An Intelligent Diabetes Diagnostic Prediction System Using Ensemble Classifier

[1]Ogundele Israel .O, [2]Sokunbi Michael .A, [3]Akinsola Adeniyi .F, [4]Akinade Abigail .O, [5]Adebayo Adeniran .A

Abstract— Healthcare sector contains large and sensitive data that needs to be handled with care. Diabetes Mellitus is one of the growing extremely fatal diseases all over the world. Globally, the rate of diabetes is rapidly increasing and according to International Diabetes Federation (IDF) Diabetes Atlas, the global diabetes prevalence in 2019 is estimated to be 9.3% (463 million people), expected to rise to 10.2% (578 million) by 2030 and 10.9% (700 million) by 2045. Currently, sub-Saharan Africa is estimated to possess 20 million people with diabetes. Nigeria has the lowest possible number of citizenry with diabetes with an estimated 3.9 million humans (or an extrapolated incidence of 4.99%) of the adult population aged 20-79-year-old. Due to this present challenges, medical professionals want a reliable prediction system that can better diagnose Diabetes at the most efficient time. An ensemble classifier has been applied for the prediction of diabetes disease which have been proven to give a better accurate result than that of single classifier. The Pima Indian Diabetes Dataset (PIDD) of 768 records was acquired from UCI (University of California Irvine) machine learning repository. The SCI-KIT software was employed as mining tool for diagnosing diabetes. The programming language used was python and Django framework, MYSQL was used for the database. The application created in the research was web-based for easy accessibility anywhere anytime. This paper focuses on improving the level of health and wellbeing of an individual based on the risk factors by providing early diagnosis of the diseases in order to predict the likelihood of having the disease. Ensemble classifier is used for the prediction because it's proven to give better accuracy than a single classifier. The ensemble classifier considered are Decision Tree, Naïve Bayes, K-Nearest Neighbour. The result of the ensemble classifier give 94.48% accuracy which outperformed a single classifier in the diagnosis prediction of diabetes diseases in order to reduce the morbidity rate based on risk factors thereby increasing the quality of life.

Index Terms— Dataset, Data Mining, Diseases, Ensemble Classifier, Machine Learning, Medical Diagnosis, Prediction.

1.0 INTRODUCTION

Diabetes is one of the deadliest diseases in the world. It is not only the most effective ailment but also the creator of different types of sicknesses like coronary heart attack, blindness, kidney illnesses, etc. it is a continual disorder that takes place while the pancreas is not able to produce insulin, or when the body cannot produce actual use of the insulin. Generally, diabetes is characterized as existing in two major forms: (a) insulin dependent (Type I) [1] and (b) noninsulin-dependent (Type II) [2]. The latter appears to be the more common, accounting for 80% of all cases [3]. It is one in all important NON-Communicable Diseases (NCDs) that is hastily attracting the attention of the international medical community, culminating in a United Nations political announcement on NCDs in September 2011 with follow-up assembly on Political Declaration of the High-level meeting of the General Assembly on the Prevention and Control of NCDs in May 2013. Globally, the weight of diabetes is rapidly increasing. It was reported by International Diabetes Federation (IDF) Diabetes Atlas, that 9.3% (463 million people) had the diseases in 2019, this will increase to 10.2% (578 million) by 2030 and 10.9% (700 million) by 2045 [4]. Most of the people having diabetes, do not know they have it and this is common to urban (10.8%) than rural areas (7.2%), and in low-income (10.4%) than high-income nation (4.0%). Currently, sub-Saharan Africa is estimated to possess 20 million people with diabetes, about 62% aren't recognized and therefore the range is predicted to exceed 41.4 million by 2035 or an amplify of 109.1%. In Sub-Saharan Africa, Nigeria has the lowest possible number of citizenry with diabetes with an estimated 3.9 million humans (or an extrapolated incidence of 4.99%) of the adult population aged 20-79-year-old. Further, in terms of morbidity, diabetes contributes to the development of coronary heart disease, renal disease, pneumonia, bacteremia, and tuberculosis (TB). It is regarded that people with diabetes are 3 instances more probably to strengthen tuberculosis and approximately 15% of TB globally is thinking to have background diabetes as a predisposing factor. This state of affairs of the double burden of sickness specially in creating countries put diabetes to compete for sources as nicely as political dedication.

Studies carried out in Nigeria indicated that the prevalence of diabetes ranged from low stage of 0.8% among adults in rural highland dwellers to over 7% in Lagos with an average of 2.2% nationally. As already pointed out, the sixth version of IDF (International Diabetes Foundation) diabetes Atlas, suggests that Nigeria is the leading country with variety of people with diabetes. In 2013, there was a related death of diabetes of 105,091 and 3.9million people with the diseases and this will increase by 125,000 between 2010 to 2030 even though the superiority of 4.99% is far much less than that of Reunion (15.38%), Seychelles (12.11%), Gabon (10.71%), Zimbabwe (9.73%), and South Africa (9.27%); in addition, there are still approximately 1.8 million Nigerians with undiagnosed diabetes [4]. Based on the statistics of people living with diabetes, it shows that there is a serious need to strengthen the predicting fashions that assists in

determining such a sickness as it is turning into one of the international hazards. To advance such solution, analyzing the already handy big diabetic information units to find out some brilliant information with the use of ensemble classifier is a step in the right direction.

Data mining has the potential to extract hidden understanding from a big quantity of data relating to diabetes. Because of that, it has a sizable position in diabetes research, now more than ever. Various researchers have used ensemble classifier for prediction some diseases such as heart diseases and this has shown a better performance accuracy than using a single technique. The main aim of this work is to see how to reduce the death rate of the diabetes patient by providing early detection and prediction of the diseases using the attributes associated with it.

This research therefore focused on developing an ensemble classifier algorithm that is capable of predicting diabetes patient earlier enough for immediate attention by the healthcare practitioner thereby increasing the quality of life.

2.0 LITERATURE REVIEW

A variety of computer-controlled models were used to predict or identify diabetes. Such models either tried to categorize patients into insulin and non-insulin, or predict blood surge output from patients. Most scientific specialists have realized that there is a notable relationship between patient's signs and indications with some continual illnesses and the blood sugar charge [5]. In designing high-performance computer-aided analysis systems, enhancing the accuracies of the computing device mastering algorithms is imperative and ensemble data-mining strategies (EDMM), mastering algorithms having an aggregate of more than one base fashions which are the most suggested techniques [6].

Nowadays, a large number of people are suffering from different human disorders such as diabetes, cancer, cardio, neuro digestive, and psychological disorders. A huge amount of data related to medical diagnosis is required, so the system can classify the whole data to make predictions about the diseases and their treatments. Machine learning techniques can be used in diagnosing different diseases viz. diabetes, breast cancer, heart problems, skin problems, Alzheimer's disease etc [7]. Diabetes mellitus is regarded as a continual disorder in which the body of affected person is incapable to produce, use and save glucose, which is a structure of sugar. This is a lifelong disorder that affects the ability of human's body to utilize the electricity located in meals [7]. Though various AI methods have been used for diabetes prognosis and classification with the aid of researchers, the most necessary strategies are K-mean classifier, Fuzzy Logic systems, Support Vector Machine, Artificial Neural Network (ANN), and many others [8], which requires countless strategies and algorithms to extract hidden facts from biomedical datasets. A research work was once carried out to classify diabetes instances and they bought 78.4% classification accuracy with 10-fold cross-validation (FCV) the usage of Evolving Self-Organizing Map [9].

2.1 CLASSIFICATION TECHNIQUES OF DATA MINING

Classification is one of the data mining techniques used to predict diseases based on the attributes of the datasets. The datasets are classified into training and test data and then used to classify new data [10]. Classification is the most commonly used method for diagnosis, estimation and optimization in the healthcare industry. Some of the most common classification methods used are:

Decision Tree (DT): DT build into a tree model structure in the database. This is used in machine learning for prediction, classification and detection by discovering knowledge from a set of data. Clinical data are used to further predict the likelihood that a person can have it in years to come or not [11].

Naives Bayes Classifier: Naïve Bayes is a data mining classification method that is used as a classifier. If a sample belongs to a specific class, this classifier is used for probability estimation. Naïve Bayes' output is high precision and fastest to train results. Typically, it used on very large datasets. The Naïve Bayes Algorithm is a sequential probabilistic algorithm, following estimation, execution, classification and prediction steps.

K-Nearest Neighbour (K-NN): This uses methods for classification and regression and is an easy method to use. The performance depends on whether the classification or regression uses k-NN.

Bayesian Classifier: [12] Classifier is used in healthcare. This classifier is based on probability to diagnose medical data and further predict or classify. Knowledge can be discovered from a set of data [13]. Data that are continuous or categorical can be used to identify the labels.

Artificial Neural Networks (ANN): ANN is a biological neural network with a special parallel learning and information processing functionality. The relation and strength of the network defined linear and non-linear of the activation function. The network comes in layer, which are the input, output and hidden layers. Input layers takes in the data, that is been process by the hidden layer and generate information through the output layer for decision making layer [14].

Support Vector Machine (SVM): its use for classification and regression to determine data in a controlled method of learning. By classifying, it splits into hyperplane. Implementation is simpler; it splits into two hyperlane / line grou ps. SVM can automate processes that make it more effective and efficient. This is mostly applicable to healthcare sector for identification, prediction and classification [15]. SVM also perform on linear programming and mathematical functions to solve problems [16].

2.1 RELATED WORK

Several other research studies have contributed to the

development and improvement of diabetes diagnostic and prediction but limited to some challenges in early detection and prediction of this diseases.

Table 1: Related works of various researcher

References	Year	Title	Classifier Techniques	Research Gap
[17]	2019	Prediction of Diabetes using Ensemble Techniques	Voting ensemble classifier	The researcher only considered voting based ensemble techniques to evaluate different classifier
[18]	2018	An Ensemble Classifier for the Prediction of Heart Disease	KNN, Decision Tree, Naïve Bayes	The research was limited to heart diseases
[19]	2017	Predicting Diabetes in Medical Datasets using Machine Learning Techniques	Decision tree, k-NN and Naïve Bayes classifier.	The application can't work anywhere at any time. It was restricted to only the medical personnel.
[20]	2014	Decision trees and multi-level ensemble classifiers for neurological diagnostics	Decision trees and multi-level ensemble classifiers	Detection of diabetes at earlier stage was not considered.
[21]	2011	A fuzzy classification system based on Ant Colony Optimization for diabetes disease diagnosis	Fuzzy classification system based on Ant Colony Optimization.	Diabetes Patients are only diagnosed but cannot detect diabetes patients at early stage.

From Table 4, various researchers acknowledged the fact that there is a demonstrated need for prediction and diagnosis in Diabetes patient. Equally and importantly accepted is the need to early predict patient with this diseases better healthcare services.

3.0 METHODOLOGY

The system diagnose patient for Diabetes using ensemble classifiers; Decision tree, Naïve Bayes and KNN was used to predict and diagnose patient for diabetes. Datasets collected from UCI was extracted and pre-processed to remove noise, and then the selected classifiers were applied to it into class label which is then further evaluated with some performance measures. Figure 1, shows the system model for the development of an early diabetes diagnosis prediction system of a patient.

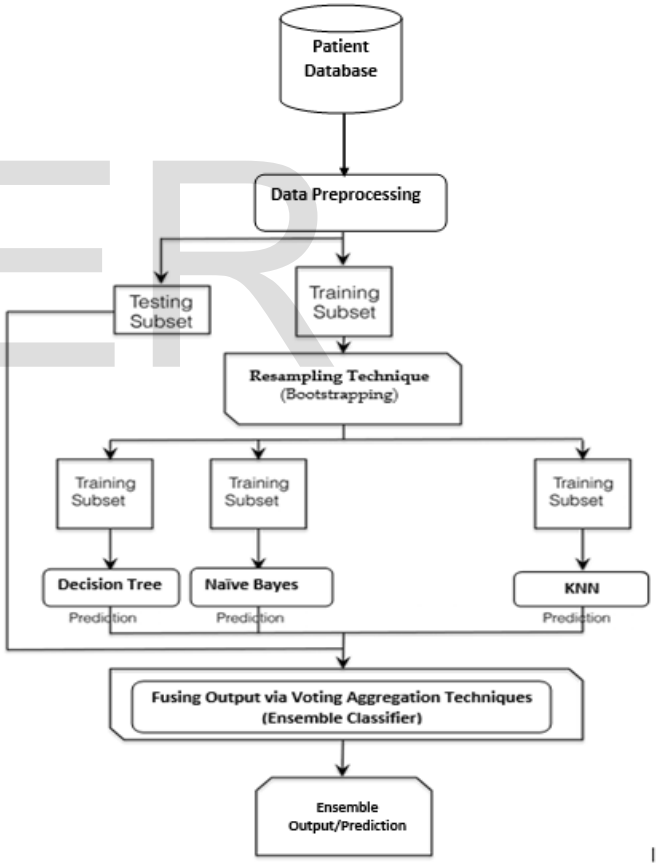

Figure 1: System model for the Development of an intelligent Diabetes Diagnosis prediction System

3.1 Ensemble Classifier

An ensemble classifier combines different machine learning classifier techniques and then perform voting aggregation on the data for prediction or classification. This research paper have considered three classifiers, which are Decision Tree, Naive Bayes and KNN. Voting aggregation is performed on the three classifiers to generate a fusing output in order to

predict the diseases accurately.

The ensemble of classifiers is performed using majority voting rule which is suitable for datasets having two class labels(0 and 1)Majority voting also known as plurality voting uses the technique of highest number of votes for classifying a new set of data instances. We have the voting classifier to be applied on the three techniques to predict the class labels. The class label output will be the most significant of the class labels that are predicted by the single classifier techniques. The ensemble classifier has a better performance in combining the strength of each classifier which leaving out the weakness of individual and better increase the accuracy of the system model in predicting diabetes diseases.

3.2 Data Selection and Processing

In computing device learning data mining and, data collection is a technique where the most applicable information is chosen from a particular area to extract informative values and to promote studying in that field. In this study, we used diabetes dataset with nine attributes to predict a patient's symptoms of gestational diabetes. This dataset is a sample dataset from the UCI server. Based on historic statistics saved in the dataset, such as age, physique mass index, blood stress and the range of pregnancies, the classifiers are skilled to decide whether the diabetes test is tremendous or terrible for an individual.

The complete details of all the eight attributes are listed in Table 2 below.

Table 2: PIMA Dataset Description.

S/No	Attribute/Description	Type
1	Number of times pregnant	Numeric
2	Plasma glucose concentration	Numeric
3	Blood pressure(Diastolic)	Numeric
4	Triceps skin fold thickness(mm)	Numeric
5	2-Hourseruminsulin	Numeric
6	Body mass index(kg/m2)	Numeric
7	Diabetes pedigree function	Numeric
8	Age (years)	Numeric
9	Class Variable (True or False)	Nominal

Data pre-processing is a technique of machine learning that comprises of converting raw data into a logical or comprehensible format. The real world data is mostly incomplete, inconsistent, unreliable, redundant and having missing values etc. The dataset contains 768 records and 9 parameters. 44 records were removed during data cleaning process due to inappropriate value recorded in the dataset. Cases like skin thickness less than 10mm, glucose level less than 1 and blood pressure less than 1 was considered as inappropriate data from the dataset. We then have 80% of the data, to have a total of 724 clean records which were used for training the Machine Learning Model while the remaining 20% was used for testing. The set of pertinent feature vector fed to the classifier help it learn more accurately in a shorter span of time.

3.3 System Design & Modelling

The developed solution is constructed on Django internet framework. Django is a high-level net platform for Python that promotes rapid improvement and clean, pragmatic architecture. Built by using skilled developers, it takes care of a great deal of the problem of Web development. It is an open source. Other technologies used in the system development include HTML5 (Hyper Text Markup Language), CCS3 (Cascading Style Sheets) JavaScript and MySQL

Patient Diagnose: This interface provides patients with an interface to input their biodata information for diagnosis. Figure 2, shows the diabetes diagnosis interface for the user input and result analysis.

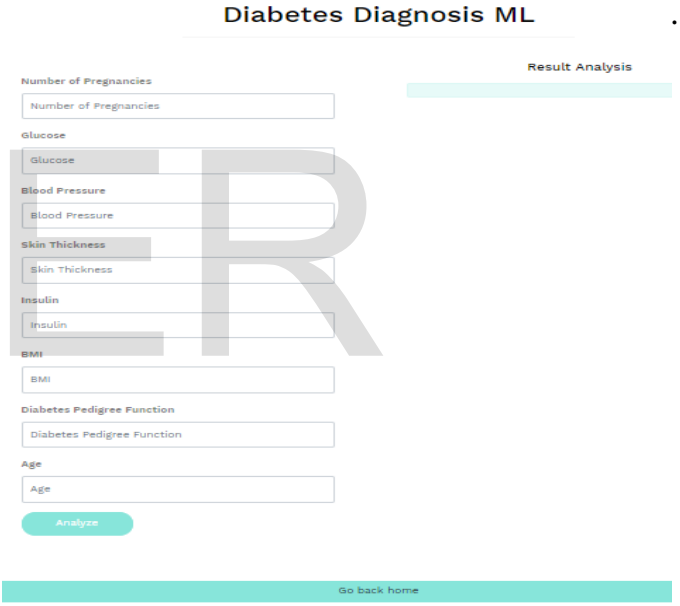

Figure 2: Patient Input interface

Response Chat: Patient can get prompt response chat from the system to solve some immediate problems related to diabetes

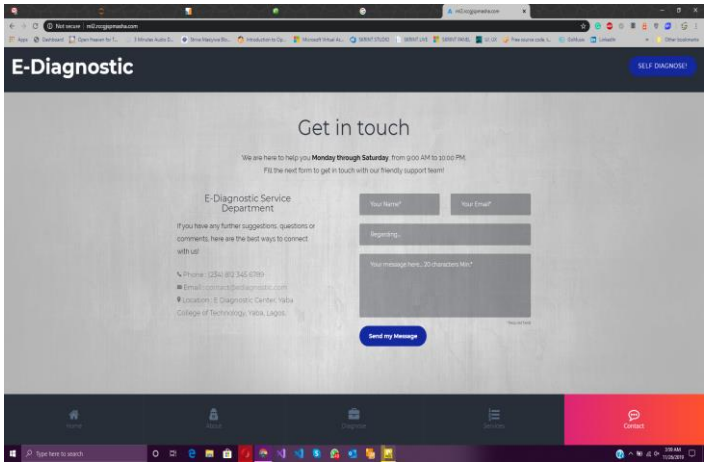

Figure 3: User quick response interface

Self-Diagnosis Page: Patient can input some parameter related to the diagnosis of the diabetes to know level of severity.

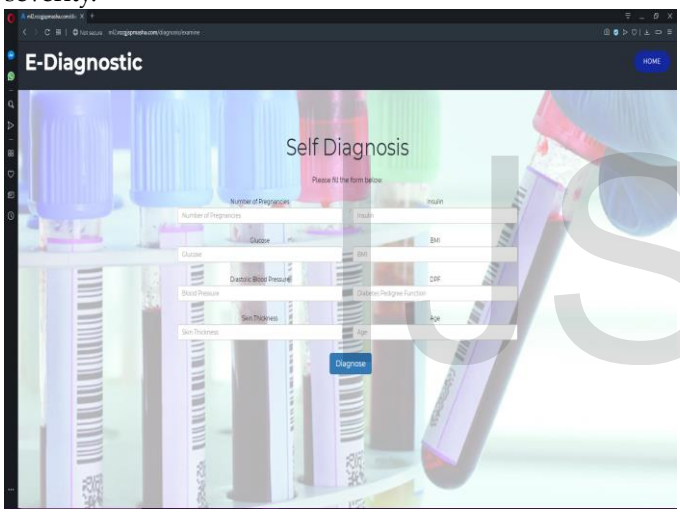

Figure 4: E-Diagnosis Interface

3.4 Sequence Diagram

This shows the interactions of the user with the application and the machine learning Engine. Figure 5 explains the activities that takes place in diagnosis and result of the analyzed symptoms. The application serves as an intermediary between the user and the Machine Learning Engine to takes in the feature parameters and output to generate from it.

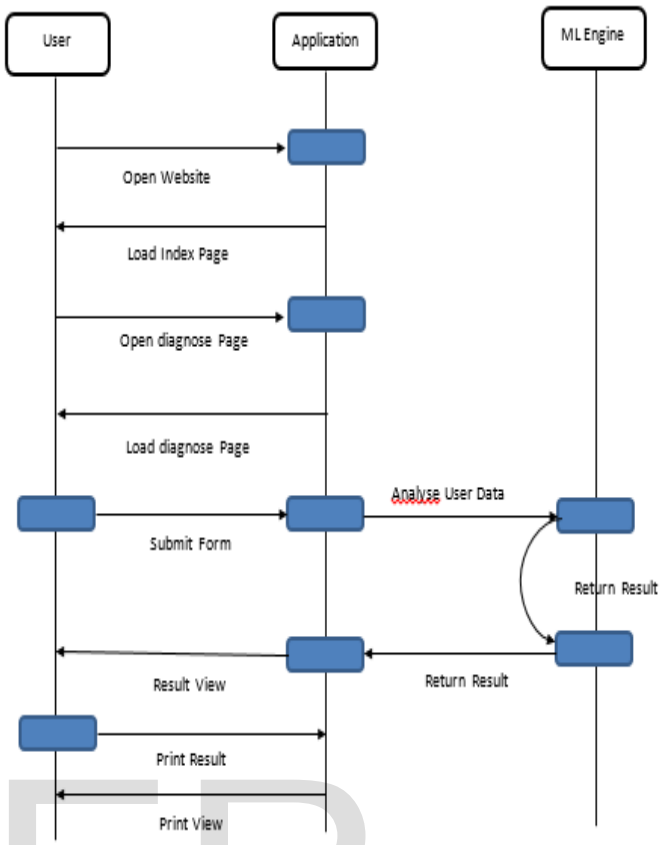

Figure 5: Intelligent Diabetes Diagnosis System Sequence Diagram

4.0 RESULT AND DISCUSSION

The solution was designed and fully implemented using Pima Indians Diabetes Database provided by the UCI Machine Learning Repository (famous repository for machine learning data sets) for training and testing. After several test performed on the designed model in Jupiter notebook. Figure 6 below shows that sample of the dataset from PIMA UCI Machine learning repository data.

s/n	Pregnan	Glucose	BloodPr	SkinThic	Insulin	BMI	Diabetes	Age	Outcome
1	6	148	72	35	0	33.6	0.627	50	1
2	1	85	66	29	0	26.6	0.351	31	0
3	8	183	64	0	0	23.3	0.672	32	1
4	1	89	66	23	94	28.1	0.167	21	0
5	0	137	40	35	168	43.1	2.288	33	1
6	5	116	74	0	0	25.6	0.201	30	0
7	3	78	50	32	88	31	0.248	26	1
8	10	115	0	0	0	35.3	0.134	29	0
9	2	197	70	45	543	30.5	0.158	53	1
10	8	125	96	0	0	0	0.232	54	1
11	4	110	92	0	0	37.6	0.191	30	0
12	10	168	74	0	0	38	0.537	34	1
13	10	139	80	0	0	27.1	1.441	57	0
14	1	189	60	23	846	30.1	0.398	59	1
15	5	166	72	19	175	25.8	0.587	51	1
16	7	100	0	0	0	30	0.484	32	1
17	0	118	84	47	230	45.8	0.551	31	1
18	7	107	74	0	0	29.6	0.254	31	1
19	1	103	30	38	83	43.3	0.183	33	0
20	1	115	70	30	96	34.6	0.529	32	1
21	3	126	88	41	235	39.3	0.704	27	0
22	8	99	84	0	0	35.4	0.388	50	0
23	7	196	90	0	0	39.8	0.451	41	1
24	9	119	80	35	0	29	0.263	29	1
25	11	143	94	33	146	36.6	0.254	51	1
26	10	125	70	26	115	31.1	0.205	41	1
27	7	147	76	0	0	39.4	0.257	43	1
28	1	97	66	15	140	23.2	0.487	22	0
29	13	145	82	19	110	22.2	0.245	57	0
30	5	117	92	0	0	34.1	0.337	38	0
31	5	109	75	26	0	36	0.546	60	0

Figure 6: Sample of the dataset

The prepared model was integrated into a Python Web Framework, Django Framework and hosted on a server for test. Machine Learning systems generates a model or pattern for calculating the result by going through the dataset used during training and get smarter on getting exposed to more clean data. Since the model deducts its pattern through the algorithm by analyzing how the model makes its prediction and how various algorithms combined to generate the desired result.

4.1 System Generated Result Screenshot

Figure 7 shows the result of the diagnosis using the ensemble classifier to show the level of severity of the patient

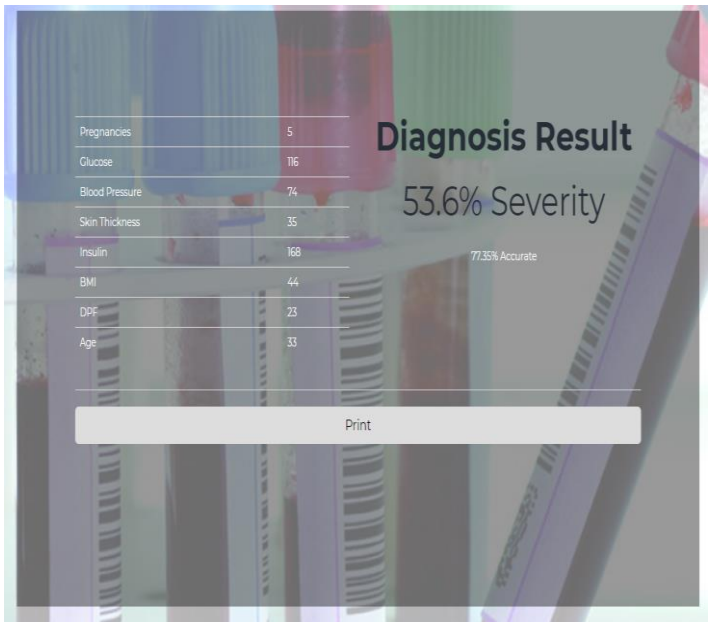

Figure 7: Diagnosis Result Interface

The diagnose result page show the key information needed for deducting the patient or application user state of health in respect to diabetes status. The left pane of the result shows the parameters inputted by the user. These parameters are passed to the Machine Learning Model to make its prediction and return the result. The right pane of the result shows the parameter returned by the Machine Learning Model after analyzing those parameters provided by the user. This shows the Severity level of the patient when it's been diagnosed using the parameters in the left pane. The higher the severity percentage, the higher the probability of diabetes. The Accuracy Percentage shows the Machine Learning Prediction. This can be improved by training the model with more clean data.

4.2 Evaluation and Result findings

This study was limited to three performance measure which include accuracy, Specificity and sensitivity. Ensemble learning can be inferred by integrating multiple models to help improve the performance of machine learning. This methodology makes it possible to produce better predictive results than a single model.

True positives (TP): An individual is predicted to have the disease and the result generate is true

True negatives (TN): An individual is predicted without the disease and the result generate is false

False positives (FP): An individual is predicted to have the disease and the result generate is false

False negatives (FN): An individual is predicted without the disease and the result generate is true

True Positive, True Negative, False Positive, False Negative are parameters used to evaluate the performance of classifier against specificity, sensitivity and F1 score. Sensitivity

identifies the people having diabetes disease, Specificity correctly identifies the people without the disease, F1-score measures the performance of the test for positive class and Precision measures the number of correct positive results divided by the number of positive results predicted by the classifier.

The formula used to calculate the Accuracy, F1-Score, Sensitivity, and Precision are showed below:

Accuracy =

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

F1-Score =

$$\frac{2TP}{2TP+FP+FN}$$

Sensitivity =

$$\frac{TP}{TP+FN}$$

Precision =

$$\frac{TP}{TP+FP}$$

The research evaluated the classifiers' that were considered in this work with the proposed ensemble classifier. The comparison of the evaluation with the classifiers' is showed in Table 3 below.

Table 3: Evaluation Performance Result of diabetes diseases

Classifier Techniques	TP	FN	FP	TN	Accuracy	F1-Score	Sensitivity	Precision
Decision Tree	240	50	67	223	79.82	80.40	82.76	78.17
Naïve Bayes	255	49	35	241	85.52	85.86	83.88	87.93
KNN	255	55	35	235	84.48	85.00	82.26	87.93
Proposed Ensemble Classifier	280	20	12	268	94.48	94.59	93.33	95.89

In this study, we used classification algorithms Naïve Bayes, Decision Trees and k-NN as ensemble classifier for predicting diabetes. The result showed that ensemble classifier is having the higher rate of accuracy about 94.48% followed by Naive Bayes (85.2%), KNN (84.48%) and Decision Tree (79.82%). Hence, an ensemble classifier approach helps in the prediction of diabetes disease with greater accuracy than that of individual classifiers considered in this study.

5.0 CONCLUSION

The significance of this research lies in the use of ensemble classifier to diagnose diabetes. Several existing systems focus on ML (Machine Learning) Model or at most two but Ensemble learning helps improve machine learning results by combining several models. This approach allows the production of better predictive performance compared to a single model implemented in other solutions. Based on this research work we can conclude that using an ensemble classifier generates a more accurate result than that individual classifier. This paper therefore provides a solution to early predict diabetes patient using ensemble classifier, this will help the health care practitioner to identify people at higher risk of having a diabetes disease in order to reduce the morbidity rate based on risk factors thereby increasing the quality of life.

ACKNOWLEDGMENT

The authors are indeed grateful to Yaba College of Technology for the provision of an enabling working environment. We acknowledge the HOD (Mrs. Adetoba B.T) of Computer Technology Department, Yaba College of Technology and colleagues in the Department. Thank you.

REFERENCES

[1] C. S. Frandsen, T. F. Dejgaard, and S. Madsbad, "Non-insulin drugs to treat hyperglycaemia in type 1 diabetes mellitus," *Lancet Diabetes Endocrinol.*, vol. 4, no. 9, pp. 766–780, 2016.

[2] F. M. Silva, C. K. Kramer, J. C. de Almeida, T. Steemburgo, J. L. Gross, and M. J. Azevedo, "Fiber intake and glycemic control in patients with type 2 diabetes mellitus: a systematic review with meta-analysis of randomized controlled trials," *Nutr. Rev.*, vol. 71, no. 12, pp. 790–801, 2013.

[3] C. A. Waldron, S. K. El-Mofty, and D. R. Gnepp, "Tumors of the intraoral minor salivary glands: a demographic and histologic study of 426 cases," *Oral Surgery, Oral Med. Oral Pathol.*, vol. 66, no. 3, pp. 323–333, 1988.

[4] N. Cho, J.E. Shaw, S. Karuranga, Y. Huang, J.D. da Rocha Fernandes, A.W. Ohlrogge, & B. Malanda, "IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045". *Diabetes research and clinical practice*, vol.*138*, pp. 271-281 2018

[5] T. A. Rashid, S. M. Abdullah, and R. M. Abdullah, "An intelligent approach for diabetes classification, prediction and description," in *Innovations in Bio-Inspired Computing and Applications*, Springer, 2016, pp. 323–335.

[6] P. K. Srimani and M. S. Koti, "Medical diagnosis using ensemble classifiers-a novel machine-learning approach," *J. Adv. Comput.*, vol. 1, pp. 9–27, 2013.

[7] S. Vanaja and K. Rameshkumar, "Performance Analysis of Classification Algorithms on Medical Diagnoses-a Survey.," *JCS*, vol. 11, no. 1, pp. 30–52, 2015.

[8] A. G. Karegowda, V. Punya, M. A. Jayaram, and A. S. Manjunath, "Rule based classification for diabetic patients using cascaded k-means and decision tree C4. 5," *Int. J. Comput. Appl.*, vol. 45, no. 12, pp. 45–50, 2012.

[9] A. Feizollah, N. B. Anuar, R. Salleh, and A. W. A. Wahab, "A review on feature selection in mobile malware detection," *Digit. Investig.*, vol. 13, pp. 22–37, 2015.

[10] P. Rani and K. Kaur, "Improve the Efficiency of Classification Algorithm in Data Mining." Lovely Professional University, 2017.

[11] N. Sharma and H. Om, "Data mining models for predicting oral cancer survivability," *Netw. Model. Anal. Heal. Informatics Bioinforma.*, vol. 2, no. 4, pp. 285–295, 2013.

[12] R. Armañanzas, C. Bielza, K. R. Chaudhuri, P. Martinez-Martin, and P. Larrañaga, "Unveiling relevant non-motor Parkinson's disease severity symptoms using a machine learning approach," *Artif. Intell. Med.*, vol. 58, no. 3, pp. 195–202, 2013.

[13] B. Zheng, S. W. Yoon, and S. S. Lam, "Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms," *Expert Syst. Appl.*, vol. 41, no. 4, pp. 1476–1482, 2014.

[14] S. Gupta, D. Kumar, and A. Sharma, "Data mining classification techniques applied for breast cancer diagnosis and prognosis," *Indian J. Comput. Sci. Eng.*, vol. 2, no. 2, pp. 188–195, 2011.

[15] P. J. García-Laencina, P. H. Abreu, M. H. Abreu, and N. Afonoso, "Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values," *Comput. Biol. Med.*, vol. 59, pp. 125–133, 2015.

[16] B. Devi, K. Rao, S. Setty, and M. Rao, "Disaster prediction system using IBM SPSS data mining tool," *Int. J. Eng. Trends Technol.*, vol. 4, pp. 3352–3357, 2013.

[17] Prema N. S., Varshith V. and Yogeswar J. "Prediction of diabetes using ensemble techniques" International Journal of Recent Technology and Engineering (IJRTE), 2277-3878, Volume-7, Issue-6S4, April 2019

[18] R. A. Kurian and K. S. Lakshmi, "An ensemble classifier for the prediction of heart disease," *Int. J. Sci. Res. Comput. Sci.*, vol. 3, no. 6, pp. 25–31, 2018.

[19] U. A. Zia and N. Khan, "Predicting diabetes in medical datasets using machine learning techniques," *Int. J. Sci. Eng. Res. Vol.*, vol. 8, no. 5, 2017.

[20] H. Jelinek, J. Abawajy, A. Kelarev, M. Chowdhury, and A. Stranieri, "Decision trees and multi-level ensemble classifiers for neurological diagnostics," *Aust. J. Med. Sci.*, vol. 1, no. 1, pp. 1–12, 2014.

[21] M. F. Ganji and M. S. Abadeh, "A fuzzy classification system based on Ant Colony Optimization for diabetes disease diagnosis," *Expert Syst. Appl.*, vol. 38, no. 12, pp. 14650–14659, 2011.