Evaluation of Machine Learning Algorithms for Early-stage Prediction of Diabetes

[1]Aminu Uba and [2]Abubakar Muhammad Miyim

[1]Department of Computer Science
Federal University Dutse

[2]Department of Information Technology
Federal University Dutse

Corresponding Author: aminuubaa@gmail.com

Abstract
Diabetes is a chronic and long-term health condition characterized by a lack of the required insulin in the body or the body's inability to utilize the produced insulin. It is usually asymptotic at an early stage. As a killer disease, diabetes is capable of causing different life-threatening complications from eye disorder to renal and kidney failure, etc. To avoid certain complications that may occur if the disease is left untreated for a very long time there is a need to develop an effective diagnostic strategy to detect the disease at an early stage. Early diagnosis will help expedite the treatment cost and life-threatening complications associated with this disease. Machine Learning (ML) techniques have been developed to help clinicians diagnose different kinds of life-threatening diseases. In this work, we employed the use of machine learning model for the early onset prediction of diabetes using the ensemble method. To achieve this, six different base classifiers were identified, trained, and evaluated to select the best performing classifiers to form an ensemble. The performance of the base classifiers and our proposed ensemble method was compared. The result shows that our proposed ensemble method outperforms all the base classifiers and the state-of-the-art result on diabetes prediction with an accuracy of 0.98 and an area under the curve (AUC) of 0.97. This means that our ensemble model was able to clearly distinguish between the two classes in our dataset.

Keywords: Machine learning, Diabetes, Ensemble, KNN, SVM, Decision Tree, Random Forest

INTRODUCTION

Diabetes mellitus is a very common disease characterized by excess sugar (glucose) in the bloodstream as a result of the pancreas' failure to produce enough insulin or the cells in the body have become resistant to insulin. This disease which is prevalent affects the ability of the human body to utilize the energy present in food [1].

According to World Health Organization, in 2014, 8.5% of adults aged 18 years and above had diabetes, while 1.6 million death is caused by diabetes. However, in lower-middle-income countries, the premature mortality rate across both periods increased [10]. This means that diabetes prevalence has been rising

more rapidly in low and middle-income countries than in high-income countries.

In Africa south of the Sahara, the number of people living with diabetes in the region according to the International Diabetes Federation (IDF) report is estimated at more than 19 million people [5]. This is expected to increase to about 143% by the year 2045, the highest predicted increase compared to other regions. Among the African countries, Nigeria has the second-largest number of people living with diabetes which is estimated at 2.7 million after South Africa with 4.6 million.

Globally, is a very well-known disease that poses a very crucial challenge in both developed and developing countries. It is generally categorized into three distinct types, Type I diabetes accounts for about 95% of all diabetes cases, and occurs as a result of the pancreas' fails to produce the required insulin. Type II diabetes occurs when the body is unable to utilize the produced insulin properly due to insulin resistance. Another type of diabetes is Gestational diabetes which mostly affects women who are not diabetic but were diagnosed with high presence of glucose level during/after pregnancy.

Machine learning techniques have proven to be one of the most widely used methods of data mining techniques to diagnose and detect various diseases such as diabetes, malaria, cancer, etc. This paper is meant to evaluate the effectiveness of individual machine learning algorithms and an ensemble method, to develop a generalized framework for early-stage diabetes prediction.

RELATED WORKS

In their work titled "A Predictive Model for Diabetes Using Machine Learning Techniques" [2] compare three machine learning algorithms, Decision tree, K-Nearest Neighbor, and Artificial Neural Network. The result indicated that Artificial Neural Network produced the highest accuracy of 97.40% followed by the decision tree algorithm with an accuracy of 96.10% and then the K-Nearest Neighbors algorithm with 88.31% accuracy. The algorithms perform relatively well but one of the downsides of their work is that most of the risk factors can't be solely relied upon to predict the early onset of diabetes as they failed to capture most of the symptoms associated with the early onset of the disease.

Islam et al. [6] propose four (4) Machine Learning models to predict the likelihood of Diabetes at the Early Stage. They found that among the four ML models, the Random Forest classifier with a cross-validation accuracy of 97.4 stood out to be the highest and therefore outperform the other three models, which set the scene for early-stage prediction of diabetes.

Xue et al. [13] compared three different machine learning algorithms SVM, Naïve Bayes, and LightGBM for diabetes prediction in their research titled "Research on Diabetes Prediction Method Based on Machine Learning". Support Vector Machine was adjudged to be the best with an accuracy of

96.54% followed by the Naïve Bayes with an accuracy of 93.27% and LightGBM with an accuracy of 88.46% was the least.

In the work of Salum et al. [9] three (3) different machine learning models SVM, KNN, and Decision Tree were compared. SVM was reported to have an outstanding result of 90.23% accuracy, followed by KNN with 75.97% and Decision Tree with 75.32% accuracy after a considerable amount of tuning. However, this research employed the use of PID (Pima Indians Diabetes Dataset) for training and testing the machine learning classifiers.

Research Gap

Though researchers have started to work on early diabetes detection, it is still early to state that much has been done in this area. However, to develop a generalized framework for early diabetes prediction, it is interesting to employ an ensemble learning algorithm for some selected machine learning classifiers for the early prediction of diabetes.

METHODOLOGY

The existing method of diabetes prediction mostly relies on using a single model to make predictions. One of the problems with this approach is that, we are not too sure if the single model would be the best predictor and whether it can be able to generalize well to unseen data remains another hurdle. In this work, a machine learning techniques using the ensemble method were used to develop a generalized framework for the early-stage prediction of diabetes

The method begins with data collection and preprocessing, model development, selection, and evaluation. The selected classifiers and the ensemble learning algorithm are then trained and tested using 10 -fold cross-validation. The results are computed and evaluated to identify the best learning algorithm for early diabetes prediction.

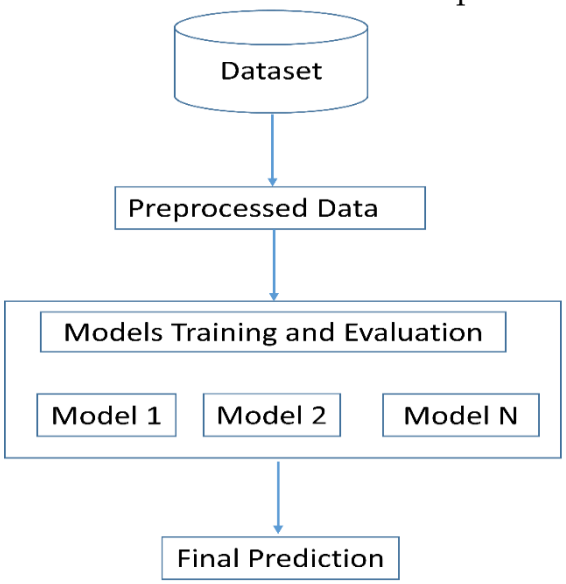

Figure 1 Overview of the proposed method

Data Collection

In this research, we used the publicly accessible diabetes dataset from the University of California Irvine Machine learning repository. The dataset consists of 520 samples and 17 attributes; one numeric attribute named age and 16 categorical attributes and one output attribute named class. Class attributes contain two classes, positive and negative indicating diabetic and non-diabetic patients respectively.

Table 1 Dataset Description

S/N	Attribute(s)	Values
1	Age	16 – 90
2	Gender	Male, 0. Female
3	Polyuria	Yes, 0. No
4	Polydipsia	Yes, 0. No
5	Sudden weight loss	Yes, 0. No
6	Weakness	Yes, 0. No
7	Polyphagia	Yes, 0. No
8	Genital Thrush	Yes, 0. No
9	Visual blurring	Yes, 0. No
10	Itching	Yes, 0. No
11	Irritability	Yes, 0. No
12	Delayed healing	Yes, 0. No
13	Partial paresis	Yes, 0. No
14	Muscle stiffness	Yes, 0. No
15	Alopecia	Yes, 0. No
16	Obesity	Yes, 0. No
17	Class	Positive, 0. Negative

Data Pre-processing

Machine learning algorithms are usually data-sensitive, if we need a better performance accuracy from the learning algorithms we need to feed the algorithms with the right data. In this work, different data processing techniques like handling missing values, target class transformation, principal component analysis, feature scaling, and data oversampling were applied to our dataset.

Models Development

The selection of these base-level classifiers depends on the performance accuracy and how they're widely used in the context of disease prediction. In this work, six different machine learning classifiers were trained and tested out of which two best performing models were selected to form the ensemble. Following is a short description of the base classifiers.

Logistic Regression

Logistic regression allows us to classify our data by modeling our prediction, that is $P(y = 1 \mid x;\theta)$ using a logistic function dependent on x, θ.

K-Nearest Neighbor

K-Nearest neighbor's classifier works off the assumption that examples close in distance to one another belong to the same class. It uses this assumption to classify examples by looking at the k nearest neighbors for every example and assigning the most common class of its neighbors [3].

Support Vector Machine
We use both linear and kernel SVM models as part of our research. SVM classifier seeks to find a hyperplane separating the labeled data such that the widest margin is created between the labeled data point. One appealing characteristic of SVM is its ability to construct hyperplanes in high to indefinite dimensional spaces [8].

Naïve Bayes
Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying the Bayes Theorem with strong (naïve) independent assumptions between the features. Naïve Bayes assumes that the value of a particular feature is independent of the value of any other feature, given the class variable.

Decision Tree
Decision algorithms learn by making the best split of features in the training set into subsets based on the homogeneity of each new subset. The process goes on until a stopping condition is reached. Two popular measures used to control how the decision tree split node i.e Gini index and entropy were used to evaluate the performance of the tree [7].

Random Forest
Random forest works by establishing a group of decision trees where each decision tree classifies the example, and the class with the most decision tree prediction becomes the model prediction for the example. The significance of random forest lies in the fact that the individual tree models as they are uncorrelated, allow for the model to cover the errors of others [14].

Ensemble Method
An ensemble method was used to evaluate its effectiveness in the early-stage prediction of diabetes. The ensemble is a technique that combines multiple learning algorithms, often referred to as base learners. Each of these base learners or individual learning algorithms solves the same problem independently to obtain a combined model with accurate and reliable decisions that could not have been achieved by using a single model [4].

RESULT AND DISCUSSIONS
In this work, 10-fold cross-validation was used to check the stability of the model's performance. Different performance measures such as Accuracy, Precision, Recall, and F1 score were used to determine the effectiveness of the proposed approach. Six different base classifiers were evaluated to select the best candidates for the ensemble. As indicated in Table 4.1 the performance of the base classifiers was compared to the proposed ensemble method. It could also be seen that three individual classifiers, Random Forest, Support Vector Machine, and Decision tree distinguished themselves from the rest of the classifiers with an accuracy of 0.97, 0.97, and 0.95 respectively.

Two best performing models Random forest and Support Vector Machine were selected to form the ensemble. The result shows that our proposed ensemble method outperforms all the other classifiers with an accuracy of 0.98. An improvement over the current state-of-the-art result in early-stage prediction of diabetes. Figure 4.1 below

shows the graphical representation of the performance accuracy comparison between individual base classifiers and the proposed ensemble method.

For better generalization and to further validate the performance of the base classifiers and the proposed ensemble method, Area under Curve (AUC) was used to compare how well the classifiers were able to differentiate between classes and also how they can perform when presented with new unseen data.

Table 2 Comparison between Classifiers and Ensemble Method

SN	Models	Accuracy
1	Logistic Regression	0.882768
2	K-Nearest Neighbor	0.888688
3	Support Vector Machine	0.974548
4	Naïve Bayes	0.863424
5	Random Forest	0.976508
6	Decision Tree	0.951056
7	**Ensemble (SVM+RF)**	**0.982391**

AUC is an important metric that measures the classifier's ability to distinguish between classes. The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes. Figure 4.2 below shows the ROC_AUC Curve scores of the base classifiers and the ensemble method. It could be seen that the proposed ensemble method and the Random Forest Classifier have the highest AUC score of 0.97 respectively. This means that these two models have better generalization ability compared to the other models. Furthermore, other models perform relatively well with AUC scores ranging from 0.94 to 0.96 except the Naïve Bayes Classifier which has an AUC score of 0.88.

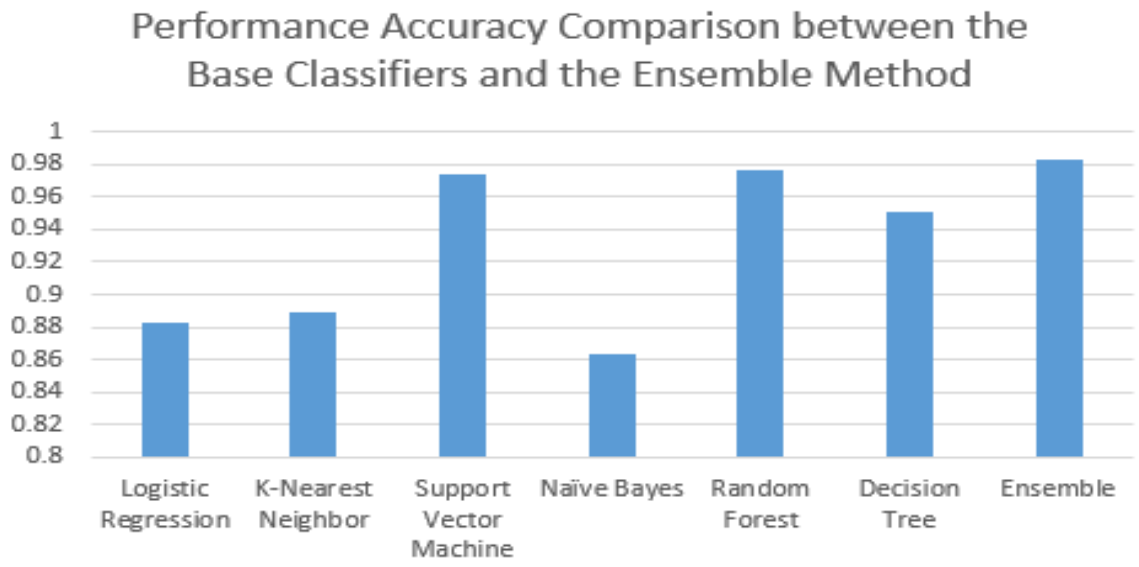

Figure 2 Comparison between base Classifiers and Ensemble Method

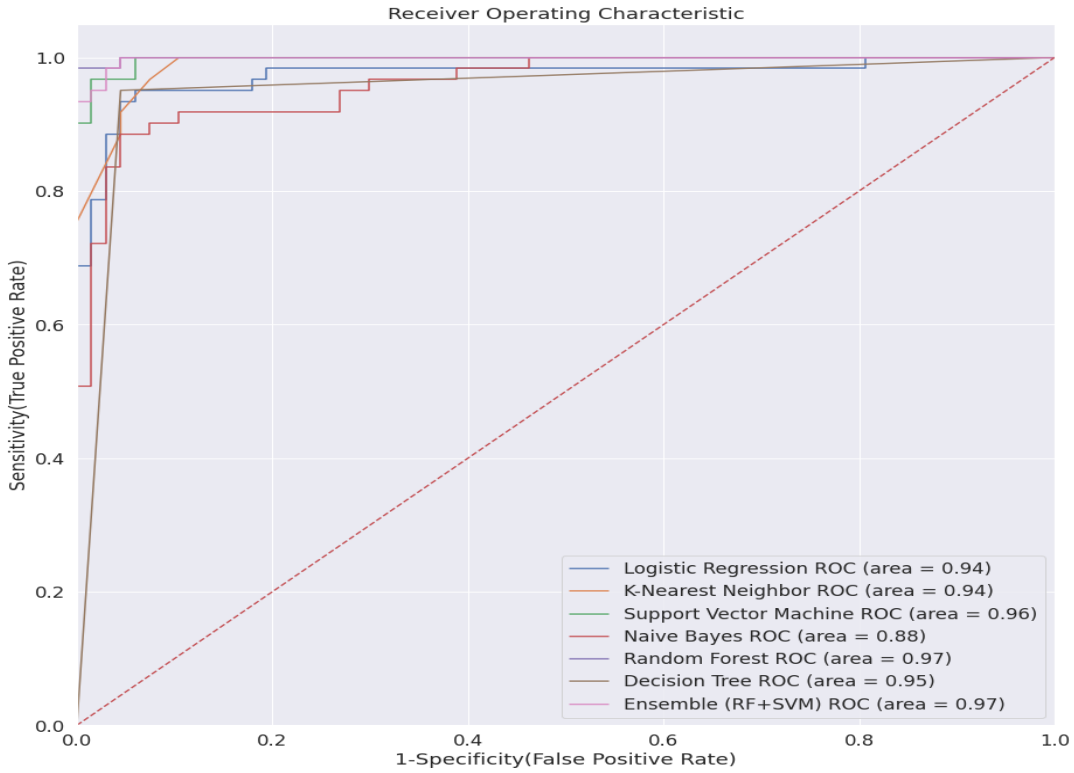

Figure 3 ROC_AUC Curve

CONCLUSION

The focus of this research is on evaluating machine learning classifiers as well as ensemble classifiers for early-stage prediction of diabetes. For this purpose, individual learning classifiers often called base classifiers including Logistic Regression, K-Nearest Neighbor, Support Vector Machine, Naïve Bayes, Decision Tree, and Random Forest was evaluated. Different base classifiers with high-performance accuracy were assembled to evaluate our proposed ensemble classifier. It is observed that combining Random Forest and Support Vector Machine using a voting classifier ensemble outperformed all the other base classifiers with an accuracy of 0.98.

In the future, other machine learning techniques like Artificial Neural Network, and Multi-Layer Perceptron should be evaluated. Similarly, other ensemble techniques like Bagging and Boosting should also be analyzed.

It is also suggested that more data should be collected and used for training. This is because, in training machine learning algorithms, it is often the case that getting more data significantly improves the performance of a learning algorithm.

REFERENCES

[1] Azrar M. A (2018) Data Mining Model Comparison for Diabetes Prediction. International Journal of Advanced Computer Science and Applications, 320-323.

[2] Evwiekpaefe A.E., Abdulkadir, N. (2021) A Predictive Model for Diabetes using Machine Learning Techniques. *The*

African Journal of Information Systems. ISSN 1936-0282

[3] Harrison, O. (2019) Machine Learning Basics with the K-Nearest Neighbors Algorithm. Medium. https://towardsdatascience.com/machine-learning-basics-with- the-k-nearest-neighbors-algorithm-6a6e71d01761.

[4] Hosni M., Abnane I., Idri A., Carillo de Gea J.M., Fernandes Aleman J.A. (2019)

Reviewing Ensemble Classification Method in Breast Cancer. Computer Methods and Program in Biomedicine. 77 pp 89-112.

[5] IDF (2020) What is Diabetes. Retrieved from International Diabetes Federation. http://www.idf.org/aboutdiabetes/what-is-diabetes.html

[6] Islam et al. (2020) Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques. Computer Vision and Machine Intelligence in Medical Image Analysis, Advances in Intelligent System and Computing 922. Doi:https//10.1007/978-981-13-8798-2_12

[7] Nizam, F. A. (2021) Decision Tree Classification Algorithm

[8] Ng, A., Ma, T. (2020) Lecture Notes on Kernel Method and Support Vector.

Machines. Stanford University, Stanford CA

[9] Salum A. H., Malaserene I., Anny L.A. (2020) Diabetes Mellitus Prediction using

Classification Techniques. International Journal of Innovative Technology and

Exploring Engineering (IJITEE)

[10] WHO (2020) Diabetes factsheet. Retrieved from World Health Organization.

http://who.int/newsroom/factsheet/detail/diabetes

[11] Nnamoko N., Hussain A., and England D. (2018) Predicting Diabetes Onset: An Ensemble Supervised Learning Approach. IEEE 978-1-5090-6017-7/18/$31.00

[12] Pradhan, N., Rani, G., Dhaka, V. S., & Poonia, R. C. (2020). Diabetes prediction using artificial neural network. *Deep Learning Techniques for Biomedical and Health Informatics*, *121*, 327–339. https://doi.org/10.1016/B978-0-12-819061-6.00014-8

[13] Xue J., Min F., Ma F. (2020) Research on Diabetes Prediction Method based on Machine Learning.

[14] Yiu, T. (2019) Understanding Random Forest. Medium. https://towardsdatascience.com/understanding-random-forest-58381e0602d2