Disease Prediction And Analysis

Sneha Singhal , Shreyansh Tyagi, Shweta Rani , Mr. Ankit Verma(Associate Professor)

KIET Group of Institutions, Ghaziabad

Abstract

Nowadays , people are suffering by various diseases namely chest pain, blood pressure , blood sugar, heart failures, etc. As, elderly people stay alone at their homes when their children are out for work, so in case of emergency immediate medication is not provided. The early prediction of the disease leads to reduce in the chances of mishappening and increases the chances of survival of patients by taking appropriate actions. Human health is wealth, there is nothing more valuable than the good health. So, proper healthcare monitoring is must. In Rural areas, health monitoring is very challenging. Most of the people dies due to lack of medical facilities or because doctor is not available instantly. As a result, it's become more important than ever to predict and prevent heart disease . Data-driven solutions for predicting heart diseases can

improve the overall research and prevention process, allowing more people to live a healthy lifestyle. Machine learning comes into play at this point. Heart disease prognosis is aided by machine learning, and the forecasts are fairly accurate.

This project includes data processing and analysis of a heart disease patient's dataset. Different algorithms, such as KNN, Decision Tree, Logistic Regression, Random Forest, and others, were used to train the models and provide predictions. After disease prediction, the alert signal is provided so that patient can be attained. The proposed research was implemented on PYTHON applying different machine learning (ML) algorithms and showed the accuracy of different algorithms

used. Random Forest provides the result with highest i.e. 95% accuracy.

Keywords : *Machine Learning(ML), Disease Prediction, Random Forest, KNN, Decision Tree, Logistic regression*

1. Introduction

In recent years, as we are moving towards urbanization people are adopting inactive lifestyle, unhealthy way of eating which leads obesity and various heart related problems . In 2020, the World Health Organization analyzed in increase in heart diseases by fifty seven percent [1]. As heart diseases leads to crucial problems which emphasizes the need of disease prediction. The prediction of disease in early stage leads to reduce the risk. Doctors and nurses gives their best to save people's life but they are not present at all times. There is a lack of medical facility at remote villages, in such scenario machines can be helpful [2]. There are various Machine Learning models for prediction and analysis. Our primary goal is to find the model with highest accuracy among all the available models. In this project, various Machine Learning algorithms are being used which includes Random Forest, KNN, Decision Tree, etc. Machine Learning apply various optimization techniques on the past collected data and conclude the decision. Machine Learning models are also helpful for doctors and early prediction of the disease leads to reduction in financial pressure[5]. According to Yuan [6], to achieve accurate decision the data set used must be of good quality, only then it is possible to get unbiased results. To predict the results from a dataset or to conclude a decision is a challenging task, therefore Machine Learning algorithms are helpful as ML has advanced computational methodologies that are used to discover meaningful and hidden results from the dataset by proper training of dataset and testing of dataset and then fitting the dataset into the various machine learning models. The aim of this study is to test the efficiency of various machine learning algorithms and in the end best machine learning algorithm among all the algorithms used will be concluded.

Below is the workflow of the proposed work.

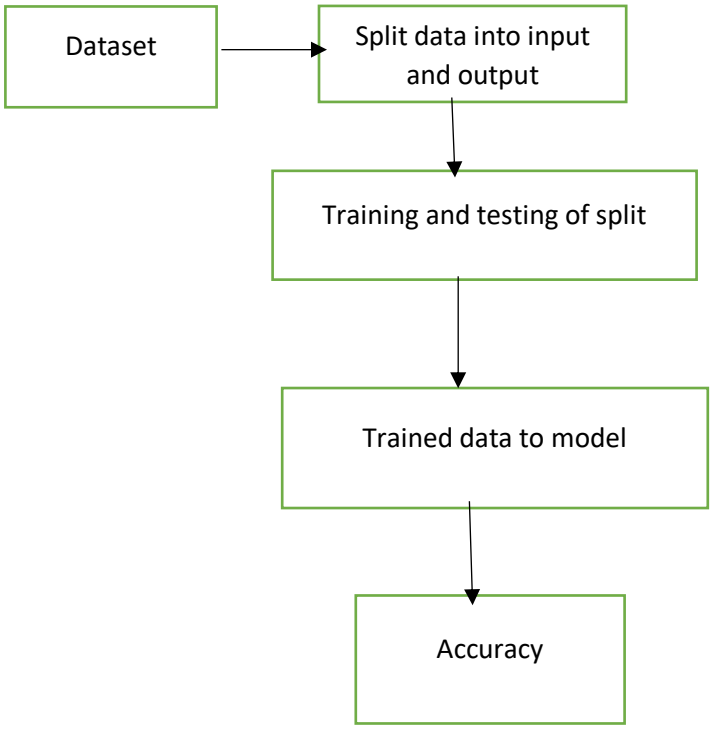

Fig. 1 : Workflow

2. Literature Review

Machine Learning (ML) is a field of study which is used in training the machine by providing various algorithms. Machine Learning helps in providing advanced methodologies which is helpful in healthcare industry by providing accurate results of disease predictions and analysis. There are three major types of Machine Learning : Supervised Machine Learning(SML), Unsupervised Machine Learning(UML) and Reinforcement Machine Learning. In this study , various machine learning algorithms are being used and their accuracy is checked by applying the algorithms on a dataset. Every model have different accuracy and it lies between 80-91%. It shows every model has accuracy greater than 80%. Only k-nearest neighbour has accuracy less than 80 percent i.e 67.21% . Random Forest has the highest accuracy. According to Sreevalli[7] , in predicting the diseases, Random Forest takes less time and less cost.

2.1 Dataset information

age denotes age of person

sex denotes whether person is male or female, 0 for female and 1 for male

cp denotes chest pain type, 0 for typical angina, 1 for atypical angina, 2 for non-anginal ,asymptomatic and 3 for pain

chol: serum cholesterol in milligrammes per deciliter

trestbps stands for resting blood pressure.

resting electrocardiographic findings (resecg) (values 0,1,2)

thalach: reached maximal heart rate

fbs denotes fasting blood sugar i.e greater than 120 mg/dl

oldpeak denotes ST depression caused by exercise when compared to rest.

Incline denotes the incline of the peak exercise portion ST

Exang : stands for exercise-induced angina.

ca denotes number of important vessels (0-3) coloured by flourosopy

thal : 3 for normal; 6 for fixed defect and 7 for reversible defect

2.2 Dataset Sample

	age	sex	cp	trestbps	chol	fbs	restecg
264	54	1	0	110	206	0	0
192	54	1	0	120	188	0	1
84	42	0	0	102	265	0	0
208	49	1	2	120	188	0	1
201	60	1	0	125	258	0	0

2.3 Analyzing features

2.3.1. Analyzing sex feature

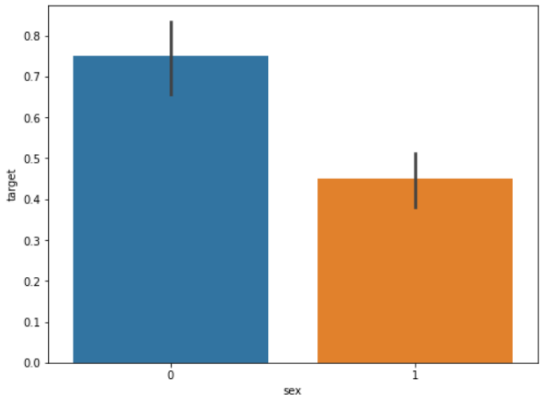

Fig. 2 Sex feature analysis graph

We notice that females has higher chances of heart diseases than males.

2.3.2. Analyzing chest pain feature

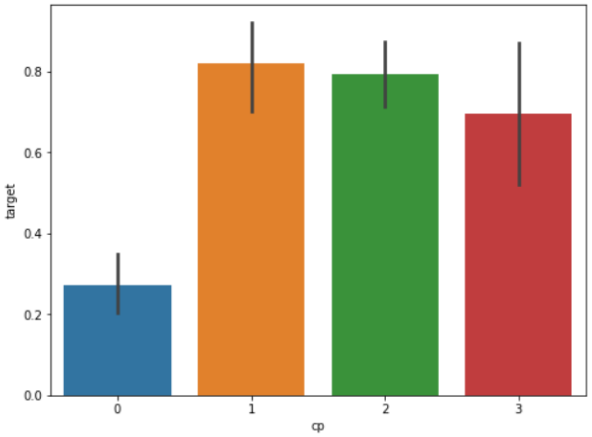

Fig 3. Chest pain analysis graph

Here, we notice that chest pain typical angina i.e type 0 has less chances of heart disease.

2.3.3 Analyzing fbs feature

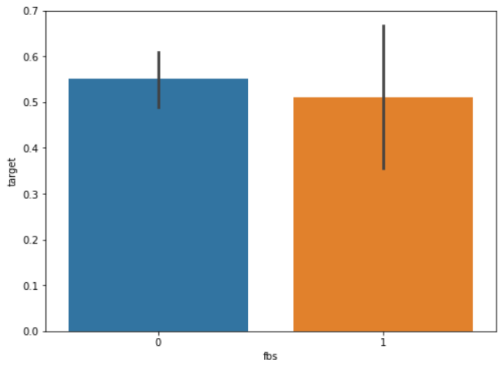

Fig 4 . fbs analysis graph

Nothing extraordinary here, this feature cannot make much difference.

2.3.4 Analyzing restecg feature

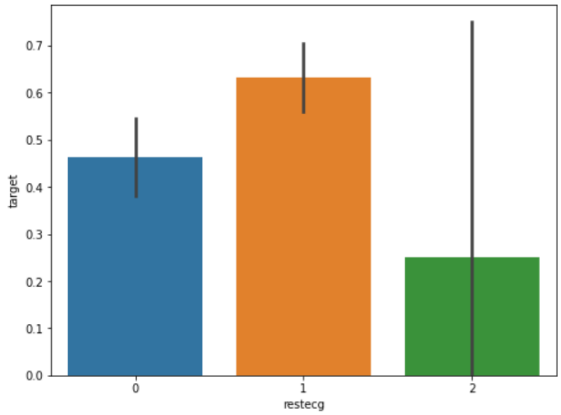

Fig 5. Restecg analysis graph

Restecg of type 2 has lesser risk to heart diseases.

2.3.5 Analyzing exang feature

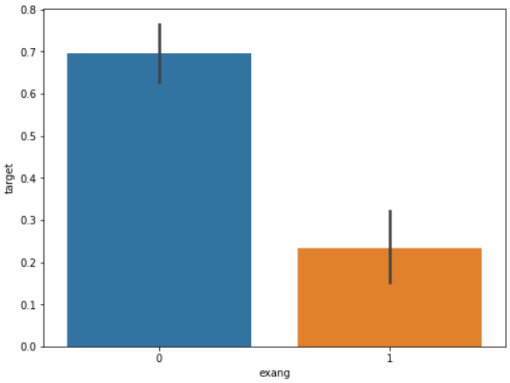

Fig 6 . exang analysis graph

Type 1 has lesser risk.

3. Methodologies

3.1 Logistic Regression (LR)

Logistic Regression(LR) is the supervised form of Machine Learning. It classifies records of dataset based on input values. It can be discrete or categorical. It is a probabilistic based statistical model. Logistic Regression work well when the dataset which is used is can be separated linearly[8].

Logistic Sigmoid Function (LSF):

P(r) = 1/(1+e^(-r))

3.1.1. Function used :

from sklearn.linear_model import LogisticRegression

```
logis = LogisticRegression()
logis.fit(X_train,Y_train)
Y_prediction_logis = logis.predict(X_test)
Y_prediction_logis.shape
```

To calculate score:

```
score_of_logis = round(accuracy_score(Y_prediction_logis,Y_test)*100,2)
```

3.2 Naïve Bayes(NB)

Based on Bayes Probability theorem, Naïve Bayes is a classification machine learning algorithm.

Naïve Bayes can also work on small data set, large dataset is not required for accurate results.

3.2.1 Function used

```
from sklearn.naive_bayes import GaussianNB

gaussianb = GaussianNB()
gaussianb.fit(X_train,Y_train)
Y_prediction_gaussianb = gaussianb.predict(X_test)
```

To calculate score :

```
score_of_gaussianb = round(accuracy_score(Y_prediction_gaussianb,Y_test)*100,2)
```

3.3 Support Vector Machine(SVM)

This is a supervised machine learning algorithm and can be used for both classification and regression analysis . It can also perform non- linear classification. This techniques works by drawing margins between the classes and the goal is to maximize the distance between margin and classes which leads to reduction in classification error.[9]

3.3.1 Function used

```
from sklearn import svmachine
svector = svmachine.SVC(kernel='linear')
svector.fit(X_train, Y_train)
Y_prediction_svmachine = svector.predict(X_test)
Y_prediction_svmachine.shape
```

To calculate score:

```
score_of_svmachine = round(accuracy_score(Y_prediction_svmachine,Y_test)*100,2)
```

3.4 K-Nearest Neighbour(KNN)

It is a supervised machine learning algorithm and is used to solve both regression and classification problems. The Problem with

KNN alogrithm is that it becomes slow with the increase in data size[9]. As the study shows, it also provides less accuracy i.e only sixty seven percent accuracy is achieved.

3.4.1 Function used

```
from sklearn.neighbors import
KNeighborsClassifier
kneighboursclassifier =
KNeighborsClassifier(n_neighbors=7)
kneighboursclassifier.fit(X_train,Y_train)
Y_prediction_ kneighboursclassifier =
kneighboursclassifier.predict(X_test)
Y_prediction_ kneighboursclassifier.shape
```

To calculate score :

```
score_ kneighboursclassifier =
round(accuracy_score(Y_prediction_
kneighboursclassifier,Y_test)*100,2)
```

3.5 Decision Tree(DT)

Decision tree uses the method of tree for its prediction. It has a root node which gets spilt in different input features and then again splits based on another features and so on. The weight of last node predicts the value. In decision tree, every node represents a different choice and further the result.

3.5.1 Function used

```
from sklearn.tree import
DecisionTreeClassifier
maximum_accuracy = 0
        for x in range(200):
decisiontree=DecisionTreeClassifie
r(random_state=x)
 decisiontree.fit(X_train,Y_train)

Y_pred_decisiontree=decisiontree.
predict(X_test)

current_accuracy=round(accuracy_
score(Y_prediction_decisiontree,Y
_test)*100,2)

if(current_accuracy>maximum_acc
uracy):
maximum_accuracy =
current_accuracy
 best_x = x
 decisiontree =
DecisionTreeClassifier(random_sta
te=best_x)
 decisiontree.fit(X_train,Y_train)
   Y_prediction_decisiontree=
decisiontree.predict(X_test)
print(Y_prediction_decisiontree.sh
ape)
```

To calculate score:

```
        score_decisiontree =
round(accuracy_score(Y_predictio
n_decisiontree,Y_test)*100,2)
```

3.6 Random Forest(RF)

A supervised classification and regression technique , Random Forest provides results with higher accuracy. It minimizes the overfitting problem . It uses multiple decision tree in parallel and it is known as parallel ensembling and is more accurate and precise in comparison to single model of decision tree.

3.6.1 Function used:

```
from sklearn.ensemble import RandomForestClassifier
maximum_accuracy = 0
for x in range(2000):
    random_forest= RandomForestClassifier(random_state=x)
random_forest.fit(X_train,Y_train)
Y_prediction_random_forest= random_forest.predict(X_test)
current_accuracy=round(accuracy_score(Y_prediction_random_forest,Y_test)*100,2)

if(current_accuracy>maximum_accuracy):
maximum_accuracy= current_accuracy
best_x = x
 random_forest = RandomForestClassifier(random_state=best_x)
 random_forest.fit(X_train,Y_train)
 Y_prediction_random_forest = random_forest.predict(X_test)
 Y_prediction_random_forest.shape
```

To calculate score:

```
score_rf = round(accuracy_score(Y_prediction_random_forest,Y_test)*100,2)
```

3.7 XGBoost(XGB)

XGBoost model is similar to Random Forest model and generates a final decision tree based on individual or single model. It uses a gradient function which aims to minimize loss function. It reduces over-fitting by computing second order gradient of loss function[8]. It can handle large dataset effectively.

3.7.1 Function used

```
import xgboost as xgbmodel

xgb= xgbmodel.XGBClassifier(objective= "binary:logistic", random_state=42)
xgb.fit(X_train, Y_train)
Y_prediction_xgbmodel = xgb.predict(X_test)
Y_prediction_xgbmodel.shape
```

To Calculate score:

```
score_of_xgb = round(accuracy_score(Y_prediction_xgbmodel,Y_test)*100,2)
```

3.8 Neural Network(NN)

Neural networks mimic the human brain and it is an algorithm that tries to solve a problem using steps that a human brain will do. Neural networks are adaptable to changes so that if there is any need to make a change, then there will be no need to redesign , only necessary changes are sufficient.

3.8.1 Function used

```
from keras.models import Sequential
from keras.layers import Dense

# https://stats.stackexchange.com/a/136542 helped a lot in avoiding overfitting

sequential_model = Sequential()
sequential_model.add(Dense(11,activation='relu',input_dim=13))
sequential_model.add(Dense(1,activation='sigmoid'))

sequential_model.compile(loss='binary_crossentropy',optimizer='adam',metrics=['accuracy'])
sequential_model.fit(X_train,Y_train,epochs=300)
Y_prediction_neural_network = sequential_model.predict(X_test)
Y_prediction_neural_network.shape
rounded = [round(x[0]) for x in Y_prediction_neural_network]

Y_prediction_neural_network = rounded
```

To calculate Score:

```
score_of_neural_network=round(accuracy_score(Y_prediction_neural_network,Y_test)*100,2)
```

4. Result

The result of the study is shown in the below table which clearly shows that Random Forest(RF) has the highest accuracy and KNN has the lowest accuracy. Every algorithm has accuracy more than eighty percent only KNN has 67 percent accuracy.

Models used	Accuracy(in %)
Logistic Regression (LR)	85.25
Naïve Bayes(NB)	85.25

Support Vector Machine (SVM)	81.97
K-Nearest Neighbour (KNN)	67.21
Decision Tree(DT)	81.97
Random Forest(RF)	95.16
XGBoost(XGB)	85.25
Neural Network(NN)	81.97

Table 1 : Models with their accuracy

Below is the graph of the result achieved.

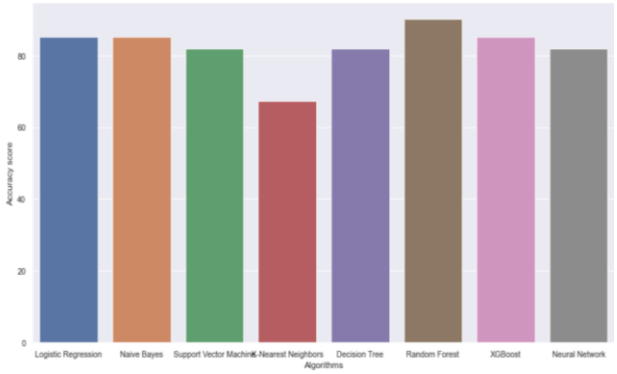

Fig 7 : Graph of Result achieved

5.Conclusion

This study predicts the various diseases based on various features like age, sex, chest pain, blood sugar, etc using several types of machine learning algorithms available. Each and every machine algorithm provides great results but Random forest gives the highest accuracy. There were some feature like fbs which does not provide any difference. Once the disease is predicted using different machine learning algorithms, further medication can be provided easily.

The future scope of this study is to use this proposed study in a sensor based device which can use sensors to sense the pulses and send an alert signal or mail to the patient family members so that immediate medication/prescription can be provided to the patient and the rate of big mishappening can be reduced.

References

1. Ajay D Wasan , Alex M Dressler, Andrea G Gillman. “A narrative review of data collection and analysis guidelines for comparative

effectiveness research in chronic pain using patient-reported outcomes and electronic health records",2019.

2. Rinkal Keniya , Aman Khakharia ,Vruddhi Shah ,Vrushabh Gada ,Ruchi Manjalkar , Tirth Thaker , Mahesh Warang ,Ninad Mehendale. "Disease Prediction from various symptoms using machine learning",
3. S. Mohan, C. Thirumalai, G. Srivastava, Effective heart disease prediction using hybrid machine learning techniques,IEEE Access 7, 81542 (2019).
4. Y. Khourdi_, M. Bahaj, Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization,Int. J. Intell. Eng. Syst. 12(1), 242 (2019).
5. Marouane Fethi Ferjani . "Disease Prediction using Machine Learning",2020.
6. F. Q. Yuan, "Critical issues of applying machine learning to condition monitoring for failure diagnosis," in 2016 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), 2016, pp. 1903–1907.
7. R.D.H.D.P. Sreevalli, K.P.M. Asia, Prediction of diseases using random forest classi_cation algorithm
8. Iqbal H. Sarker. "Machine Learning: Algorithms, Real-World Applications and Research Directions", 2021.
9. Batta Mahesh. "Machine Learning Algorithms- A Review", 2018.
10. W. Richert, L. P. Coelho, "Building Machine Learning Systems with Python", Packt Publishing Ltd., ISBN 978-1-78216-140-0
11. J. M. Keller, M. R. Gray, J. A. Givens Jr., "A Fuzzy K-Nearest Neighbor Algorithm", IEEE Transactions on Systems, Man and Cybernetics, Vol. SMC-15, No. 4, August 1985
12. https://www.geeksforgeeks.org/machine-learning/
13. S. Marsland, Machine learning: an algorithmic perspective. CRC press, 2015.
14. M. Bkassiny, Y. Li, and S. K. Jayaweera, "A survey on machine learning techniques in cognitive radios," IEEE Communications Surveys & Tutorials, vol. 15, no. 3, pp. 1136–1159, Oct. 2012.
15. https://en.wikipedia.org/wiki/Instance-based_learning

16. R. S. Sutton, “Introduction: The Challenge of Reinforcement Learning”, Machine Learning, 8, Page 225-227, Kluwer Academic Publishers, Boston, 1992
17. P. Harrington, “Machine Learning in action”, Manning Publications Co., Shelter Island, New York, 2012
18. A. Mir, S.N. Dhage, in 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA) (IEEE, 2018), pp. 1{6
19. Y. Khourdi_, M. Bahaj, Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization, Int. J. Intell. Eng. Syst. 12(1), 242 (2019)
20. S. Vijayarani, S. Dhayanand, Liver disease prediction using svm and naive bayes algorithms, International Journal of Science, Engineering and Technology Research (IJSETR) 4(4), 816 (2015)
21. S. Mohan, C. Thirumalai, G. Srivastava, Effective heart disease prediction using hybrid machine learning techniques, IEEE Access 7, 81542 (2019)
22. T.V. Sriram, M.V. Rao, G.S. Narayana, D. Kaladhar, T.P.R. Vital, Intelligent parkinson disease prediction using machine learning algorithms, International Journal of Engineering and Innovative Technology (IJEIT) 3(3), 1568 (2013)
23. A.S. Monto, S. Gravenstein, M. Elliott, M. Colopy, J. Schweinle, Clinical signs and symptoms predicting influenza infection, Archives of internal medicine 160(21), 3243 (2000)
24 R.D.H.D.P. Sreevalli, K.P.M. Asia, Prediction of diseases using random forest classification algorithm
25 . D.R. Langbehn, R.R. Brinkman, D. Falush, J.S. Paulsen,M. Hayden, an International Huntington's Disease Collaborative Group, A new model for prediction of the age of onset and penetrance for huntington's disease based on cag length, Clinical genetics 65(4), 267 (2004).
26. G. Battineni, G. G. Sagaro, N. Chinatalapudi, and F Amenta, “Applications of machine learning predictive models in the chronic disease diagnosis,” Journal of Personalized Medicine, vol. 10, no. 2, p. 21, 2020.
27. B. Manjulatha and P. Suresh, “An ensemble model for predicting chronic diseases using machine learning algorithms,” in Smart Computing Techniques and Applications, pp. 337–345, Springer, New York, NY, USA, 2021.

28. C.-H. Jen, C.-C. Wang, B. C. Jiang, Y.-H. Chu, and M.-S. Chen, "Application of classification techniques on development an early-warning system for chronic illnesses," Expert Systems with Applications, vol. 39, no. 10, pp. 8852–8858, 2012.
29. D. Gupta, S. Khare, and A. Aggarwal, "A method to predict diagnostic codes for chronic diseases using machine learning techniques," in Proceedings of the 2016 International Conference on Computing, Communication and Automation (ICCCA), pp. 281–287, IEEE, Greater Noida, India, April 2016.
30. M. Chen, H. Yixue, K. Hwang, L Wang, and L. Wang, "Disease prediction by machine learning over big data from healthcare communities," Ieee Access, vol. 5, pp. 8869–8879, 2017.
31. R. Ge, R. Zhang, and P Wang, "Prediction of chronic diseases with multi-label neural network," IEEE Access, vol. 8, pp. 138210–138216, 2020.
32. H. MacLeod, S. Yang, O. Kim, C. Kay, and S. Natarajan, "Identifying rare diseases from behavioural data: a machine learning approach," in Proceedings of the 2016 IEEE First International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE), pp. 130–139, IEEE, Washington, DC, USA, June 2016.
33. M. A. Myszczynska, P. N. Ojamies, A. M. B. Lacoste et al., "Applications of machine learning to diagnosis and treatment of neurodegenerative diseases," Nature Reviews Neurology, vol. 16, no. 8, pp. 440–456, 2020.
34. I. Preethi and K. Dharmarajan, "Diagnosis of chronic disease in a predictive model using machine learning algorithm," in Proceedings of the 2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE), pp. 191–96, IEEE, Bengaluru, India, October 2020.
35. J. Wiens and E. S Shenoy, "Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology," Clinical Infectious Diseases, vol. 66, no. 1, pp. 149–153, 2018.
36. S. Swaminathan, K. Qirko, T. Smith et al., "A machine learning approach to triaging patients with chronic obstructive pulmonary disease," PLoS One, vol. 12, no. 11, Article ID e0188532, 2017.
37. Z. Wang, J. W. Chung, X. Jiang, Y. Cui, M. Wang, and A. Zheng, "Machine learning-based prediction system for chronic kidney disease using associative classification technique," International

Journal of Engineering & Technology, vol. 7, pp. 1161–1167, 2018.
38. A. Kumar and A. Pathak, “A machine learning model for early prediction of multiple diseases to cure lives,” Turkish Journal of Computer and Mathematics Education (TURCOMAT), vol. 12, no. 6, pp. 4013–4023, 2021.
39. C. Kalaiselvi, “Diagnosing of heart diseases using average K-nearest neighbor algorithm of data mining,” in Proceedings of the 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), pp. 3099–3103, IEEE, New Delhi, India, March 2016.
40. D. Jain and V. Singh, “Feature selection and classification systems for chronic disease prediction: a review,” Egyptian Informatics Journal, vol. 19, no. 3, pp. 179–189, 2018.