Elinvo (Electronics, Informatics, and Vocational Education)
Vol 10, No. 2, 2025, pp. 205-215
ISSN 2580-6424 (printed) | ISSN 2477-2399 (online)

Assessment of Diabetes Classification Using Ensemble Learning Methods: Bagging, Random Forest, and Extreme Gradient Boosting

Jonas de Deus Guterres[1*], Fatchul Arifin[2]

[1,2]Yogyakarta State University, Yogyakarta, Indonesia
[1] jonas.2022@student.uny.ac.id*; [2] fatchul@uny.ac.id;
* corresponding author

Article Info

Article history:
Received September 23, 2025
Revised December 01, 2025
Accepted December 03, 2025
Available Online December, 2025

Keywords:
Diabetes classification; ensemble learning; random forest; XGBoost; bagging

Abstract

Diabetes mellitus is a chronic disease with increasing global prevalence, making accurate and reliable prediction an important research challenge in healthcare analytics. Although numerous machine learning techniques have been applied to diabetes classification, previous studies have reported inconsistent and moderate predictive performance, particularly when using limited datasets and single classifiers. This study addresses this problem by conducting a structured performance assessment of ensemble learning methods for diabetes classification using the Pima Indians Diabetes Dataset. Three ensemble algorithms—Bagging, Random Forest, and Extreme Gradient Boosting (XGBoost)—were evaluated under identical experimental conditions. The assessment investigated the impact of feature selection based on Pearson correlation compared to using all available features, along with systematic hyperparameter tuning and five-fold cross-validation. The model's efficacy was assessed through the application of accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC). The findings collectively underscore the efficacy of ensemble learning methodologies, demonstrating their capacity to yield precise classification results. Models that incorporated the complete feature set consistently outperformed those that utilized a subset of features. The Bagging classifier exhibited superior performance relative to the other techniques under investigation, achieving an AUC of 0.87 and a testing accuracy of 0.83. Although Random Forest also demonstrated commendable performance, XGBoost exhibited signs of overfitting, notwithstanding its high training accuracy. Consequently, these results indicate that an effective ensemble-based diabetes classification method has been identified within the experimental parameters of this study.

INTRODUCTION

Diabetes mellitus is one of the most common chronic diseases worldwide. It remains a significant issue among health systems of society in general [1]. In case of uncontrolled diabetes, cardiovascular diseases, kidney failure, and other serious sequelae were the complications [2], and precise diagnosis of diabetes at the earliest stage is a must to prevent major complications or death.

Machine learning-based classification models have been applied in many fields on diverse topics, for example, educational data mining, student outcome prediction, etc., based on decision trees and ensemble learning to improve the prediction capacity [3], [4].Recently, machine learning (ML) methods have been widely used in healthcare to perform disease prediction and clinical choices of

patient treatment and classification of diabetes [5], [6]. These techniques are effective for studying complex multivariate or multidimensional data and can identify hidden trends that may be overlooked by normal statistics or single classifiers [5], [7].

In public-facing medical data, for example, the Pima Indians Diabetes Dataset (PIDD) has been widely used to evaluate diabetes prediction models [8]. The dataset's relatively small sample size, moderate class imbalance, and weak to moderate correlations between the clinical characteristics of the disease and the target labels are concerning. Many of the machine learning algorithms are already well known to be excellent predictors, but the attributes of their inheritance and model design are associated with difficulties in classifying the diabetes task and predicting variables [9]. So, much of previous research has shown reasonable accuracy with a necessity for strong, predictable, stable models [10], [11], in our case, and, of course, with a load requiring the robustness of models.

Kumari et al. applied soft voting ensemble with Random Forest, logistic regression, and Naïve Bayes classifiers, for instance, and achieved an accuracy of around 79% [12]. Although the combined classifiers showed that performance achieved was still limited. Similarly, Saxena et al. studied feature selection and classification in earlier studies and showed that Random Forest performs with the best accuracy (79.8%) compared to other models showing relatively poor performance [9]. Febrian et al. observed that, in addition, most of the classifiers widely employed, e.g., K-Nearest Neighbor and Naïve Bayes, have lower than 80% accuracy [10], reinforcing the consistent issue of the performance of each learning algorithm. To overcome these drawbacks, ensemble learning is a recent subfield that focuses on the latest developments from recent years. Ensemble is defined as involving multiple base learners and features a unique set of actions collected from disparate sources and combined into a single framework for accurate predictions and the most stable generalizations [13]. This is in line with what Laila et al. demonstrated: that ensemble-based architectures outperform single-classifier models for the prediction of early-stage diabetic risk and accurately characterize the nonlinear relationships in clinical features [5]. Hasan et al. [7] showed that using ensemble learning is robust to compact and noisy medical data. However, the majority of studies do not provide integrated knowledge processes on how the attributes of the dataset, like feature correlation, preprocessing approach, and hyperparameter selection, contribute to the extracted results and only indicate the estimated performance.

In fact, most of the studies of diabetes prediction contain patchiness in model feature selection and data processing. This basic Pearson correlation is one of the most popular parameters for exploratory validity and interpretability, and little or no additional work was done to explain how weak features to target variable correlations affect ensemble model performance [9]. Because, based on several recent works, ensemble methods (e.g., bagging and random forest) perform incredibly well with a single feature and a weak predictor with very high variance, mainly attributable to the ensemble processes of weak learners [13], [14], and they greatly enhance the generalization. This property is crucial to the case of PIDD sets, where no feature is truly effective at prediction; it should also dominate.

Similarly, an effective use of Extreme Gradient Boosting (XGBoost) due to its capacity to adapt regularization, consider nonlinear interaction of features, and model fit, as well as optimize model performance, gained popularity [15]. Other studies have also shown the competitive performance of the XGBoost approach on structured classification-based tasks [16]. However, XGBoost is hyperparameter sensitive; poor selection will often lead to overfitting when the training set is in short supply [17]. These results show much focus on the systematic optimization of hyperparameters for the performance evaluation of ensemble learning methods.

While diabetes prediction has been widely studied, there is limited literature available to assess it in two general categories [18]. In brief, the first class addresses system-level research [19] and offers potential uses for monitoring algorithms, but there is no algorithmic comparison. In the second half, it designs the algorithm and introduces performance measurement in a much more basic way and explains

the research problem or the rationale behind the methodology (like feature search and normalization approaches) [17]. As a result, no complete comparative study has yet been performed to compare various ensemble classifiers, in particular in the single experimental setting.

Thus, there is a corresponding research gap for doing a structured performance measurement that links dataset properties directly with the behavior of ensemble models. Without something else to make, simply building an early detection engine or user interface is deployable, for which only extensive comparative analysis is required that captures the efficacy of different ensemble training methods on similar datasets under continuous preprocessing, feature selection, and validation processes. An analysis of such a type is vital for identifying the best ensemble methods and for providing methodological insights for future system development and clinical use.

Thus, this paper presents the work by measuring the performance of ensemble learning methods to classify diabetes and how it addresses those limitations. Specifically, this project investigates the ensemble techniques—bagging, random forest, and extreme gradient boosting (XGBoost)—on the Pima Indians Diabetes Dataset to estimate the comparative performance of the three ensemble techniques to one another. This evaluation is systematically considered and discussed in relation to the effects of feature correlation, feature selection from the full set of features, and hyperparameter tuning on classification performance. This evaluation method aims to find out the ensemble method that is preferable for evaluation measures with a lot of metrics, using the same standard preprocessing as k-fold cross-validation in the evaluation.

The predictive performance has been terrific. Our evaluation thus tackles the inconsistent accuracy observed in prior studies and is intended to validate whether ensemble learning can improve diabetes classification as long as the setup is correct. In the future, I expect to assist in the development of a systematic model of ensemble-based diabetes prediction and also guide the formulation of more stable machine learning models for decision support systems for health needs.

METHODS

This study aims to compare how well different ensemble learning methods classify diabetes. Instead of creating a system for practical use, the goal is to evaluate and compare how well ensemble algorithms predict outcomes under controlled, identical experimental conditions. The experiments were run using Python in the Google Colaboratory environment. The data used came from the publicly available Pima Indians Diabetes Dataset (PIDD), which is available on the Kaggle platform at https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database. Figure 1 shows the overall process of the proposed methodology.

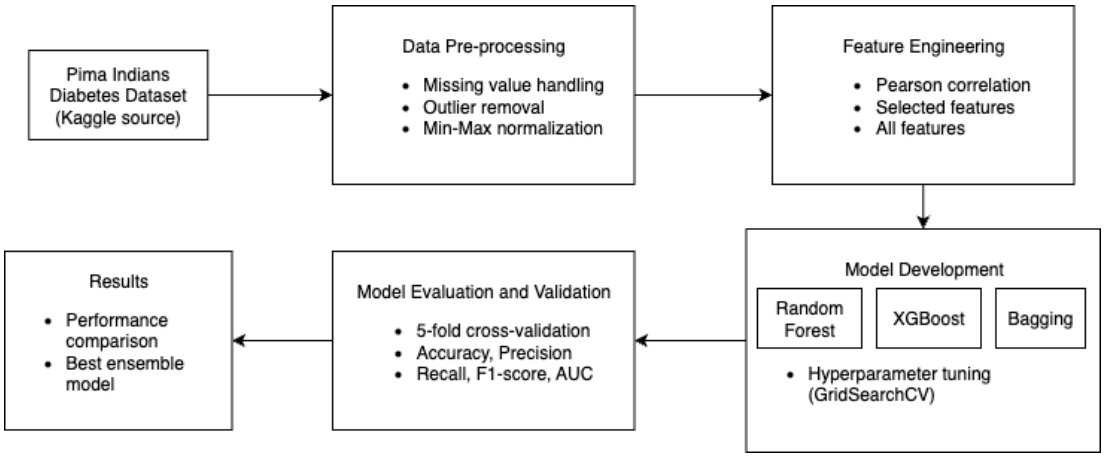


Figure 1. Workflow of the proposed methodology

ISSN 2580-6424 (printed) | ISSN 2477-2399 (online)

Proposed Ensemble Method

We focused on three ensemble learning methods: Random Forest, Extreme Gradient Boosting (XGBoost), and Bagging. Such algorithms have previously demonstrated superior performance on small datasets, reducing overfitting and identifying nonlinear patterns in feature correlations.

The Random Forest (RF) ensemble method builds multiple decision trees using bootstrap samples and uses the cumulative predictions to reduce variance and improve generalization [20]. RF is well suited for solving overfitting problems that frequently exist in individual decision tree models, particularly in classification problems that require structured datasets [14]. In addition, Random Forest utilizes a feature importance mechanism that examines whether features have any significant contribution by detecting the increase in prediction error in out-of-bag samples after permuting feature values [14].

Chen and Guestrin proposed an ensemble method, based on gradient boosting, called Extreme Gradient Boosting (XGBoost), to improve both learning time and prediction accuracy [15]. XGBoost builds decision trees in a sequential manner, with each of the new models making an effort to correct the mistakes introduced by the previous model by decreasing a loss function that was specified [21]. The final prediction is made through consolidating the weights derived from many weak learners, where it can detect complex nonlinear dependencies [14]. Moreover, regularization techniques are used in XGBoost to achieve precision and overcome the issues regarding model complexity and overfitting—important issues in relatively small datasets [14].

Bagging, an ensemble meta-classifier, makes several bootstrap samples from the data, while the predictions are combined to produce the output [20]. Bagging stabilizes the model and reduces the variance by averaging predictions at multiple base learners, specifically on small datasets with noisy features [22]. Despite its simplicity, bagging can be successfully implemented in classification tasks, particularly when every individual learner is sensitive to data variation [23].

Bagging and random forest are essential ensemble methods in reducing variance by combining multiple predictors trained on bootstrap samples [21], [23]. These techniques have attracted considerable attention in the field of machine learning for both medicine and healthcare, especially due to their robustness and tolerance to noisy high-dimensional datasets [6].

Dataset Description

We used the Pima Indians Diabetes Dataset, which represents clinical data from 21–81-year-old female patients. This dataset has 8 numeric features (pregnancies, glucose concentration, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, and age), and one binary target variable (1 or 0), namely diabetes. With 768 instances in this dataset (268 positive, 500 negative samples), it indicates a moderate class imbalance. Afterward, the dataset was divided into training and testing (80:20) to use for model testing. There was no data augmentation and no synthetic data. This procedure was done to respect the originality of data distribution and to enable a fair comparison with previous studies that employed the same dataset. Data augmentation techniques are identified as extension points in future work.

Data Pre-processing

Preprocessing the data improved data quality and model reliability. Many attributes in the dataset had zero values, which are physiologically implausible, notably for insulin and skin thickness features. These were considered as outliers and the samples for four affected individuals were discarded. Consequently, the dataset size was reduced to 764 instances. As described in (1), min–max normalization was defined to scale all numerical features into the range [0, 1]. This method of

normalization was chosen as it preserves the original data distribution while preventing features with larger numeric ranges from dominating the learning process. For tree-based ensemble methods, min–max normalization is well suited and thus allows for a consistent comparison between classifiers.

$$x' = a\ \frac{(x-\min(x))(b-a)}{\max(x)-\min(x)} \quad (1)$$

Where x' is the normalized value x, x is the original variable value being normalized, $\min(x)$ is the minimum value in the dataset, $\max(x)$ is the maximum value in the dataset; a and b are the lower and upper limits of the normalization range.

While the z-score standardization and other normalization methods were also explored, min–max normalization was ultimately chosen. The approach was selected due to its simplicity, usability, and its frequent application in medical machine learning studies which deal with clinical variables that have constrained states in clinical data.

Table 1 shows the influence of mix-max scaling on the normalization of these selected dimensions, as well as a comparison of how these variables compared with those of others after a mix-max scaler was added to them.

Table 1. Normalization Data

	Raw Data				Normalization Data			
index	Pregnancies	Glucose	BMI	Age	Pregnancies	Glucose	BMI	Age
count	764.0	764.0	764.0	764.0	764.0	764.0	764.0	764.0
mean	3.86	121.34	32.42	33.19	0.23	0.5	0.29	0.2
std	3.37	30.13	6.85	11.7	0.2	0.19	0.14	0.2
min	0.0	44.0	18.2	21.0	0.0	0.0	0.0	0.0
25%	1.0	99.0	27.5	24.0	0.06	0.35	0.19	0.05
50%	3.0	117.0	32.0	29.0	0.18	0.47	0.28	0.13
75%	6.0	140.0	36.52	41.0	0.35	0.62	0.37	0.33
max	17.0	199.0	67.1	81.0	1.0	1.0	1.0	1.0

Feature Engineering

Feature selection greatly improves model interpretability and generalization, particularly for medical datasets; redundant details may be harmful to improving modeling [24], [25]. In addition, a Pearson correlation analysis was performed to confirm the linear relationship between each input feature and the response variable. Pearson's correlation coefficients vary from −1 to 1; the nearest they can attain to zero is usually an indication of weak correlation among variables. The correlational analysis was succeeded by the construction of model experiments: (a) model training where the feature being selected has a higher correlation value and (b) model training with all features. Therefore, the general goodness of fit in ensemble learning can be systematically checked: is the feature selection a better fit than the feature matching?

Other feature selection methods such as Spearman correlation, mutual information, and wrapper-based methods were not used, but Pearson correlation was selected, which is computationally efficient and optimal for continuous numerical features. Various techniques to handle more complex feature selection are therefore proposed to be examined in the future.

$$r = \frac{\sum(x_i-\overline{x})(y_i-\overline{y}))}{\sqrt{\sum(x_i-\overline{x})^2\sum((y_i-\overline{y})^2)}} \quad (2)$$

Where r is the Pearson correlation coefficient, x_i is the individual data points in the variable x, $\overline{x}$ is the mean of all x-values, y_i is the individual data points in the variable y, $\overline{y}$ is the mean of all y-values.

Hyperparameter Tuning

The hyperparameter tuning was performed to maximize model performance and ensure fair comparisons of classifiers. A grid search was conducted, providing various sets of salient hyperparameters for each ensemble method and its corresponding combinations of them. Parameter ranges were selected based on the most frequently found ranges that had already been proposed in the published works (Table 2).

Table 2. Hyperparameter Tuning

Model	Parameter	Range
Random Forest Classifier	Criterion	[‘Gini’, ‘entropy’]
	Max Depth	[3,5,10]
	Max Features	[‘auto’, ‘sqrt’,’log2’]
	Min Samples Split	[2,5,10]
	N Estimators	[200,500,800]
XGBoost Classifier	Column Sample by Level	[0.1,0.4,1.0]
	Column Sample by Tree	[0.1,0.4,1.0]
	Learning rate	[0.01,0.1,0.2,0.3]
	Max Depth	[3,5,6,10,15,20]
	N Estimators	[50,100,150]
	Sub Sample	[0.1,0.5,1.0]
Bagging Classifier	Max Features	[0.5,1.0,1.5,2.5,3.5,5.0]
	Max Samples	[0.1,0.2,0.5,10.0]
	N Estimators	[25,50,100]

Evaluation Metrics

Model accuracy (ACC), precision (PR), recall (RC), F1-score (F1), and the area under the receiver operating characteristic curve (AUC) were used. Most of these measures are essentially summaries of the classification in different scenarios, especially so for datasets of classes with class-imbalanced data. The receiver operating characteristic (ROC) curve and area under the curve (AUC) are widely used as performance metrics on binary classification problems and have been gaining popularity in the medical classification arena with respect to class imbalances [26]. In similar fashion, k-fold cross-validation is applied to verify a stable performance estimation and reduce sampling bias [27].

$$ACC = \frac{TP+TN}{TP+FP+TN+FN} \tag{3}$$

$$PR = \frac{TP}{TP+FP} \tag{4}$$

$$RC = \frac{TP}{TP+FN} \tag{5}$$

$$F1 = 2 \times \frac{(PR \times RC)}{(PR+RC)} \tag{6}$$

$$AUC = \sum_{i=1}^{n-1} \frac{(FPR_{i+1}-FPR_i)\times(PPR_{i+1}-PPR_i)}{2} \tag{7}$$

Where,

- FPR_i and TPR_i = False Positive rate and True Positive Rate, which are

$$FPR = \frac{FP}{FP+TN} \tag{8}$$

and,

$$TPR = \frac{TP}{TP+FN}. \tag{9}$$

Where FP is false positive, TN is true negative, TP is true positive, and FN is false negative.

- n = the total number of points on the receiver operating characteristic (ROC) curve.

80:20 was used to split the dataset into training and testing set. A 5-fold cross-validation (k = 5) for the training data set was done to allow for valid validation robustness and reduce performance estimation variation [27].

RESULT AND DISCUSSION

Correlation Analysis and Selected Features

Pearson correlations between each feature and the target label were shown in Figure 2. Data show that all features have correlations < 0.50. With an index of zero to no correlation at a value of 0.20, as shown in Table 3, the four features selected from the feature selection experiment were glucose (0.47), body mass index (BMI) (0.29), age (0.24), and pregnancies (0.22). Fewer correlations occurred on other features, and such correlations were discarded for the selected feature.

Table 3. Selected Features Based on Pearson Correlation

Feature	Correlation Value
Glucose	0.47
BMI	0.29
Age	0.24
Pregnancies	0.22

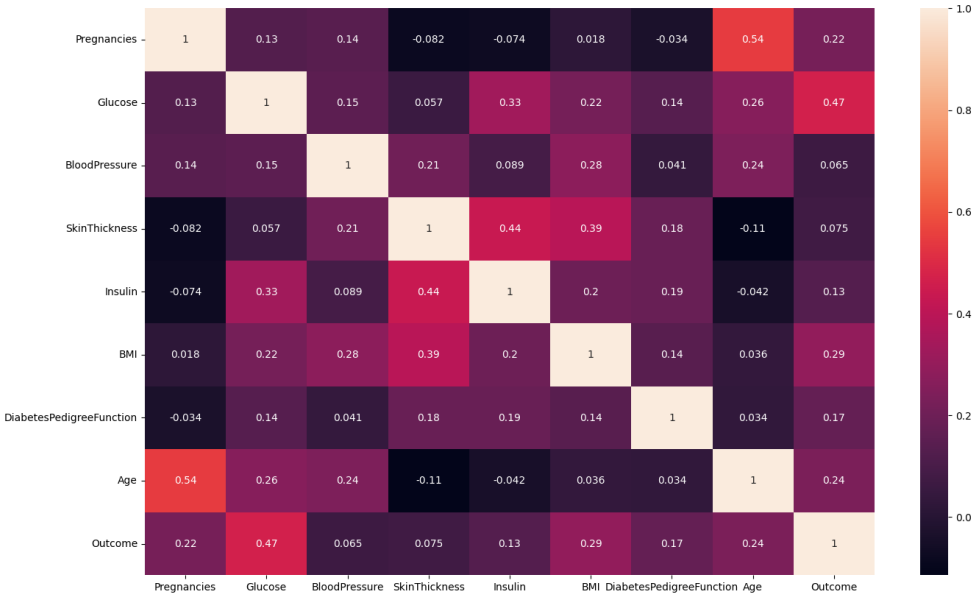


Figure 2. Correlation between features

Feature Configuration and Model Setup

Two feature configurations were proposed for each ensemble algorithm to assess the effect of feature selection, chosen features (SF) and all features (AF). Table 4 provides a summary of feature usage configuration for each model.

Table 4. Feature Configuration for Each Ensemble Model

Model ID	Algorithm	Feature Set
RF-SF	Random Forest	Selected Features
RF-AF	Random Forest	All Features
XGB-SF	XGBoost	Selected Features
XGB-AF	XGBoost	All Features
BAG-SF	Bagging	Selected Features
BAG-AF	Bagging	All Features

Training Performance After Hyperparameter Tuning

Hyperparameter tuning for each ensemble algorithm was done independently using Grid Search with fivefold cross-validation. The optimal hyperparameter configurations obtained are shown in Table 5.

Table 5. Hyperparameter Tuning Result

Algorithm	Parameter	Range
Random Forest Classifier	Criterion, Max Depth, Max Features, Min Samples Split, N Estimators	Entropy, 10, auto, 2, 500
Xgboost Classifier	Column Sample by Level, Column Sample by Tree, Learning rate, Max Depth, N Estimators, Sub Sample	0.5, 0.9, 0.1, 6, 50, 0.8
Bagging Classifier	Max Features, Max Samples, N Estimators	1.0, 0.2, 50

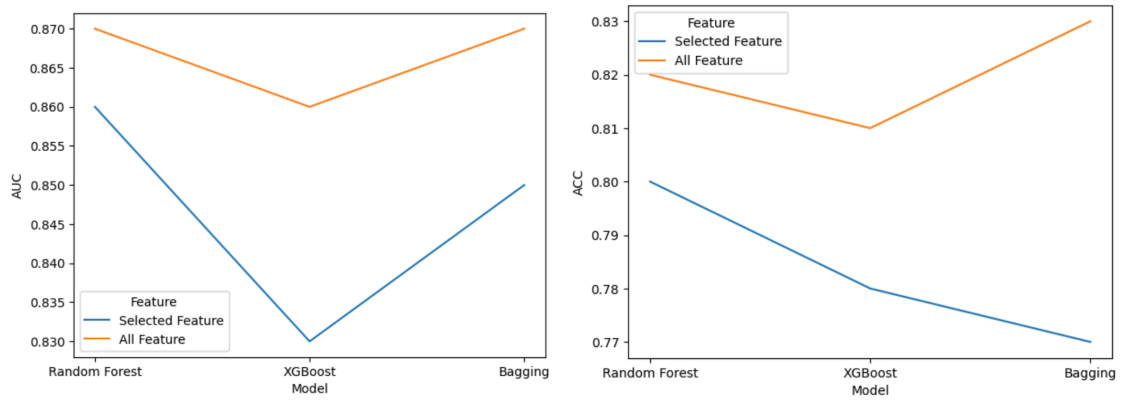
Table 6 displays the results after tuning hyperparameters for training performance. Two different feature configurations of XGBoost models achieved the highest training accuracy to a maximum of 1.00. Random Forest models then obtained 0.98 (RF-SF) and 0.99 (RF-AF) accuracies. Bagging models performed poorly on training accuracy (0.82 and 0.84, respectively, for BAG-SF and BAG-AF). Hyperparameter tuning provided slight gains in F1-score among all the models, suggesting the ensemble methods is relatively robust to parameter variation.

Table 6. Training Performance Result

Model ID	PR	RC	F1	AUC	ACC
RF-SF	0.97	1.00	0.98	0.97	0.98
RF-AF	0.99	1.00	0.99	0.99	0.99
XGB-SF	1.00	1.00	1.00	1.00	1.00
XGB-AF	1.00	1.00	1.00	1.00	1.00
BAG-SF	0.83	0.91	0.87	0.79	0.82
BAG-AF	0.85	0.92	0.88	0.81	0.84

Testing Performance and Comparison with Previous Studies

Finally, the model was tested on the testing dataset. The generalized model was evaluated with two performance metrics, AUC and accuracy (ACC).



(a) (b)

Figure 3. Comparison of the score metrics for two types of features (a) in AUC, (b) in ACC

Among the methods studied in 3(a), Random Forest and Bagging have a marginally larger AUC than that of XGBoost. Moreover, it can be seen in Figure 3(b) that an all-feature-encompassing configuration is always superior to a selected-feature-inclusive one, Bagging attaining the best accuracy and XGBoost the weakest.

The performance statistics of tests are displayed in Table 7. All ensemble algorithms show that models trained with all features outperform models trained with a selected set of features. The best-performing model under this test was the bagging model with all features (BAG-AF), with a significant test accuracy of 0.83 and an AUC of 0.87. Random Forest with all features (RF-AF) had similar accuracy and AUC of 0.82 and 0.87, respectively. XGBoost models showed somewhat lower testing accuracy while exhibiting perfect performance on training data, revealing possible overfitting.

Table 7. Testing Data Result

Model ID	PR	RC	F1	AUC	ACC
RF-SF	0.89	0.82	0.85	0.86	0.80
RF-AF	0.89	0.85	0.87	0.87	0.82
XGB-SF	0.88	0.79	0.84	0.83	0.78
XGB-AF	0.91	0.81	0.86	0.86	0.81
BAG-SF	0.85	0.81	0.83	0.85	0.77
BAG-AF	0.88	0.88	0.88	0.87	0.83

According to a comparison with previous methods, illustrated in Table 8, the proposed approach performed significantly better compared with previous studies carried out from 2021 to 2023 in terms of accuracy and AUC values.

Table 8. Comparison of The Proposed Method with Previous Studies

Author	Best Method	PR	RC	F1	AUC	ACC
Kumari et al. (2021) [12]	Soft Voting Classifier (Random Forest, Logistic Regression and Naïve Bayes)	0.73	0.70	0.72	0.81	0.79
Saxena et al. (2022) [9]	Random Forest	0.71	0.80	-	0.84	0.80
Febrian et al. (2023) [10]	Naïve bayes	0.73	0.71	-	-	0.76
Proposed Method (AF)	Bagging (BAG-AF)	0.88	0.88	0.88	0.87	0.83
	Xgboost (XGB-AF)	0.91	0.81	0.86	0.86	0.81
	Random Forest (RF-AF)	0.89	0.85	0.87	0.87	0.82

CONCLUSION

The current study had to resolve the problem of varying and moderate performance seen in the previous diabetes classification studies, which depended on the Pima Indians Diabetes Dataset. Several previous works have employed machine learning algorithms without determining how feature setting, ensemble strategies, and fine-tuning of hyperparameters lead to classification performance under similar experimental conditions. This paper aims to tackle the problem by having an intensive performance evaluation of three ensemble learning methods (bagging, random forest, and extreme gradient boosting, or XGBoost) in an organized fashion for the two feature configurations (e.g., selected features based on Pearson correlation and all current features). All the models were assessed according to the same preprocessing approach, hyperparameter tuning, and validation approach to enable fair comparisons. The empirical results confirm that ensemble learning models deliver reliable performance on diabetes classification, with the models trained using all features outperforming those trained using few features. The Bagging classifier (which outperformed the others in the analysis group) demonstrated the highest testing accuracy of 0.83 (0.87 AUC). Even though XGBoost performed completely in our training data, we can also say its performance test results are lower than those of Random Forest, which is stable and competitive. Tuning hyperparameters gave modest improvements in performances, which proved the performances of the model ensemble methods evaluated here are comparatively robust to parameter changes. It is inferred that the issue of determining a suitable method of ensemble learning for diabetes classification in a controlled way could be solved in this study. The findings clearly demonstrate that ensemble methods, such as bagging with all aspects, provide a significantly superior and homogeneous performance compared to previous studies. While our study investigates the Pima Indians Diabetes

Dataset, the ensemble learning-based methods that we considered were not specific to the dataset but are applicable to other structured medical datasets with similar traits. Hence, results indicate that ensemble methods like bagging and random forest are better suited to use with very small sample sizes and minimal feature correlation. Additional studies should be done on alternative feature selection techniques, data augmentation techniques, or even datasets to increase the generalization and practicality of the models in real clinical settings.

ACKNOWLEDGMENT

The author was grateful to the lecturers of Department of Electronics and Informatics Engineering Education, Faculty of Engineering, Universitas Negeri Yogyakarta for advising and financially providing him in the writing of the paper.

REFERENCES

[1] World Health Organization, *Global Report on Diabetes*, Geneva, Switzerland: WHO Press, 2021.
[2] B. Hidayat, R. V. Ramadani, A. Rudijanto, P. Soewondo, K. Suastika, and J. Y. S. Ng, "Direct medical cost of type 2 diabetes mellitus and its associated complications in Indonesia," *Value in Health Regional Issues*, vol. 28, pp. 82–89, 2022, doi: 10.1016/j.vhri.2021.04.006.
[3] I. Gunawan and A. P. Widyassari, "Integrating psychological stress indicators with academic data for student dropout prediction: A decision tree and expert system approach," *Elinvo (Electronics, Informatics and Vocational Education)*, vol. 10, no. 2, pp. 131–146, 2025, doi: 10.21831/elinvo.v10i2.89031.
[4] M. A. S. Pawitra, H.-C. Hung, and H. Jati, "A machine learning approach to predicting on-time graduation in Indonesian higher education," *Elinvo (Electronics, Informatics and Vocational Education)*, vol. 9, no. 2, pp. 294–308, 2024, doi: 10.21831/elinvo.v9i2.77052.
[5] U. E. Laila, K. Mahboob, A. W. Khan, F. Khan, and T. Whangbo, "An ensemble approach to predict early-stage diabetes risk using machine learning: An empirical study," *Sensors*, vol. 22, no. 14, p. 5247, 2022, doi: 10.3390/s22145247.
[6] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, "Deep EHR: A survey of recent advances in deep learning techniques for electronic health record analysis," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 5, pp. 1589–1604, 2018, doi: 10.1109/JBHI.2017.2767063.
[7] M. K. Hasan, M. A. Alam, D. Das, E. Hossain, and M. Hasan, "Diabetes prediction using ensembling of different machine learning classifiers," *IEEE Access*, vol. 8, pp. 76516–76531, 2020, doi: 10.1109/ACCESS.2020.2989857.
[8] Kaggle, "Pima Indians Diabetes Database," 2023. [Online]. Available: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database
[9] R. Saxena, S. K. Sharma, M. Gupta, and G. C. Sampada, "A novel approach for feature selection and classification of diabetes mellitus," *Computational Intelligence and Neuroscience*, vol. 2022, 2022, doi: 10.1155/2022/3820360.
[10] M. E. Febrian, F. X. Ferdinan, G. P. Sendani, K. M. Suryanigrum, and R. Yunanda, "Diabetes prediction using supervised machine learning," *Procedia Computer Science*, vol. 216, pp. 21–30, 2023, doi: 10.1016/j.procs.2022.12.107.
[11] S. S. M. Rahman, A. H. M. Kamal, and M. M. Rahman, "A comparative study of machine learning algorithms for diabetes prediction," *Journal of Healthcare Engineering*, vol. 2021, 2021, doi: 10.1155/2021/5542400.
[12] S. Kumari, D. Kumar, and M. Mittal, "An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier," *International Journal of Cognitive Computing in Engineering*, vol. 2, pp. 40–46, 2021, doi: 10.1016/j.ijcce.2021.01.001.
[13] D. C. Yadav and S. Pal, "An experimental study of diversity of diabetes disease features by bagging and boosting ensemble method with rule-based machine learning classifier algorithms," *SN Computer Science*, vol. 2, no. 1, p. 50, 2021, doi: 10.1007/s42979-020-00425-5.
[14] S. K. Kiangala and Z. Wang, "An effective adaptive customization framework for small manufacturing plants using extreme gradient boosting and random forest ensemble learning algorithms," *Machine Learning with Applications*, vol. 4, p. 100024, 2021, doi: 10.1016/j.mlwa.2021.100024.
[15] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[16] T. Wang, Y. Bian, Y. Zhang, and X. Hou, "Classification of earthquakes, explosions and mining-induced earthquakes based on XGBoost algorithm," *Computers & Geosciences*, vol. 170, p. 105242, 2023, doi: 10.1016/j.cageo.2022.105242.
[17] T. Shaikhina, D. Lowe, S. Daga, D. Briggs, R. Higgins, and N. Khovanova, "Decision tree and random forest models for outcome prediction in medical datasets," *Biomedical Signal Processing and Control*, vol. 52, pp. 456–462, 2019, doi: 10.1016/j.bspc.2017.01.012.
[18] A. S. Sadiq, S. M. Abdulazeez, and D. M. Zeebaree, "A comparative study of machine learning classifiers for diabetes disease prediction," *Technology Reports of Kansai University*, vol. 62, no. 3, pp. 1511–1523, 2020.
[19] C. V. Raghavendran, G. N. Satish, N. S. L. K. Kurumeti, and S. M. Basha, "An analysis on classification models to predict possibility for type 2 diabetes of a patient," in *Innovative Data Communication Technologies and Application (ICIDCA 2021)*, Singapore: Springer, 2022, pp. 181–196, doi: 10.1007/978-981-16-7167-8_14.
[20] R. Kasarda, N. Moravčíková, G. Mészáros, M. Simčič, and D. Zaborski, "Classification of cattle breeds based on the random forest approach," *Livestock Science*, vol. 267, p. 105143, 2023, doi: 10.1016/j.livsci.2022.105143.
[21] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
[22] G. O. Anyanwu, C. I. Nwakanma, J. M. Lee, and D. S. Kim, "Novel hyper-tuned ensemble random forest algorithm for the detection of false basic safety messages in internet of vehicles," *ICT Express*, vol. 9, no. 1, pp. 122–129, 2023, doi: 10.1016/j.icte.2022.06.003.
[23] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996, doi: 10.1007/BF00058655.
[24] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.
[25] M. H. Al-Timemy, R. S. Khudhair, and A. K. Al-Mashhadani, "Diabetes mellitus prediction using machine learning techniques," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 2, pp. 416–423, 2021, doi: 10.14569/IJACSA.2021.0120251.
[26] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006, doi: 10.1016/j.patrec.2005.10.010.
[27] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. 14th Int. Joint Conf. Artificial Intelligence (IJCAI)*, 1995, pp. 1137–1143.