Multi Disease Prediction System using Random Forest Algorithm in Healthcare System

Dr.R.Shanthakumari[1]
Associate Professor
Department of Information Technology
Kongu Engineering College
Perundurai, Erode, Tamil Nadu.
shanabiraki@gmail.com

Dr.C.Nalini[2]
Professor
Department of Information Technology
Kongu Engineering College
Perundurai, Erode, Tamil Nadu.
nalini@kongu.ac.in

Mr.S.Vinothkumar[3]
Assistant Professor (Sr.G)
Department of Information Technology
Kongu Engineering College
Perundurai, Erode, Tamil Nadu.
vinoths.it@kongu.edu

Dr.E.M.Roopadevi[4]
Associate Professor
Department of Information Technology
Kongu Engineering College
Perundurai, Erode, Tamil Nadu.
roopadevi@kongu.ac.in

Mr.B.Govindaraj[5]
B.Tech-Final Year
Department of Information Technology
Kongu Engineering College
Perundurai, Erode, Tamil Nadu.
govi1012000@gmail.com

2022 International Mobile and Embedded Technology Conference (MECON) | 978-1-6654-2020-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/MECON53876.2022.9752432

Abstract:- **Health is one of the important factors to be considered by an individual. With the increasing number of diseases and the population, medical practitioners find it hard to diagnose many numbers of diseases and predict whether the individual is suffering from the disease or not, over intensive population growth. Here comes the need for dynamic Health care systems, which are established to meet the health requirements of the population. Such systems are built with technology, health care, and data, that have to be processed in a smart, efficient, and precise course of action. Prediction system is the best technology that can meet the level of expectation in this field. There are many models proposed for single disease identification. However, very little is proposed concerning multiple disease identification. The main aim of the disease prediction model which identifies the multiple disease possibility by analyzing the health record of the patient. We consider the diseases such as Heart disease, Diabetes, and Kidney disease using some of the basic parameters such as Pulse Rate, Cholesterol, Blood Pressure, Heart Rate, etc., and also the risk factors associated with the disease can be found using prediction model with good accuracy and Precision. Many models are created by python pickling method and compared by applying it to multiple data sets and using different performance measures.**

Keywords: Data preprocessing, Feature Selection, Multiple Diseases, Prediction model.

I. INTRODUCTION

While working with the analysis of existing systems in the division of health care analysis systems, it is unequivocal that one disease is predicted once. In article [1,2,5,10], diabetes is diagnosed, article [1,2,4,9] heart disease is diagnosed and in article [2,6,7,8] kidney disease is diagnosed. Most of the articles focus on a specific disease once at a time. If an organization needs to analyze their patient's health report, there arises a need to deploy many models for each disease. The approach which is made use in the existing systems is practicable for one disease but not for more than one disease. The mortality rate is increased since the precise disease is not diagnosed properly. Even though the patient recovered from one disease may also suffer from other diseases. Few existing models use some specific parameters to analyze the disease which will not look into the possible disease that is caused due to the effect of the previous disease. If we take diabetes, there is a probability of the diseases like hearing loss, heart disease, and dementia. In this model, we consider the diagnosis of heart disease, diabetes, and kidney disease. Also, other diseases like skin-related diseases, COVID-19, and other diseases can be included in the model. The analysis of the system is designed adaptable in such a way that many diseases can be included subsequently. The appropriate model file of that disease should be loaded to the UI. When building a new model to predict disease, the programmer has to devise a python pickling technique to carry through the behavior of the model. While using TKinter UI, the developer loads the file that is pickled to retrieve the behavior of the model. If the user needs to analyze the health condition of the patient, either they can predict a specific disease or utilizing the report which comprises of the relevant features considered to diagnose the disease. The main motto is to reduce the increasing mortality rate by alerting the patients beforehand according to their status of health.

II. LITERATURE SURVEY

The vast majority of current studies focused on a common illness. When a user wants to analyze diabetes, they must use one model, and when they try to analyze heart disease, they must use another model. This is a lengthy procedure. Furthermore, if a user has several illnesses but the current method can only predict one of them, then there is a risk of death. The key goal of this article is to create a multiple disease prediction model, so the deep learning and machine learning approaches employed are briefly discussed. Different deep learning and machine learning techniques are employed to analyze diabetes, heart disease prediction, and cancer detection. The patient's condition is determined using the random forest algorithm and a variety of other algorithms. The accuracy of logistic regression for diabetes diagnosis is 92 percent, Random forest for heart disease classification is 95 percent, and SVM for cancer detection is 96 percent [1].

This system was tested with a reduced collection of features from the Chronic Kidney Disease, Diabetes, and Heart Disease datasets deployed in an optimized SVM-Radial bias kernel process, and it was also examined to other machine learning strategies in R studio, including Decision tree, SVM-Polynomial, SVM-Linear, and Random forest. Accuracy, specificity, sensitivity, and misclassification rate have all been used to test the efficiency of these machine learning techniques. According to the findings of the trial, the enhanced SVM-Radial bias kernel strategy achieves precision of 89.9%, 98.7%, and 98.3%, in the Heart Disease, Diabetes, and Chronic Kidney Disease datasets, respectively [2]. As in articles [1,2, 3] diseases are considered and 7 classification algorithms are used then the algorithm with the highest accuracy will be deployed in the final decision support predictive model.[4]-[10] papers have been stated that disease prediction with different kinds of algorithms and different environments.

III. PROPOSED METHODOLOGY

In the health-care sector, the latest research proposes a method for forecasting various diseases. The method has dealt with Heart Disease, Diabetes, and Kidney Disease from the UCI dataset. Standards of Medical Datasets considered in Heart Disease, Diabetes and Kidney Disease respectively, as a research step. Figure 1 explains the Multiple disease prediction system. The proposed predictive system constitutes of three sub modules as, Data pre-processing module, Feature selection module and Predictive system module.

A. Dataset Pre-processing

Data preprocessing is essential in data analytics because it removes unnecessary and noisy data. Unwanted data and missing data are often collected by data sources. The data pre-processing is done to process the missing values. Numerical missing values are filled using statistical measures i.e., mean of each attribute. Categorical missing values are filled using mode value of the classified attribute.

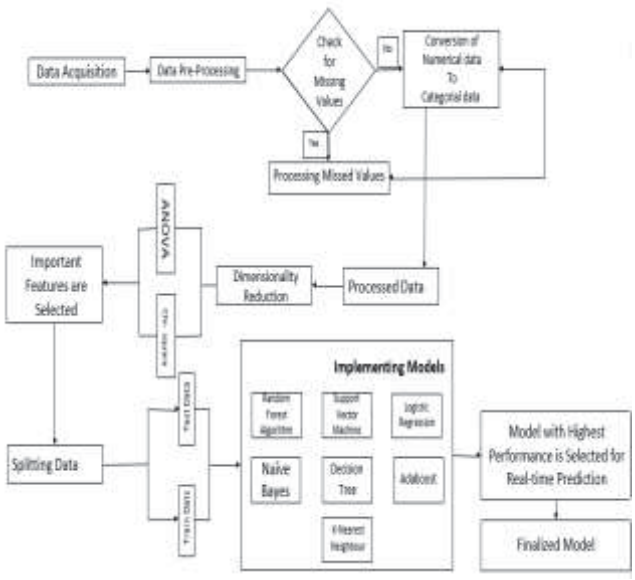

Fig. 1 Proposed System for Multiple Disease Prediction

A.1 Conversion of Numerical Data to Categorical Data

The dataset is converted from numerical to categorical data in order to obtain balanced and normalized dataset. The code snap for the conversion of numerical to categorical data for the diabetes, heart disease, and kidney disease dataset are as picturized in Figure 2,3 and 4 severally.

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	6	3	0	3	0	2	1
0	4	6	3	2	3	1	0	0
0	7	0	0	0	4	3	1	1
0	6	6	4	5	4	1	0	1
0	6	6	4	9	4	2	0	0

Fig. 2 First Five Rows of Diabetes Dataset after Pre-processing and conversion

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
1	1	0	1	2	0	1	7	0	1	2	2	3	1
2	1	0	2	1	1	0	6	1	3	0	0	3	1
3	1	0	2	1	0	1	4	1	2	0	0	3	1
2	1	0	2	1	0	1	6	0	0	2	1	3	1
2	0	0	1	4	1	1	2	0	1	1	3	2	1

Fig. 3 First Five Rows of Heart Disease Dataset after Pre-processing and conversion

age	bp	sg	al	su	rbc	...	pcv	wc	rc	htn	dm	cad	appet	pe	ane	classification
2	2	3	1	0	0	...	3	1	5	1	1	0	0	0	0	0
0	0	3	4	0	0	...	3	1	4	0	0	0	0	0	0	0
3	2	1	2	3	0	...	2	1	4	0	1	0	1	0	1	0
2	1	0	4	0	0	...	2	1	3	1	0	0	1	1	1	0
2	2	1	2	0	0	...	2	1	4	0	0	0	0	0	0	0

Fig. 4 First Five Rows of Kidney Disease Dataset after Pre-processing and Conversion

B. Feature Selection

It is a technique of identifying the most important features in a dataset. The data source generated data with a large number of attributes. Few of the features in the dataset will not impact the performance of the prediction system. Only necessary features will consider for the Decision-making system, it will reduce the computational complexity. The following algorithms are applied on the pre-processed data to trace out the best features from the dataset.

1.Chi-Squared: It is a feature selection algorithm that provides the best features from the dataset. This is applied only for categorical data. It is used to test the independence of two events.

2. ANOVA: It stands for ANalysis Of VAriance. It is a feature selection algorithm that helps to gain information about the dependent and independent variables.

Figure 5.depicts the feature selection output for all three datasets. The Heart disease dataset features with considerable score value and common to both the scripts results are selected for classification namely: Exang, Cp, Ca, Oldpeak and Thalach. The Diabetes dataset features selected for classification are: Glucose, Age, Pregnancies, Insulin and BMI. The Kidney disease dataset features selected for the classification are: Hemo, sg, htn, dm, al and bgr.

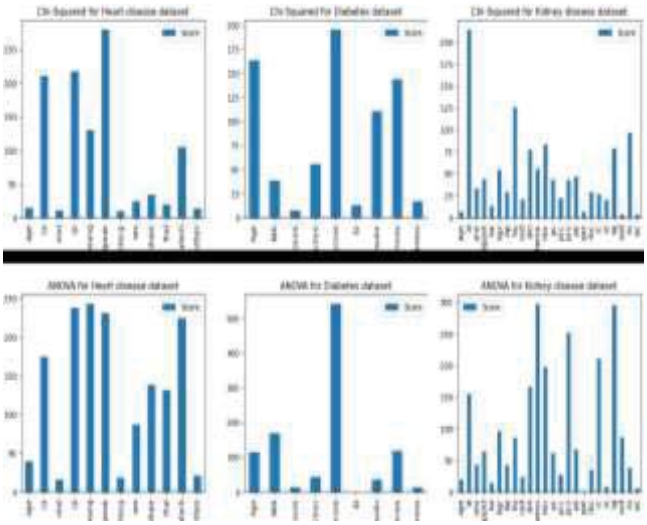

Fig. 5 Feature Selection Output for All Three Datasets

C. Prediction System

Among these three datasets, only kidney disease dataset had missing values. These are processed with mode values for numerical and categorical attributes respectively. All the numerical features are then converted to categorical. Thus, all the datasets are done with pre-processing and normalization.

With proper pre-processing been done, the datasets are now fit to train the machine learning algorithms. For dimensionality reduction, these pre-processed datasets are applied to Chi-Squared and ANOVA feature selection algorithms. The best attributes that are ideally close to its intrinsic dimensions are obtained with greater scores. Only those attributes with top scores and common to both the feature selection algorithms' results are retained for proceeding with algorithms' performance analysis.

1) C.1 Classification Models

The algorithms applied are 5 base classifiers (SVM, KNN, Decision Tree, Naïve Bayes and Logistic Regression), and 2 Ensemble techniques (Adaboost, Random Forest). Support Vector Machine: SVM is an algorithm which is used mainly as a classifier and also for regression and outlier detection. K Nearest Neighbors: KNN is a supervised machine learning technique that is used for classification. Based on the K nearest neighbor's and the distance between them the new input class will belong to. Decision Tree: It is a supervised machine learning technique which is used for classification problems. The classifier is a tree structure that consist of internal nodes, branch and leaf nodes. The internal nodes represent the features of the dataset, branches represent the decision rules and leaf nodes represent the outcome.

Naïve-Bayes: The Naïve-Bayes algorithm is classification algorithm that uses Bayes theorem to predict the probability of different classes based on various attributes. Logistic Regression: It is a classification algorithm used to predict the probability of the target variable.

Adaboost: Adaboost is an ensemble technique. It is used for binary classification and it enhances the performance of any machine learning algorithm. Random Forest: The Random Forest algorithm is a classification algorithm that is similar to decision tree. It adds randomness to the model. It searches for the best features from the random subset of features. A random forest is made up of a large number of independent decision trees that function together as an ensemble. Each tree in the random forest generates a class prediction, and the class that has the highest ballots becomes the prediction of our model. The basic idea behind random forest is a clear but effective one.

Initially, Random forest, SVM, KNN, Decision Tree, Naïve Bayes, Logistic Regression and Adaboost algorithms are trained with the 80% of each disease dataset. After testing the trained model with other 20% of each dataset, all the previously mentioned performance measures are calculated. All the performance measures are calculated and tabulated for all these 7 algorithms for all 3 datasets individually. Final model for disease prediction is recommended, with comparatively best performance measures tabulated. All these stages are represented diagrammatically.

After tracing out the best algorithm as Random forest algorithm for each dataset, those algorithms are now trained with the respective entire dataset. These trained models are saved individually as pickle files. Once the user enters the health records, these pickled files are loaded (un-pickling) and the input values are supplied to respective loaded models for prediction. All three models predict the disease with given records of specific fields. Their prediction results are shown in the same module. Interpreting the results, we conclude with using Random Forest Algorithm for proposing the predictive system. The proposed model is depicted in Figure 6.

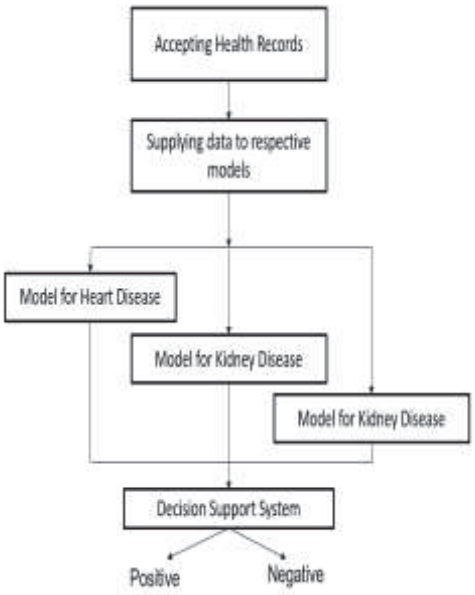

Fig. 6 Integration of Multiple Dataset

2) C.2 Python Pickling Technique for Random Forest Algorithm

After the model is built, the model is saved in binary form into a pickle file which can later be loaded while analyzing real-time data for prediction of disease Example:

```
# For Diabetes

d1=pd.read_excel('Diab_Categ.xlsx')
```

```
X=d1[['Glucose','Age','Pregnancies','BMI','Insulin']]
Y = d1.iloc[:,-1]
classifier=RandomForestClassifier(n_estimators=10)
classifier.fit(X,Y)
pickle.dump(classifier, open ('DiabPred.pickle', 'wb'))
```

and similarly, for heart disease and kidney disease models.

3) C.3 Pickle File Loading to Use the Model

The pickle files are loaded and deployed in the GUI. It is used to predict multiple disease and provide combined results.

```
import pickle as pickle
diabetes_pickle_file_saving=pickle.load(open('DiabetesPred
iction.pickle', 'rb'))
heart_pickle_file_saving
=pickle.load(open('HeartPrediction.pickle', 'rb'))
kidney_pickle_file_saving
=pickle.load(open('KidneyPrediction.pickle', 'rb'))
```

IV. PERFORMANCE ANALYSIS

Data set acquisition is the first step in building disease prediction model. For constructing a multi-disease prediction model, 3 diseases in particular, three different datasets are acquired from UCI machine learning repository. All three datasets differed with respect to the dimensions, characteristics and the type of attributes. In this work, three different datasets for three diseases is used. The datasets are Heart disease dataset, Kidney Disease and Diabetes dataset. The Diabetes dataset consist of 9 columns with outcome as the dependent variable. The Diabetes dataset is given Figure 7. The Heart Disease dataset consist of 14 columns with target as dependent variable. The heart disease dataset is shown in Figure 8. The kidney Disease dataset consist of 26 columns with classification as the dependent variable. The kidney disease dataset is shown in Figure 9. Train and test sets are prepared in accordance with industry requirements. The data is split into 80 percent for training and 20 percent for testing using the Scikit learn train test split process.

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
2	138	62	35	0	33.6	0.127	47	1
0	84	82	31	125	38.2	0.233	23	0
0	145	0	0	0	44.2	0.630	31	1
0	135	68	42	250	42.3	0.365	24	1
1	139	62	41	480	40.7	0.536	21	0

Fig. 7 First Five Rows of Diabetes Dataset

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
52	1	0	125	212	0	1	168	0	1.0	2	2	3	0
53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
61	1	0	148	203	0	1	161	0	0.0	2	1	3	0
62	0	0	138	294	1	1	106	0	1.9	1	3	2	0

Fig. 8 First Five Rows of Heart Disease Dataset

age	bp	sg	al	su	rbc	pc	...	pcv	wc	rc	htn	dm	cad	appet	pe	ane	classification
48	80	1.020	1	0	0	0	...	44	7800	5.2	1	1	0	0	0	0	0
7	50	1.020	4	0	0	0	...	38	6000	4.0	0	0	0	0	0	0	0
62	80	1.010	2	3	0	0	...	31	7500	4.0	0	1	0	1	0	1	0
48	70	1.005	4	0	0	1	...	32	6700	3.9	1	0	0	1	1	1	0
51	80	1.010	2	0	0	0	...	35	7300	4.6	0	0	0	0	0	0	0

Fig. 9 First Five Rows of Kidney Disease Dataset

For comparing the performance of various algorithms on all 3 datasets, the following measures are calculated as shown in Table 1, Table 2, and Table 3. The graphical representation of the performance analysis is illustrated in Figure 10, 11 and 12 respectively.

1. Accuracy
2. R-Squared
3. Time of Execution (in Secs)
4. Precision
5. Recall
6. F1-Score

TABLE I. PERFORMANCE ANALYSIS OF DIABETES PREDICTION MODEL

Algorithms	Accuracy (%)	Time of Execution (Sec)	R-Squared (%)	Precision (%)	Recall (%)	F1-Score (%)
RFA	92.3	0.37	0.64	94	91	92
SVM	79.75	0.127	0.08	78	76	76
LR	79	0.048	0.047	76	74	75
KNN	80.5	0.134	-1.19	46	46	46
NB	79	0.019	0.046	77	76	77
AB	82	0.61	0.182	80	80	80
DT	80	0.012	-0.046	75	69	72

TABLE II. PERFORMANCE ANALYSIS OF HEART DISEASE PREDICTION MODEL

Algorithms	Accuracy (%)	Time of Execution (Sec)	R-Squared (%)	Precision (%)	Recall (%)	F1-Score (%)
RFA	98.05	0.14	92	98	98	98
SVM	84.74	0.033	38	85	84	85
LR	86.36	0.036	45	87	86	86
KNN	82.47	0.087	45	87	86	86
NB	83.44	0.014	33	84	83	83
AB	85.39	0.49	41	86	85	86
DT	91.56	0.012	66	92	92	92

TABLE III. PERFORMANCE ANALYSIS OF KIDNEY DISEASE PREDICTION MODEL

Algorithms	Accuracy (%)	Time of Execution (Sec)	R-Squared (%)	Precision (%)	Recall (%)	F1-Score (%)
RFA	**99.17**	**0.13**	**0.96**	**100**	**98**	**99**
SVM	**98.17**	**0.019**	**0.96**	**99**	**99**	**99**
LR	**98.33**	**0.045**	**0.93**	**98**	**98**	**98**
KNN	**98.34**	**0.042**	**0.93**	**98**	**98**	**98**
NB	**91.67**	**0.016**	**0.625**	**90**	**94**	**92**

AB	99.17	0.54	0.96	99	100	99
DT	95	0.011	0.775	94	97	96

D. *Comparison of Different Prediction System for Multiple Disease*

The classification algorithm applied to the best features of the dataset after feature selection is done has produced the results are depicted in Figure 10, 11 and 12. These Figures shows the results on accuracy, time, R-squared values. The confusion matrix tabulation is illustrated in Table 1, Table 2 and Table 3 show the various results obtained by applying the classification algorithm on the 3 different datasets namely heart disease, Kidney disease and diabetes. From these tabulations, we can conclude that Random Forest Algorithm has resulted with best performance measures analysis.

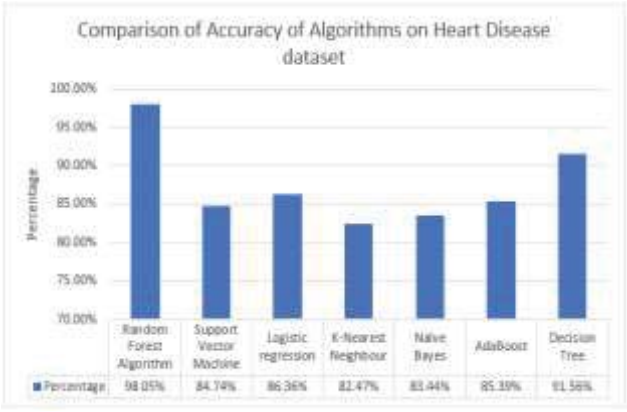

Fig. 10 Comparison of Accuracy of Algorithms on Heart Disease Dataset

For Heart disease: It has an accuracy of 98.05%, precision as 100%, Recall as 96%, F1-Score as 98%.

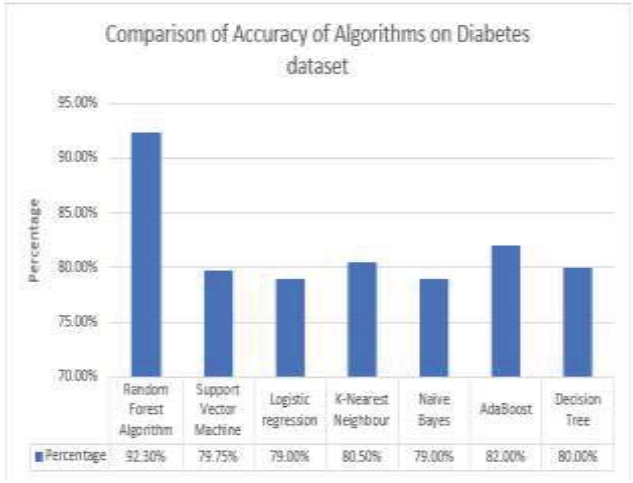

Fig. 11 Comparison of Accuracy of Algorithms on Diabetes Dataset

For Diabetes: It has an accuracy of 92.30%, precision as 94%, Recall as 91%, F1-Score as 92%. For Kidney Disease: It has an accuracy of 99.17%, precision as 100%, Recall as 99%, F1-Score as 99%.

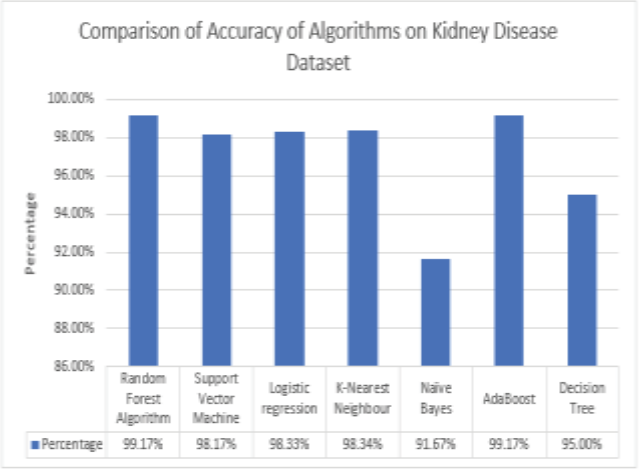

Fig. 12 Comparison of Accuracy of Algorithms on Kidney Dataset

V. CONCLUSION

Multiple disease prediction model is the system in which the users are allowed to input their health records that they have obtained from various prescribed tests' reports. Disease predictions with this model can be considered as an intermediate step between taking up the prescribed health check-up test and consultation with the medical practitioners. Thereupon, not all the patients will have a need to go for consultation. Compared with the existing system, the model that has been proposed is more flexible and efficient, as the prediction of all three diseases is done with single set of inputs and the users are given a choice, either to predict with the features they are most interested in or with the system recommended features. The benefit of multiple disease prediction model is that it can predict the occurrence probability of various diseases in advance, thereby reducing the mortality ratio. The proposed model has overcome one of the demerits of the existing model, which is prediction can be done with single set of input data with higher flexibility and it is fixed with 3 types of diseases.

In future, the developers can add any type and number of diseases with better scalability on datasets. If the dataset, which is going to be collected in the future, has records of patients from various birth places all over the world, then its trained model will be more efficient than the proposed one. Moreover, the larger the dataset, more will be the benefit to be gained. As a result of that, the model may be recommended to more users irrespective of geographical locations and its conditions.

VI. REFERENCES

[1] Akkem, Y, "Multi Disease Prediction Model by using Machine Learning and Flask API", IEEE Conference Record #48766; IEEE Xplore ISBN: 978-1-7281-5371-1, (ICCES 2020). Diabetes, heart, cancer, diabetic retinopathy

[2] Karthikeyan. H., Menakadevi. T, "Multi-disease prediction model using improved SVM-radial bias technique in healthcare monitoring system" ,Journal of Ambient Intelligence and Humanized Computing, Springer Germany, 2020

[3] Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes—2020 American Diabetes Association Diabetes Care 2020; 43 (Suppl. 1): S14–S31 | https://doi.org/10.2337/dc20-S002.

[4] Chait. R. S., Sulabha. S., Dangare., Apte, "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques", International Journal of Computer Applications (0975 1.888), Volume 47 ,No.10, June 2012.

[5] Dhiraj. D., Gajanan. P., Ektaa. M, " Designing Disease Prediction Model Using Machine Learning Approach", Proceedings of the Third International Conference on Computing Methodologies and Communication (ICCMC 2019), IEEE Xplore Part Number: CFP19K25-ART; ISBN: 978-1-5386-7808-4, 2019.

[6] Dr. Vijayarani. S, Mr. Dhayanand. S, " Data Mining Classification Algorithms for Kidney Disease Prediction, International Journal on Cybernetics & Informatics (IJCI) ,Vol. 4, No. 4, August 2015 DOI: 10.5121/ijci.2015.4402 Tamil Nadu (2015).

[7] Dr. Vijayarani. S, Mr. Dhayanand. S,"Kidney Disease Prediction Using SVM and ANN Algorithms, International Journal of Computing and Business Research (IJCBR) ISSN (Online), 2229-6166, Volume 6, Issue 2, 2015

[8] Rajadevi. R., Roopa devi. E. M., Shanthakumari. R.: Features Selection for Predicting Heart Disease using Black Hole Optimization Algorithm and XGBoost Classifier, International Conference on Computation Communication Informatics (ICCCI 2021), 2021

[9] Misir. R, Mitra. M, Samanta. R. K., "A reduced set of features for chronic kidney disease prediction, J Pathol Inf 1:8–24 ,2017

[10] Fontecha. J, González. I, Bravo. J, "A usability study of a mHealth system for diabetes self-management based on framework analysis and usability problem taxonomy methods", J Ambient Intell Hum Comput,Vol 1,pp1–11, 2019