ML based Interactive Disease Prediction Model

D. Sharathchandra[#1], M. Raghu Ram[*2]

[#]*PG student, Dept. of EIE*, Kakatiya Institute of Technology & Science, Warangal, India
[*]*Assoc. Prof., Dept. of EIE,* Kakatiya Institute of Technology & Science, Warangal, India
[1]*sharathdevaram11@gmail.com*
[2]mrr.eie@kitsw.ac.in

***Abstract*-The application of Machine learning algorithms to predict diseases is one of the finest methodology to reduce heavy work load on doctors and related medical staff. Based on the World Health Organization (WHO) report, about 85% heart disease deaths are due to Heart Attacks and Heart Strokes. In India the average death rate due to cardiovascular diseases is about 272 per 10,000 population which is greater than global average of 235 per 10,000 population. From the recent survey results, which was released by the Union Ministry of Family and Health Welfare (MoFHW), the Diabetes disease positive ratio is gradually increasing in India. 11.5 percent people were tested positive for Diabetes among urban and rural Indians who are with age 45 and above. Even there is availability of wide range of treatment methods of heart stroke patients & diabetes, Heart attack with Diabetes is the major cause of death in all parts of rural and urban areas of entire India. There are several factors causing heart and diabetes problems which include Age, Gender, Blood Pressure, Glucose levels, Skin thickness and Insulin. These are easily measured in primary care facility centres. The accurate estimation and analysis of heart & diabetes disease patients reports data may help in predicting future heart problems including diabetes. Globally, the application of computerized machine learning methods to predict future problems is in trend now. The Health Monitoring Departments and Fields uses machine learning algorithms to predict and analyse in a wider way to solve problems in fraction of seconds. From the famous proverb "Prevention is Better Than Cure", if we apply this to medico and health field we can save people from major Heart Diseases (HD's) along with Diabetes. The proposed Dual disease prediction technique is user interactive based method. The proposed method observe inputs from the end user with realistic data to predict heart and diabetes disease. In the presented work, we used Logistic regression model (LR) and Support vector machine (SVM) model for prediction of diseases. The proposed model works with 85 and 78 percent accuracy in prediction of heart and diabetes diseases respectively.**

Keywords—Heart Disease, Diabetes, Machine Learning, Logistic Regression, Support Vector Machine

I. INTRODUCTION

In generally machine learning is subset of artificial intelligence (widely known as AI) and the AI is a special field in the computer science approaches. The main goal of machine learning is to reduce burden on programmers. Machine learning models works with data's which are provided by user or initial programmer and the machine analyses our input data set and produces our desired output. The term cardiovascular disease [1] is combination of blood circular system and heart disease.

978-1-6654-5883-2/22/$31.00©2022 IEEE

Unhealthy diet, heavy usage of tobacco, alcohol consumption, hyper tension, physical inactivity and glucose levels in blood are the major reasons for heart attacks and strokes. Most of these are easily measured in every primary health care centres across India. The accurate analysis of factors which are responsible for heart disease prevents future attacks and strokes. Diabetes disease is the main reason for heart stroke and attacks. The conversion failure from glucose to energy in blood cells results as extra unwanted sugar in blood system. Due to this the functioning of blood stream throughout the body may get disturbed.The application of machine learning algorithms [2] to predict future attacks and strokes reduces work load on medical staff. It is the new approach to detect disease with the help of machine learning algorithms. These methods will reduce time and increase faster disease identification. In the section II, details of proposed method is presented and in section III results obtained using proposed methodology is presented with conclusion in section IV.

II. PROPOSED METHOD

The proposed module entitled "User Interactive-Dual Disease Detection Using Machine Learning" have several stages to develop. The complete details of the work from collection of data sets to the end prediction model are explained in this unit. The block diagram of the project is shown in Fig. 1.

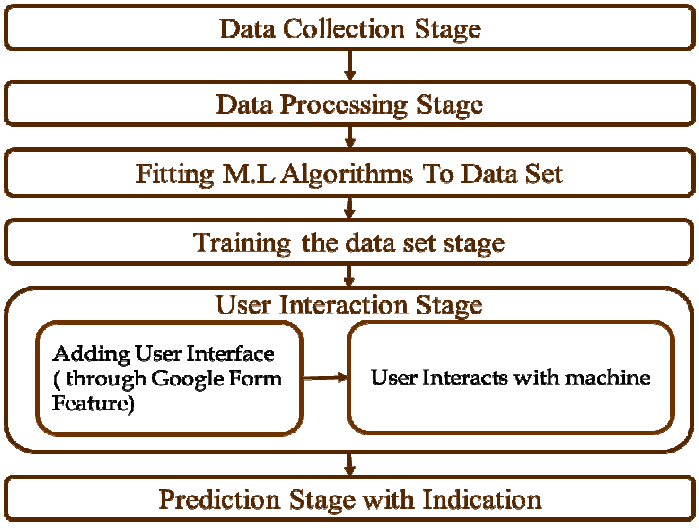
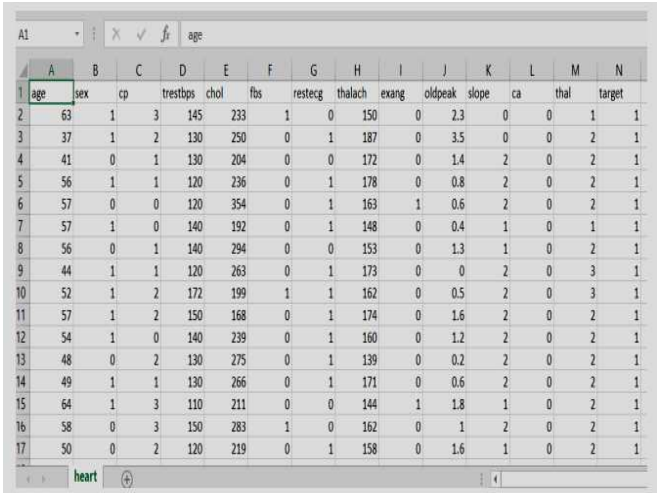


Fig.1 Block diagram of the proposed algorithm

(i) Data collection stage: The heart disease data and diabetes data are collected form UCI ML Repositoryand DataWorld webserver. These data sets are available in comma separated values format. We usedthese data sets andtrained machine

2022 IEEE Delhi Section Conference (DELCON) | 978-1-6654-5883-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/DELCON54057.2022.9752947

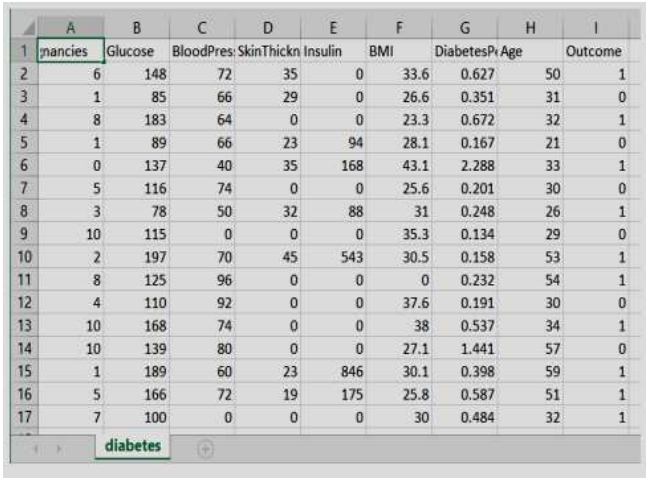
with machine learning algorithms to predict with new input data. This heart data set consists of 13 features and one output.The features of heart data set are Gender, Age, Chest pain type, Cholesterol, Fasting blood sugar levels, Electrocardiographic Results, Maximum heart rate achieved, Exercise Induced Angina, ST depression induced by exercise relative to rest, Slope of Peak exercise ST segment, Number of major vessels (0–3) coloured by fluoroscopy and finally Thalassemia Value. The output is classified as if it is 0 i.e. no heart disease, if it is 1, the person have heart disease. The pictorial view of data set is depicted in Fig. 2 and Fig. 3.



age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
44	1	1	120	263	0	1	173	0	0	2	0	3	1
52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
57	1	2	150	168	0	1	174	0	1.6	2	0	2	1
54	1	0	140	239	0	1	160	0	1.2	2	0	2	1
48	0	2	130	275	0	1	139	0	0.2	2	0	2	1
49	1	1	130	266	0	1	171	0	0.6	2	0	2	1
64	1	3	110	211	0	0	144	1	1.8	1	0	2	1
58	0	3	150	283	1	0	162	0	1	2	0	2	1
50	0	2	120	219	0	1	158	0	1.6	1	0	2	1

Fig.2 Heart data set details

The features of diabetes data set are Number of times Pregnant, Plasma glucose concentration in a 2h oral glucose tolerance test, Triceps skin fold thickness, Insulin (2h serum insulin), Body mass index, Diabetes Pedigree Function and finally Age. The 6 output is classified as 0 if the person without diabetes disease otherwise it is represented as 1.



nancies	Glucose	BloodPres	SkinThickn	Insulin	BMI	DiabetesP	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0
10	168	74	0	0	38	0.537	34	1
10	139	80	0	0	27.1	1.441	57	0
1	189	60	23	846	30.1	0.398	59	1
5	166	72	19	175	25.8	0.587	51	1
7	100	0	0	0	30	0.484	32	1

Fig.3 Diabetes data set details

(ii) Data processing stage: The collected data from the previous stage are processed in this stage. Python panda library is used for data processing and analysing. The misfiled data and unwanted data values are removed from data set. This will help in better fitting of algorithms to the data set with higher accuracy score. The details of data processing results are presented in Fig. 4 to Fig. 5.

```
# getting some info about the diabetes data
diabetes_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Pregnancies              768 non-null    int64
 1   Glucose                  768 non-null    int64
 2   BloodPressure            768 non-null    int64
 3   SkinThickness            768 non-null    int64
 4   Insulin                  768 non-null    int64
 5   BMI                      768 non-null    float64
 6   DiabetesPedigreeFunction 768 non-null    float64
 7   Age                      768 non-null    int64
 8   Outcome                  768 non-null    int64
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

```
# checking for missing values in heart data
heart_data.isnull().sum()
```

```
age         0
sex         0
cp          0
trestbps    0
chol        0
fbs         0
restecg     0
thalach     0
exang       0
oldpeak     0
slope       0
ca          0
thal        0
target      0
dtype: int64
```

Fig.4 Data processing of heart and diabetes data

(iii) Logistic regression algorithm: Logistic regression algorithm [4] is supervised regression algorithm. It predicts True or False or 1 or 0 or Yes or No type values. The logistic regression function works with S shaped sigmoid function. It predicts values in between 0 and 1. The mathematical representation of Logistic Regression is as follows:

The plane equation will be:

$$OUTPUT = ((SLOPE\ OF\ THE\ PLANE) * DATA\ INPUT) + INTERCEPT$$

The fitting logistic regression model to the heart and diabetes data sets is established writing python code using machine learning libraries. The function used for fitting algorithm is “LogisticRegression ()”.

Support vector machine **Error! Reference source not found.**[5][5]**Error! Reference source not found.** model algorithm: The algorithm Support Vector Machine is classification based algorithm. It can also use for regression methods. It uses kernel trick for prediction of the output. Support Vector Machines are indicated as SVM Model in this work. The mathematical representation is possible with hyperplane methods.

$$w.x + b = \pm 1 \ or \ w.x - b = \mp 1$$

Where X id data point, W is the slope of the plane, B is the intercept. The fitting support vector machine[3] model to the heart and diabetes data sets is established writing python code using machine learning libraries. The function used for fitting algorithm is

“svm.SVC (kernel='linear')”

(iv) Training the data sets: In this stage user can divide the datasets into train and test data sets in order to train data set for prediction. User has chance to divide data in their own ratio. The results of the work are presented in Fig.6 to Fig.14.

Fig.5 Training and test data- User choice

(v) User interaction stage: In this user interaction stage, user interacts with machine to provide particular details regarding heart[5]**Error! Reference source not found.Error! Reference source not found.Error! Reference source not found.** and diabetes diseases. In this stage we developed scroll bars to provide details. As the scroll bar position changes, the value also changes. The changed value is taken as input user detail. We developed a total of 20 scroll bars to interact with machine for providing inputs. In order to provide the details the machine has to present some indications. Based on these indications user provide their particulars through scroll bars. All these indications are provided in red colour.

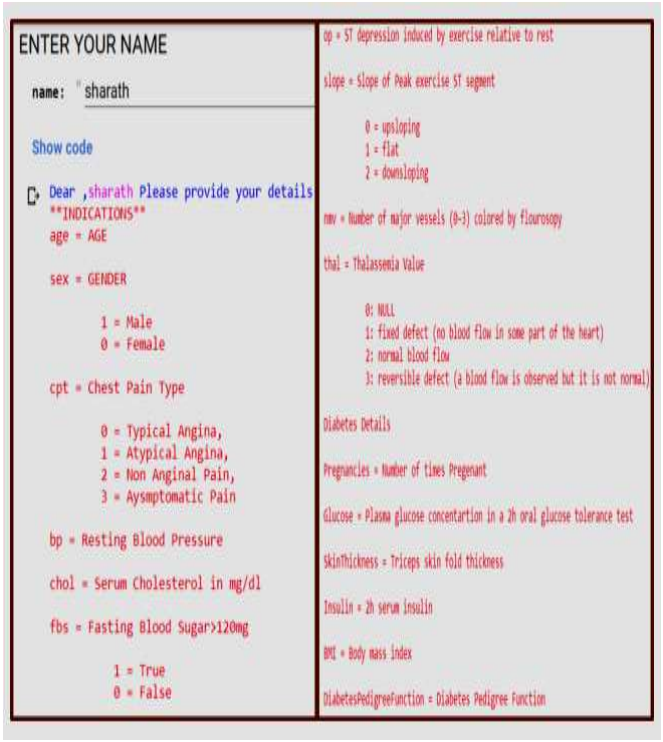


Fig.6 User indications (*Red coloured text*)

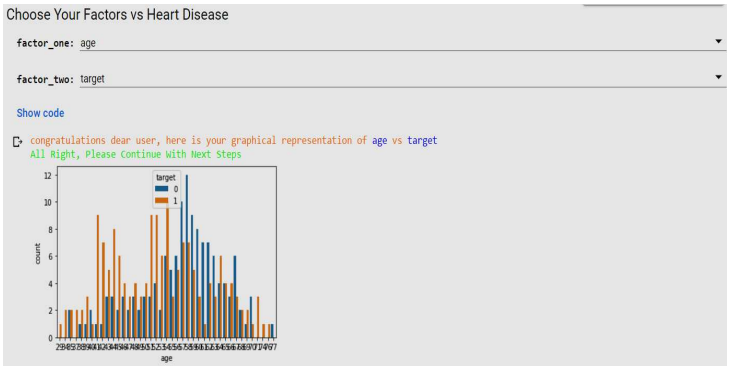


Fig.7 Graph representing age versus heart disease outcome with user interaction

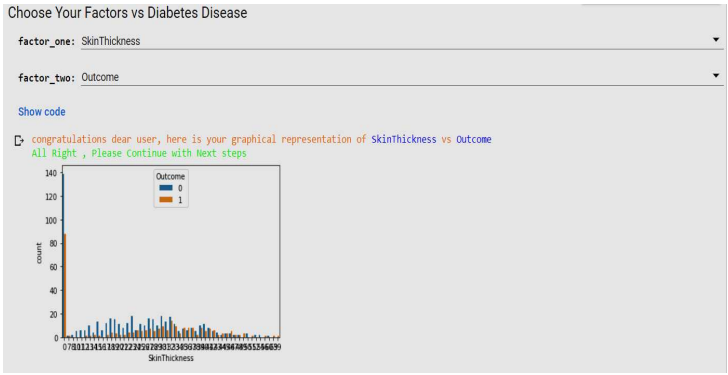


Fig.8 Graph representing skin thickness versus diabetes disease outcome with user interaction

The scroll bars are developed with the help of google form feature. These scroll bars are helpful in providing inputs by the end user. We used twenty scroll bars in order to provide details.

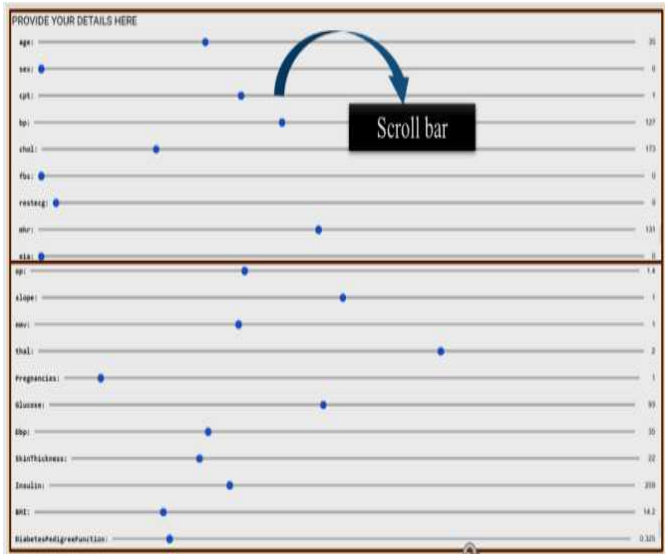


Fig.9 User interaction through scroll bars

(vi) Prediction stage: This is the final stage of this project. In this user has choice to predict is result with two algorithms. The first one is Logistic Regression Algorithm indicated as LR Model, the second one is Support Vector Machine Model indicated as SVM Model. User has a chance to select

any one from these two to predict heart disease and diabetes disease.

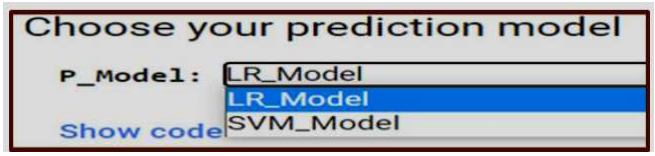


Fig.10 User selection for desired prediction model

III. RESULTS

In the presented work, the user has a chance of selecting prediction model between LR Model and SVM Model. It is developed with the help of google form drop down menu feature. This google form feature is similar to tkinter module in python to create a user interface.

(a) Predicted outputs: In this presented work, we predicted diseases with two models. The first one is Heart disease prediction and the second one is Diabetes disease prediction. So the results are presented in a combined manner. The heart and diabetes diseases are predicted with user selected model, and user provided inputs through scrollbars.

(b) Confusion matrix: The confusion matrix is also called as error matrix. It a matrix used to calculate the performance of algorithm. With the help of confusion matrix, we identify the true positive, true negative, false positive and false negative numbers.

Fig.11 Dual disease prediction as negative (green) with two models

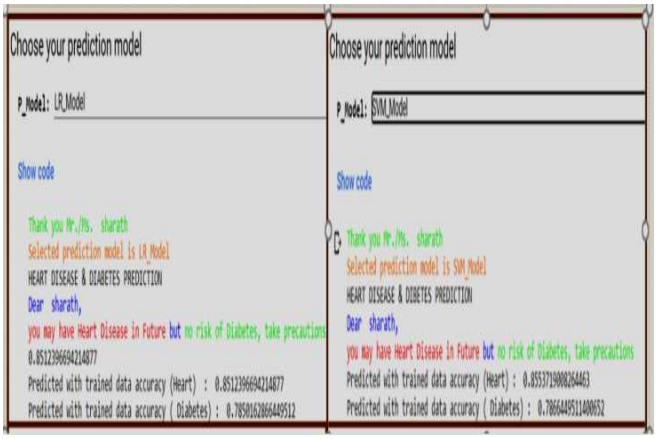


Fig.12 Heart disease as positive (red) and diabetes as negative prediction (green)

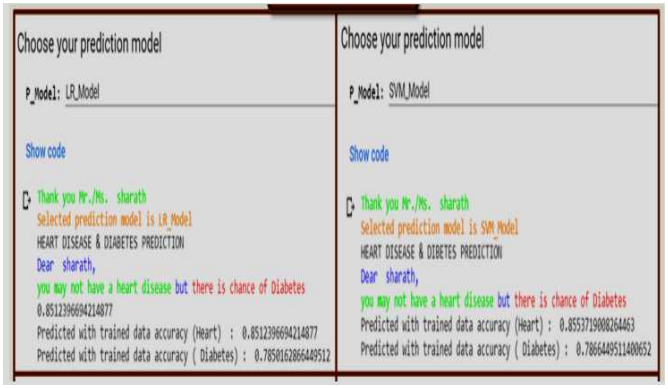


Fig.13 Heart disease as negative (green) and diabetes as positive prediction (red)

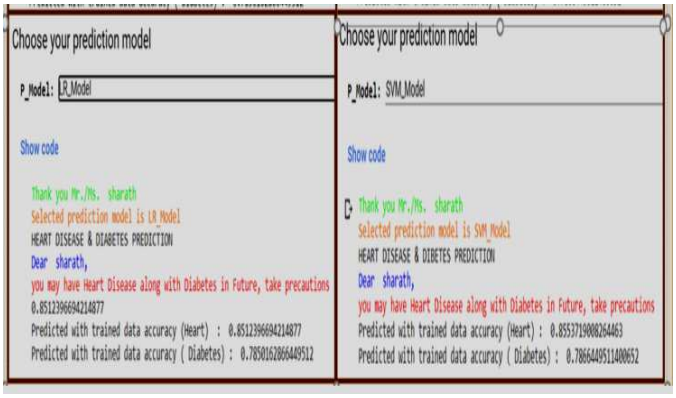


Fig.14 Dual disease prediction as positive (red) with two models

$$Confision\ Matrix = \begin{pmatrix} True\ Positive & False\ Positive \\ False\ Negative & True\ Negative \end{pmatrix}$$

Accuracy score: The accuracy for 2 models with two data sets calculated as accuracy. The accuracy score computations are presented in Table I.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Table I. Accuracy score

Type	LR MODEL	SVM MODEL
Heart Disease	0.8512396694214877	0.8533719008264463
Diabetes Disease	0.7850162866449512	0.7866449511400652

IV. CONCLUSION

The presented work on "User Interactive-Dual Disease Detection" is a user interface based project module. The end user will identify every detail about working of the project. With the help of this project user can predict their future disease status. The scroll bars provided are user convenient and easy to operate. User has a chance to shuffle between prediction models. The Logistic Regression Model and Support Vector Model works with better accuracy score. These are developed with limited data sets, we consider larger data sets the accuracy score will be improved. The Heart and Diabetes Diseases are predicted with 85 and 78 percent accuracy with LR and SVM models with minor changes in accuracy score.

V. FUTURE SCOPE

This work may extendable up to multiple disease detection with multiple user interfaces with multiple ML algorithms for faster disease prediction considering larger data set values

REFERENCES

[1] M. Saw *etal.*, "Estimation of prediction for getting heart disease using logistic regression model of machine learning," in proc. *IEEE Int. Conf. Comput. Commun. Inform,* Coimbatore, India, 2020, pp. 1-6.

[2] Sharma H and Rizvi M, "Prediction of heart disease using machine learning algorithms: A survey", *Int. J. on Recent Innov. Trends Comput. Commun.,* vol. 5, no. 8, pp. 99-104, Aug. 2017.

[3] Ajit S and Mehul P. Barot, "Study of heart disease diagnosis by comparing various classification algorithms", *Int. J. Eng. Adv. Technol.,* vol.8, no.2S2, Jan. 2019.

[4] Avinash G and Pavan Kumar T, "Heart disease prediction using effective machine learning techniques", *Int. J. Recent Technol. Eng.,* vol.8, no.184, Jun. 201

[5] Sarath Babu *etal.*, "Heart disease diagnosis using data mining technique," in proc. *IEEE Int. Conf. Electron. Commun. Aerospace Technol.* Coimbatore, India, 2017, pp. 1-4.

[6] Jaymin Patel *etal.*, "Heart disease prediction using machine learning and data mining technique," *Int. J. Comput. Sci. Commun.,* vol. 7, no. 1, pp. 129-137, Mar. 2016.

[7] L. Liu, "Research on logistic regression algorithm of breast cancer diagnose data by machine learning," in proc. *Int. Conf. Robots Intell. Syst.*, 2018, pp. 157-160.

[8] Nidhi Bhatia and Kiran Jyoti, "An analysis of heart disease prediction using different data mining techniques," *Int. J. Eng. Research Technol.*, vol. 1, no. 8, Oct. 2012.

[9] M. Anbarasi *etal.*, "Enhanced prediction of heart disease with feature subset selection using genetic algorithm," *Int. J. Eng. Sci. Technol.*, vol. 2, no. 10, pp. 5370-5376, 2010.

[10] N. Deepika, "Association rules for classification of heart attack patients", *Int. J. Adv. Eng. Sci. Technol.,* vol. 11, no. 2, pp. 253-257, 2011.