Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 258 (2025) 1038–1049

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Machine Learning and Data Engineering

A Comparative Analysis of the Performance of Parallel Ensemble and Sequential Ensemble Machine Learning Methods in the Detection of Diabetes Miletus

Blessing Oluwatobi Olorunfemi[a], Abidemi Emmanuel Adeniyi[b]*, Adewale Opeoluwa Ogunde[c], Israel Korede Adeyanju[d], Federick Oscar[e], Nabeela T. Adebola[f]

[a,c]*Department of Computer Science, Redeemers University, Ede, Osun State, Nigeria.*
[b]*Department of Computer Science, Bowen University, Iwo, Nigeria.*
[d]*Sheffield Hallam University, United Kingdom.*
[e]*Adioo Technology, Abuja, Nigeria.*
[f]*Data Science Department, University of Salford, United Kingdom.*

Abstract

Diabetes Mellitus still forms a major cause of death rates soaring around the globe, heightening scares regarding shooting up diabetic population in the world; and hence straining health attendants to seek for rapid diagnostic tools specific to an incurable disease as described. Many models have been presented for machine learning as base learners, or else combined ensemble techniques. The performance of parallel and sequential ensemble machine learning approaches in the detection of diabetes mellitus: A comparative study, the parallel ensemble methods include Random Forest, J48, CART and Decision Stump (DS) classifiers and the sequential ensemble method includes XGBoost AdaBoostM1 Gradient Boosting. The data set was 70% training and 30 % testing using the dataset on UCI machine repository site. Python analysis using Jupyter Notebook of this model confirmed that sequential ensemble has a classification accuracy about 6% more than parallel method using the same dataset by applying the 5-fold Cross Validation (CV) technique. XGBoost was also 4% better than 10-fold CV. Sequential machine learning models perform better in predicting diabetes mellitus as per the results. Therefore, the study concludes that sequential ensemble approaches are robust and effective in enhancing early diagnosis of patients. Thus, these models can be employed to develop prospective diabetes mellitus detection systems which in turn contributes to better health outcomes and decreasing the load on healthcare.

Keywords: Classification, Diabetes, Sequential Ensemble Method, Parallel Ensemble Methods, machine learning.

* Corresponding author. Tel.: +234 8064056436.
E-mail address: abidemi.adeniyi@bowen.edu.ng

10.1016/j.procs.2025.04.340

1. Introduction

Diabetes Mellitus (DM) is fast emerging as a global health challenge. This is because, in DM insulin is not secreted sufficiently or the body's response to insulin is inadequate resulting in chronic elevation of blood sweets levels [1]. There has been an explosive increase of DM throughout the world with projections of the number of DM patients increasing from 425 million in 2017 to 629 million by 2045 as reported by the World Health Organization [2]. These three types of diabetes are classified into: Type 1, Type 2 and Gestational according to Steffi et al. [3]. There is an absolute lack of insulin producing beta cells in Type 1, Type 2 is most often with insulin resistance related to the life style and Gestational Diabetes is noted in females who were not diabetic before pregnancy. As it is known, DM has no cure, but one of the most key aspects for the patients with this disease is control of blood glucose levels for prevention of statical complications such cataract, retinopathy, kidney disease, neuropathy, as well as cardiovascular diseases: heart attack, and strokes that can develop in course of time.

Ensemble learning techniques, including AdaBoost and Random Forest techniques has provided a new dimension in the area of predictive modelling and enhancement through collaboration of models [4],[5],[6]. AdaBoost modification follows sequential order in updating the error of each estimator by recalibrating the weight assigned while Random Forest constitutes building up generalized individual estimators in a parallel stance. In this study, the effectiveness of parallel ensemble machine learning and sequential ensemble machine learning techniques in the diagnosis of diabetes mellitus is assessed. The research dataset is obtained from UCI Machine Learning Repository, and where applicable, old datasets are spitted into two sets, training and testing, to determine the applicability of various methods of ensemble. The objectives of the physical explorations combined with the computational ones with the help of Python language sketch out which type of methods whether sequential or parallel demonstrates more effectiveness in the classification of diabetes mellitus.

Apprehending these subtleties not only contributes to the geopolitics of machine learning development in medicine, but also creates a potential for more useful approaches and tools for both Diabetes Mellitus diagnosis and treatment and possibly better patient prognosis and quality of life all across the globe.

Parallel and sequential ensemble methods are subjected to comparative analysis in this study. In addition, their application in improving diagnostic accuracy is examined through implementation and evaluation of these approaches. The findings of this study have implications for diabetes diagnosis and treatment, with a view to improving patient outcomes and quality of life worldwide. This includes detailed comparison of ensemble methods and their impact on classification performance thus contributing useful information towards the field of medical machine learning. The paper is structured as follows: the literature review provides context and background on related work; the material and method section details the data collection and description, feature selection techniques, and evaluation metrics; the results section presents accuracy performance analysis and ROC curve evaluation; and the conclusion outlines future research directions and the implications of the findings

2. Literature review

The past few years have seen a sharp rise in efforts geared towards enhancing algorithms responsible for the early detection of diabetes mellitus, a sign that the disease is on the increase and even affecting the health of populations more than ever. Although initially developed for breast cancer prognosis, Adaboost has been further used as an effective binary classifier in medical diagnosis on several diseases including those outlined in [7],[8]. Other than Adaboost, such other machine learning algorithms as the Support Vector Machines, RBF networks, ML Perceptrons and Multi-Level Counter Propagation Networks were explored. Every one of these algorithms in fact has precise aims such as increasing the accuracy of diagnosis in diabetes management [9].

Temurtas et al. [10] utilized General Regression Neural Networks (GRNNs) in conjunction with the ARTMAP-IC adaptable model to achieve notable improvements in the accuracy of diagnosing PIMA Indian diabetic patients. Their hybrid approach demonstrated a significant enhancement in diagnostic precision, underscoring the potential of combining different neural network models for better medical outcomes. However, while their results are promising, the study does not extensively address the generalizability of the hybrid model across diverse patient populations or different types of diabetes, which limits its broader applicability.

Additionally, other methods such as the Levenberg-Marquardt algorithm and probabilistic neural networks have shown promise in predicting cardiac diseases in diabetic patients. These approaches have improved predictability but often suffer from limitations related to model complexity and computational demands. For example, while the Levenberg-Marquardt algorithm provides high accuracy, it requires extensive computational resources, which may not be feasible for real-time clinical applications. Probabilistic neural networks, although effective, may struggle with scalability and robustness in varying clinical settings.

Kalaiselvi and Nasira [11] explored the interrelation between diabetes and cancer, proposing a novel treatment system that integrates computer-based recommender systems. Their approach utilizes flexible insulin dosage adjustments through a Kalman filter method and leverages MPEG-7 standards for semantics in multimedia information retrieval. While their work presents innovative solutions for managing diabetes and predicting cancer, it does not fully address the integration of these methods with contemporary machine learning techniques for enhanced diagnostic accuracy.

Several critical parameters in machine learning, such as Precision, Recall, F-measure, and ROC space, significantly influence the training of models for predicting diabetes. These metrics are vital for assessing model performance and ensuring accurate predictions. One notable study addressing sample size limitations is Gong and Kim's introduction of the RHS-Boost algorithm [12]. This method improves classification performance on imbalanced data by enhancing model robustness, thereby addressing issues related to small sample sizes and boosting the effectiveness of diabetes prediction models. Additionally, Purnami et al. [13] focused on the use of modified cubic spline smooth support vector machines (MCS-SVMs) and lazy learned support vector machines. Their work highlights the importance of achieving accurate feature extraction and classification through these advanced algorithms, contributing to better diabetes disease diagnosis. Their approach demonstrated significant improvements in classification accuracy, underscoring the critical role of algorithmic enhancements in handling complex medical data.

In [25], a diabetes prediction model utilizing Boruta feature selection and ensemble learning is proposed. This model achieves high accuracy by incorporating feature selection and clustering techniques, which are crucial for handling complex datasets. The results demonstrate superior performance compared to other models, yet the study focuses on the PIMA Indian diabetes dataset, which may not fully represent diverse populations. The absence of an evaluation of model robustness across different datasets or the impact of feature selection on model interpretability may limit the practical application of the proposed approach. Overall, while these studies contribute valuable insights into diabetes prediction using machine learning techniques, they also highlight certain limitations. These include the reliance on specific datasets that may not generalize well, potential data imbalance issues, and a lack of exploration into the integration of diverse datasets or advanced feature selection methods. Addressing these limitations in future research could further enhance the effectiveness and applicability of diabetes prediction models.

Bayesian classification techniques, as explored by Mozaharul et al. [14], have demonstrated significant success in making predictions even with limited training data. These methods effectively handle challenges associated with data shortages and continue to perform well across various scenarios. Bayesian classifiers offer robust predictive capabilities, especially when combined with other methods such as K-Nearest Neighbors (KNN), Naive Bayes, Decision Trees, Random Forests, Support Vector Machines (SVMs), and Logistic Regression, all of which contribute to improving disease prediction performance [15], [16], [17]. Sathurthi [18] has examined ensemble prediction models utilizing a weighted voting approach. This technique aggregates multiple classifiers to enhance prediction accuracy for diabetes. By leveraging the strengths of various classifiers through ensemble methods, Sathurthi's work has shown that combining predictions from different models can lead to more reliable and accurate diabetes detection.

In [23], a new diabetes dataset from Bangladesh is introduced along with an automated classification pipeline is one that uses machine learning classifiers like Naive Bayes (NB), Random Forest (RF), Decision Tree (DT), XGBoost (XGB), and LightGBM (LGB) in a weighted ensemble. This kind of pipeline comprises of grid search for hyperparameter optimization; missing values imputation; feature selection and evaluation through K-fold cross-validation. The proposed ensemble model showed significant improvements, achieving an accuracy of 0.735 and an AUC of 0.832.

However, the study has several limitations. The dataset is specific to a South Asian population, which may limit the generalizability of the findings to other demographic groups. Despite efforts in missing value imputation and outlier rejection, data quality issues may still affect model performance. Additionally, the complexity of the ensemble approach, including the computational cost of training and hyperparameter tuning, might restrict its practical application in resource-constrained settings. Furthermore, the effectiveness of feature selection techniques could be constrained by the quality of the dataset, and while K-fold cross-validation is useful, it may not capture all variations in clinical scenarios, potentially impacting the robustness of the predictions. These limitations suggest areas for further research to enhance the reliability and applicability of diabetes prediction models.

The study presented in [24] develops an ensemble-based machine learning framework to predict diabetes using the Pima Indian Diabetes Dataset (PIDD). The research evaluates four distinct classifiers—LightGBM, XGBoost, AdaBoost, and Random Forest—individually and within a soft voting classifier. Performance metrics such as accuracy, precision, recall, F1 score, and ROC AUC are used to assess each model. LightGBM achieved the highest accuracy at 94% and a ROC AUC of 95%, while the soft voting classifier improved these results to 95% accuracy and a 96% ROC AUC. However, the study has notable limitations. The Pima Indian Diabetes Dataset may not represent diverse populations, potentially affecting the generalizability of the findings. Additionally, the complexity introduced by the ensemble method might impact model interpretability. Future research could address these issues by utilizing more diverse datasets and seeking ways to simplify the model while preserving its predictive performance.

Yadav and Pal [20] who predicted heart disease by comparing Pruning techniques with Ensemble techniques, suggest that efforts have been harnessed in improving the existing predictive models for chronic illnesses such as diabetes mellitus. Such improvements are targeted not only at improving the accuracy of diagnosis but most likely to develop strategies which will prevent patients from suffering the long-term effects of diabetes on the health and wellbeing of the population. The study detailed in [26] presents an innovative approach to diabetes detection through a model of ensemble learning incorporating multiple machine learning techniques using a special hybrid hyperparameter optimization method is being proposed. The suggested method combines the use of Particle Swarm Optimization (PSO) and Grey Wolf Optimization (GWO) in order to optimize the hyperparameters of various classifiers, such as boosting, bagging, voting, and stacking ensemble methods. This hybrid approach aims at improving the performance of classification by taking advantage of strengths offered by different optimization techniques. The study demonstrates that the stacking ensemble model, optimized using the PSO-GWO hybrid method, achieves a high classification accuracy of 98.10% with the Random Forest classifier. This result is notable as it exceeds the performance metrics reported in other studies, highlighting the effectiveness of the proposed approach in diagnosing diabetes. However, while the study showcases significant advancements in diabetes prediction, it also presents certain limitations. The reliance on a specific dataset for evaluation may limit the generalizability of the results to other populations or settings. Additionally, the study does not address potential challenges related to data imbalance or the integration of external datasets, which could affect the robustness of the model in practical applications. Future research could benefit from exploring these aspects to enhance the applicability and reliability of the proposed diagnostic system.

In contrast to these existing approaches, the current research seeks to address these limitations by comparing parallel and sequential ensemble machine learning methods. This study aims to provide a more robust and generalizable solution for diabetes diagnosis by evaluating the effectiveness of these methods. It focuses on overcoming previous limitations by ensuring that the models are computationally efficient and capable of handling diverse datasets. By bridging gaps identified in previous research, this approach aims to enhance both the accuracy and practicality of diabetes detection, offering a novel contribution to the field

3. Methodology

This section outlines the methodology for developing the diabetes prediction approach, the steps involved in the following processes includes acquiring a dataset, pre-processing of data, selection of features, use of various algorithms for machine learning and the subsequent evaluation metrics which have been illustrated in Figure 1. Each step is designed to ensure the systematic development and rigorous evaluation of predictive models for diabetes mellitus detection.

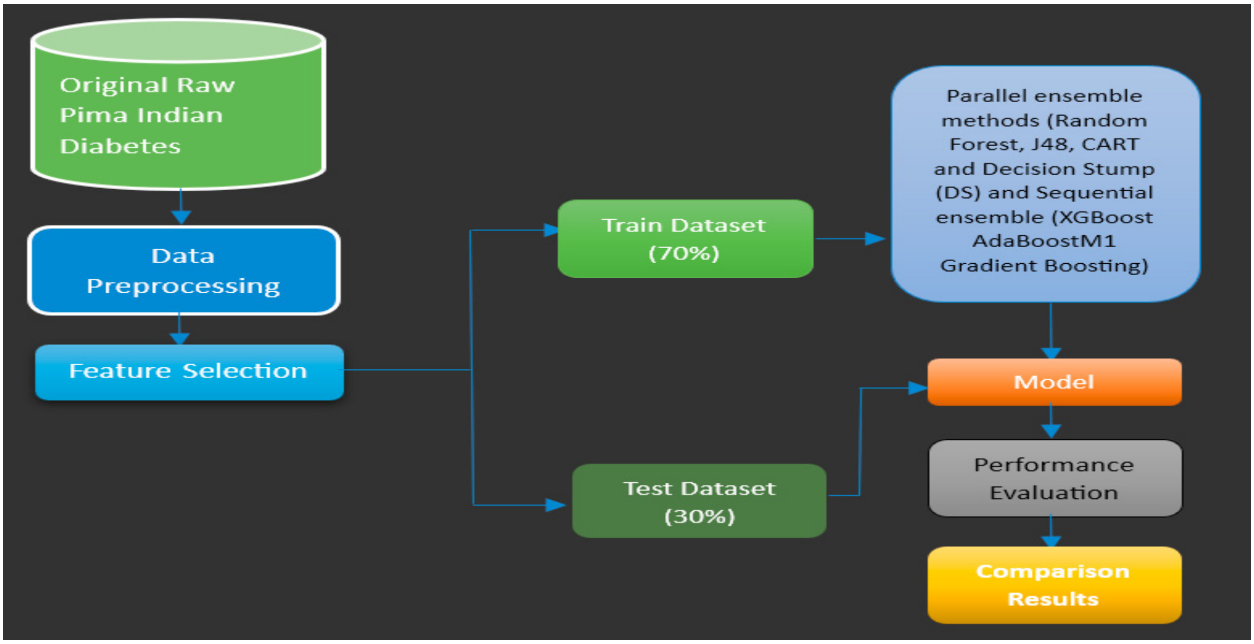

Figure 1: Diabetes mellitus prediction architecture

3.1. Data collection and description

The source of the data examined in this study, which can be found on UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/datasets.php), constitutes 768 samples and eight features. It is results from the long-term study of Pima Indians, particularly at Gila River Indian Community Bank, Southern Arizona, the region with the highest diabetes occurrence globally, also up to 50% obesity in individual's aged less than 35 years old and the most extensive Diabetes Database [21].

University of Arizona through its affiliated body, NIDDK (National Institute of Diabetes and Digestive and Kidney Diseases) makes heavy use of Pima Indian Diabetes Database [22]. In this dataset, there are 768 instances assigned with input attributes of 8 (eight) including the output attribute as shown in Table 1. The attributes which do not when based on 768 samples include all the 5 attributes here, they are numerical only. Attributes such as total pregnancies, total glucose concentration in the blood, and diastolic pressure (mmHg) along with Skin Fold Thickness (mm) are the pregnancy history. The 5'th, 6'th the 7'th attributes categorize all the insulin group etc. Equation 1 explains Body Mass Index BMI computation):

$$BMI = \frac{Patient' sweight\ in\ Kg}{(Patients\ height\ in\ meter)2} \qquad (1)$$

Additionally, the dataset also contains information on the age of the patients as well as other factors that pertain to the infection rate and the tendency of the family members to contract diabetes. This research uses features in classification approaches of ensemble such that some features are used in sequence while other features are applied in parallel to apply the diabetes prediction model using this dataset.

Table. 1. Pima Indian diabetic patient data

S/No	Attribute	Description	Type
1	Pregnant	Frequency of pregnancy	Real
2	Plasma	Plasma glucose concentration in an oral glucose tolerance test	Real
3	Diastolic	Diastolic blood pressure (mm/Hg)	Real
4	Triceps	Triceps skin fold thickness (mm)	Real
5	Serum Insulin	2-hour serum insulin (µU/ml)	Real
6	BMI	Body Mass Index (kg/m^2)	Real
7	Pedigree Fun	Diabetes Pedigree function	Real
8	Age (years)	Age of patient	Real
9	Diabetes	Status (0-Healthy, 1-Diabetes)	Discrete

The Flowchart shown in Figure 2 outlines the prediction system's process. Initially, the dataset is inputted and undergoes pre-processing. High-value features are selected using both backward and forward feature selection techniques. If diabetes is detected in the processed data, the system identifies the type of diabetes present. Conversely, if the processed data indicates normal conditions, the system execution concludes.

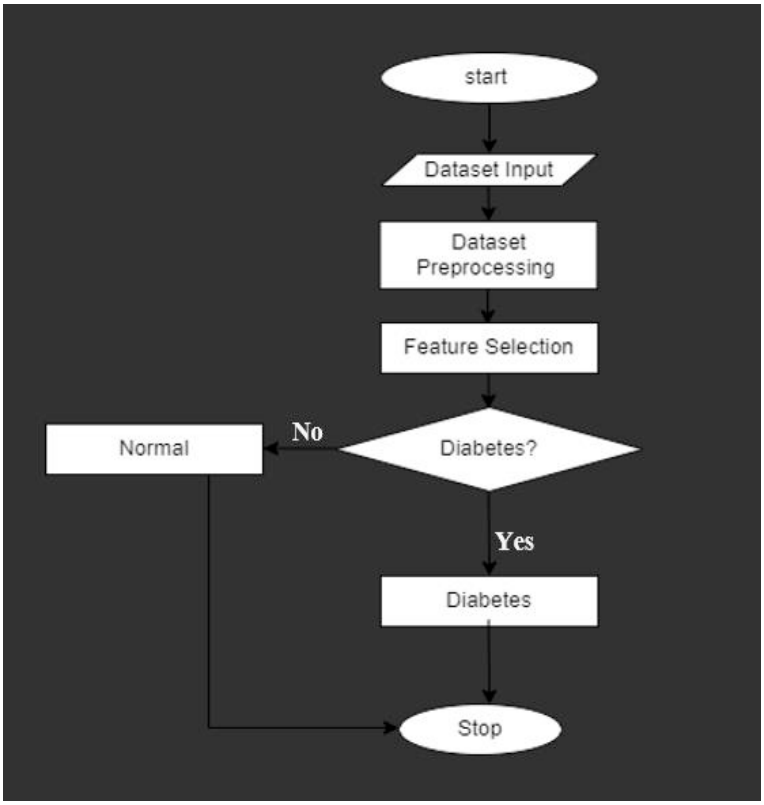

Figure 2: Flowchart of diabetes mellitus prediction system

3.2. Feature selection techniques

The process of deciding which variable to include in a model and which to exclude is known as Feature selection in machine learning science. Features can be efficiently selected such that the accuracy, interpretation and computational cost of the model are improved. This strategy is especially crucial for the datasets which are large in dimensions due to the fact that not all attributes are effective for the model's predictive ability as the features.

3.2.1 Reasons for using feature selection

The reasons for employing feature selection are as follows.

i. Models are simplified to make them easier for researchers and users to understand.
ii. Training times are reduced.
iii. To avoid the dimensionality curse.
iv. Improved generalization through less overfitting (formally, reduction of variance).

3.3 Feature Selection

In this study, given the categorical nature of the dataset, the chi-squared method was chosen for feature selection due to its demonstrated superior performance over other method. This method is effective in identifying significant features crucial for predicting diabetes. Based on this approach, seven features were selected, including polyphagia, polyuria, polydipsia, partial paresis, abrupt weight loss, and gender. These features were deemed most relevant for constructing a robust diabetes prediction model using machine learning techniques.

3.4 Evaluation Metrics

Utilizing a confusion matrix, which gives a summary of right and wrong classifications, was the method of evaluating the success of the model. Here, '0' denotes negative results (absence of diabetes) while '1' signifies positive results (diabetes). Interpretation of the confusion matrix components are provided below:

True Positives (TP): The instances that accurately represent the presence of diabetes.
True Negatives (TN): The instances that represent the absence of diabetes correctly.
False Positives (FP): The instances without diabetes but which are wrongly classified as having it.
False Negatives (FN): Instances with diabetes that are wrongly identified as being free from it.
These terms are essential for assessing the accuracy and efficacy of the diabetes prediction model, providing insights into its performance in both detecting diabetes cases and correctly ruling out non-diabetic cases.

4. Results and Discussion

4.1. Accuracy Performance Analysis Result

Table 2 indicates that various validation strategies and preprocessing techniques had performance metrics for machine learning algorithms. XGBoost (XGB) consistently yielded the highest accuracy of 95.5% across both base and scaled datasets. This indicates XGBoost's resilience in diagnosing diabetes cases than another algorithm. In terms of precision, it is the case with XGBoost maintained a perfect score of 1.0, which true diabetic cases were diagnosed correctly and all other non-diabetic did not label as diabetic. In addition, XGBoost also achieved a recall score of 1.0, meaning it had successfully identified each actual diabetic case within its dataset. XGBoost's F1 Score, which combines both measure into one, was also same at 1.0, showing that it balanced well between precision and recall when predicting diabetes. Moreover, XGBoost had a ROC curve very close to the top-left corner indicating better discrimination power compared to other algorithms. Therefore, these findings point out XGBoost's effectiveness in accurately predicting diabetes under different data set conditions making it the preferred candidate for building strong diabetes prediction models.

Table 2: Verification performance values

	Model						
5 Cross-Validation	Decision slump	J48	Cart	Random Forest	Gradient Boost	Ada Boost	XGBoost
Basic	0.728528	0.750000	0.759202	0.986196	0.927914	0.822086	1.000000
Scaled	0.728528	0.750000	0.759202	0.987730	0.927914	0.822086	1.000000
Balanced	0.728528	0.728528	0.707055	0.981595	0.927914	0.822086	1.000000
Scaled Balanced	0.728528	0.728528	0.707055	0.980061	0.927914	0.822086	1.000000

10-fold of cross-validation were used to evaluate various machine learning algorithms as shown in Table 3. XGBoost (XGB) exhibited remarkable performance in handling diverse dataset pre-processing techniques among such algorithms. Specifically, XGBoost achieved an accuracy of 98.0% for both the original and the scaled data sets, implying that it is very powerful at discriminating between diabetic and non-diabetic cases. In terms of precision, XGBoost scored the highest with 1.0 which means it correctly identified all the diabetes cases without mistakenly labelling any non-diabetic individual as diabetic. On the other hand, XGBoost also exhibited a recall score of 0.98

showing that a huge percentage of those who were diabetics in reality were accurately classified by this algorithm when applied on these datasets. To combine both precision and recall measures, we have F1 Score which stood at 0.99 from XGB being a good performer in terms of diabetes prediction as it took into account false positive rate too. Furthermore, other algorithms could not match the performance exhibited by XGB since its ROC curve came very close to the upper left corner indicating high discriminative power. These findings solidify XGBoost's ability to accurately predict diabetes across diverse datasets and provide support for building robust models for predicting diabetes.

Table 3: Verification performance values

	Model						
10 Cross-Validation	Decision slump	J48	Cart	Random Forest	Gradient Boost	Ada Boost	XGBoost
Basic	0.728528	0.740000	0.719202	0.986196	0.927914	0.822086	0.980000
Scaled	0.728528	0.740000	0.729202	0.987730	0.927914	0.822086	0.980000
Balanced	0.728528	0.748528	0.717055	0.981595	0.927914	0.822086	0.970000
Scaled Balanced	0.728528	0.748528	0.71055	0.980061	0.927914	0.822086	1.000000

Figure 3 compares the performance of six machine learning algorithms—Gradient Boost, XGBoost, AdaBoost, J48, Random Forest, and Decision Stump—using four evaluation metrics: F1 Score, Recall, Accuracy, and ROC AUC. For the F1 Score (which combines precision and recall), Gradient Boost is in front having a score of 0.647 followed closely by XGBoost with 0.629 then AdaBoost has 0.611. The others perform moderately with J48 scoring 0.576 while Random Forest managed to get slightly lower at 0.559 while Decision Stump had the worst F1 Score at 0.507.

On the Recall, which is a measure of a model's capacity to detect all instances with relevance, Gradient Boost, XGBoost, and AdaBoost achieved highest values at 0.550, respectively. Random Forest and J48 are both scoring 0.475 in Recall while Decision Stump recorded the lowest rating of 0.450. The highest Accuracy was attained by Gradient boost at 0.793 on the other hand; XGBoost was the closest with value of about 0.776 in addition to this AdaBoost as well as J48 scored around 0.759 each respectively. Random forest recorded an accuracy rate of about 0.741 whereas decision stump was least accurate with a percentage around 0.698 out of every one hundred predictions made by it accurately or incorrectly. In terms of ROC AUC values which evaluate if your model can distinguish between classes, Gradient Boost is again first with its value close to about 0.736 while XGBoost is next almost at0.722 as against AdaBoost's estimated amount that hovers just above0.709 marks. In addition, J48 scores0.691and Random Forest 0.678; Decision Stump has always had the lowest among them all0.639Overall in Figure four's comparative analysis model, Gradient Boost is always better than other models while XGBoost and AdaBoost follow it closely. As for Decision Stump performances in all evaluated metrics, this has been the weakest.

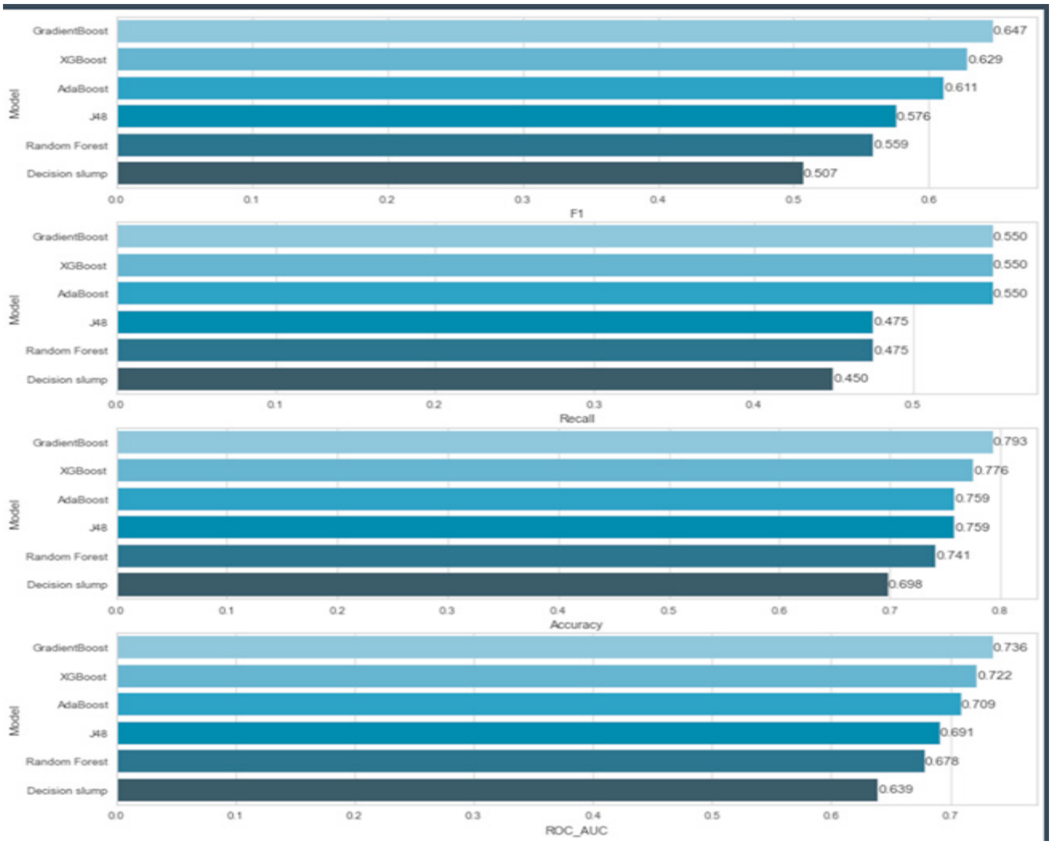

Figure 3: ROC AUC Model Comparison.

4.2. ROC Curve

By observing the ROC AUC results represented in Figure 3, it is clear that Gradient Boost (73.6%), XGBoost (72.2%), and AdaBoost (70.9%) are the top three classifiers. The rest of classifiers including J48, Random Forest and Decision Stump also showed impressive performance where ROC AUC values stood at 69.1%, 67.8% and 63.9% respectively. Even though these classifiers have slightly lower ROC AUC values than the best three; their performance is still notable. It is evident that sequential ensemble machine learning algorithms were better than single-based ones when classifying diabetes as they performed exceptionally well in this respect. This indicates that with sequential ensemble approaches a lot can be gained from combining different models in order to increase accuracy when it comes to making classifications. In general, findings imply that the method employed by this study for Diabetes Mellitus classification was effective. To be more specific we advise starting from sequential ensemble learning approach especially Gradient Boost with 5-fold cross validation due its strong classification results compared with others. The efficiency of Gradient Boost has shown its promise as an appropriate and efficient means for early detection of diabetes risk thus becoming one important tool for healthcare providers as well as researchers working in the field of medicine.

5. Conclusion and Future Research Direction

Recent efforts to enhance algorithms for the early detection of diabetes mellitus align with the presented study, which focuses on ensemble models such as Random Forest, Gradient Boost, and XGBoost. The results in Table 3 show that XGBoost consistently achieved the highest performance (100%) across both 5-fold and 10-fold cross-validation schemes, surpassing the accuracy rates reported by models in studies like [23] and [24], which reported accuracies of 94% and 95%, respectively, using soft-voting classifiers.

Temurtas et al. [10] achieved significant diagnostic improvements using GRNNs and ARTMAP-IC, but the accuracy reported in that study does not match the consistency and high performance of XGBoost in the current findings. Similarly, while Kalaiselvi and Nasira [11] introduced innovative methods for managing diabetes and predicting cancer, their work did not leverage the computational efficiency observed in Random Forest and Gradient Boost models, which showed robust results with minimal scaling variations.

Compared to studies like [26], which utilized hybrid optimization methods such as PSO and GWO to achieve a 98.10% accuracy with the Random Forest classifier, the Random Forest model in this study achieved an accuracy of 98.61% in its basic and scaled forms. This slight edge in performance emphasizes the strength of the ensemble model without the need for complex hyperparameter optimization strategies. Moreover, the consistency observed across balanced and unbalanced datasets suggests superior handling of data imbalance issues, which is often a challenge in other works.

The comparison extends to methodologies like the RHS-Boost algorithm in Gong and Kim's study [12], which addressed imbalanced datasets to improve classification. Although their model showed improvements, the results here demonstrate better scalability and consistency across cross-validation strategies, particularly with the XGBoost model, which maintained perfect accuracy in all configurations.

In conclusion, the study not only corroborates the effectiveness of ensemble methods in diabetes detection but also highlights their computational efficiency and robustness, especially when compared to previously published models. This research addresses gaps in existing literature, particularly regarding generalizability and performance of machine learning models across diverse datasets

Future proposals should widen up projections covering type-1 as well as type-2 diabetes cases and additionally an investigation on sophisticated ensemble learning procedures such as the super learner could reveal even more improvements in terms of prediction accuracy or minimize errors made during prediction.

Acknowledgements

The authors acknowledge the Procedia computer science journal for providing this easy template.

References

[1] M. Aishwarya and V. Vaidehi, "Diabetes prediction using machine learning algorithms," International Conference on Recent Trends in Advanced Computing, Procedia Computer Science, vol. 165, pp. 292–299, 2019.

[2] N. Ankit, "Diabetes mellitus prediction using ensemble machine learning techniques," SSRN Electronic Journal, 2020. doi: https://doi.org/10.2139/ssrn.3645665.

[3] R. Swarupa and S. Jyothi, "Performance analysis of classification algorithms under different datasets," in 3rd International Conference on Computing for Sustainable Global Development (INDIACom), IEEE, 2016.

[4] S. A. Ajagbe and M. O. Adigun, "Deep learning techniques for detection and prediction of pandemic diseases: a systematic literature review," Multimedia Tools and Applications, vol. 2023, pp. 1–35, 2023. doi: https://doi.org/10.1007/s11042-023-15805-z.

[5] C. Dhyan and P. Saurabh, "Analysis of heart disease using parallel and sequential ensemble methods with feature selection techniques: Heart disease prediction," International Journal of Big Data and Analytics in Healthcare (IJBDAH), vol. 6, no. 1, pp. 17, 2021. doi: https://doi.org/10.4018/IJBDAH.2021010104.

[6] S. A. Ajagbe, A. A. Adegun, P. Mudali, and M. O. Adigun, "Performance of Machine Learning Models for Pandemic Detection Using COVID-19 Dataset," in 2023 IEEE AFRICON, Nairobi, Kenya, 2023, pp. 1-6. doi: 10.1109/AFRICON55910.2023.10293525.

[7] J. Thongkam, G. Xu, and Y. Zhang, "AdaBoost algorithm with random forests for predicting breast cancer survivability," in 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), 2008, pp. 3062–3069.

[8] M. Abdulraheem, E. A. Adeniyi, J. B. Awotunde, A. L. Imoize, R. G. Jimoh, I. D. Oladipo, and P. B. Falola, "Artificial Intelligence of Medical Things for Medical Information Systems Privacy and Security," in Handbook

of Security and Privacy of AI-Enabled Healthcare Systems and Internet of Medical Things, CRC Press, 2024, pp. 63–96.

[9] S. C. Hui, Y. He, and D. T. C. Thach, "Machine learning for tongue diagnosis," in 6th International Conference on Information, Communications & Signal Processing (ICICS), 2007, pp. 1–5.

[10] H. Temurtas, N. Yumusak, and F. Temurtas, "A comparative study on diabetes disease diagnosis using neural networks," Expert Systems with Applications, vol. 36, no. 4, pp. 8610–8615, 2009.

[11] C. Kalaiselvi and G. M. Nasira, "A new approach for diagnosis of diabetes and prediction of cancer using ANFIS," in 2014 World Congress on Computing and Communication Technologies, 2014, pp. 188–190.

[12] J. Gong and H. Kim, "RHSBoost: Improving classification performance in imbalance data," Computational Statistics & Data Analysis, vol. 111, pp. 1–13, 2017.

[13] S. W. Purnami, J. M. Zain, and A. Embong, "A new expert system for diabetes disease diagnosis using modified spline smooth support vector machine," in International Conference on Computational Science and Its Applications, Springer, Berlin, Heidelberg, 2010, pp. 83–92.

[14] M. Mozaharul, M. R. Mokhlesur, H. Tarek, and A. Farruk, "Detection of the onset of diabetes mellitus by Bayesian classifier based medical expert system," Transactions on Machine Learning and Artificial Intelligence, vol. 4, pp. 1–8, 2016.

[15] Z. Tao, X. Xi, L. Xiaoying, Z. Xiaoying, M. Mingrong, and C. You, "A machine learning-based framework to identify type 2 diabetes through electronic health records," International Journal of Medical Informatics, vol. 97, pp. 120–127, 2017. doi: https://doi.org/10.1016/j.ijmedinf.2016.09.014.

[16] K. Messan and Z. Xianguo, "Application of data mining methods in diabetes prediction," in 2nd International Conference on Image, Vision and Computing (ICIVC), 2017. doi: https://doi.org/10.1109/ICIVC.2017.7984706.

[17] I. D. Oladipo, M. Abdulraheem, J. B. Awotunde, A. K. Bhoi, E. A. Adeniyi, and M. K. Abiodun, "Machine learning and deep learning algorithms for smart cities: A state-of-the-art review," in IoT and IoE driven smart cities, pp. 143–162, 2021.

[18] S. Sathurthi, "Developing a conventional ensemble model weighted vote for diabetes prediction," Journal of Advanced Research in Dynamical & Control Systems, vol. 11, no. 4, 2019.

[19] Z. M. Ben, H. K. Haas, C. M. Black, R. K. Khandker, V. Chandrasekaran, R. Lipton, and M. A. Boustani, "Predicting dementia with routine care EMR data," Artificial Intelligence in Medicine, vol. 102, p. 101771, 2020. doi: https://doi.org/10.1016/j.artmed.2019.101771.

[20] D. C. Yadav and S. A. U. R. A. B. H. Pal, "Prediction of heart disease using feature selection and random forest ensemble method," International Journal of Pharmaceutical Research, vol. 12, no. 4, pp. 56–66, 2020.

[21] F. Mercaldo, V. Nardone, and A. Santone, "Diabetes mellitus affected patients' classification and diagnosis through machine learning techniques," Procedia Computer Science, vol. 112, pp. 2519–2528, 2017.

[22] R. S. Rasooly, B. Akolkar, L. M. Spain, M. H. Guill, C. T. Del Vecchio, and L. E. Carroll, "The national institute of diabetes and digestive and kidney diseases central repositories: A valuable resource for nephrology research," Clinical Journal of the American Society of Nephrology, vol. 10, no. 4, pp. 710–715, 2015.

[23] A. Dutta, M. K. Hasan, M. Ahmad, M. A. Awal, M. A. Islam, M. Masud, and H. Meshref, "Early prediction of diabetes using an ensemble of machine learning models," International Journal of Environmental Research and Public Health, vol. 19, no. 19, p. 12378, 2022. doi: https://doi.org/10.3390/ijerph191912378.

[24] N. H. Taz, A. Islam, and I. Mahmud, "A comparative analysis of ensemble-based machine learning techniques for diabetes identification," IEEE, 2022. [Online]. Available: https://www.ieee.org. [Accessed: 5 Sep. 2024].

[25] H. Zhou, Y. Xin, and S. Li, "A diabetes prediction model based on Boruta feature selection and ensemble learning," BMC Bioinformatics, vol. 24, no. 224, 2023. [Online]. Available: https://www.biomedcentral.com/articles/10.1186/s12859-023-05505-9. [Accessed: 5 Sep. 2024].

[26] H. Ulutas, R. B. Günay, and M. E. Sahin, "Detecting diabetes in an ensemble model using a unique PSO-GWO hybrid approach to hyperparameter optimization," Neural Computing and Applications, 2024. doi: https://doi.org/10.1007/s00500-024-06548-5.