UNIOSUN Journal of Engineering and Environmental Sciences. Vol. 3 No. 1. March. 2021
DOI: 10.36108/ujees/1202.30.0160

# Development of a Diabetes Melitus Detection and Prediction Model Using Light Gradient Boosting Machine and K-Nearest Neighbour

**Omodunbi, B. A., Okomba, N.S., Olaniyan, O.M., Esan, A. and Adewa, T. A.**

**Abstract:** Diabetes mellitus is a health disorder that occurs when the blood sugar level becomes extremely high due to body resistance in producing the required amount of insulin. The aliment happens to be among the major causes of death in Nigeria and the world at large. This study was carried out to detect diabetes mellitus by developing a hybrid model that comprises of two machine learning model namely Light Gradient Boosting Machine (LGBM) and K-Nearest Neighbor (KNN). This research is aimed at developing a machine learning model for detecting the occurrence of diabetes in patients. The performance metrics employed in evaluating the finding for this study are Receiver Operating Characteristics (ROC) Curve, Five-fold Cross-validation, precision, and accuracy score. The proposed system had an accuracy of 91% and the area under the Receiver Operating Characteristic Curve was 93%. The experimental result shows that the prediction accuracy of the hybrid model is better than traditional machine learning.

**Keywords:** Diabetes disease, Prediction, Machine-learning algorithm, Light Gradient Boosting, K-Nearest Neighbor.

## I. Introduction

Diabetes is a chronic disease with the ability to kill and shorten the lives of human beings. The characteristic of diabetes is that blood glucose is higher than the normal level, which is caused by defective insulin secretion or its impaired biological effects, or both. Diabetes can be divided into two categories, type 1 diabetes (T1D) and type 2 diabetes (T2D). Type 1 diabetes is common among the young, mostly less than 30 years old while type 2 is common among the old. The typical clinical symptoms are increased thirst and frequent urination, high blood glucose levels [1]. The classification and detection of disease is an aspect of the biomedical study that is important due to the growing rate of diverse kinds of disease globally. Diabetes mellitus is a major cause of other deadly ailments such as heart and kidney diseases [2]. Statistics reveals that there are over 425 million [3] individuals that are suffering from this disease, which has led to the death of a lot of people. The detection of diabetes is very vital to the patient since it will aid the patient to take necessary precautionary measure for proper management of their health. Machine learning is an aspect of artificial intelligence that involves instructing the machine with the aid of an algorithm learning from streams of data without human intervention [4]. The objective of machine learning disease classification models is to produce an accurate and precise diagnosis of diseases, which enables quick and less laborious diagnosis of patients at an early stage of the ailment. Algorithms such as logistic regression, Artificial Neural Network (ANN), and decision tree has been used in the past for detection and diagnosis of Diabetes Mellitus. Machine learning algorithm can be used to perform specific task by applying knowledge (pattern) learnt from a given dataset.

The detection of diabetes mellitus is an active field of research and many discoveries have

**Omodunbi, B. A., Okomba, N.S., Olaniyan, O.M., Esan, A. and Adewa, T. A.**
**(**Department of Computer Engineering, Federal University Oye Ekiti, Ekiti State, Nigeria.)
**Corresponding Author**: nnamdi.okomba@fuoye.edu.ng
**Phone Number**:+234..
**Submitted**: 26-Nov-2020
**Accepted**: 4-Mar-2021

been done in this scope of research, below are some of the works previously done.

Anuja *et al*. proposed the use of a Support Vector Machine for the detection of diabetes mellitus. 10-fold classification was performed with support vector machine classifier. The model performance was evaluated using Accuracy score, Sensitivity and Specificity as performance metrics with the values of 78%, 80% and 76.5% respectively [5].

Santhanam and Padmavathi presented a machine learning model for diabetes diagnosis, Support Vector Machine was used, while the K-means and Genetic algorithm were used for Dimensional reduction to enhance model performance. 10-fold Cross validation was carried out on the dataset for model Evaluation [6].

Aiswarya *et al* attempted to find a better and efficient way of detecting diabetes mellitus. Using diabetes dataset with a cross validation approach, the result obtained in the study revealed that accuracy of 74.8% was obtained with J48 whereas the Naïve Bayes gives a 79.5% level of accuracy by using 70:30 split[7].

Aminul and Nusrat proposed the prediction of onset diabetes with the use of a machine learning algorithm. The machine learning algorithms used are Logistic Regression, Naïve Bayes and Random Forest. PIMA diabetes dataset was applied for the model training. Logistic regression had a maximum level of accuracy (78.01%) and an Area Under the Curve of 0. 833[8].

Minyechil *et al*, presented machine learning models for analyzing and predicting diabetes. Support Vector Machine, Decision Tree and ensemble models which consisted of three independent models (support vector machine, Decision Tree and Naïve Bayes) were used. The ensemble model had the highest accuracy of 90.36% [9].

Aishwarya *et al*. presented the performance evaluation of five machine learning model for detecting diabetes. The machine learning models used are K-Nearest Neighbor, Decision Tree, Naïve Bayes, Support Vector Machine, Logistic Regression and Random forest, the pima diabetes dataset was also employed for training all the models. The evaluated performance of these models revealed that Logistic Regression had the best accuracy with 77.6% accuracy and 73.6 % area under the receiver operating characteristic curve. The accuracy was quite low and might lead to high misclassification rates in a real-world scenario [10].

Abdulhakim et al. proposed a machine learning model for detecting diabetes in patients using three-machine learning models viz: Support Vector Machine, K-nearest Neighbor and Decision tree. The Pima dataset was employed in training the models, and the results of the study proved that SVM outperforms decision tree and KNN, with the highest accuracy of 90.23%. A major drawback of this work is the use of accuracy as the only metric for evaluating the model performance [11].

## II. Materials and Methods

The tools used for developing this machine learning classifier is python in Jupiter lab, a lot of libraries in sklearn were adopted during the data preprocessing and modelling stage [12]. Flask was used for deploying the developed model. In this study, the diabetes dataset adopted for training the model has eight features, these features are commonly used to predict the symptoms of diabetes also known as diabetes indicators.

The features in the dataset are age, body mass index, blood pressure and number of times pregnant, the dataset consists of eight

attributes and 768 cases. The source of this dataset is UCI machine learning repository available online [13]. The complete details of all the eight attributes are listed in Table 1.

Table 1: Description of Diabetes Dataset

| S/N | FEATURE | DESCRIPTION |
|---|---|---|
| 1 | Pregnancies | the number of times the patient has been pregnant |
| 2 | Glucose | the blood glucose level on testing |
| 3 | Blood pressure | the diastolic blood pressure |
| 4 | Skin Thickness | the skin fold thickness of the triceps |
| 5 | Insulin | the amount of insulin in a 2 hour serum test |
| 6 | BMI. | the body mass index (BMI) |
| 7 | Diabetes Pedigree Function | the family history of the patient |
| 8 | Age | the age of the patient |
| 9 | Outcome | if the person is predicted to have diabetes or not |

## A. Machine Learning Algorithm

The two machine learning algorithms used in this research are ensemble machine learning Algorithms (Light Gradient Boosting Algorithm) and K Nearest Neighbor (KNN)

### i. Light Gradient Boosting

Light GBM is a gradient boosting algorithm built on decision tree algorithms with high-performance. It is applied in classification, ranking and many additional machine-learning problems. It divides the tree leaf wise with the best fit while other boosting algorithms divide the tree depth wise or level wise relatively differing from leaf-wise. In this light, the LGBM algorithm has the capability of decreasing loss which later results in improved and efficient model exactness. LGBM has a fast training speed, good accuracy, high efficiency and low memory usage. It also supports parallel learning and is capable of handling large-scale data [14].

### ii. K-Nearest Neighbors (KNN)

KNN is a simple and popular machine learning technique which is majorly used for classification tasks [15]. KNN is a type of instance-based learning sometimes identified as a sluggish learner, which essentially aims at approximating the local functions while all evaluation is deferred until classification. It can be helpful to allot weights to the contributions of each of the neighbours. K-Nearest Neighbor is simple to implement, robust to error in the input data and can learn non-linear boundaries.

## B. Model Development Process

The proposed combined model boosts the accuracy of the machine learning classifier, with the ability to predict the current and future medical condition of a patient. The framework is shown in figure 1. It comprises of some important phases which are discussed hereafter:

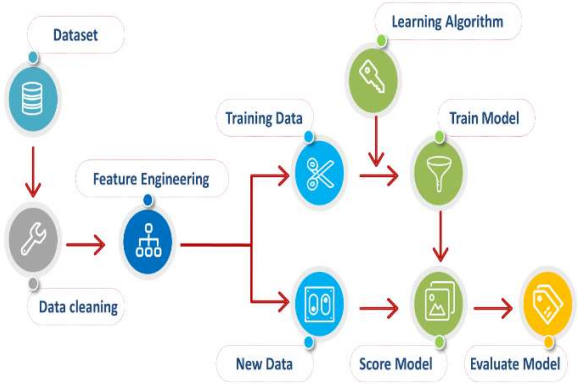


Figure 1: Model Development Process

### i. Data Preprocessing and Data cleaning

Data cleaning is the process of preparing data for analysis by removing or modifying data that is incorrect, incomplete, irrelevant, duplicated, or improperly formatted. The initial step in data cleaning is observing the rate of missing/invalid values in the dataset. The diabetes dataset has five features such as

blood pressure, skin thickness, glucose, insulin and BMI with "0" as a value. These values have to be treated, to amplify the ability of the model learning from the dataset. So, the missing values were replaced with the average values of each column.

### ii. Exploratory Data Analysis

Exploratory Data Analysis (EDA) ensures that the data is valid and does not involve any imbalances in the dataset. EDA also helps to provide data-driven insights before developing a predictive model. In this study, python and pandas library were majorly used for exploratory data analysis with the objective of exploring the diabetes dataset, create visual distributions, identify and eliminate outliers and also to uncover correlations between two features in the dataset.

### iii. Feature Correlation

A **correlation matrix** shows the correlation coefficients that exist between different features in a tabular form. For this dataset, the correlation coefficients between each feature were calculated, Figure 2 denotes the correlation coefficients between the features. There are no redundant features present in this dataset because none of the coefficient value is greater than 0.5. Figure 2 shows the correction between features in the diabetes dataset.

### iv. Feature Engineering

Feature Engineering deals with extracting important features from the dataset before fitting it into machine learning models. Two important features were extracted.

Feature 1: Body Mass Index (BMI) Descriptor: If the BMI Index value is:

a) Less than 18.5: underweight and possibly malnourished.
b) 18.5 to 24: healthy weight range for young and middle-aged adults.
c) 25.0 to 29.9: overweight.
d) Above 30: obese.

Feature 2: Insulin Indicative Range: If insulin level (2-Hour serum insulin (mu U/ml)) is >= 16 and <= 166, it is considered as normal range else it is Abnormal.

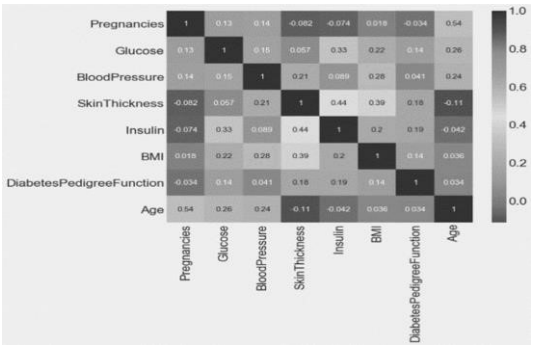


Figure 2: Image displaying the Spearman Correlation Coefficients

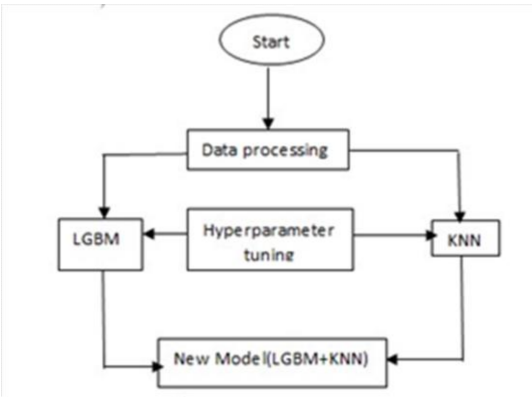


Figure 3: The Process Flow of the Model Aggregation

### v. Model Aggregation

The two models (LGBM and K-Nearest Neighbour) were aggregated with the use of a soft voting classifier, in which the probability vector for each predicted class (for all classifiers) are added up and averaged. The winning class is therefore considered as the highest value. It predicts the class labels based on the predicted probabilities p for classifier, figure 3 shows the model aggregation process,

The soft voting classifier can be represented mathematically as

$$y = a \quad \max_i \sum_{j=1}^{m} w_j p_{ij} \qquad (1)$$

where $w_j$ is the weight that can be assigned to the jth classifier.

e) **Hyperparameter Tuning and Model Generalization**

Random Search Cross Validation was used for finding the best hyperparameter values. Random search is a method that makes use of random combinations of the hyperparameters to search for the best solution. Figure 4 presents a list of hyperparameters and their individual range of values.

```
param_test = {'learning_rate' : [0.01, 0.02, 0.03, 0.04, 0.05, 0.08, 0.1, 0.2, 0.3, 0.4],
              'n_estimators' : [100, 200, 300, 400, 500, 600, 800, 1000, 1500, 2000],
              'num_leaves': sp_randint(6, 50),
              'min_child_samples': sp_randint(100, 500),
              'min_child_weight': [1e-5, 1e-3, 1e-2, 1e-1, 1, 1e1, 1e2, 1e3, 1e4],
              'subsample': sp_uniform(loc=0.2, scale=0.8),
              'max_depth': [-1, 1, 2, 3, 4, 5, 6, 7],
              'colsample_bytree': sp_uniform(loc=0.4, scale=0.6),
              'reg_alpha': [0, 1e-1, 1, 2, 5, 7, 10, 50, 100],
              'reg_lambda': [0, 1e-1, 1, 5, 10, 20, 50, 100]}
```

Figure 4: Sample screenshot of the hyperparameter list and their individual range of values

The list of hyperparameters and their importance in the model development process

i. learning_rate: This hyperparameter determines the performance of each tree on the final result.
ii. n_estimators : This hyperparameter gives the number of tree to be produced
iii. num_leaves: it gives the amount of leaves in a tree, the default value is 31
iv. reg_alpha: regularization

## C. Model Evaluation

The following machine learning metrics are used in this work to evaluate the model performance.

***Accuracy (Acc):*** is one of the machine learning metrics which gives the proportion of corrected labelled samples by the model that it is being tested and is calculated as

$$\text{Accuracy} = \frac{True\ Positive + True\ Negative}{True\ Positive + T \quad Negative + Fa \quad Negative + Fals \quad Positive} \qquad (2)$$

***Precision (P):*** is the amount of positive predictions divided by the whole number of positive class values predicted

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + Fa \quad Positive} \qquad (3)$$

***The F score*:** which is sometimes referred to as the F1 score or F measure, is a measure of the assessments of the model accuracy.

$$F_1 = 2.\frac{precion \times recall}{precion + recall} \qquad (4)$$

## D. Receiver Operating Characteristics

Receiver Operating Characteristic (ROC) Curve is a way to compare diagnostic tests. It is a plot of the true positive rate versus the false positive rate. The *Area Under Curve* value lies between 0.5 to 1 where 0.5 denotes a bad classifier and 1 denotes an excellent classifier. By analogy, the closer the Area under the curve to one, the better the model performance at distinguishing between patients with the disease and no disease[16].

# III. Results and Discussion

The developed model was evaluated and validated through different machine learning methods and metrics to determine how efficient and effective the developed model

can perform and to check if the model generalizes or not.

### A. Model Cross Validation with K-Folds

The performance of the model was validated with 5-fold cross-validation. The sub-sample with the highest accuracy is 94% and the sub-sample with the least accuracy is 87.7%, but the accuracy of the model is 91% and a receiver operating characteristic curve of 93.4%. Table 2 shows the fivefold cross-validation result while Figure 6 depicts the Receiver Operating Characteristics curve with 0.933 AUC.

Table 2: 5-fold Cross Validation results for LGBM+KNN

| Fold | Accuracy | precision | Recall | F1 Score | ROC AUC |
|---|---|---|---|---|---|
| 1 | 0.896 | 0.896 | 0.796 | 0.843 | 0.922 |
| 2 | 0.877 | 0.797 | 0.87 | 0.832 | 0.918 |
| 3 | 0.916 | 0.902 | 0.852 | 0.876 | 0.937 |
| 4 | 0.902 | 0.88 | 0.83 | 0.854 | 0.94 |
| 5 | 0.941 | 0.893 | 0.943 | 0.917 | 0.953 |
| Mean | 0.906 | 0.873 | 0.858 | 0.865 | 0.934 |
| STD | 0.021 | 0.039 | 0.049 | 0.03 | 0.013 |

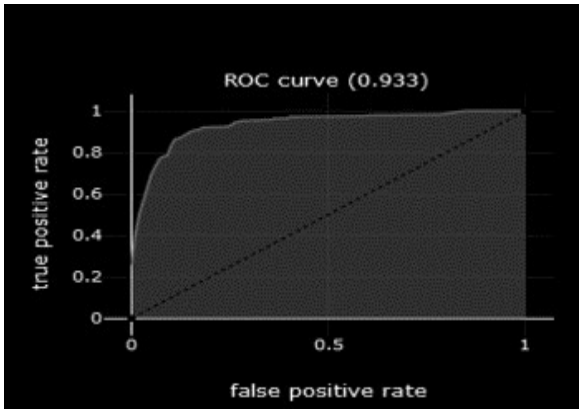


Figure 6: Receiver Operating Characteristics

When compared with other models based on accuracy, the developed LGBM+KNN model had a higher accuracy than the other models under consideration; this is as shown in table 3.

Table 3: Performance comparison with other Models

| Algorithm | Accuracy |
|---|---|
| Logistic Regression | 87% |
| Decision Tree | 83% |
| k-Nearest Neighbor | 82% |
| Random Forest | 85% |
| Multi-Layer Perceptron | 84% |
| Support Vector Machine | 89% |
| Naïve Bayes | 81.8% |
| LightGBM | 89% |
| Hybridized Model(LGBM+KNN) | 91% |

### B. Model Deployment

The developed machine learning model was deployed on a web platform for usability with a flask framework. The system takes four values as input and gives out a categorical value, which can be either Positive (Diabetic) or Negative (Not Diabetic), the web platform was tested with a real-life scenarios and was confirmed working, figure 7 reveals the front-end interface that can be used by a medical practitioner or by an individual.

## IV. Conclusion

In recent times, Diabetes mellitus is one of the leading causes of premature death globally. Hence, the need for early detection and diagnosis. In this study, a machine learning model was developed to predict and detect diabetes mellitus. This was done by combining a gradient boosting algorithm known as the light gradient boosting algorithm (LGBM) with K-Nearest Neighbor to form LGBM+KNN hybrid. The hybridized model incorporates a learning and a predictive process. The model achieved a prediction accuracy of 91%, with a precision of 87% and the area under the receiver operating characteristic curve of 93%. In addition, a web application was developed to classify the patients based on their diabetes level by collecting real-time data from various health care facilities. This application can assist a physician in managing patients with high risk of developing diabetes.


Figure 7: Model User Interface

## References

[1] Iancu, I., Mota, M. and Iancu, E. "Method for the Analysing of Blood Glucose Dynamics in Diabetes Mellitus Patients", *In Proceedings of the 2008 IEEE International Conference on Automation*, Quality and Testing, Robotics, Cluj-Napoca, 2008. doi: 10.1109/AQTR.2008.4588883

[2] Krasteva, A., Panov, V., Krasteva, A., Kisselova, A. and Krastev, Z. "Oral Cavity and Systemic Diseases", *Diabetes Mellitus. Biotechnol. Biotechnol. Equip.* vol. 25, 2011, pp. 2183–2186, doi: 10.5504/BBEQ.2011.0022

[3] https://www.who.int/health-topics/diabetes

[4] Shiliang Sun "A Survey of Multi-view Machine Learning", *Neural Comput. Applic*, vol. 23, no. 7–8, 2013, pp. 2031–2038.

[5] Anuja, K.V. and Chitra, R. "Classification Of Diabetes Disease Using Support Vector Machine", *International Journal of Engineering Research and Applications*, vol. 3, Issue 2, 2013, pp. 1797-1801.

[6] Santhanam, T. and M.S Padmavathi, T.M. "Application of K-means and Genetic Algorithms for Dimension Reduction by Integrating SVM for Diabetes Diagnosis", *Procedia Computer Science*, vol. 47, 2015 pp. 76 – 83.

[7] Aiswarya, l., Jeyalatha, S. and Sumbaly R. "Diagnosis of Diabetes using Classification Mining Techniques", *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, vol.5, no.1, 2015, pp. 1-14.

[8] Aminul, I. and Nusrat, J. "Prediction of Onset Diabetes using Machine Learning Techniques", *International Journal of Computer Applications*, vol. 180, no. 5,2017, pp. 7-11.

[9] Rahul, J., Preeti, M. and Minyechil A. "Analysis and Prediction of Diabetes Mellitus using Machine Learning Algorithm", *International Journal of Pure and Applied Mathematics*, vol. 118, no. 9, 2018, pp. 871-878

[10] Aishwarya M. and Vaidehi V. "Diabetes Prediction using Machine Learning Algorithms**",** *International Journal of Computer Science and Information Technologies,* vol. 7, no. 3, 2016, pp. 1174-1179

[11] Abdulhakim, S.H., Malaserene, I. and Leema, A.A. "Diabetes Mellitus Prediction using Classification Techniques", *International Journal of Innovative Technology and Exploring Engineering (IJITEE),* vol. 9, Issue 5, 2020

[12] Richert, W. and Coelho, L.P. "Building Machine Learning Systems with Python", *Packet Publishing Ltd.,* ISBN 978-1-78216-140-0

[13] UCI repository of bioinformatics Databases, Website: http://www.ics.uci.edu/~mlearn/MLRepository.html.

[14] https://www.analyticsvidhya.com/blog/2017/06/which-algorithm-takes-the-crown-light-gbm-vs-xgboost/

[15] Duda, R.O. and Hart, P.E. "Pattern Classification and Scene Analysis", 1973