Performances Analysis Of Multiple Diseases Prediction Systems Using Multiple Machine Learning Models

1st Nang Seint Seint Soe
Faculty of Computer Science
University of Computer Studies(Taunggyi)
Taunggyi, Myanmar
seintseintsoe@ucstgi.edu.mm

2nd Nan Saw Kalayar
Faculty of Computer Science
University of Computer Studies(Taunggyi)
Taunggyi, Myanmar
sawkalayar@ucstgi.edu.mm

***Abstract*— There are many health care centres in the world with advanced diagnostic tools. Although having this tool some patients could not receive suitable treatments and may suffer to death. Therefore, this paper proposed a system to predict two diseases (Diabetes and Chronic Kidney Disease) by using four classification techniques (Decision Tree, Gaussian Naive Bayes, Support Vector Machine and Random Forest). As a result, numerous lives can be saved by early detection. The accuracy of each algorithm is calculated and matched with each other to find the better one for prediction. The attributes of datasets used in this system are taken from the Kaggle Website but some additional new attribute (creatinine level) are added into the diabetes dataset according to world health organization (WHO) guide line as the contribution of this paper. Chronic Kidney Dataset in Kaggle, they used 400 instances, but in our Chronic Kidney Dataset, there are total all 1000 instances are generated and used.**

***Keywords*— *Decision Tree (DT), Gaussian Naive Bayes GNB), Support Vector Machine (SVM), Random Forest (RF), World Health Organization (WHO), Chronic Kidney Disease (CKD*)

I. INTRODUCTION

Today, in the world, there are many health care centers using the latest developments medical material. Although having this up to date material, some patient could not have a chance to access this material. The reason behind this is time because most medical system have a little time and it is difficult for them to control their time. Therefore, by using machine learning tools, this paper proposed a system that recognized patients who are challenging diseases like Chronic Kidney and Diabetes disease at an early step. As an advantage, the correct actions could be given in time to them.

The four machine learning prediction techniques are used in this paper. They are Decision Tree, Gaussian Naive Bayes, Support Vector Machine (SVM) and Random Forest classifiers. The attributes of diabetes dataset and chronic kidney disease dataset used in this paper are reference from the Kaggle Website but some new additional attribute (creatinine level) are added into the diabetes dataset according to world health organization (WHO) guide line as the contribution of this paper.

For the diabetes dataset, it is obviously seen that when predicting the original dataset taking from Kaggle, the accuracy score is low. When we add the new attribute creatine and then make classification, the system gets the better accuracy results.

For predicting the Chronic Kidney Disease, our own Chronic Kidney dataset is created and the attributes in the dataset are the same as the original Chronic Kidney Disease dataset from Kaggle. Therefore, in this system, own CKD dataset is generated and used with 1000 instances. The dataset is not including missing values. As the experimental results, the prediction accuracy using our own dataset received the better accuracy results than the original dataset from Kaggle which have 400 instances.

II. RELATED WORK

D. Mandem, B. Prajna proposed the signs provided by the user as input and gives the probability of the disease as an output of disease. Two kinds of diseases prediction are done by applying the random forest classifier [2].

T. Ture, A. Sawant, R. Singh, C. Patil created the system that recognized patients with core diseases like Heart disease, Kidney disease and Diabetes disease at early condition. They used Random Forest classifier to predict the three diseases [4].

B.S. Ahamed, M.S. Arya, A.O.V. Nancy proposed a system to predict diabetes mellitus disease. For predicting their system, they used three machine learning model such as Light Gradient Boosting Machine, Gradient Boosting classifier, and Random Forest Classifier. In their system the accuracy of each classifier is calculated and compared the accuracy results [1].

S. Reshma, S. Shaji, S.R. Ajina, S.R.V. Priya, A. Janisha used Ant Colony Optimization (ACO) method and Support Vector Machine (SVM) classifier to forecast whether the person is having CKD or not by using minimum number of features [3].

III. PROPOSED SYSTEM DESIGN

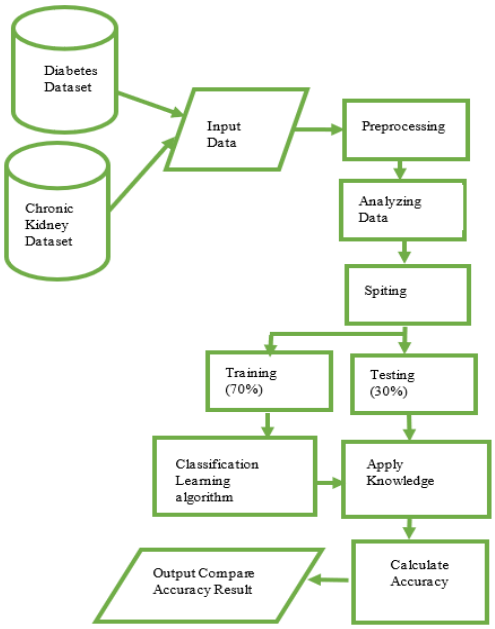

Fig 1. Proposed System Design

The proposed system consists of the following works. First the data from the datasets are input into the system. In the second stage, preprocessing step and analyzing the data steps are done. After passing that stage, the data are split into training and testing stage. 70% of data in dataset are used as a training data and the rest are used as a testing data and then predict the dataset using the four classification models. The accuracy result getting by each model are compared and suggest the better accuracy result which model get.

IV. DATASETS AND DATA PREPROCESSING

A. *Diabetes Dataset*

The system predicted two disease datasets. They are Diabetes Dataset and Chronic Kidney Dataset. For predicting the Diabetes disease, the dataset is taking from the Kaggle Website, UCI machine learning repository. This dataset has eight attributes exclude the target values. The eight features are Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree-Function and Age. As the contribution of this paper the new feature creatinine level is added into the diabetes dataset. The range of creatinine level for women who have not diabetes is between 0.6 to 1.1 mg/dL (53 to 97.2 umol/L). Thus, in diabetes dataset, there are total all ten attributes including the target value. In the diabetes dataset using in this system, there have 768 cases, 268 are predicted as having diabetes and the residual 500 are predicted as non-diabetes.

B. *Chronic Kidney Dataset*

For predicting the Chronic Kidney Disease, the proposed system generated own Chronic Kidney dataset and the attributes used in the dataset are the same as the original Chronic Kidney Disease dataset from Kaggle Website.

In Chronic Kidney Disease dataset, 24 features are used. They are age, blood pressure, specific gravity, albumin, sugar, red blood cells, pus cell, pus cell clumps, bacteria, blood glucose random, blood urea, serum creatinine, sodium, potassium, hemoglobin, packed cell volume, white blood cell count, red blood cell count, hypertension, diabetes mellitus, coronary artery disease, appetite, pedal edema and anemia.

In that dataset there are total all 25 attributes including the target value. 11 attributes of which are numerical and 14 attributes of which are nominal. From Kaggle Website, the original dataset of Chronic Kidney Disease dataset includes total all 400 instances. In this system, own chronic kidney dataset is generated with 1000 instances. The range of the attribute's values are references as CKD dataset from the Kaggle. The feature selection is performed to determine the most relevant attributes for detecting CKD and rank them according to their predictability. There are multiple attributes that appear to be correlated. Example hemoglobin, age and blood pressure which demands for a correlation analysis. The dataset is not including missing values.

C. *Data Preprocessing*

For implementing the system, first the data from the dataset are putting into the system. In the original Chronic Kidney Disease dataset which is taking from Kaggle consists of missing value. Thus, the dataset must be cleaned up as the preprocess step to increase the prediction accuracy. The missing values are replaced with mean for numeric attributes and most frequent value for nominal variables. After that Encode the labels for the nominal attribute's values. And then calculate the correlation between the attributes. For the prediction process, each dataset is split into two groups. The first 30% group is used for testing while the left 70% is used for the training process.

After splitting stage, data standardization is done, built the model and calculate the accuracy result for each model.

V. METHODOLOGY

The four machine learning prediction techniques are used in this paper. They are Decision Tree, Gaussian Naive Bayes, Support Vector Machine (SVM) and Random Forest.

A. *Decision Tree Classifier*

In decision tree classifier, it all initiates with root node and similar a tree it branches off into a sub tree. Based on some conditions the decisions are made and it branches off. Decision nodes are used to make decision taken and it branches off. But the leaf nodes are the outcome of the result occupied and there is no more branch. Basically, the decision is occupied built on the reply yes or no.

B. *Gussian Naïve Bayes Classifier*

Gaussian Naive Bayes (GNB) is a classification method based on the probabilistic method and Gaussian distribution. Gaussian Naive Bayes suppose that each parameter (also called features or predictors) has an independent capacity of

forecasting the output variable. The grouping of the prediction for all parameters is the ending prediction, that returns a probability reliant on variable to be classified in each group. The last classification is assigned to the group with the higher probability.

C. Support Vector Machine

The system in this paper used linear Support Vector Machine. The used of SVM algorithm is to make the greatest line or choice border that can separate n-dimensional space into classes. This best decision border is called a hyperplane. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors.

D. Random Forest Classifier

The random forest algorithm is working as the following steps.

Step 1: The algorithm picks random samples from the dataset created.

Step 2: The algorithm will choose tree for each sample selected. Then it will develop a prediction outcome from each decision tree created.

Step 3: Voting will be done for every predicted outcome.

Step 4: And lastly, the process will choice the highest voted prediction outcome as the last prediction.

VI. Performance Analysis and Conclusion

A. Performance Analysis

The performance of the machine learning models was calculated with the help of confusion matrix, which was utilized to display the specific results. A confusion matrix shows the prediction summary in matrix form. It represents how many predictions are right and improper per class. Figure Fig. 2 show a confusion matrix for binary classification case.

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \quad (1)$$

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

$$F1\ score = 2 \text{ x } \frac{precision \text{ x } recall}{precision+reacall} \quad (4)$$

		Predicted CKD	
		ckd	***not ckd***
Ground Truth	***ckd***	TP	FP
	not ckd	FN	TN

Fig 2. Confusion Matrix

The system is implemented with two kinds of dataset. The first one, the proposed system add new attribute to the diabetes dataset and then compare the accuracy result of the original diabetes dataset and the new diabetes dataset. The accuracy results can be seen as Table I. In the original dataset, 8 attributes are used and the prediction accuracy result is low. When add a new attributes creatinine, and make prediction with 9 attributes, the better prediction accuracy results are achieved.

TABLE I. Comparison Of Performance Analysis For Old Diabetes Dataset And New Diabetes Dataset

Class-ifier	**Accuracy (%)**		**Precision (%)**		**Recall (%)**		**F1-Score (%)**	
	8 Attributes	***9 Attributes***	***8 Attributes***	***9 Attributes***	***8 Attributes***	***9 Attributes***	***8 Attributes***	***9 Attributes***
DT	0.69	1.00	0.75	1.00	0.79	1.00	0.77	1.00
GNB	0.77	0.98	0.80	0.98	0.88	0.99	0.84	0.99
SVM	0.74	0.96	0.76	0.94	0.90	1.00	0.83	0.97
RF	0.77	1.00	0.83	1.00	0.86	1.00	0.84	1.00

In the second work, the system compared the CKD dataset from Kaggle (400 instances) and new own generated CKD Dataset with 1000 instances. The more instances are used the better accuracy result are received as the following Table II.

TABLE II. Comparison Of Performance Analysis For Old CKD Dataset And New CKD Dataset

Class-ifier	**Accuracy(%)**		**Precision (%)**		**Recall (%)**		**F1 Score (%)**	
	Data 400	***Data 1000***	***Data 400***	***Data 1000***	***Data 400***	***Data 1000***	***Data 400***	***Data 1000***
DT	0.98	1.00	0.99	1.00	0.99	1.00	0.99	1.00
GNB	0.96	1.00	0.99	1.00	0.95	1.00	0.97	1.00
SVM	0.98	1.00	0.97	1.00	1.00	1.00	0.99	1.00
RF	0.98	1.00	0.97	1.00	1.00	1.00	0.99	1.00

B. Conclusion

This system used four classification techniques to predict the two diseases. For the diabetes disease prediction, the proposed system proved that the more attributes are used, the better accuracy result is achieved. For the chronic kidney disease prediction, the more instances are used in dataset, the better accuracy results are received. For future work, the more diseases will be added into the system and make prediction with more machine learning classifiers.

Acknowledgment

Thank you all who support me at every corner to do this research.

References

[1] B.S. Ahamed, M.S. Arya, A.O.V. Nancy " Diabetes Mellitus Disease Prediction Using Machine Learning Classifiers with Oversampling and Feature Augmentation", pp 1-14, vol. 2022, Article ID 9220560, 2022.

[2] D. Mandem, B. Prajna " Multi Disease Prediction System", vol. 8, pp. 504-509, November 2021.

[3] S. Reshma, S. Shaji, S.R. Ajina, S.R.V. Priya, A. Janisha, "Chronic Kidney Disease Prediction using Machine Learning", pp 137-140, vol.9, July 2020.

[4] T. Ture, A. Sawant, R. Singh, C. Patil, "Multiple Disease Prediction System", vol. 11. pp. 1238-1244, March 2023 . T. Ture, A. Sawant, R. Singh, C. Patil, "Multiple Disease Prediction System", vol. 11. pp. 1238-1244, March 2023