyoti Raghatwan, Heart disease prediction and detection, Int. J. Res. Appl. Sci. Eng. Technol. (IJRASET) (ISSN: 2321-9653) 7 (IV) (2019) IC Value: 45.98.

[23] Shadab Adam Pattekari, Asma Parveen, Prediction system for heart disease using naive Bayes, Int. J. Adv. Comput. Math. Sci. (ISSN: 2230-9624) 3 (3) (2012) 290–294.

[24] Garima Singh, Kiran Bagwe, Shivani Shanbhag, Shraddha Singh, Sulochana Devi, Heart disease prediction using Naïve Bayes, Int. Res. J. Eng. Technol. (IRJET) (ISSN: 2395-0072) 04 (03) (2017) e-ISSN: 2395-0056.

[25] Mrs.G. Subbalakshmi, Mr.K. Ramesh, Mr.M. Chinna Rao, Decision support in heart disease prediction system using naive Bayes, Indian J. Comput. Sci. Eng. (IJCSE) (ISSN: 0976-5166) 2 (2) (2011).

[26] Mohammad A.M. Abushariah, Assal A.M. Alqudah, Omar Y. Adwan, Rana M.M. Yousef, Automatic heart disease diagnosis system based on artificial neural network (ANN) and adaptive neuro-fuzzy inference systems (ANFIS) approaches, J. Softw. Eng. Appl. 2014 (7) (2014) 1055–1064, http://dx.doi.org/10.4236/jsea.2014.

[27] Soumonos Mukherjee, Anshul Sharma, Intelligent heart disease prediction using neural network, Int. J. Recent Technol. Eng. (IJRTE) (ISSN: 2277-3878) 7 (5) (2019).

[28] Rachana Deshmukh, Omeshwari Bhomle, Apurva Chimote, Shubhada Lunge, Shruti Dekate, Payal Gourkhede, Sonali Rangari, Heart disease prediction using artificial neural network, Int. J. Adv. Res. Comput. Commun. Eng. 8 (1) (2019).

[29] Sameh Ghwanmeh, Applying advanced NN-based decision support scheme for heart diseases diagnosis, Int. J. Comput. Appl. (0975–8887) 44 (2) (2012).

[30] Abhale Babasaheb Annasaheb Vijay Kumar Verma, Prediction for heart disease problem based on most suitable recommendation, Int. J. Eng. Res. Manag. Technol. (ISSN: 2348-4039) 3 (4) (2016).

[31] Sayali Ambekar, Rashmi Phalnikar, Disease risk prediction by using convolutional neural network, in: 2018 Fourth International Conference on Computing, Communication Control and Automation (ICCUBEA), 16, 2018, http://dx.doi.org/10.1109/ICCUBEA.2018.8697423, IEEE Explore.

[32] Ying An, Nengjun Huang, Xianlai Chen, FangXiang Wu, Jianxin Wang, High risk prediction of cardiovascular diseases via attention based deep neural networks, IEEE/ACM Trans. Comput. Biol. Bioinform. (2019) http://dx.doi.org/10.1109/TCBB.2019.2935059.

[33] K. Subhadra, B. Vikas, Neural network based intelligent system for predicting heart disease, Int. J. Innov. Technol. Explor. Eng. (IJITEE) (ISSN: 2278-3075) 8 (5) (2019).

Healthcare Analytics 2 (2022) 100032

Contents lists available at ScienceDirect

Healthcare Analytics

journal homepage: www.elsevier.com/locate/health

A predictive analytics approach for stroke prediction using machine learning and neural networks

Soumyabrata Dev [a,b,*], Hewei Wang [c,d], Chidozie Shamrock Nwosu [e], Nishtha Jain [a], Bharadwaj Veeravalli [f], Deepu John [g]

[a] ADAPT SFI Research Centre, Dublin, Ireland
[b] School of Computer Science, University College Dublin, Ireland
[c] Beijing University of Technology, Beijing, China
[d] Beijing-Dublin International College, Beijing, China
[e] National College of Ireland, Dublin, Ireland
[f] Department of Electrical and Computer Engineering, National University of Singapore, Singapore
[g] School of Electrical and Electronic Engineering, University College Dublin, Ireland

ARTICLE INFO

Keywords:
predictive analytics
machine learning
neural network
electronic health records
stroke

ABSTRACT

The negative impact of stroke in society has led to concerted efforts to improve the management and diagnosis of stroke. With an increased synergy between technology and medical diagnosis, caregivers create opportunities for better patient management by systematically mining and archiving the patients' medical records. Therefore, it is vital to study the interdependency of these risk factors in patients' health records and understand their relative contribution to stroke prediction. This paper systematically analyzes the various factors in electronic health records for effective stroke prediction. Using various statistical techniques and principal component analysis, we identify the most important factors for stroke prediction. We conclude that age, heart disease, average glucose level, and hypertension are the most important factors for detecting stroke in patients. Furthermore, a perceptron neural network using these four attributes provides the highest accuracy rate and lowest miss rate compared to using all available input features and other benchmarking algorithms. As the dataset is highly imbalanced concerning the occurrence of stroke, we report our results on a balanced dataset created via sub-sampling techniques.

1. Introduction

We have witnessed amazing developments in the field of medicine with the aid of technology [1]. With the advent of annotated dataset of medical records, we can now use data mining techniques to identify trends in the dataset. Such analysis has helped the medical practitioners to make an accurate prognosis of any medical conditions. It has led to an improved healthcare conditions and reduced treatment costs. The use of data mining techniques in medical records have great impact on the fields of healthcare and bio-medicine [2,3]. This assists the medical practitioners to identify the onset of disease at an earlier stage. We are particularly interested in stroke, and to identify the key factors that are associated with its occurrence.

Several studies [4–7] have analysed the importance of lifestyle types, medical records of patients on the probability of the patients to develop stroke. Further, machine learning models are also now employed to predict the occurrence of stroke [8,9]. However, there is no study that attempts to analyse all the conditions related to patient, and identify the key factors necessary for stroke prediction. In this paper, we attempt to bridge this gap by providing a systematic analysis of the various patient records for the purpose of stroke prediction. Using a publicly available dataset of 29072 patients' records, we identify the key factors that are necessary for stroke prediction. We use principal component analysis (PCA) to transform the higher dimensional feature space into a lower dimension subspace, and understand the relative importance of each input attributes. We also benchmark several popular machine-learning based classification algorithms on the dataset of patient records.

The main contributions of this paper are as follows – (a) we provide a detailed understanding of the various risk factors for stroke prediction. We analyse the various factors present in Electronic Health Record (EHR) records of patients, and identify the most important factors necessary for stroke prediction; (b) we also use dimensionality

* Corresponding author at: School of Computer Science, University College Dublin, Ireland.
E-mail address: soumyabrata.dev@ucd.ie (S. Dev).

[1] In the spirit of reproducible research, the code and data to reproduce the results in this manuscript are available online here: https://github.com/Soumyabrata/EHR-features.

https://doi.org/10.1016/j.health.2022.100032
Received 28 December 2021; Received in revised form 4 February 2022; Accepted 9 February 2022
Available online xxxx

reduction technique to identify patterns in low-dimension subspace of the feature space; and (c) we benchmark popular machine learning models for stroke prediction in a publicly available dataset. We follow the spirit of reproducible research, and therefore the source code of all simulations used in this paper are available online.[1]

The structure of the paper is as follows. The remaining part of Section 1 provides an overview of the related work, and describes the dataset used in our study. Section 2 covers the correlation analysis and feature importance analysis. The results from Principal Component Analysis are explained in Section 3. The data mining algorithms used for predictive modelling and their performance on the dataset is detailed in Section 4. Finally, Section 5 concludes the paper and discusses future work.

1.1. Related work

Existing works in the literature have investigated various aspects of stroke prediction. Jeena et al. provides a study of various risk factors to understand the probability of stroke [8]. It used a regression-based approach to identify the relation between a factor and its corresponding impact on stroke. In Hanifa and Raja [9], an improved accuracy for predicting stroke risk was achieved using radial basis function and polynomial functions applied in a non-linear support vector classification model. The risk factors identified in this work were divided into four groups — demographic, lifestyle, medical/clinical and functional. Similarly, Luk et al. studied 878 Chinese subjects to understand if age has an impact on stroke rehabilitation outcomes [10]. Min et al. in [11] developed an algorithm for predicting stroke from potentially modifiable risk factors. Singh and Choudhary in [12] have used decision tree algorithm on Cardiovascular Health Study (CHS) dataset for predicting stroke in patients. A deep learning model based on a feed-forward multi-layer artificial neural network was also studied in [13] to predict stroke. Similar work was explored in [14–16] for building an intelligent system to predict stroke from patient records. Hung et al. in [17] compared deep learning models and machine learning models for stroke prediction from electronic medical claims database. In addition to conventional stroke prediction, Li et al. in [18] used machine learning approaches for predicting ischaemic stroke and thromboembolism in atrial fibrillation.

The results from the various techniques are indicative of the fact that multiple factors can affect the results of any conducted study. These various factors include the way the data was collected, the selected features, the approach used in cleaning the data, imputation of missing values, randomness and standardization of the data will have an impact on the outcome of any study carried. Therefore, it is important for the researchers to identify how the different input factors in an electronic health record are related to each other, and how they impact the final stroke prediction accuracy.

Studies in related areas [3,19] demonstrate that identifying the important features impacts the final performance of machine learning framework. It is important for us to identify the perfect combination of features, instead of using all the available features in the feature space. As indicated in [3], redundant attributes and/or totally irrelevant attributes to a class should be identified and removed before the use of a classification algorithm. Therefore, it is essential for data mining practitioners in healthcare to identify how the risk factors captured in electronic health records are inter-dependent, and how they impact the accuracy of stroke prediction independently.

1.2. Electronic health records dataset

An Electronic Health Record (EHR) also known as Electronic Medical Record (EMR), is a repository of information for a patient. It is an automated, computer readable storage of the medical status of a patient that is keyed in by qualified medical practitioners. The records contain vitals, diagnosis or medical exam results of a patient. The future of medical diagnosis looks promising with the optimal use of EHR. The use of EHR increased from 12.5% to 75.5% in US Hospitals between 2009 and 2014 as indicated by the statistics recorded in [20].

For our study, we use a dataset of electronic health records released by McKinsey & Company as a part of their healthcare hackathon challenge.[2] The dataset is available from Kaggle,[3] a public data repository for datasets. The dataset contains the EHR records of 29072 patients. It has a total of 11 input attributes, and 1 output feature. The output response is a binary state, indicating if the patient has suffered a stroke or not. The remaining 11 input features in EHR are: patient identifier, gender (G), age (A), binary status if the patient is suffering from hypertension (HT) or not, binary status if the patient is suffering from heart disease (HD) or not, marital status (M), occupation type (W), residence (urban/rural) type (RT), average glucose level (AG), body mass index (BMI), and patient's smoking status (SS). The dataset is highly unbalanced with respect to the occurrence of stroke events; most of the records in the EHR dataset belong to cases that have not suffered from stroke. The publisher of the dataset has ensured that the ethical requirements related to this data are ensured to the highest standards. In the subsequent discussion of this paper, we will exclude the patient identifier as one of the input feature. We will consider the remaining 10 input features, and 1 response variable, in our study and analysis.

2. Analysing electronic health records

In this section, we provide an analysis of electronic health records dataset. We perform correlation analysis of the features. We use the entire dataset of EHR records to perform such analysis on the input features of the EHR records. Correlation analysis is useful for feature selection in the following manner: if two features have very high correlation, one of them can be ignored in the prediction of occurrence of stroke as it does not contribute any additional knowledge to the prediction model. Moreover, we evaluate the behaviour of the features individually and in a group to gaze the importance of each individual feature in predicting the occurrence of a stroke. A systematic analysis of the input feature space is an integral part for stroke prediction. It is important to find the optimal and minimal set of predictive features to reduce the computational cost of modelling and efficient archival of EHR records. This paves us the path for clinicians to record *only* those features in the EHR records that are most efficient for stroke prediction.

2.1. Correlation between features

We use Pearson's correlation coefficient to generate Fig. 1, which shows the correlation between different patient attributes. The strength of the linear relationship between any two features of the patient's electronic health data will be determined by this correlation value. We have used a colourmap in Fig. 1, such that the blue colour represents positive correlation, while red is negative. The deeper the colour and larger the circle size, the higher is the correlation between the two patient attributes.

As is intuitive, the correlation of an attribute with itself is unity. There is a significant correlation between a patient's marital status and their age with 0.5 correlation index. There is also a positive correlation between patient's age and the type of their work with 0.38 correlation index, whether they suffer from hypertension and heart disease or not and their average glucose level. This correlation of patient's age with other attributes seems intuitive, as most ailments occur in an ageing population. The type of residence of patient is not correlated with any other attribute. Patient's type of work has a positive correlation with their marital status with 0.35 correlation index.

[2] https://datahack.analyticsvidhya.com/contest/mckinsey-analytics-online-hackathon/.

[3] https://www.kaggle.com/fedesoriano/stroke-prediction-dataset.

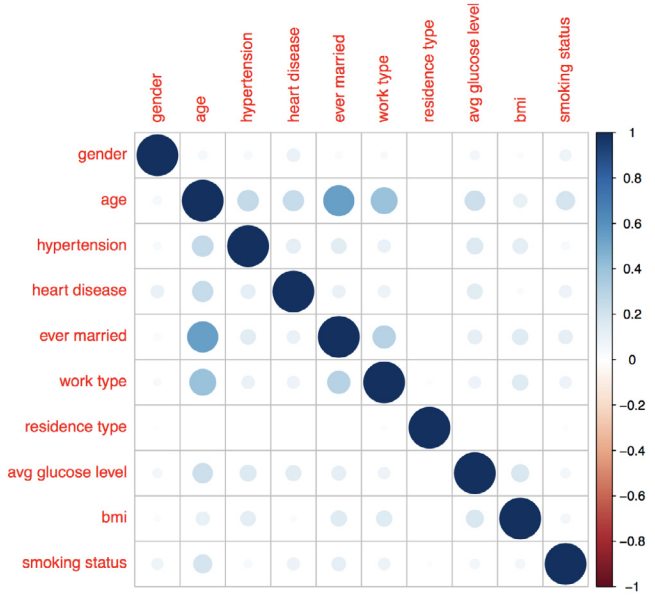


Fig. 1. Correlation matrix for patient attributes in the dataset. These attributes are gender, age, status 0/1 if patient is suffering from hypertension, status 0/1 if patient is suffering from heart disease, marital status, work type, residence type, average glucose level, body mass index and patient's smoking status.

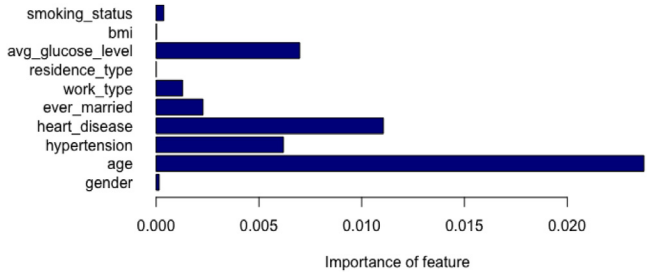


Fig. 2. Importance of patient attributes in predicting the occurrence of stroke with a Linear Vector Quantization (LVQ) model.

In summary, the correlation matrix shown in Fig. 1 tells us that none of the features are highly correlated with each other. Thus, each feature might have their individual contribution towards stroke prediction. The next two subsections analyse the importance of an individual feature for stroke prediction.

2.2. Individual features for stroke prediction

Fig. 2 shows the importance of each patient's attribute in predicting the occurrence of stroke using a Learning Vector Quantization (LVQ) model. The relative importance of a patient's attribute is measured by the increase in the model's prediction error due to that attribute. We use the varImp method from the R `caret` package[4] to compute this relative feature importance. As Fig. 2 illustrates, patient's age (A) is the feature with highest importance in predicting the occurrence of stroke. The other features with high importance are presence of heart disease (HD), patient's average glucose level (AG) and presence of hypertension.

The analysis described above shows patient's age (A) has a comparatively higher importance by itself, yet a combination of different features may improve prediction because they are not correlated with each other. Furthermore, we also compute the $CHADS_2$ score for the EHR records. CHADS2 score is a stroke risk score for non valvular atrial

[4] http://topepo.github.io/caret/index.html.

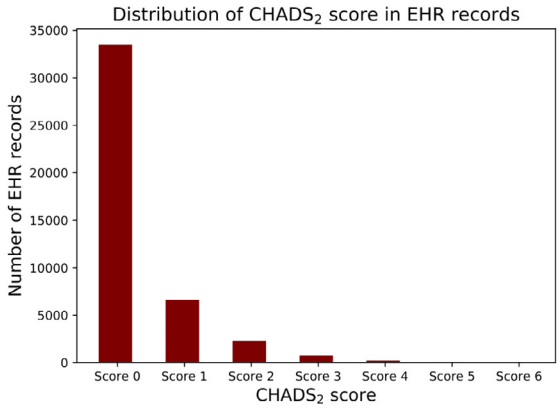


Fig. 3. We show the distribution of $CHADS_2$ score in the EHR records dataset. We observe that most of the $CHADS_2$ score values are low.

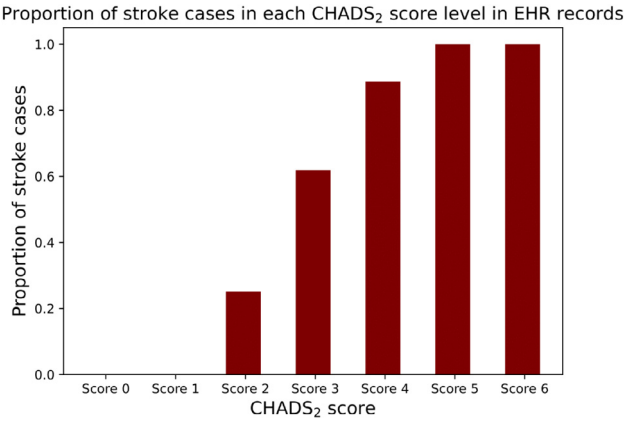


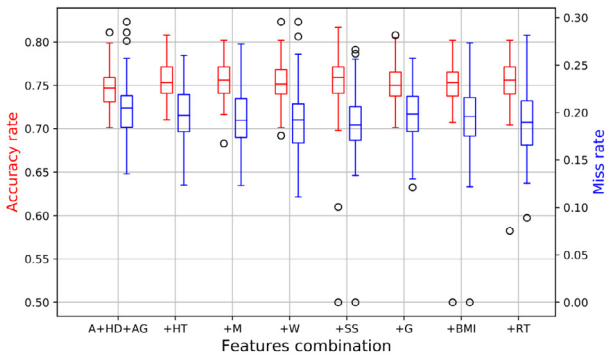
Fig. 4. The proportion of cases with a stroke event as predicted by $CHADS_2$ for each score level respectively.

fibrillation, where C represents congestive heart failure or impairment of left ventricular function, H represents whether the patient has hypertension, A stands for age, D stands for diabetes, S stands for stroke, or transient ischaemic attack, history of thromboembolism. Fig. 3 shows the distribution of the $CHADS_2$ score for our dataset. We observe that most of the EHR observations have a low $CHADS_2$ score. Fig. 4 shows the proportion of cases with a stroke event as predicted by CHADS2 for each score level. We observe that the larger the CHADS2 score, the higher is the occurrence of stroke cases. Combined with Figs. 3 and 4, we find that most people having $CHADS_2$ score of 1 or 2, have a low probability of stroke. We observe that only a small number of people with $CHADS_2$ score greater than 2 have a higher probability of occurrence of stroke.

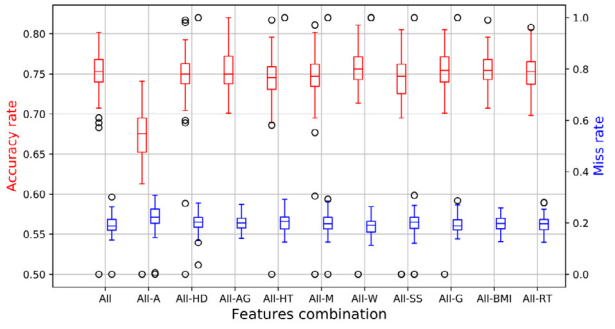
2.3. Selection of optimum features for stroke prediction

In the previous section, we used the features one at a time and saw that there are only few features which have higher importance in stroke prediction. In this section, we use all features then subsequently remove one feature at a time or add one feature at a time and further analyse the importance of an individual feature in stroke prediction. We use neural network algorithm to produce the results. We use a perceptron neural network for this experiment. We use the entire dataset of EHR records to perform the feature analysis.

Fig. 5(a) shows the results for subsequently adding one feature at a time. The first result corresponds to using features A, HD and AG. These features are chosen as from Fig. 2, we observe that these three features show higher importance compared to others. When



(a) Adding features



(b) Removing features

Fig. 5. We measure the accuracy rate and miss rate of stroke prediction when (a) when features are added one at a time, and (b) individual features are deleted one at a time. The box plots represent the distribution of the metrics, computed from 100 experiments.

HT is added to this pool, we see that the accuracy rate is slightly improved and miss rate is slightly decreased. Note that HT is the fourth important variable from the analysis of Fig. 2. Subsequently, when other features are added one by one, the accuracy rate and miss rate show very slight variation. The accuracy and miss rate are almost same for all the remaining configurations. Therefore, it suggests that the four features: A, HD, AG and HT can be optimum features as there is no improvement offered by addition of other features.

Fig. 5(b) shows the results for deleting one feature at a time from the pool of all features. Here we can observe that the accuracy rate is significantly affected when the feature A is removed from the pool. This is inline with our earlier discussion which showed that A has highest score amongst all features (ref to Fig. 2). There is no much significant changes when other features are removed from the pool.

3. Principal component analysis

In this section, we analyse the variance in the dataset using Principal Component Analysis (PCA). In this multivariate analysis, the dataset is transformed into a set of values of linearly uncorrelated variables called principal components such that maximum variance is extracted from the variables. These principal components act as summaries of the features of the dataset. These new basis functions do not have a physical interpretation. However, these new basis functions are linear combinations of the original feature vectors. In this work, we do not restrict ourselves on the feature analysis using traditional feature elimination techniques. However, we use dimensionality reduction technique to transform the high-dimensional feature space onto 2-dimensional feature space to understand the inter-relation amongst the feature space. PCA can be used to reduce the feature space for predictive modelling if the first few components capture most of the variance in the data.

We analyse the 10 dimensional patient attribute in the lower dimension subspace using PCA. We provide a brief primer on principal component analysis and mathematically formulate our problem statement.

Let us suppose that $\mathbf{X}$ is the variable matrix of dimension $m \times n$. In this case, m indicates the total number of input attributes in EHR, and n is the total number of patient records in the dataset. Therefore, $m = 10$ and $n = 29072$ in this analysis for stroke prediction. We vectorize the individual features f_{1-10} from the matrix $\mathbf{X}$, into corresponding $\widetilde{\mathbf{v}}_j \in \mathbb{R}^{mn\times 1}$ where $j = 1, 2, \ldots, 10$. Finally, the $\widetilde{\mathbf{v}}_j$ features are stacked together to create the matrix $\hat{\mathbf{X}} \in \mathbb{R}^{mn\times 10}$:

$$\hat{\mathbf{X}} = [\widetilde{\mathbf{v}}_1, \widetilde{\mathbf{v}}_2, \widetilde{\mathbf{v}}_3, \ldots, \widetilde{\mathbf{v}}_{10}]. \tag{1}$$

We perform the PCA on the normalized matrix of $\hat{\mathbf{X}}$, that is normalized using the corresponding means $\bar{v_j}$ and standard deviations σ_{v_j} of the individual features. The normalized matrix $\ddot{\mathbf{X}}$ is represented as:

$$\ddot{\mathbf{X}} = \left[\frac{\widetilde{\mathbf{v}}_1 - \bar{v_1}}{\sigma_{v_1}}, \frac{\widetilde{\mathbf{v}}_2 - \bar{v_2}}{\sigma_{v_2}}, \ldots, \frac{\widetilde{\mathbf{v}}_j - \bar{v_j}}{\sigma_{v_j}}, \ldots, \frac{\widetilde{\mathbf{v}_{10}} - \bar{v_{10}}}{\sigma_{v_{10}}}\right]. \tag{2}$$

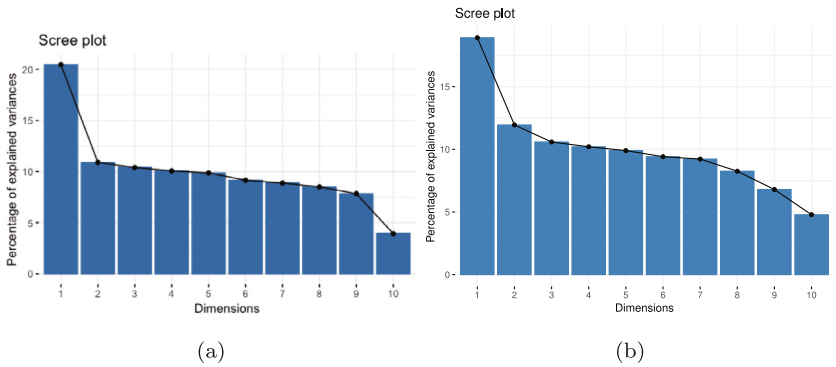
We interpret how the results from PCA are related to predictive variables or features represented by patient attributes and the individual observations represented by the medical health records. We study the relation between the first two principal components and the individual input variables. We also study the importance of the first two principal components for a given observation. We use the guide by Abdi and Williams [21] for this study.

3.1. Variance explained by principal components

A scree plot is used to select the components which explain most variability in the dataset, generally 80% or more variance. Fig. 6(a) shows the percentage of variance in the dataset explained by the different principal components in the original dataset. Here, we can observe that the variance explained by different principal components are very low. Out of 10 principal components, 8 are needed to explain variance of 88.2%. Balanced dataset means that the dataset is balanced with respect to the stroke labels using random sampling. The original dataset is unbalanced because there are more samples possessing negative stroke labels, as compared to positive label stroke samples. We make it balanced by considering all the positive stroke samples, and then randomly picking equal number of negative stroke samples from the rest. This will make a balanced dataset with equal number of positive and negative stroke samples.

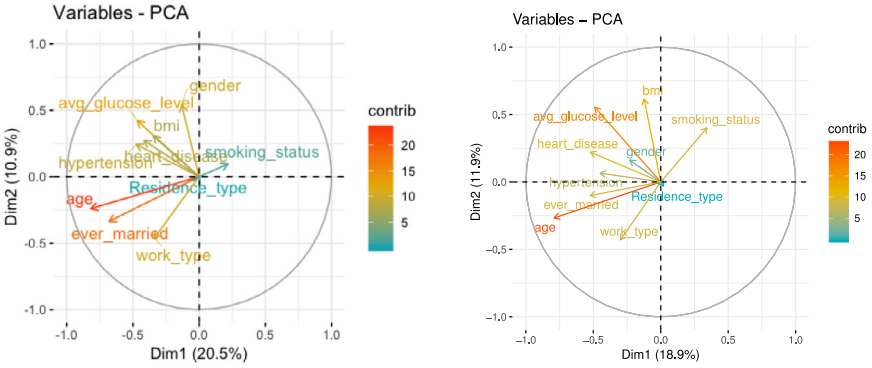
All principal components are orthogonal to each other and hence uncorrelated. Therefore, each individual PC can be useful to explain a unique phenomenon. The distributed variance in Fig. 6 indicates that the different principal components are explaining different underlying phenomenon. These phenomenon can be analysed based on the variable loadings. Variable loadings are the contribution of different variables to an individual principal component. We have included the scree plot for the balanced dataset as well in Fig. 6(b). This is useful for researchers to understand the impact of unbalanced nature of the dataset on the subspace representation. Table 1 shows the contribution of each variable towards first two principal components. The sum of the squares of all loadings for an individual principal component equals unity. Therefore, if the variable loading crosses a threshold value of $\sqrt{1/10} = 0.31$, it indicates that the variable has a strong contribution towards the principal component. In Table 1, the variables that cross the threshold are in bold. Here we observe that variables like A, AG, HT and M have strong contribution towards the first principal component. The variable HD also shows significant contribution towards it. From earlier discussions, we saw that these features actually are important from stroke prediction point of view. Therefore, it indicates that the first principal component (which has stronger loadings from these variables) might be useful in predicting the stroke.

In the following sections, we will assess the role of these principal components in stroke prediction.



(a) (b)

Fig. 6. Percentage of variance explained by different principal components. We show the scree plot for (a) original, (b) balanced datasets.



(a) (b)

Fig. 7. Biplot representation of the input attributes on the first two principal components. We show the biplot for (a) original, (b) balanced datasets.

Table 1
We check the contributions of the different features in the first and second principal components. We report the absolute values of the different loading factors.

Features	PC_1	PC_2
gender	0.092	**0.516**
age	**0.571**	0.230
hypertension	**0.331**	0.232
heart_disease	0.290	0.261
ever_married	**0.475**	**0.322**
work_type	0.236	**0.442**
residence_type	0.001	0.003
avg_glucose_level	**0.326**	**0.403**
bmi	0.242	0.293
smoking_status	0.152	0.090

3.2. Relation between principal components with patient attributes

Fig. 7 describes the biplot representation. It shows how the different input attributes are correlated with other, and also depend on the first and second principal components. The x-axis and y-axis indicate the first and second principal components respectively. The each vectors represent an input attribute, and its length indicate its importance. We observe that the average glucose level and the heart disease are correlated to each other. The age has the biggest contribution in the first two principal component. We also observe that the orientation of the different feature vectors in the two-dimensional feature space is the same as the unbalanced dataset.

Fig. 7(a) illustrates that the contribution of patient's residence type is minimum to the two principal components. We also observe that age and status of marriage are correlated with each other, and have a high contribution to the first principal component. The smoking status of a patient and its average glucose level are orthogonal to each other, indicating that they provide different information to the feature space. However, smoking status and age point opposite to each other, indicating that they provide similar information, but have a diverging characteristics. We also compute the biplot for the EHR records on the balanced dataset. We show the corresponding biplot in Fig. 7(b). We observe that the average glucose level and the heart disease are correlated to each other. The age has the biggest contribution in the first two principal component. We also observe that the orientation of the different feature vectors in the two-dimensional feature space is the same as the unbalanced dataset.

3.3. Relation of principal components with individual records

Finally, we also check the relation of individual patient record observations on the first two principal components. In Fig. 8, each dot represents a patient record observation.

Fig. 8(a) shows the importance of principal components onto each of the records. This is generally represented by the *cos2* measure, indicating the squared distance from the origin. This indicates that observations that possess a high *cos2* value can be represented by the principal components, as opposed to observations with a low *cos2* value. We observe that a significant number of points are clustered near the origin, which cannot be represented completely by the principal components. Therefore, the principal component can indicate only a part of the entire information in the feature space.

Fig. 8(b) colour codes the patient records based on the status of the stroke. As the dataset is highly unbalanced, we observe that most of the observations are colour coded with negative status of stroke. We observe a few observations with positive status of stroke (observations with 1 label). However, these positive stroke observations are not located in clusters. This indicates that higher-dimensional features are necessary to separate the observations.

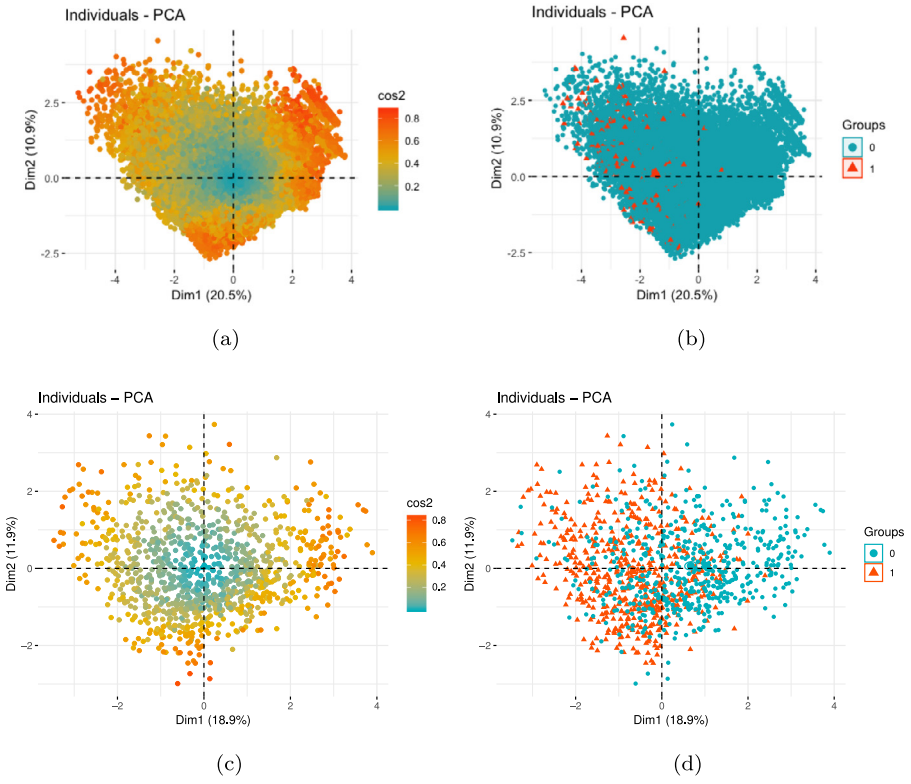


Fig. 8. Subspace representation of the different patient records in reference to the first two principal components. The observations are colour coded based on (a) *cos2* measure indicating the importance of principal components on the observation; and (b) status of stroke. We also use balanced dataset and observe the sub-space representation in (c) and (d).

We also compute the subspace representation of the different features for the balanced EHR dataset. We show it in Fig. 8(c) and (d). We observe that the observations with *stroke* and *no stroke* are scattered throughout the two principal axes. The observations with similar labels are not clustered together. Therefore, the features of the EHR records are important for efficient stroke prediction.

3.4. Discussion

This section discussed how principal component analysis can assist in a clear understanding of the original feature space of patient records. We showed that the first two principal components can cumulatively capture only 31.4% of the total variance in the input feature space. More so, first eight components can explain only about 88% of the total variance. Moreover, we studied the contribution of different patient attributes to the first two principal components. We could see that the two components do not represent the health records data perfectly. We also looked at the contribution of the first two principal components in the representation of individual health records. We could see that some health records can be represented by the two components, but some cannot. Thus, all principal components are needed to have a good representation of the variance in medical records. We cannot get a significant reduction of feature space for predictive modelling without a significant loss of variance in the data. Hence, we use all the principal components for predictive modelling of stroke occurrence.

In the next section, we compare the state of art machine learning classification techniques for predicting the occurrence of stroke in a patient's medical record. As discussed, we use ten patient attributes as input features to the models.

4. Stroke prediction

We provide a detailed analysis of various benchmarking algorithms in stroke prediction in this section. We benchmark three popular classification approaches — neural network (NN), decision tree (DT) and random forest (RF) for the purpose of stroke prediction from patient attributes. The decision tree model is one of the popular binary classification algorithm. This method involves building a tree-like decision process with several condition tests, and then applying the tree to the medical record dataset. Each node in this tree represents a test, and the branches correspond to the outcome of the test. The leaf nodes finally represent the class labels. The pruning ability of such algorithm makes it flexible and accurate, which is required in medical diagnosis. We also benchmark the dataset on random forest approach. The flexibility and ease of use of the random forest algorithm coupled with its consistency in producing good results, even with minimal tuning of the hyper-parameters makes this algorithm valuable in this application. The possibilities of over-fitting are limited by the number of trees existent in the forest. Moreover, random forest can also provide adequate indicators on the way it assigns significance to each of these input variables. We also benchmark the performance of a 2-layer shallow neural network. Artificial neural networks are quite popular these days, and they offer competitive results. We implement the feed-forward multi-layer perceptron model using the `nnet` R package.

4.1. Benchmarking using all features

Our dataset contains a total of 29072 medical records. Out of this, only 548 records belong to patients with stroke condition, and the remaining 28524 records have no stroke condition. This is a highly unbalanced dataset. This creates problem in using this data directly for training any machine-learning models. Therefore, we use random downsampling technique to reduce the adverse impact of the unbalanced nature of the dataset. We refer the 548 records as the minority class, and the remaining 28524 records with no stroke condition as the majority class. Subsequently, we create a dataset of 1096 observations, that consists of 548 minority samples and 548 majority samples. This balanced dataset is created by considering all the 548 minority samples,

Table 2
Performance evaluation of neural network, decision tree and random forest on our dataset of electronic medical records. We compare their performance for three different cases – (a) using all the original features, (b) using the PCA-transformed data of the first two principal components, and (c) using the PCA-transformed data of the first eight components. We choose 8 components, as 8 components are necessary to cumulatively contain more than 80% of the explained variance. We report the average value of precision, recall, F-score, accuracy, miss rate and fall-out rate, based on 100 experiments.

Features	Way	Precision	Recall	F-score	Accuracy	Miss rate	Fall-out rate
Original features (All)	DT	0.75	0.74	0.74	0.74	0.17	0.24
	RF	0.74	0.73	0.73	0.74	0.18	0.25
	NN	0.80	0.74	0.77	0.77	0.16	0.18
	CNN	0.74	0.72	0.73	0.74	0.17	0.24
	SVM	0.67	0.68	0.68	0.68	0.23	0.32
	LASSO	0.78	0.72	0.75	0.76	0.19	0.20
	ElasticNet	0.79	0.71	0.75	0.76	0.19	0.19
Original features (A+HD+AG+HT)	DT	0.78	0.71	0.74	0.75	0.20	0.21
	RF	0.76	0.74	0.75	0.75	0.18	0.24
	NN	0.78	0.71	0.74	0.75	0.19	0.20
PCA features (PC1 and PC2)	DT	0.78	0.65	0.71	0.73	0.24	0.19
	RF	0.71	0.68	0.69	0.69	0.23	0.28
	NN	0.77	0.67	0.72	0.74	0.22	0.20
PCA features (PC1 till PC8)	DT	0.75	0.68	0.72	0.73	0.21	0.23
	RF	0.73	0.69	0.71	0.72	0.21	0.25
	NN	0.80	0.68	0.73	0.75	0.21	0.17

Table 3
Accuracy variance for NN, SVM, LASSO and ElasticNet.

Features	Way	Precision	Accuracy variance
Original features (All)	NN	0.80	0.000377
	SVM	0.67	0.000470
	LASSO	0.78	0.000380
	ElasticNet	0.79	0.000467

Table 4
Hyperparameters of the CNN model.

Layers	In channels/Out channels	Kernel, Stride, Padding	Activation functions
Conv1	1/16	3, 1, 1	ReLu
Conv2	16/8	2, 1, 0	ReLu
Layers	In features/Out features	Kernel, Stride, Padding	Activation functions
Linear1	32/16	–	ReLu
Linear2	16/1	–	Sigmoid

and the remaining 548 majority samples are selected randomly from the 28524 patient records. All the three machine learning models are trained on this balanced dataset of 1096 observations.

In our experiment, another deep learning approach, the convolutional neural network (CNN) is implemented for the prediction of stroke. In our configuration, the number of hidden layers is four while the first two layers are convolutional layers and the last two layers are linear layers, the hyperparameters of the CNN model is given in Table 4. We use the same train and test split for CNN training and testing procedure, the ten inputs features are reshaped into 1 * 2 * 5 for inputs.

We also calculate accuracy variance for the benchmarking methods. Table 3 shows the experiments of accuracy variance for NN, SVM, LASSO and ElasticNet.

4.2. Benchmarking using top four features

Here we present results for stroke prediction when all the features are used and when only 4 features (A, HD, AG and HT) are used. These features are selected based on our earlier discussions. In addition to the features, we also show results for stroke prediction when principal components are used as the input. Since we observed that almost 8 principal components are needed to explain a variance of greater than 80%, we present results for both the cases when only first 2 principal components are used and when all the components are used. Table 2 shows the evaluation metrics for all the different configurations. In order to remove sampling bias, we perform 100 random downsampling experiments. The ratio of the number of training observations and testing observations is 70: 30. Table 2 reports the average values for all the approaches.

We observe that, among different machine learning approaches,[5] neural network works with better accuracy for different feature combinations. When we compare the neural network results for cases when all features are used and when only 4 features are used, we do not observe a significant improvement of all features over 4 features. Therefore, we can get a good stroke prediction accuracy of up to 78% with a low miss rate of 19% by using only 4 features (A, HD, AG and HT). This result might not be sufficient to guide treatment and prevention measures on an individual level but it will assist in supporting allocation and resources on a population and/or cohort level.

Here, for the case when the principal components are used as inputs, we observe that there is only slight improvement in accuracy of neural network, when all components are used compared to the first two components only. This is inline with our earlier discussion, where we illustrated that the first principal component can be important from the stroke prediction point of view. The variables that contributed most to the first component have shown good stroke prediction possibility. Therefore, use of only first two components have similar results compared to the case when all components are used.

When we compare the principal component results to the case when actual features are used, it can be observed that the accuracy of neural networks for both cases are comparable. However, if we look at the miss rate, the miss rates are higher for the case of principal components. The miss rate is an important evaluation metric as we would want to be able to detect all the strokes without a fail. We expect miss rate to be as low as possible and a slight degradation in the miss rate value is important for us for further consideration. Therefore, our analysis suggests that the best possible results for stroke prediction can be achieved by using neural network with 4 important features (A, HD, AG and HT) as input.

Finally, we illustrate the distribution of the accuracy values, by using the top 4 features — age, heart disease, average glucose level, hypertension from the dataset. We perform the experiments 100 times to remove any sampling bias in the training and testing sets. Fig. 9

[5] We do not benchmark our approaches against the McKinsey Kaggle challenge winner, as the model code is not publicly available.

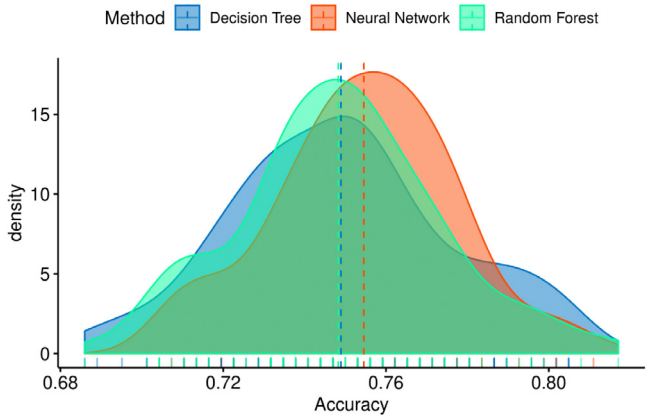


Fig. 9. Histogram distribution of classification accuracies obtained from the 100 experiments using top 4 features — age, heart disease, average glucose level, hypertension for the benchmarking algorithms.

illustrates this. We observe that most of them have similar performance, with their mean overlapping around similar values. We also compute the variance of the accuracies of the benchmarking methods for the 100 random sub-sampling observations. The variance of decision tree, neural network and random forest are 0.00073, 0.00049, and 0.00061 respectively. This indicates that there is no sampling bias involved in the benchmarking results.

5. Conclusion and future work

In this paper, we presented a detailed analysis of patients' attributes in electronic health record for stroke prediction. We systematically analysed different features. We performed feature correlation analysis and a step wise analysis for choosing an optimum set of features. We found that the different features are not well-correlated and a combination of only 4 features (A, HD, HT and AG) might have good contribution towards stroke prediction. Additionally, we performed principal component analysis. The analysis showed that almost all principal components are needed to explain a higher variance. The variable loadings however showed that the first principal component which has the highest variance might explain the underlying phenomenon of stroke prediction. Finally, three machine learning algorithms were implemented on a set of different features and principal components configurations. We found that neural network works the best with a feature combination of A, HD, HT and AG. The accuracy and miss rate for this combination are 78% and 19% respectively.

We have seen promising results from using just 4 features. The accuracy of the perceptron model cannot be improved further for primarily two reasons: lack of additional discriminatory feature set; and lack of additional dataset. We observed that most of the existing features in the EHR dataset are highly correlated to each other, and therefore do not add any additional information to the original feature space. Furthermore, a larger dataset will enable us to train our deep neural networks more efficiently. We plan to collect institutional data in our planned future work. The systematic analysis of the different features in the electronic health records will assist the clinicians in effective archival of the records. Instead of recording and storing all the features, the data management team can archive *only* those features that are essential for stroke prediction. Thus, in future, we plan to integrate the electronic records dataset with background knowledge on different diseases and drugs using Semantic Web technologies [22,23]. Knowledge graph technologies [23,24] can be used in order to publish the electronic health records in an interoperable manner to the research community. The added background knowledge from other datasets can also possibly improve the accuracy of stroke prediction models as well. We intend to collect our institutional dataset for further benchmarking of these machine learning methods for stroke prediction. We also plan to perform external validation of our proposed method, as a part of our upcoming planned work.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was conducted with the financial support of Science Foundation Ireland under Grant Agreement No. 13/RC/2106_P2 at the ADAPT SFI Research Centre at University College Dublin. ADAPT, the SFI Research Centre for AI-Driven Digital Content Technology, is funded by Science Foundation Ireland through the SFI Research Centres Programme.

References

[1] G. Sivapalan, K. Nundy, S. Dev, B. Cardiff, J. Deepu, ANNet: A lightweight neural network for ECG anomaly detection in IoT edge sensors, IEEE Transactions on Biomedical Circuits and Systems (2) (2022).
[2] H.C. Koh, G. Tan, et al., Data mining applications in healthcare, J. Healthc. Inf. Manage. 19 (2) (2011) 65.
[3] I. Yoo, P. Alafaireet, M. Marinov, K. Pena-Hernandez, R. Gopidi, J.-F. Chang, L. Hua, Data mining in healthcare and biomedicine: a survey of the literature, J. Med. Syst. 36 (4) (2012) 2431–2448.
[4] J.F. Meschia, C. Bushnell, B. Boden-Albala, L.T. Braun, D.M. Bravata, S. Chaturvedi, M.A Creager, R.H. Eckel, M.S. Elkind, M. Fornage, et al., Guidelines for the primary prevention of stroke: a statement for healthcare professionals from the American heart association/American stroke association, Stroke 45 (12) (2014) 3754–3832.
[5] P. Harmsen, G. Lappas, A. Rosengren, L. Wilhelmsen, Long-term risk factors for stroke: twenty-eight years of follow-up of 7457 middle-aged men in goteborg, sweden, Stroke 37 (7) (2006) 1663–1667.
[6] C.S. Nwosu, S. Dev, P. Bhardwaj, B. Veeravalli, D. John, Predicting stroke from electronic health records, in: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2019, pp. 5704–5707.
[7] M.S. Pathan, Z. Jianbiao, D. John, A. Nag, S. Dev, Identifying stroke indicators using rough sets, IEEE Access 8 (2020) 210318–210327.
[8] R.S. Jeena, S. Kumar, Stroke prediction using SVM, in: Proc. International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), 2016, pp. 600–602, http://dx.doi.org/10.1109/ICCICCT.2016.7988020.
[9] S.-M. Hanifa, K. Raja-S, Stroke risk prediction through non-linear support vector classification models, Int. J. Adv. Res. Comput. Sci. 1 (3) (2010).
[10] J.K. Luk, R.T. Cheung, S. Ho, L. Li, Does age predict outcome in stroke rehabilitation? A study of 878 Chinese subjects, Cerebrovasc. Dis. 21 (4) (2006) 229–234.
[11] S.N. Min, S.J. Park, D.J. Kim, M. Subramaniyam, K.-S. Lee, Development of an algorithm for stroke prediction: a national health insurance database study in Korea, Eur. Neurol. 79 (3–4) (2018) 214–220.
[12] M.S. Singh, P. Choudhary, Stroke prediction using artificial intelligence, in: 2017 8th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON), IEEE, 2017, pp. 158–161.
[13] P. Chantamit-o, Prediction of stroke disease using deep learning model.
[14] A. Khosla, Y. Cao, C.C.-Y. Lin, H.-K. Chiu, J. Hu, H. Lee, An integrated machine learning approach to stroke prediction, in: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, 2010, pp. 183–192.
[15] C.-Y. Hung, C.-H. Lin, T.-H. Lan, G.-S. Peng, C.-C. Lee, Development of an intelligent decision support system for ischemic stroke risk assessment in a population-based electronic health record database, PLoS One 14 (3) (2019) e0213007.
[16] D. Teoh, Towards stroke prediction using electronic health records, BMC Med. Inform. Decis. Mak. 18 (1) (2018) 1–11.
[17] C.-Y. Hung, W.-C. Chen, P.-T. Lai, C.-H. Lin, C.-C. Lee, Comparing deep neural network and other machine learning algorithms for stroke prediction in a large-scale population-based electronic medical claims database, in: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2017, pp. 3110–3113.

[18] X. Li, H. Liu, X. Du, P. Zhang, G. Hu, G. Xie, S. Guo, M. Xu, X. Xie, Integrated machine learning approaches for predicting ischemic stroke and thromboembolism in atrial fibrillation, in: AMIA Annual Symposium Proceedings, 2016, American Medical Informatics Association, 2016, p. 799.

[19] S. García, J. Luengo, F. Herrera, Tutorial on practical tips of the most influential data preprocessing algorithms in data mining, Knowl.-Based Syst. 98 (2016) 1–29.

[20] B.A. Goldstein, A.M. Navar, M.J. Pencina, J. Ioannidis, Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review, J. Am. Med. Inform. Assoc. 24 (1) (2017) 198–208.

[21] H. Abdi, L.J. Williams, Principal component analysis, Wiley Interdiscip. Rev. Comput. Stat. 2 (4) (2010) 433–459.

[22] B. Tilahun, T. Kauppinen, C. Keßler, F. Fritz, Design and development of a linked open data-based health information representation and visualization system: potentials and preliminary evaluation, JMIR Med. Inform. 2 (2) (2014).

[23] F. Orlandi, A. Meehan, M. Hossari, S. Dev, D. O'Sullivan, T. AlSkaif, Interlinking heterogeneous data for smart energy systems, in: 2019 International Conference on Smart Energy Systems and Technologies (SEST), IEEE, 2019, pp. 1–6.

[24] J. Wu, F. Orlandi, I. Gollini, E. Pisoni, S. Dev, Uplifting air quality data using knowledge graph, in: 2021 Photonics & Electromagnetics Research Symposium (PIERS), IEEE, 2021, pp. 2347–2350.

Diabetes & Heart Disease Prediction Using Machine Learning

BhaveshDhande[1], Kartik Bamble[2], Sahil Chavan[3], Tabassum Maktum[4]

[1,2,3,4]Ramrao Adik Institute of Technology, D Y Patil Deemed to be University, Navi Mumbai, India

Abstract One of the root causes of mortality in today's world is the culmination of several heart disease and diabetes illnesses. In clinical data analysis, predicting multiple diseases is a significant challenge. The machine learning approach has proved to be functional in assisting in the decision-making and governing of large amounts of data generated by the healthcare field. The various experiments scratch the surface of machine learning to predict different diseases. The papers present a novel method for identifying significant features using machine learning techniques, which improves the diagnosis of multi-purpose disease prediction. The different features and many well-known classification methods are used to implement the prediction model to predict the heart disease and diabetes. The proposed method utilizes ensemble approach for achieving a higher degree of accuracy rates for by using classification algorithms and feature selection methods. The proposed method implements voting classifier that has sigmoid SVC, AdaBoost, and Decision tree algorithms. The paper also implements the traditional classifiers and presents the comparison of different models in terms of accuracy. The web application is also developed for users to avail its services very easily and make it convenient for their use, particularly in the prediction of heart and diabetes collectively.

Keywords: Machine Learning, classification, feature selection, prediction, heart disease, diabetes

1. Introduction

The answer to calculate risk through diseases via model-based prediction is very difficult. Due to this, the examination of several medical datasets and their forecasting using soft computing is very handy and a cheaper way for professionals in the healthcare industry. These techniques include exploratory analysis and constructive models that support the professional in statistical decision-making, which is a massive requirement of the medical industry. The high blood pressure, high cholesterol, diabetes, peculiar pulse rate, and various other risk factors make it harder to detect illnesses [1]. The severity of heart disease in humans was determined using various data mining and neural network methodologies. The decision trees, genetic algorithm, Naive Bayes, and K-nearest neighbor algorithm are all used to classify the severity of the disease. Because the features of their problems are complex, the diseases must be treated with caution. Failure to do so can reduce the effectiveness of organs or lead to early death. The various metabolic diseases are identified using a medical science approach and raw data mining. The data mining techniques with classification plays a vital role in predicting heart disease and data research [2]. The random forest is an ensemble machine learning algorithm. It is perhaps the most popular and widely used machine learning algorithm with excellent performance across a wide range of classification and regression predictive modeling problems. Also, random forest approach has a brute way of tuning parameters that helps in easier feature selection [3]. The process of aggregating the votes for class labels from individual models and predicting the class with the most votes is called a hard voting ensemble. Whereas, in a soft voting ensemble, the predicted probabilities for class labels are added up, and the class label with the highest sum probability is predicted. The voting ensemble method assumes that all the models have equal contributions in predictions, which is a drawback because some models give better results in some scenarios and poor in others. The Adaptive Boosting (AdaBoost) algorithm is a boosting technique in machine learning that is employed as an ensemble method. The weights are re-allocated to each instance, with higher weights applied to improperly identified instances. This is termed as adaptive boosting. In supervised learning, boosting is used to reduce bias and variation. It is based on the concept of sequential learning [4]. The linear SVM algorithm is used for linearly separable data, which implies that if a dataset can be categorized into two classes using only a single straight line, it is called linearly separable data, and the classifier employed is called linear SVM. Under the supervised learning approach, one of the most prominent machine learning algorithms is logistic regression. It is a method for predicting a categorically dependent variable from a set of independent factors and to describe data and explain the connection between one dependent binary variable and one or more independent variables. As a result, the result must be a discrete or categorical value such as Yes or No, 0 or 1, true or false, and so on. But, instead of giving exact values like 0 and 1, it delivers probabilistic values that are somewhere between 0 and 1. In order to

predict and diagnose the recurrence of cardiovascular disease, ensemble learning incorporates the model approaches of five classifiers, including support vector machine, artificial neural network, Nave Bayesian, regression analysis, and random forest [7]. Dataset Cleveland's cardiovascular data records were retrieved from the UCI repository, and for Diabetes, the PIMA dataset has been used. The findings of the experiments showed that an ensemble model is a superior strategy in terms of diagnostic performance predictability and accuracy.

In this paper an ensemble approach is proposed to detect the heart disease and diabetes. The different algorithms combined include ADA-boost, decision tree and sigmoid SVC. The classifiers are combined by varying their weights. The major contribution of the paper is as follows:

- Provide a new approach to concealed patterns in the medical data.
- To predict the chance of heart disease with the highest accuracy of prediction.
- To predict the chance of diabetics with the highest accuracy of prediction.
- Error rate compression for the results found to make it relatively exact in accuracy.

The rest of the paper is organized as follows:
The section 2 gives the survey of existing systems for predicting heart diseases and diabetes. The section 3 demonstrates the proposed system along with the ensemble approach. The results of the proposed system are presented in section 4. Finally, the conclusion and future work is expounded in Section 5.

2. Literature Survey

The most general causes of mortality on the planet have been heart and diabetic problems. Furthermore, today, the prediction of the same or even hinting at a minute probability of it is a problem that needs a solution. In the medical field, machine learning has paved its purpose by helping make choices and predict by training over large amounts of data existing in the form of datasets.

The study in [1] represent that diabetes mellitus and hypertension were moderately associated while cardiovascular diseases are strongly associated with severity and mortality for COVID-19. The paper helps to gain relation between diabetics and heart diseases and create a link or gain experience to handle data for both diseases at the same time as it gives an idea of immunity prediction of COVID through the data of diabetics and heart. The quantitative estimate of severity outcomes and or deaths in COVID-19 patients was performed with Comprehensive Meta-Analysis Software (CMA) version 3.0. In paper [2] the K-means clustering algorithm is used for predicting heart diseases and analysis is carried out using visualization tool tableau. The Cleveland heart disease raw dataset with 76 features of 303 patients was pre-processed with exploratory data analysis which narrowed down the dataset to 209 records and 7 important features. The study includes 4 types of chest pain with age, maximum heart rate, and chest pain type which are considered as vital features in prediction. In paper [3] HRFLM method is proposed that stands for union of Linear Method (LM) and Random Forest (RF), which boosts efficiency by improving selection. The study involves pre-processing of Cleveland UCI repository with use of R rattle GUI (Feature Selection and Classification modelling) which provides an easy-to-use visual graphics, working environment for the user of the dataset, and building the predictive analytics. The several approaches are presented in [4] for predicting heart diabetes. The methodology with logistic regression provides 96% precision. This was the first paper that observed the study of more than one dataset and competes between algorithms, with pipeline affected to 98.8% fidelity using Adaptive Boost classifier. In paper [5] the difficulties in the diabetic analysis were convened in relation to the COVID-19 rate. The conclusion derived from the study is that different categories of diabetes have a unique effect on the percentage of mortality rate. The paper [6] Bhavesh Dhandeproposes an approach to predict diabetes mellitus by applying machine learning techniques. The paper concludes that minimum redundancy maximum relevant approach is better than principal component analysis. It cements random forests as a better algorithm than others. The two datasets, Luzhou and Pima were utilized, with 80.84% and 77.21% accuracies fetched respectively. The ensemble approach with various classification algorithms such as KNN, Adaptive boost (AB) Gradient boost (XB), decision tree and random forest is proposed in [7]. Based on the analysis of different algorithms, it can be concluded that the proposed system on this research edged are under cover (AUC) promisingly. The perfect couple for prediction turned out to be an ensemble of (AB+XB) classifiers. In paper [8], for predictive data mining for medical diagnosis various techniques such as KNN, Neural Networks, Bayesian classification, Classification based on clustering, Decision Tree, etc. are used. An overview of all prediction models is studied in this research paper. According to a performance study of data mining algorithms Naïve Bayes, Decision Tree, KNN provide the highest accuracy rate. The Weka 3.6.0 tool was used for conducting the research. In paper [9], the methodology is for finding out the best algorithm to extract best features from the medical dataset. Whenever, data collection is followed by data pre-processing, data mining, and pattern evaluation, the suitable and highest accuracy is achieved. The data extraction was performed with the WEKA software tool then compared using predictive accuracy, ROC curve, ROC value. The approach for prediction of heart disease by applying various

algorithms like ANN, random forest, SVM is presented in [10]. By using 3-fold cross-validation along with SVM algorithm maximum accuracy achieved was 83.17%. The application of decision tree algorithm with 37 splits and 6 leaf nodes led to an accuracy of 79.12%, and when used with 5-fold cross-validation technique accuracy achieved was 79.54%. By using random forest algorithm, the accuracy achieved was 85.81%, which is maximum as compared to all other algorithms. The method presented in [11] utilizes XGBoost, AdaBoost, gradient boosting, extra trees, light gradient boosting Lightgbm, SGDC, Nu SVM algorithms for prediction of cardiovascular effects. The data pre-processing was performed on UCI repository and Framingham dataset. The data pre-processing using the Multiple Imputation Chain Equation model for filling inexistent values proved to be efficient way of data pre-processing and the accuracy of 95.83% was obtained by using the stacking algorithm. In [12] research was carried out using three methods KNN, Neural Networks and SVM on real dataset of Algerian people. Neural Network algorithm gave highest accuracy of 93%. [13] paper predicted diabetes based on different human body attributes. Study showed that Body Mass Index (BMI) and growing age are major factors in the development of risk for diabetes.

The limitations of existing system are as follows: The prediction of possibilities is not accurate for disease aggregated inputs and hence thereby cannot handle enormous datasets for patient records effectively. As the machine learning approach is based on predicting outcomes using existing data, we cannot be sure of whether it will apply to the current specimen on which it would be experimented, because evolution is a constant. To add up, poor results on very small datasets causes frequent overfitting occurrences. To conclude, these previous studies focus on the particular impacts of specific machine learning techniques and not on the optimization of these techniques using optimized methods.

3. Proposed Methodology

The paper presents a methodology to predict heart diseases and diabetes by applying machine learning techniques and perform classification by using ensemble classifiers over the datasets involved. This will provide an easy glance at the data involved to build analysis. Thus, the proposed methodology will then materialize into a model evaluation phase based on performance, only after bits like feature selection and data pre-processing. By considering the features extracted, the proposed method will frame an idea of the presence of either disease. Furthermore, all the basic models will be analyzed by ensemble techniques for the prediction. Innovative models like the Random Forest and Adaptive Booster will be added, which will further be evaluated using voting classifiers with weights and without weights.

3.1. System Architecture

The process of making new observations or categorizations from the given available data with the help of supervised learning approach is known as classification. The training dataset is hatched within several classifiers and passed through them.
These 1-n classifications are then used through a combined powerful algorithm. This algorithm then further analyzes the data by considering the ensemble considered model. This ensemble model has a build of various boosters/classifiers/outliers to improve results. Then, after building a model with the available test dataset, the prediction of any random data set can be performed. The system architecture of the proposed method is shown in Figure 1. The working of the proposed method is as follows:

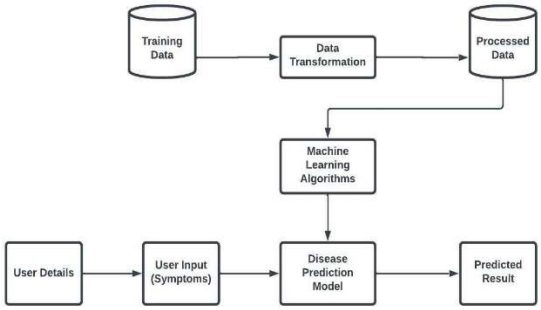


Figure 1: System Architecture

Training Data is undergoing the transformation to generate useful information i.e., processed data by various data analysis and preprocessing using various basic methods. After obtaining processed data, the algorithms can be implemented on the data involved. User then interacts by giving input which is symptoms to the application as a feed to predict results. Then, the algorithms work upon comparisons through various models in the disease prediction model to fetch the required Predicted Result.

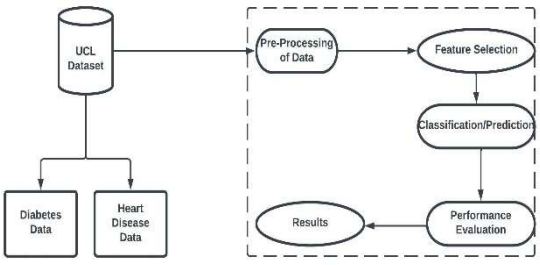


Figure 2: Dataset Handling

The paper considers the two datasets i.e., Heart and Diabetics Datasets. The advantage here is that since both have been taken from the UCL repository and belong to the same result, we can club their results.

The data from these datasets is preprocessed as per the need of exploratory data analysis and the right information is extracted. The process of feature selection is then applied, where the most important features and the impactful ones are thoroughly selected by using required classifiers and outliers.
Further, the different classification models are built using random forest, Ada boost, KNN, logistic regression and decision tree approach to predict the disease. Further ahead, these results are then performance evaluated to check their accuracies and precision, to obtain the final declarative results with their legitimacy decided. After calculating the accuracy of all the above-mentioned models, we selected the top-performing models which are sigmoid SVC, AdaBoost, and Decision tree, and combined them into a single voting classifier.For diabetes prediction after calculating all model accuracies we created a voting classifier with the following weights follows KNN - 2, Decision Tree – 1, Logistic Regression – 2, and Random Forest – 2.

4. Implementation Details& Results

The proposed method is implemented by the use of the environment provided by Google Colab and key Python libraries. The various ways in which data has been analyzed are listed as follows: Plots of Attributes for Detailed Analysis, Voting Classifier: Pair, Scatter Plots, KDE/Target Visualization Plot, Correlation Matrix, Box/Pair Plotting, Plotting Comparison of Models Before & After Standardization. The various models in which data has been analyzed are listed as follows: Ensemble Classifiers, XGBoost, Decision Tree, Logistic Regression, KNN, Random Forest, AdaBoost, Sigmoid SVC, Polynomial SVC, RBF SVC & Linear SVC.

Before beginning EDA, we double-checked the dataset's dimension and variable data types, as well as looked for null values. Diabetes affects about a third of the persons in the dataset, and this division of the 'Outcome' will help our algorithms forecast more accurately for both classes (1 and 0). Co-relational Matrix is plotted to compute relation between the features used for prediction. Co-relation between the features is directly proportional to the accuracy of prediction. The co-relation between the features is improved by scaling all the feature in a specific range of values.

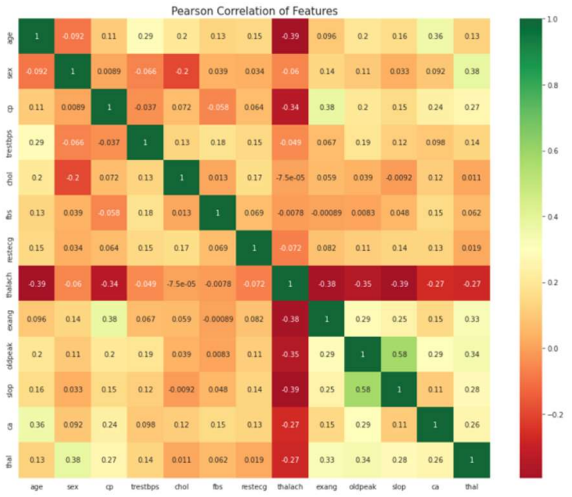


Figure 3: Heart Correlation Matrix

Out of the raw data of 303 patients, 6 had null values in either ca or thal; for this being so few instances of missing data, those data points were just dropped, making a total of 297 data points. The original labels, ranging from 0, no heart disease, to 4, the most advanced stage of heart disease, was redesigned to range from 0, no heart disease, and the original values of 1, 2, 3, and 4 were squished into a single category, 1, "presence of heart disease." In these figures, we are extracting data from the dataset using dat.info so that the dataset can be studied in an analytical manner.

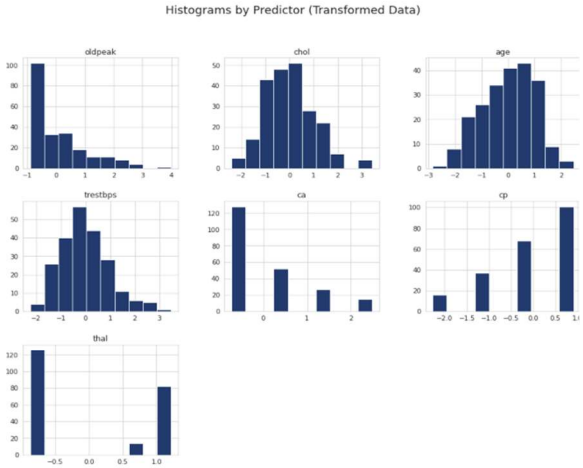


Figure 4: Heart Histogram Attribute Analysis

After selecting the top 7 features for prediction using the random forest classifier. From correlation matrix (Figure 3) and the histogram (Figure 4), we can observe that the features are not correlated to each other. Hence, we scaled all the values between -2 to 2. After data preprocessing the accuracies and recall of most of the algorithm has increased significantly.

Table 1: Model Accuracies for Heart (Before Parameter Tuning)

Algorithm	Accuracy	Recall
Random Forest	84.00%	0.8571
Logistic Regression	85.33%	0.8857
XGBoost	80.00%	0.8000
Decision Tree	74.67%	0.7714
AdaBoost	76.00%	0.8285
KNN	84.00%	0.8285
SVC	81.33%	0.8571

Table 2: Model Accuracies for Heart (After Parameter Tuning)

Algorithm	Accuracy	Recall
Random Forest	84.00%	0.8571
Logistic Regression	85.33%	0.8857
Decision Tree	85.33%	0.8571
AdaBoost	84.00%	0.8285
KNN	84.00%	0.8285
Linear SVC	85.33%	0.8571
RBF SVC	82.67%	0.8857
Sigmoid SVC	88.00%	0.8571
Polynomial SVC	81.33%	0.8571

Comparing Table 1 and Table 2, most models are seeing a lot of improvement all models are up on recall score; the KNN model's accuracy increased from 59% to 84%; the SVC's accuracy increased from 55% to 81% with the exception of XGBoost, which saw a 4% decrease.We have selected the three top-performing models: the sigmoid kernel from the SVCs, AdaBoost from the boosters, and the decision tree over the random forest, for a total of 3 individual models. We combined them into a single voting classifier below. The accuracy which we achieved was 88.57 %. This ensemble method does as well as two of the individual models, the decision tree and AdaBoost, but not as well the sigmoid SVC

To demonstrate the link between different variables, we produced a correlation matrix shown as Figure 5. This graph enables us to see which characteristics have a high correlation with others and hence eliminate them from the model; however, none of the variables had a high correlation with any of the others.The dataset has 8 features which are Pregnancies, Glucose, blood pressure, skin thickness, Insulin, BMI, Diabetes Pedigree Function, Age.

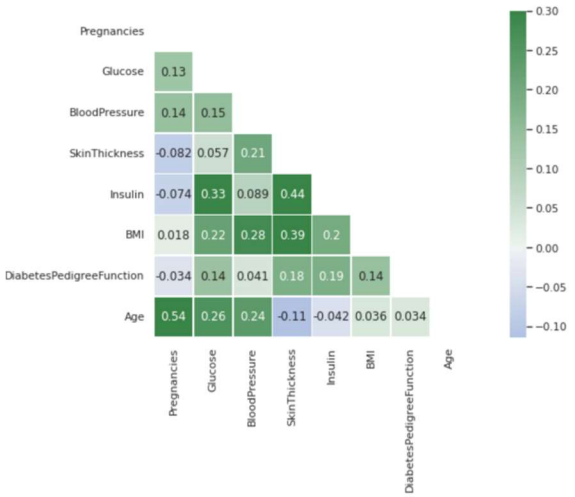


Figure 5: Diabetes Co-relational Matrix

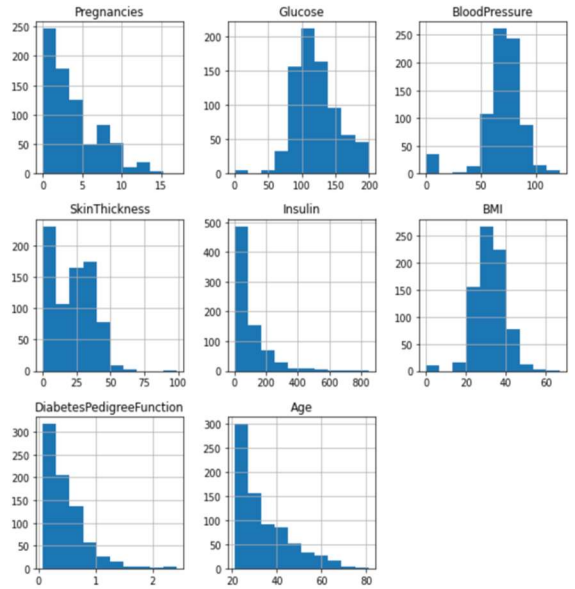


Figure 6: Diabetes Histogram Attribute Analysis

We have developed prediction models using the following classifiers Random Forest, Adaptive Boosting, KNN, Logistic Regression and Decision Tree.

Table 3: Model Accuracies for Diabetic Prediction

Model	Accuracy
Random Forest	72.73%
KNN	77.93%
Logistic Regression	77.92%
Decision Tree	74.46%
AdaBoost	72.73%
VC (without weights)	80.52%
VC (with weights)	80.95%

In Table 3, the two ensemble models used are Adaptive Booster (Boost) & Random Forest (Mean)

- as the foundation. By the combination of different classifiers using Voting Classifier with and without weights, precision can be enhanced. The precision boosts to 80.95% and 80.52% for voting classifiers with or without weights respectively. The same weights were applied on other foundations with the hefty weights delegated for the effective models. The weights used in voting classifier are as follows KNN - 2, Decision Tree – 1, Logistic Regression – 2 and Random Forest – 2.

5. Conclusion

The heart disease and diabetes can be synonymously aggregated to drive a patient's conclusions. In this paper sufficient exploratory analysis and pre-analysis of normalized models has been carried out to understand the need of these predictions using ensemble technique. The system promises to handle and correlate both events of heart and diabetes to drive to quicker prediction using machine learning concepts .For heart disease, it can be concluded that the Voting Classifier of Decision Tree, Sigmoid SVC, and Adaboost has the highest accuracy of 88.57 % and for diabetes, the voting classifier has an accuracy of 80.95 %. The proposed methodology can be extended since it has a scope to conclude immunity of a patient from COVID through the study conducted. The development of a robust model with the help of automated feature selection to work on possibility of COVID through the analysis of both diseases can be carried out in future.

References

[1]Bianca de Almeida-Pititto, Patrícia M. Dualib, Lenita Zajdenverg, JoanaRodrigues Dantas,Filipe Dias de Souza , Melanie Rodacki and Marcello Casaccia Bertoluci on behalf of Brazilian Diabetes Society Study Group (SBD), “Severity And Mortality Of COVID 19”, In Patients With Diabetes, Hypertension And CardiovascularDisease: A Meta-analysis, Diabetology & Metabolic Syndrome Research, (2020)

[2] R. Indrakumari, T. Poongodi, Soumya Ranjan Jena, “Heart Disease Prediction using Exploratory Data Analysis, International Conference” on Smart Sustainable Intelligent Computingand Applications under (ICITETM 2020)

[3] Senthilkumar Mohan 1, Chandrasegar Thirumalai1, And Gautam Srivastava 2,3, (Member,IEEE), “Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques”,(IEEE ACCESS)

[4] Aishwarya Majumdar, Dr. Vaidehi , “Diabetes Prediction using Machine Learning Algorithms, International Conference” On Recent Trends In Advanced Computing (2019, ICRTAC2019)

[5] Emma Barron, Chirag Bakhai, Partha Kar, Andy Weaver, Dominique Bradley, Hassan Ismail, Peter Knighton, Naomi Holman, Kamlesh Khunti, Naveed Sattar, Nicholas J Wareham,Bob Young, Jonathan Valabhji, Associations of type 1 and type 2 diabetes with COVID-19 related mortality in England: a whole-population study, (Lancet Diabetes Endocrinol 2020)

[6] Quan Zou, Kaiyang Qu , Yamei Luo, Dehui Yin, Ying Ju,Hua Tang, “Predicting Diabetes Mellitus With Machine Learning Techniques,” Bioinformatics and Computational Biology, (Frontier Genetics Journal, 2018)

[7] Md. Kamrul Hasan , Md. Ashraful Alam , Dola Das, Eklas Hossain, (Senior Member, IEEE), And Mahmudul Hasan, “Diabetes Prediction Using Ensembling of Different MachineLearning Classifiers”, (IEEE ACCESS 2020)

[8] Jyoti Soni, Ujma Ansari, Dipesh Sharma, Sunita Soni, Predictive Daa Mining for Medical Diagnosis”, International Journal of Computer Applications, (Volume 17, #8,2011)

[9] Himanshu Sharma, MA Rizvi, “Prediction of Heart Disease using Machine Learning Algorithms: A Survey” ,International Journal on Recent and Innovation Trends in Computing and Communication (2016)

[10] Baban U. Rindhe, Nikita Ahire, Rupali Patil, Shweta Gagare, Manisha Darade, “Heart Disease Prediction Using Machine Learning”, IJARSCT, (2021)

[11] Nabaouia Louridi, Samira Douzi, Bouabid El Ouahidi, “Machine Learning-Based Identifcation Of Patients With A Cardiovascular Defect”, Journal of Big Data (2021)

[12] Dhai Eddine Salhi, Abdelkamel Tari, M-Tahar Kechadi, “Using Machine learning for heart disease prediction”, researchgate.net.

[13] Minakshi R. Rajput ,Sushant S. Khedgikar, “Diabetes prediction and analysis using medical attributes: A machine learning approach”, Journal of Xian University of Architecture and Technology.

A Review on Heart Disease Prediction using Machine Learning and Data Analytics Approach

M. Marimuthu
Assistant Professor
Coimbatore Institute of Technology
Coimbatore

M. Abinaya
UG Scholar
Coimbatore Institute of Technology
Coimbatore

K. S. Hariesh
UG Scholar
Coimbatore Institute of Technology
Coimbatore

K. Madhankumar
UG Scholar
Coimbatore Institute of Technology
Coimbatore

V. Pavithra
UG Scholar
Coimbatore Institute of Technology
Coimbatore

ABSTRACT
Heart is the next major organ comparing to brain which has more priority in Human body. It pumps the blood and supplies to all organs of the whole body. Prediction of occurrences of heart diseases in medical field is significant work. Data analytics is useful for prediction from more information and it helps medical centre to predict of various disease. Huge amount of patient related data is maintained on monthly basis. The stored data can be useful for source of predicting the occurrence of future disease. Some of the data mining and machine learning techniques are used to predict the heart disease, such as Artificial Neural Network (ANN), Decision tree, Fuzzy Logic, K-Nearest Neighbour(KNN), Naïve Bayes and Support Vector Machine (SVM). This paper provides an insight of the existing algorithm and it gives an overall summary of the existing work.

Keywords
Data mining, Heart disease, Machine learning, Medical centre.

1. INTRODUCTION
Heart disease is one of the prevalent disease that can lead to reduce the lifespan of human beings nowadays. Each year 17.5 million people are dying due to heart disease [1]. Life is dependent on component functioning of heart, because heart is necessary part of our body. Heart disease is a disease that affects on the function of heart [2]. An estimate of a person's risk for coronary heart disease is important for many aspects of health promotion and clinical medicine. A risk prediction model may be obtained through multivariate regression analysis of a longitudinal study [3]. Due to digital technologies are rapidly growing, healthcare centres store huge amount of data in their database that is very complex and challenging to analysis. Data mining techniques and machine learning algorithms play vital roles in analysis of different data in medical centres. The techniques and algorithms can be directly used on a dataset for creating some models or to draw vital conclusions, and inferences from the dataset. Common attributes used for heart disease are Age, Sex, Fasting Blood Pressure, Chest Pain type, Resting ECG(test that measures the electrical activity of the heart), Number of major vessels colored by fluoroscopy, Threst Blood Pressure (high blood pressure), Serum Cholestrol (determine the risk for developing heart disease), Thalach (maximum heart rate achieved), ST depression (finding on an electrocardiogram, trace in the ST segment is abnormally low below the baseline), painloc (chest pain location (substernal=1, otherwise=0)), Fasting blood sugar, Exang (exercise included angina), smoke, Hypertension, Food habits, weight, height and obesity[4]. Table 1 summarizes the most common types of the heart disease as follows.

Table 1 Different types of heart disease [5]

Arrhythmia	The heart beat is improper whether it may irregular, too slow or too fast.
Cardiac arrest	An unexpected loss of heart function, consciousness and breathing occur suddenly.
Congestive heart failure	The heart does not pump blood as well as it should, it is the condition of chronic.
Congenital heart disease	The heart's abnormality which develops before birth.
Coronary artery disease	The heart's major blood vessels can damage or any disease occurs in the blood vessels.
High Blood Pressure	It has a condition that the force of the blood against the artery walls is too high.
Peripheral artery disease	The narrowed blood vessels which reduce flow of blood in the limbs, is the circulatory condition.
Stroke	Interruption of blood supply occur damage to the brain.

Figure 1 depicts the parts of human heart such as Left atrium, Right atrium, Right ventricle, Left ventricle, Aorta, pulmonary vein, Pulmonary valve, Pulmonary artery, Tricuspid valve, Aortic valve, Mitral valve, Superior vena cava and Interior vena cava.

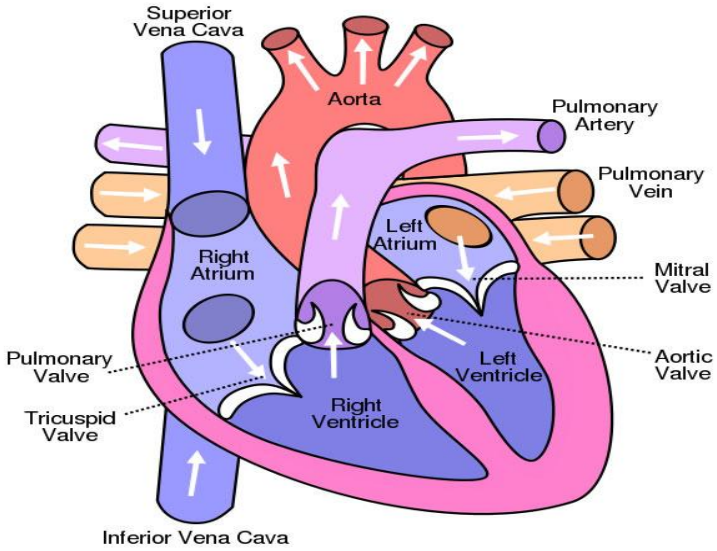


Figure 1 Human Heart [6]

This paper is organized as follows. Section 2 gives an overall literature review of the existing work. Section 3 provides a conclusion and future work.

2. LITERATURE REVIEW

There are numerous works has been done related to disease prediction systems using different data mining techniques and machine learning algorithms in medical centres.

K. Polaraju et al, [7] proposed Prediction of Heart Disease using Multiple Regression Model and it proves that Multiple Linear Regression is appropriate for predicting heart disease chance. The work is performed using training data set consists of 3000 instances with 13 different attributes which has mentioned earlier. The data set is divided into two parts that is 70% of the data are used for training and 30% used for testing. Based on the results, it is clear that the classification accuracy of Regression algorithm is better compared to other algorithms.

Marjia et al, [8] developed heart disease prediction using KStar, j48, SMO, and Bayes Net and Multilayer perception using WEKA software. Based on performance from different factor SMO and Bayes Net achieve optimum performance than KStar, Multilayer perception and J48 techniques using k-fold cross validation. The accuracy performances achieved by those algorithms are still not satisfactory. Therefore, the accuracy's performance is improved more to give better decision to diagnosis disease.

S. Seema et al,[9] focuses on techniques that can predict chronic disease by mining the data containing in historical health records using Naïve Bayes, Decision tree, Support Vector Machine(SVM) and Artificial Neural Network(ANN). A comparative study is performed on classifiers to measure the better performance on an accurate rate. From this experiment, SVM gives highest accuracy rate, whereas for diabetes Naïve Bayes gives the highest accuracy.

Ashok Kumar Dwivedi et al, [10] recommended different algorithms like Naive Bayes, Classification Tree, KNN, Logistic Regression, SVM and ANN. The Logistic Regression gives better accuracy compared to other algorithms.

MeghaShahi et al, [11] suggested Heart Disease Prediction System using Data Mining Techniques. WEKA software used for automatic diagnosis of disease and to give qualities of services in healthcare centres. The paper used various algorithms like SVM, Naïve Bayes, Association rule, KNN, ANN, and Decision Tree. The paper recommended SVM is effective and provides more accuracy as compared with other data mining algorithms.

Chala Beyene et al, [12] recommended Prediction and Analysis the occurrence of Heart Disease Using Data Mining Techniques. The main objective is to predict the occurrence of heart disease for early automatic diagnosis of the disease within result in short time. The proposed methodology is also critical in healthcare organisation with experts that have no more knowledge and skill. It uses different medical attributes such as blood sugar and heart rate, age, sex are some of the attributes are included to identify if the person has heart disease or not. Analyses of dataset are computed using WEKA software.

R. Sharmila et al, [13] proposed to use non- linear classification algorithm for heart disease prediction. It is proposed to use bigdata tools such as Hadoop Distributed File System (HDFS), Mapreduce along with SVM for prediction of heart disease with optimized attribute set. This work made an investigation on the use of different data mining techniques for predicting heart diseases. It suggests to use HDFS for storing large data in different nodes and executing the prediction algorithm using SVM in more than one node simultaneously using SVM. SVM is used in parallel fashion which yielded better computation time than sequential SVM.

Jayami Patel et al, [14] suggested heart disease prediction using data mining and machine learning algorithm. The goal of this study is to extract hidden patterns by applying data mining techniques. The best algorithm J48 based on UCI data has the highest accuracy rate compared to LMT.

Purushottam et al, [15] proposed an efficient heart disease prediction system using data mining. This system helps medical practitioner to make effective decision making based on the certain parameter. By testing and training phase a

certain parameter, it provides 86.3% accuracy in testing phase and 87.3% in training phase.

K.Gomathi et al, [16] suggested multi disease prediction using data mining techniques.Nowadays, data mining plays vital role in predicting multiple disease. By using data mining techniques the number of tests can be reduced. This paper mainly concentrates on predicting the heart disease, diabetes and breast cancer etc.,

P.Sai Chandrasekhar Reddy et al, [17] proposed Heart disease prediction using ANN algorithm in data mining. Due to increasing expenses of heart disease diagnosis disease, there was a need to develop new system which can predict heart disease. Prediction model is used to predict the condition of the patient after evaluation on the basis of various parameters like heart beat rate, blood pressure, cholesterol etc. The accuracy of the system is proved in java.

Ashwini shetty et al, [18] recommended to develop the prediction system which will diagnosis the heart disease from patient's medical dataset. 13 risk factors of input attributes have taken into account to build the system. After analysis of the data from the dataset, data cleaning and data integration was performed.

Jaymin Patel et al, [19] suggested data mining techniques and machine learning to predict heart disease. There are two objectives to predict the heart system. 1. This system not assume any knowledge in prior about the patient's records. 2. The system which chosen must be scalar to run against the large number of records. This system can be implemented using WEKA software. For testing, the classification tools and explorer mode of WEKA are used.

Boshra Brahmi et al, [20] developed different data mining techniques to evaluate the prediction and diagnosis of heart disease. The main objective is to evaluate the different classification techniques such as J48, Decision Tree, KNN, SMO and Naïve Bayes. After this, evaluating some performance in measures of accuracy, precision, sensitivity, specificity are evaluated and compared. J48 and decision tree gives the best technique for heart disease prediction.

Noura Ajam [21] recommended artificial neural network for heart disease diagnosis. Based on their ability, Feed forward Back propogation learning algorithms have used to test the model. By considering appropriate function, classification accuracy reached to 88% and 20 neurons in hidden layer. ANN shows result significantly for heart disease prediction.

Prajakta Ghadge et al, [22] suggested big data for heart attack prediction. The objective of this paper is to provide prototype using big data and data modelling techniques. It can be also used to extract patterns and relationships from database which associated with heart disease. This system consists of two databases namely, original big dataset and another is updated one. A java-file system named HDFS used to provide a user with reliable. This system can assist the healthcare practitioners to make intelligent decisions. The automation in this system would be advantageous.

S.Prabhavathi et al, [23] proposed Decision tree based Neural Fuzzy System (DNFS) technique to analyse and predict of various heart disease. This paper reviews the research on heart disease diagnosis. DNFS stand for Decision tree based Neural Fuzzy System. This research is to create an intelligent and cost effective system, and also to improve the performance of the existing system. Specifically in this paper, data mining techniques are used to enhance heart disease prediction. The result of this research shows that the SVM and neural networks results highly positive manner to predict heart disease. Still the data mining techniques are not encouraging for heart disease prediction.

Sairabi H.Mujawar et al, [24] used k-means and naïve bayes to predict heart disease. This paper is to build the system using historical heart database that gives diagnosis. 13 attributes have considered for building the system. To extract knowledge from database, data mining techniques such as clustering, classification methods can be used. 13 attributes with total of 300 records were used from the Cleveland Heart Database. This model is to predict whether the patient have heart disease or not based on the values of 13 attributes.

Sharan Monica.L et al[25] proposed an analysis of cardiovascular disease. This paper proposed data mining techniques to predict the disease. It is intend to provide the survey of current techniques to extract information from dataset and it will useful for healthcare practitioners. The performance can be obtained based on the time taken to build the decision tree for the system. The primary objective is to predict the disease with less number of attributes.

Sharma Purushottam et al, [26] proposed c45 rules and partial tree technique to predict heart disease. This paper can discover set of rules to predict the risk levels of patients based on given parameter about their health. The performance can be calculated in measures of accuracy classification, error classification, rules generated and the results. Then comparison has done using C4.5 and partial tree. The result shows that there is potential prediction and more efficient. Table 2 describes the accuracy of the heart disease with different techniques are shown below.

Table 2 A comparative study of various algorithms in literature review.

YEAR	AUTHOR	PURPOSE	TECHNIQUES USED	ACCURACY
2015	Sharma Purushottam et al,[15]	Efficient Heart Disease Prediction System using Decision Tree.	Decision tree classifier	86.3% for testing phase. 87.3% for training phase.
2015	Boshra Brahmi et al, [20]	Prediction and Diagnosis of Heart Disease by Data Mining Techniques.	J48, Naïve Bayes, KNN, SMO	J48 gives better accuracy than other three techniques.
2015	Sairabi H. Mujawar et al, [24]	Prediction of Heart Disease using Modified K-means and by using	Modified k-means algorithm, naive bayes algorithm.	Heart Disease detection=93%. Heart Disease

		Naïve Bayes.		undetection=89%.
2015	Noura Ajam et al, [21]	Heart Disease Diagnoses using Artificial Neural Network.	ANN	88%
2015	Sharma Purushottam et al, [26]	Heart Disease Prediction System Evaluation using C4.5 Rules and Partial Tree.	C4.5 rules and Naive Bayes algorithm	C4.5 gives better accuracy than Naive Bayes.
2016	Marjia et al, [8]	Prediction of Heart Disease using WEKA tool.	K Star	75%
			J48	86%
			SMO	89%
			Bayes Net	87%
			Multilayer Perception	86%
2016	S. Seema et al, [9]	Chronic Disease Prediction by mining the data.	Naïve Bayes	Highest accuracy achieved by SVM, in case of heart disease 95.56%
			Decision Tree	Highest accuracy of 73.588% achieved by Naïve Bayes in case of diabetes.
			Support Vector Machine	
2016	Ashok Kumar Dwivedi et al[10]	Evaluate the performance of different machine learning techniques for heart disease prediction.	Naïve Bayes	83%
			KNN	80%
			Logistic Regression	85%
			Classification Tree	77%
2016	K. Gomathi et al,[16]	Multi Disease Prediction using Data Mining Techniques.	Naïve Bayes	Heart Disease: 79% Diabetes: 77.6% Breast Cancer: 82.5%
			J48	Heart Disease: 77% Diabetes: 100% Breast Cancer: 75.5%
2016	Jayamin Patel et al, [19]	Heart Disease Prediction using Machine Learning and Data Mining Technique.	J48, Logistic model tree algorithm, Random forest algorithm	J48 gives 56.76% which is better than LMT algorithm of accuracy 55.75%.
2016	Ashwini Shetty A et al, [18]	Different Data Mining Approaches for Predicting Heart Disease.	WEKA tool, MATLAB. Neural Network	84%
			Hybrid Systems	89%
2016	Prajakta Ghadge et al, [22]	Intelligent Heart Disease Prediction System using	Hadoop, Mahout, Naïve bayes.	The automation of this system makes extremely

		Big Data.		advantageous.
2016	S. Prabhavathi et al, [23]	Analysis and Prediction of Various Heart Diseases using DNFS Techniques.	Decision tree, c4.5, SVM, naïve bayes.	Accuracy according to the types of heart disease. CVD Diagnosis= between 85% and 99%. CHD Diagnosis= between 82% and 92%.
2016	Sharan Monica. L et al,[25]	Analysis of CardioVasular Disease Prediction using Data Mining Techniques.	J48	91.4%
			Naïve Bayes	88.5%
			Simple CART	92.2%
2017	Jayami Patel et al,[14]	Heart disease Prediction using Machine Learning and Data mining Technique.	LMT, UCI	UCI gives better accuracy, compared to LMT.
2017	P. Sai Chandrasekhar Reddy et al, [17]	Heart disease prediction using ANN algorithm in data mining.	ANN	Accuracy proved in JAVA.
2018	Chala Bayen et al,[12]	Prediction and Analysis the occurrence of Heart Disease using data mining techniques.	J48, Naïve Bayes, Support Vector Machine.	It gives short time result which helps to give quality of services and reduce cost to individuals.
2018	R. Sharmila et al, [13]	A conceptual method to enhance the prediction of heart diseases using the data techniques.	SVM in parallel fashion	SVM provides better and efficient accuracy of 85% and 82.35%. SVM in parallel fashion gives better accuracy than sequential SVM.

3. CONCLUSION AND FUTURE WORK

By using different types of data mining and machine learning techniques to predict the occurrence of heart disease have summarized. Determine the prediction performance of each algorithm and apply the proposed system for the area it needed. Use more relevant feature selection methods to improve the accurate performance of algorithms. There are several treatment methods for patient, if they once diagnosed with the particular form of heart disease. Data mining can be of very knowledge form such suitable dataset.

In conclusion, as identified through the literature survey, believe only a marginal success is achieved in the creation of predictive model for heart disease patients and hence there is a need for combinational and more complex models to increase the accuracy of the predicting the early onset of heart disease. With the more amount of data being fed into the database the system will be very intelligent.

There are many possible improvements that could be explored to improve the scalability and accuracy of this prediction system. Due to time limitation, the following research / work need to be performed for the future. Would like to make use of testing different discretization techniques, multiple classifier voting technique and different decision tree types namely information gain and gain ratio. Willing to explore different rules such as association rule, logistic regression and clustering algorithms.

4. REFERENCES

[1] Animesh Hazra, Arkomita Mukherjee, Amit Gupta, Asmita Mukherjee, "Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review", Research Gate Publications, July 2017, pp.2137-2159.

[2] V. Krishnaiah, G. Narsimha, N. Subhash Chandra, "Heart Disease Prediction System using Data Mining Techniques and Intelligent Fuzzy Approach: A Review", International Journal of Computer Applications, February 2016.

[3] Guizhou Hu, Martin M. Root, "Building Prediction Models for Coronary Heart Disease by Synthesizing Multiple Longitudinal Research Findings", European Science of Cardiology, 10 May 2005.

[4] T.Mythili, Dev Mukherji, Nikita Padaila and Abhiram Naidu, "A Heart Disease Prediction Model using SVM-Decision Trees- Logistic Regression (SDL)", International Journal of Computer Applications, vol. 68, 16 April 2013.

[5] https://www.medicalnewstoday.com/articles/257484.php.
[6] Nimai Chand Das Adhikari, Arpana Alka, and rajat Garg, "HPPS: Heart Problem Prediction System using Machine Learning".
[7] K. Polaraju, D. Durga Prasad, "Prediction of Heart Disease using Multiple Linear Regression Model", International Journal of Engineering Development and Research Development, ISSN:2321-9939, 2017.
[8] Marjia Sultana, Afrin Haider, "Heart Disease Prediction using WEKA tool and 10-Fold cross-validation", The Institute of Electrical and Electronics Engineers, March 2017.
[9] Dr.S.Seema Shedole, Kumari Deepika, "Predictive analytics to prevent and control chronic disease", https://www.researchgate.net/punlication/316530782, January 2016.
[10] Ashok kumar Dwivedi, "Evaluate the performance of different machine learning techniques for prediction of heart disease using ten-fold cross-validation", Springer, 17 September 2016.
[11] Megha Shahi, R. Kaur Gurm, "Heart Disease Prediction System using Data Mining Techniques", Orient J. Computer Science Technology, vol.6 2017, pp.457-466.
[12] Mr. Chala Beyene, Prof. Pooja Kamat, "Survey on Prediction and Analysis the Occurrence of Heart Disease Using Data Mining Techniques", International Journal of Pure and Applied Mathematics, 2018.
[13] R. Sharmila, S. Chellammal, "A conceptual method to enhance the prediction of heart diseases using the data techniques", International Journal of Computer Science and Engineering, May 2018.
[14] Jayami Patel, Prof. Tejal Upadhay, Dr. Samir Patel, "Heart disease Prediction using Machine Learning and Data mining Technique", March 2017.
[15] Purushottam, Prof. (Dr.) Kanak Saxena, Richa Sharma, "Efficient Heart Disease Prediction System", 2016, pp.962-969.
[16] K.Gomathi, Dr.D.Shanmuga Priyaa, "Multi Disease Prediction using Data Mining Techniques", International Journal of System and Software Engineering, December 2016, pp.12-14.
[17] Mr.P.Sai Chandrasekhar Reddy, Mr.Puneet Palagi, S.Jaya, "Heart Disease Prediction using ANN Algorithm in Data Mining", International Journal of Computer Science and Mobile Computing, April 2017, pp.168-172.
[18] Ashwini Shetty A, Chandra Naik, "Different Data Mining Approaches for Predicting Heart Disease", International Journal of Innovative in Science Engineering and Technology, Vol.5, May 2016, pp.277-281.
[19] Jaymin Patel, Prof. Tejal Upadhyay, Dr.Samir Patel, "Heart Disease Prediction using Machine Learning and Data Mining Technique", International Journal of Computer Science and Communication, September 2015-March 2016, pp.129-137.
[20] Boshra Brahmi, Mirsaeid Hosseini Shirvani, "Prediction and Diagnosis of Heart Disease by Data Mining Techniques", Journals of Multidisciplinary Engineering Science and Technology, vol.2, 2 February 2015, pp.164-168.
[21] Noura Ajam, "Heart Disease Diagnoses using Artificial Neural Network", The International Insitute of Science, Technology and Education, vol.5, No.4, 2015, pp.7-11.
[22] Prajakta Ghadge, Vrushali Girme, Kajal Kokane, Prajakta Deshmukh, "Intelligent Heart Disease Prediction System using Big Data", International Journal of Recent Research in Mathematics Computer Science and Information Technology, vol.2, October 2015 - March 2016, pp.73-77.
[23] S.Prabhavathi, D.M.Chitra, "Analysis and Prediction of Various Heart Diseases using DNFS Techniques", International Journal of Innovations in Scientific and Engineering Research, vol.2, 1, January 2016, pp.1-7.
[24] Sairabi H.Mujawar, P.R.Devale, "Prediction of Heart Disease using Modified K-means and by using Naïve Bayes", International Journal of Innovative research in Computer and Communication Engineering, vol.3, October 2015, pp.10265-10273.
[25] Sharan Monica.L, Sathees Kumar.B, "Analysis of CardioVasular Disease Prediction using Data Mining Techniques", International Journal of Modern Computer Science, vol.4, 1 February 2016, pp.55-58.
[26] Sharma Purushottam, Dr. Kanak Saxena, Richa Sharma, "Heart Disease Prediction System Evaluation using C4.5 Rules and Partial Tree", Springer, Computational Intelligence in Data Mining,vol.2, 2015, pp.285-294.