Early Identification of Diabetes Mellitus using Parallel and Sequential Ensemble Methods

Kinga Mary Temidayo
Information Technology Department,
Federal University of Technology, Akure, Ondo State, Nigeria.

Bamidele Moses Kuboye
Information Technology Department,
Federal University of Technology, Akure, Ondo State, Nigeria.

Akinbami Emmanuel Ayokunle
Information technology Department,
Federal University of Technology, Akure, Ondo State, Nigeria.

ABSTRACT
Diabetes Mellitus (DM) is a chronic and rapidly increasing health condition, affecting millions worldwide due to factors such as modern lifestyles and inadequate early detection methods. Current clinical diagnostics, while effective, often fail to identify early-stage DM, resulting in delayed treatment and higher risks of severe complications. This study proposes a hybrid ensemble machine learning model that combines both parallel and sequential ensemble methods with forward and backward feature selection techniques to enhance the early prediction of DM. The ensemble methods include J48, Classification and Regression Trees (CART), Decision Stump, Random Forest for parallel ensemble methods, and Gradient Boosting, XGBoost, and AdaBoostM1 for sequential methods. The study utilized a diabetes dataset containing features such as glucose levels, blood pressure, insulin levels, and BMI, applying the ensemble models to improve prediction accuracy. The experimental results showed that Random Forest, from the parallel ensemble methods, achieved a classification accuracy of 100%, significantly outperforming individual classifiers. Similarly, Gradient Boosting, from the sequential ensemble models, also yielded 100% accuracy. The combination of these models through a voting ensemble further enhanced the system's performance, producing superior prediction results with minimal errors. The findings emphasize that combining multiple ensemble techniques with feature selection can dramatically improve predictive performance. This study contributes a robust and scalable model for real-time diabetes prediction that can assist in the timely diagnosis and management of diabetes, potentially reducing global health risks associated with this disease.

General Terms
Identification of diabetes mellitus, ensemble methods.

Keywords
Diabetes Mellitus, ensemble learning, feature selection, machine learning, Early Detection.

1. INTRODUCTION
Diabetes Mellitus (DM) is a kind of disease that prevent the body from obtaining sufficient energy from the meals eaten. It is a chronic disease characterized by unusually high blood glucose levels [1]. This is caused by a deficiency of insulin production or when there is inability to properly utilize the insulin in the body. In recent years, advancements in medical imaging and machine learning have shown great promise in improving the early detection of diabetes mellitus. Convolutional Neural Networks (CNNs), a class of deep learning models, have particularly demonstrated remarkable success in analyzing medical images and identifying patterns indicative of various diseases, including diabetes-related conditions. Early detection in disease management is crucial for enhancing treatment effectiveness and minimizing the adverse effects of diseases on individuals and society. As healthcare continues to emphasize the significance of timely diagnosis, the focus on early detection has gained increasing attention. Identifying diseases promptly allows for the administration of appropriate treatment at the earliest possible stage, thereby improving patient outcomes [2]. When diabetes is detected early, it can be managed [1]. It spreads rapidly, and thus, according to World Health Organization (2020), DM will increase by the end of 2017. It was predicted that DM would affect around 425 million people between the age of 20 to 79 years, and this figure was projected to increase to 629 million by 2045 [3].

Traditionally, the diagnosis of diabetes relies on clinical measures such as fasting blood glucose levels, oral glucose tolerance tests, as well as glycated hemoglobin (HbA1c) levels. These methods, while effective, require invasive procedures and may not always detect early-stage diabetes or pre-diabetes.

Ensemble learning-based systems, an aspect of machine learning, refers to methods that produce several models that are combined effectively to make prediction. Ensemble methods enhance the accuracy and strength of building a prediction model by combining a collection of base classifiers [4].

Both Sequential ensemble techniques and Parallel ensemble techniques are the two types of Ensemble methods. Sequential ensemble methods, such as Adaptive Boosting, create base learners in sequential order (AdaBoost). According to Yiheng and Weidong [5], there are various approaches in ensemble method, including Random Forest, Bagging, AdaBoost, XGBoost, Light, and Stacking. Experimental works reveal that the performance of ensemble learning (Ada boosting) is better than individual learners. The reliance of base learners is promoted by the consecutive production of base learners. The model's performance is improved upon by giving previously misrepresented learners larger weights. In the same ensemble techniques, base learners are generated in a similar format, e.g., random forest. The parallel production of basis learners is used in parallel approaches to support base learner independence. The independence of base learners reduces the mistake caused by the application of averages greatly.

In creation of a predictive model, feature selection is the process of minimizing the number of input variables. In minimizing the computational cost of modelling and, in some situations, improve the model's performance, it's beneficial to reduce number of input variables. As a result, this research suggests a more accurate prediction, which would be achieved by combining parallel and sequential ensemble methods with feature selection strategies to minimize the aforesaid prediction problems.

2. RELATED WORKS

[6] aims to improve the accuracy of diabetes diagnosis applications by employing artificial intelligence techniques. The research focuses on utilizing data mining and metaheuristic algorithms to enhance diagnosis with the objective to diagnose diabetes accurately, by combinationing Harmony search algorithm, genetic algorithm, and particle swarm optimization algorithm with K-means clustering, followed by K-nearest neighbor (KNN) classification. The proposed model achieved an accuracy of 91.65%, surpassing previous approaches. However, the study has limitations, primarily focusing on classification accuracy without considering interpretability of selected features or computational efficiency. [7] explores the application of artificial intelligence (AI), particularly deep learning (DL), in enhancing the diagnosis and screening of diabetic retinopathy (DR). Motivated by the substantial global burden of DR and the necessity for early detection to prevent vision loss, the authors aim to assess the effectiveness of AI techniques in detecting and grading DR from digital fundus photographs or optical coherence tomography (OCT). The authors analyze the current state of AI in DR diagnosis and screening, focusing on DL models applied to retinal images from various imaging modalities. the effectiveness of AI in detecting DR lesions,

Key results indicate that DL algorithms, particularly convolutional neural networks (CNNs), exhibit high accuracy and efficiency in detecting and grading DR, outperforming or rivaling human experts. These algorithms can differentiate between different DR stages and associated conditions like diabetic macular edema (DME). Some identified limitations include standardizing datasets, addressing regulatory considerations, and validating AI-based systems in real-world clinical.

A comprehensive review of AI applications in diabetes management was carried out by [8]. It explored FDA-approved AI/ML-based medical devices for tasks like automatic retinal screening and clinical diagnosis support, emphasizing it potential in enhancing patient self-management. Additionally, the authors evaluate ML models' performance in predicting new-onset diabetes, highlighting its promising but not yet superior results compared to conventional statistical approaches. Challenges such as overfitting and generalizability, underscoring the need for further research to optimize AI's accuracy and applicability in diabetes diagnosis, prevention, and treatment were discussed.

[9] conduct a systematic review aiming to explore the potential of machine learning (ML) and artificial intelligence (AI) in enhancing the detection, diagnosis, and self-management of Diabetes Mellitus (DM) Employing a systematic review approach, 107 relevant articles from Scopus and PubMed databases published within the last six years were selected. It analysis covers various aspects including datasets, preprocessing techniques, feature selection, ML and AI techniques, and performance metrics used in DM research. Despite thorough screening, potential limitations in the search strategy and the focus on ML and AI approaches might impact the generalizability of the findings.

To enhance the explainability of artificial intelligence (AI) applications in healthcare, with focus 0n the diagnosis of type 2 diabetes. [10] address the challenge posed by the complexity of AI technologies by applying seven explainable artificial intelligence (XAI) tools and techniques to different parts of the AI application, including input, processing, and output. These techniques encompass smart technologies, common expression, color management, local interpretable model-agnostic explanation (LIME), classification and regression trees (CART), donut charts, as well as graphical user interface (GUI). Through experimentation, the effectiveness of the approach in improving the transparency, comprehensibility, interpretability, and understandability of AI applications, particularly in diabetes diagnosis were demonstrated. However, the study's focus on diabetes diagnosis and reliance on experimental results for evaluation may limit its generalizability to other healthcare applications. Customization of XAI tools and techniques based on specific application needs could introduce additional complexity, warranting further exploration in future research.

[11] conducted a systematic review on the use of smart devices and machine learning for diabetes management. The study aimed to examine how these technologies can improve blood sugar control, predict risk events, and enhance overall patient care. The authors found that many studies utilized these technologies to address issues such as blood glucose prediction and automatic insulin dosing. Overall, the review underscores the promise of integrating smart technology and AI in improving the quality of life for diabetes patients.

In [12], Generative Pretrained Transformer (GPT) combined with association rule mining to improve the accuracy and interpretability of Type 2 diabetes mellitus (T2DM) diagnosis. It aims to improve diagnostic accuracy and give actionable insights for healthcare practitioners through this integrated approach. Utilizing the Pima Indians Diabetes dataset, it was compared against traditional machine learning models, including LightGBM, using Python and Jupyter Notebook for analysis, NiaARM for rule mining, and SHAP for interpretability. The results show that while NiaARM generated robust predictive rules, LightGBM outperformed the GPT-based model in multiple performance metrics. Disparities in GPT predictions highlighted interpretability challenges.

[13] conducted a comparative study on the use of machine learning (ML) techniques to improve the diagnosis of diabetes, motivated by the high prevalence of the disease and the need for early detection to prevent complications. The study evaluates 15 classification techniques on two datasets: a diabetic clinical dataset (DCA) from Assam in India, and the PIMA Indian diabetic dataset. Key findings include that logistic regression outperformed other algorithms in both datasets. It achieved the highest accuracy and Matthews correlation coefficient (MCC). However, the study's findings are constrained by the datasets' size and nature, potentially affecting generalizability. ML algorithm performance may vary with different datasets, impacting the conclusions. Sharma and Shah [14] likewise conducted a comprehensive review of

machine learning techniques applied in diabetes detection, driven by the escalating prevalence of diabetes mellitus and the necessity for accurate detection methods. The objectives include exploring various algorithms, including supervised, unsupervised, and reinforcement learning methods, and examining the role of deep learning models compared to traditional approaches. The review encompasses discussions on challenges such as data inadequacy, model deployment, and future prospects for enhancing detection methods.

[15] present a research endeavor motivated by the transformative impact of Internet of Things (IoT), cloud computing, and Artificial Intelligence (AI) on healthcare systems, leading to the emergence of smart healthcare. The objectives centered on designing a disease diagnosis model that diagonises heart disease and diabetes by leveraging the convergence of AI and IoT techniques. The CSO-CLSTM model achieves high accuracies, sensitivities, and specificities in diagnosing heart disease and diabetes, outperforming existing classifiers across various scenarios and datasets. However, limitations such as the CSO algorithm's slow search precision and susceptibility to local optima, along with considerations regarding data quality, computational complexity, and resource requirements, may impact the model's optimization process and implementation in real-world healthcare systems.

3. MATERIALS AND METHODS

This section discusses the report on the system setup and the machine learning techniques used in modelling the system for detecting diabetes mellitus.

3.1 Description of Proposed Methodology

For diabetic attribute correlation strength, forward and backward feature selection-based algorithms were applied. This study proposes utilizing an ensemble model for diabetes attribute correlation strength that includes forward and backward features selection-based technique. The study adopted parallel and sequential ensemble approaches. The J48 method, Classification and Regression Tree (CART), and Decision Stump (DS) were used to produce a Random Forest in the first experiment. The J48 algorithm, CART, and DS was used in the second phase of the experiment, along with three successive ensemble methods: XG Boost, AdaBoostM1, and Gradient Boosting. Average voting algorithms was used to measure the final prediction. It provided flexibility in combination strategies to achieve the maximum possible classification accuracy.

3.2 Algorithms Description

The following algorithms described in the next sub-sectors were used in this work.

a. Computation of J48

Algorithm J48 (D)
Input: a data D
Begin
Tree = {}
If (D is “pure”) || (other stopping criteria met) then terminate;
For all attribute a a ∈ *D* D do
Compute criteria of impurity function if a is splitted;
a_{best} = Best attribute according to the above-computed criteria
Tree = Create a decision node that tests a_{best} in the root
D_v = induced sub-datasets from D based on a_{best}
For all D_v do
Begin
Tree $_v$ = J48 (D_v)
Attach Tree $_v$ to the corresponding breach of Tree
End
Return Tree
End

b. Computation of Decision Stump

Input: A set of feature responses $\{f_j(x_n)\}$ extracted by applying the feature *fj* to each training sample and associated labels $\{y_1, \ldots, y_n\}$. A set of non-negative weights *{w1,…, wn}*
Output: Θ is a threshold value. Attention! P ϵ {-1, +1} is a direction value. When the mean value of the positive sample is smaller than the mean value of the negative samples, the direction value *p* is 1. Otherwise, it is -1.

$$g(f_i, p; \Theta) = \begin{cases} 1 \; if \; pfj(x) < p\theta \\ 0 \; if \; otherwise \end{cases} \quad (2)$$

e is the error of the result of classification by this weak classifier g. *e* must be smaller than 0.5
steps of algorithm
Compute the weighed mean of the positive samples and negative samples.

$$\mu P = \frac{\sum_{i-1}^{n} wjfj(xi)yi}{\sum_{i-1}^{n} wiyi}, \mu N = \frac{\sum_{i-1}^{n} wjfj(xi)yi}{\sum_{i-1}^{n} wiyi} \quad (3)$$

Set the threshold to

$$\theta = \frac{1}{2}(\mu P \pm \mu N). \quad (4)$$

Compute the error associated with the two possible values of the direction.

$$\epsilon - 1 = \sum_{i-1}^{n} wi|yi - g(fi(xi); -1; \theta)| \quad (5)$$

$$\epsilon - 1 = \sum_{i-1}^{n} wi|yi - g(fi(xi); +1; \theta)| \quad (6)$$

Set p^* = argmin and the Ep^* Pin{-1, +1}

c. CART (Classification and Regression Trees)

This algorithm repeatedly works in three main steps:

1. Find the best split for every characteristic. There are K-1 possible splits for each feature with K different values. Find the split that maximizes the criterion for splitting. The best splits are found in the resultant set of splits (one for each feature).

2. Find the optimum split for the node, find the split that maximizes the splitting criterion from the best splits from step.

3. Split node by using the best node split from Step ii, then repeat Step i till the stopping requirement is met.

As for splitting criterion, Gini’s impurity index was used, which is defined for node *t* as:

$$i(t) = \Sigma_{i=1} C(i|j)p(i|t)p(j|t) \quad (7)$$

where, $C(i|j)$ is cost for misclassifying a class j case as a class i case (in case $C(i|j) = 1$, if i $6= j$ and $C(i|j) = 0$ *if* $i = j$), $p(i|t)$ $(p(j|t)$ respectively) is probability of case in class $i(j)$ given that it falls into node t.

The Gini impurity criterion is type of reduction of impurity that is defined as:

$$\Delta i(s,t) = i(t) - pLi(tL) - pRi(tR) \quad (8)$$

Where,

$\Delta i(s, t)$ is the reduction of impurity at node t with split *s*, *pL(P.R.)* are probabilities of sending the case to the left (right) child node *to (tR)*, and *i(tL) (i(tR))* is Gini impurity measure for left (right) child node. Pruning will be utilized in conjunction with cross-validation error rate estimation to improve the decision tree's generalization. The pruning algorithm works as follows:

1. Split the training data randomly into ten folds.

2. Select the pruning level for the tree (level 0 equals to full decision tree).

3. Use nine folds for the creation of 9 new pruned trees and calculate error on the last 10th fold.

4. Repeat Step 2 until all pruning levels are used.

5. Find the minor error and use pruning level assigned to it.

6. Until pruning level is reached, delete all terminal nodes in the lowest tree level then assign the decision class to the parent node. Decision value is equal to class with a higher number of cases covered by the node.

d. Random Forest Ensemble Method

As a parallel ensemble method, the bagging meta classifier methodology can be applied with random forest. This method aims to create an uncorrelated forest of trees that range from weak to strong learners, producing more accurate results than a single tree. This study relies on the decision tree classifier, which accepts a criterion called entropy; this is used for ranking the information gain. It measures the impurity in a group of samples when a decrease in entropy is achieved, referred to as information gain. Information gain calculates the difference between the entropy before the split and after the dataset split. The decision tree algorithm implements equation below

$$Info(x) = -\Sigma_{i=1}^{n} pi\ log2\ pi \quad (9)$$

Where the probability that a random row in x belongs to a class i is given by pi.

$$InfoY\ (x) = \Sigma_{j=1}^{k} \frac{|xj|}{|x|} \times Info(xj) \quad (10)$$

$$Gain(Y) = Info(x) - Info_y\ (x_x) \quad (11)$$

where,

i. $Info(x)$ in (9) is the average number of information needed to identify the class label in *x*.

ii. $|X_j| / |x|$ In (10) represents the weight of the J^{th} partition.

iii. $InfoY\ (x)$ Is the expected information needed to classify a row from *x* based on the partitioning of *Y*.

The feature importance is calculated as the normalized total reduction of entropy by a particular feature, implying higher information gain. It is implemented using a method called feature_importances_*[]*. This method takes in each column and returns a relative value for its importance.

e. Computation of Random Forest

To generate *c classifier:*
For i = 1 to c, do
Randomly sample the training data D with replacement to produce D_i
Create a root node, N_i containing D_i
Call BuildTree(N_i)
end for
BuidTree(N):
If *N* contains instances of only one class, then
return
else
Randomly select $x_{\%}$ of the possible splitting features in *N*
Select the feature *F* with the highest information gain split on
Create f child nodes of *N*, $N_1, \ldots, N_f$, where *F* possible values $(F_1, \ldots, F_f)$
For *i*= 1 to f, do
Set the contents of N_i to D_i, where *Di* is all instance in *N* that match
F_i
Call BuildTree(N_i)
end for
end if

f. Computation of AdaboostM1

Initialization:

1. Given training data from the instance space $S = \{(x1, y1), \ldots, (xm, ym)\}$ where x_i $\epsilon\ \varkappa\ and\ yi\ \epsilon\ \curlyvee = \{-1, +1\}$.
2. Initialize the distribution $Di\ (i) = \frac{1}{m}$.

Algorithm:
For *t = 1, …, T: do*
Train a weak learner $h_t: \varkappa \rightarrow$ R using distribution D_t.
Determine – weight $\dot{\alpha}t\ of\ ht$.
Update the distribution over the training set:

$$D_{t+1}(i) = (D_{t(i)e}\text{-}\dot{\alpha}tyiht(xi))/Zt \quad (12)$$

Where *Zt* is a normalization factor chosen so that D_{t+1} will be a distribution.
End for
Final score:

$$f(x) = \sum_{t=0}^{T} \dot{\alpha}tht(x)\ and\ H(x) = sign\ (f(x)) \quad (13)$$

g. Computation of Gradient Boosting

$$F_0(x) atrg\ minp\ \sum_{t=1}^{N} L(yt, p) \quad (14)$$

For m = 1 to M, do:

$$\bar{y} = -[(\partial L(y1F(xi)))/(\partial F(xi))]_{f(x)\ =f\ m\text{-}1(x),}^{I=1,\ N} \quad (15)$$

$am = arg\ min\ a, \beta \sum_{t=1}^{N}[\tilde{y} - \beta h(xi: a)]^2$ (16)

$P_m = \arg \min_{a,\beta} \sum_{i=1}^{N} L[(yt, Fm - 1(xi) + ph(xi: am))\ 2$ (17)

$F_m(x) = F_{m-1}(x) + p_m h(x, a_m)$ (18)

End for

End Algorithm

h. Computation of XG Boost

Data: Dataset and hyperparameters
Initialize $f_0(x)$;
For k = 1,2, …, M do

Calculate $g_k = \frac{\partial L(y,f)}{\partial f}$; (19)

Calculate $h_k = \frac{\partial^2 L(y_1 f)}{\partial f^2}$; (20)

Determine the structure by choosing splits with maximized gain

A = $\frac{1}{2}[\frac{G_L^2}{H_L} + \frac{G_R^2}{H_R} + \frac{G^2}{H}]$; (21)

Determine the leaf weights w* = $-\frac{G}{H}$;

Determine the base leaner

$\hat{b}(x) = \sum_{j=1}^{T} \omega I$; (22)

Add trees

$fk\ (x) = fk - 1\ (x) + \hat{b}(x)$; (23)

End
Result:

$$f(x) = \sum_{k=0}^{M} fk(x) \quad (24)$$

3.3 Evaluation Metrics

The evaluation of the model was done using a confusion matrix. It sums up the number of correct and incorrect predictions. It is a 2 X 2-dimensional matrix that deals with binary classification. Table 3.2 shows the representation of the confusion matrix that was used for evaluation. The two classes are 0 and 1, implying negative (no diabetes) and positive (diabetes) results, respectively. The diagonal values represent accurate predictions, while the non-diagonal values indicate inaccurate predictions.

Terminologies of confusion matrix as follows:

1. True Positives [TP]: These are the positive cases that the classifier properly classified.
2. True Negatives [TN]: These are the negative cases that the classifier properly classified.
3. False Positives [FP]: These are the negative cases that were wrongly classified as positive.
4. False Negatives [FN]: These are the positive cases that were wrongly classified as unfavorable.

The model's performance was assessed using the metrics listed below.

i. Accuracy: This is based on the confusion matrix; the rate of accuracy was computed using the formula below:

$$Accuracy = \frac{TP+FP}{TP+TN+FP+FN} \quad (25)$$

ii. Precision: This is referred to as positive predictive values. It is calculated by using this formula:

$$Precision = \frac{TP}{TP+FP} \quad (26)$$

iii. Recall: This is also referred to as sensitivity; it is calculated by using this formula:

$$Recall = \frac{TP}{TP+FN} \quad (27)$$

iv. Matthews correlation coefficient (MCC) or phi coefficient

$$\text{MCC} = \frac{TP \times FP - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TP+FN)}} \quad (28)$$

4. RESULTS

In this section, the explanation of the dataset employed is done. The classification report achieved using the proposed method are presented.

4.1 Dataset

The dataset contains medical details of patients, including features such as glucose level, blood pressure, insulin level, BMI, age, and more. The target variable indicates whether a patient has diabetes. The goal of this dataset is to build and evaluate different machine learning or deep learning models to predict the onset of diabetes. National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) focuses on diabetes and kidney disease. With 768 female samples, all input attributes provide numeric data values exclusively, and the first character indicates the number of pregnant patients. The second characteristic is the glucose levels in the body. The third attribute represents the measurement of blood pressure (diastolic) in millimetres of mercury (mmHg). The fourth characteristic shows skin thickness in millimetres.

The total amount of insulin produced is described by the 5th, 6th, and 7th properties. The body mass index (BMI) defined in equation (1)

$$BMI = \frac{Patient's weight\ in\ Kg}{Patients\ height\ in\ meter} \quad (29)$$

of the infected patients and reliance on diabetes family hierarchy, respectively. The last characteristic cited denotes the present age of patients. The proposed classification techniques, parallel and sequential ensemble methods with feature selection techniques, were applied on the dataset to generate the diabetes prediction model.

4.2 Result of Proposed Algorithm

The algorithm was fed with the dataset after carrying out preprocessing steps which include replacement of NaN values with zeros, mean imputation, feature selection using the forward and backward feature selection method. After this, the parallel methods were used as base classifiers for training the model. Table 1 shows the result of the four parallel method employed, Random Forest ensemble method gave the highest accuracy, precision, recall, F1 score, Mathew correlation coefficient, AUC_ROC and AUC_PR. Decision Stump performed least.

The proposed Methodology combines both sequential and parallel ensemble method, then uses a voting classifier for making final prediction. Table 2 shows the result of the Sequential method, with Gradient boost performing best having 100% accuracy, precision, recall, F1 score, correlation coefficient, AUC_ROC and AUC_PR.

Table 1. Model Classification Report for the Parallel Ensemble Method

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	MCC (%)	AUC_ROC	AUC_PR
Decision slump	76	68	59	63	45	72	70
J48	78	78	50	61	48	71	73
CART	76	78	54	64	51	73	77
Random Forest	100	100	100	100	100	100	100

Table 2: Model Classification Report for Sequential

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	MCC (%)	AUC_ROC	AUC_PR
XG BOOST	83	81	68	74	62	79	80
Adaboosting MI	81	79	62	70	57	77	81
Gradient Boosting	100	100	100	100	100	100	100

To put the overall model's performance into comparison, 8 features were applied, namely,
pregnancies, Glucose, BloodPresure, Skin Thickness, insulin, BMI Diabetes Pedigree Function and Age. However, it was observed that Gradient boosting and Random Forest performed better than other methods for training and testing.
The model was deployed in the form of an API using sklearn Machine Learning libraries and pyqt5 for its graphical UI module in python. XG Boost from Sequential ensemble method input data with 100% accuracy. Figure 4.21 shows the interface to get the user inputs, so as to detect if the patient is diabetic or not. The code used to access the prediction interface (dia.py).

Figure 1 and figure 2 shows a prediction based on a patient's input. The patient has an high glucose level, which normally signifies a diabetic patient. The model successfully validate that the patient is diabetic.

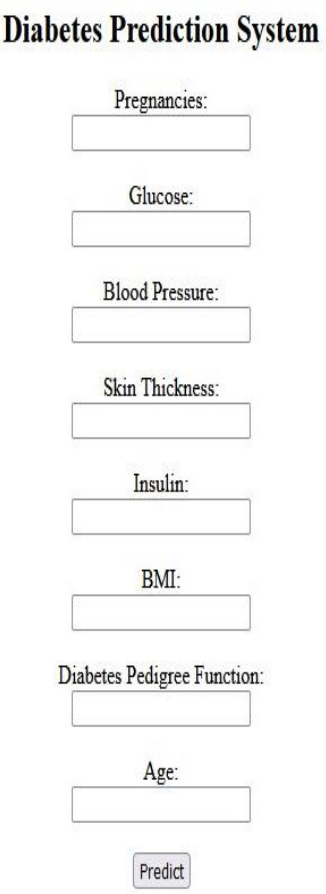

Fig 1: Patience User Interface

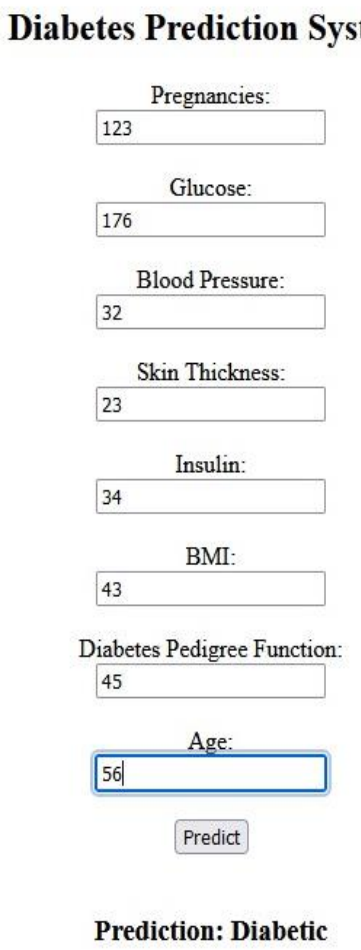

Fig 2: Interface showing a Diabetic Patient

5. CONCLUSION

The results from experiments conducted in this study showed that AdaBoostM1, XG Boost, and Gradient Boosting performed better than the base learning algorithms applied in this work. Researchers and developers can leverage on the predictive model developed in this work to make quick predictions of diabetes mellitus, which could save many lives. This work introduced a novel combination of parallel and sequential ensemble methods with feature selection techniques for the prediction of DM, which played a vital role in resolving the problems of noisy data, over/underfitting, residual errors associated with base-level models.

Feature selection methods, such as Artificial Neural Network Hybrid Ensemble and Fuzzy-based models could be considered for future tasks. The types of diabetes cannot be predicted based on the data used in the present study; future efforts would focus on predicting and determining the different types of diabetes in the human body. This has the potential to increase the accuracy of diabetes prediction.

6. REFERENCES

[1] Raihan, M. M. S., Raihan, M., and Akter, L. 2021, January). A comparative study to predict the diabetes risk using different kernels of support vector machine. In *2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST)* (pp. 547-551). IEEE.

[2] Setyawati, R., Astuti, A., Utami, T. P., Adiwijaya, S., and Hasyim, D. M. 2024. The importance of early detection in disease management. *Journal of World Future Medicine, Health and Nursing*, *2*(1), 51-63.

[3] World Health Organization. 2020. Global report on diabetes. World Health Organization. Https://apps.who.int/iris/handle/10665/204871

[4] Sathurthi, S., and Saruladha, K. 2019. Evaluation of Ensemble Prediction Techniques on Electronic Health Data. In *2019 International Conference on Intelligent Computing and Control Systems (ICCS)* (pp. 1184-1188). IEEE.

[5] Li, Y., and Chen, W. 2020. A comparative performance assessment of ensemble learning for credit scoring. *Mathematics*, *8*(10), 1756.

[6] Li, X., Zhang, J. and Safara, F. 2023. Improving the Accuracy of Diabetes Diagnosis Applications through a Hybrid Feature Selection Algorithm. Neural Process Lett 55, 153–169. https://doi.org/10.1007/s11063-021-10491-0.

[7] Huang Xuan , Wang Hui , She Chongyang , Feng Jing , Liu Xuhui , Hu Xiaofeng , Chen Li , Tao Yong. 2022. Artificial intelligence promotes the diagnosis and screening of diabetic retinopathy Frontiers in Endocrinology vol 13. https://www.frontiersin.org/journals/endocrinology/articles/10.3389/fendo.2022.946915 DOI=10.3389/fendo.2022.946915 ISSN=1664-2392.

[8] Jyotismita Chaki, S. Thillai Ganesh, S.K Cidham, S. Ananda Theertan, 2022. Machine learning and artificial intelligence-based Diabetes Mellitus detection and self-management: A systematic review. Journal of King Saud University - Computer and Information Sciences, Volume 34, Issue 6, Part B, Pages 3204-3225, ISSN 1319-1578, https://doi.org/10.1016/j.jksuci.2020.06.013. (https://www.sciencedirect.com/science/article/pii/S1319157820304134).

[9] Yu-Cheng Wang, Tin-Chih Toly Chen, and Min-Chi Chiu, 2023. A systematic approach to enhance the explainability of artificial intelligence in healthcare with application to diagnosis of diabetes, Healthcare Analytics, Volume 3,2023,100183, ISSN 27724425. https://doi.org/10.1016/j.health.100183.(https://www.sciencedirect.com/science/article/pii/S2772442523000503)

[10] Makroum, Mohammed Amine, Mehdi Adda, Abdenour Bouzouane, and Hussein Ibrahim. 2022. "Machine Learning and Smart Devices for Diabetes Management: Systematic Review" *Sensors* 22, no. 5: 1843. https://doi.org/10.3390/s22051843.

[11] Kopitar, Leon, Iztok Fister, Jr., and Gregor Stiglic. 2024. "Using Generative AI to Improve the Performance and Interpretability of Rule-Based Diagnosis of Type 2 Diabetes Mellitus" *Information* 15, no. 3: 162. https://doi.org/10.3390/info15030162.

[12] Nomura, A., Noguchi, M., and Kometani, M. 2021. Artificial Intelligence in Current Diabetes Management and Prediction. *Curr Diab Vol* **21**, 61. https://doi.org/10.1007/s11892-021-01423-2

[13] Gupta, D., Choudhury, A., and Gupta, U. 2021. Computational approach to clinical diagnosis of diabetes disease: a comparative study. *Multimed Tools Appl* **80**, 30091–30116. https://doi.org/10.1007/s11042-020-10242-8

[14] Sharma, T., and Shah, M. 2021 A comprehensive review of machine learning techniques on diabetes detection. *Vis. Comput. Ind. Biomed. Art* **4**, 30. https://doi.org/10.1186/s42492-021-00097-7.

[15] R. F. Mansour, A. E. Amraoui, I. Nouaouri, V. G. Díaz, D. Gupta and S. Kumar, 2021. Artificial Intelligence and Internet of Things Enabled Disease Diagnosis Model for Smart Healthcare Systems, in *IEEE Access*, vol. 9, pp. 45137-45146, doi: 10.1109/ACCESS.2021.3066365.