All Sciences Proceedings
http://as-proceeding.com/

3rd International Conference on Innovative Academic Studies

September 26-28, 2023 : Konya, Turkey

ICIAS
https://www.icias.net/

A hybrid model with feature selection and hyper parameters for detecting diabetes in PIMA Indian dataset

Cihan Açıkyürek[*] and Gökalp Çınarer [2]

[1] Department of Electrical and Electronics Engineering/Faculty of Engineering and Architecture, Yozgat Bozok University, Yozgat, Türkiye
[2]Departmentof Computer Engineering/Faculty of Engineering and Architecture, Yozgat Bozok University, Yozgat, Türkiye

[*](cihanackyurek@gmail.com) Email of the corresponding author

Abstract – Diabetes is a prevalent global health concern, with the timely detection of the disease playing a crucial role in treatment and prevention. Artificial Intelligence (AI) and Machine Learning (ML) algorithms have gained prominence due to their ability to analyze large datasets, aiding in disease diagnosis and treatment. This study focuses on developing accurate models for the early diagnosis of diabetes. We explored the performance of various ML algorithms, including K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Logistic Regression (LR), Extra Trees (ET), AdaBoost (AB), and Gradient Boosting (GB) while also employing different preprocessing techniques, hyperparameter tuning, XGBoost feature selection and crossover strategies. Furthermore, we tested a hybrid model using validation scenarios to assess its effectiveness. The study's outcomes revealed that the Logistic Regression algorithm achieved the highest classification accuracy, reaching 77%. This result highlights the potential of ML techniques, particularly Logistic Regression, in early diabetes diagnosis.

Keywords – Machine Learning, Diabetes Prediction, Artificial Intelligence, Hybrid Model

I. INTRODUCTION

Diabetes is a significant health problem that has a substantial impact on a large population globally. It is characterized by high blood sugar levels and can lead to significant health issue in the long term, such as heart, kidney disease, blindness, and nerve damageS [1]. Therefore, early detection of diabetes is of great importance in terms of treatment options and preventing disease progression [2].

ML algorithms have been investigated in many studies for the diagnosis and prediction of diabetes [3]. These algorithms have the potential to identify important features within the data by analyzing complex relationships. As a result, they can assist in the early management and diagnosis of diabetes. ML is a field of AI that involves analyzing large amounts of data to detect patterns and make predictions [4].

A review conducted by Masood et al. [5] addressed the use of different ML techniques in diabetes diagnosis, along with their advantages and disadvantages.

In a comparative study, the usability of different classification algorithms (Naive Bayes (NB), KNN, Decision Trees and Artificial Neural Networks (ANN)) for diabetes diagnosis was compared [6].

In another study the usability of ML algorithms for the diagnosis of gestational diabetes was examined [7]. The results indicated that gestational diabetes diagnosis could be accurately predicted using these algorithms

The aim of this study is to obtain the highest performance on ML algorithms with a hybrid model that produces the most suitable results for early diagnosis of diabetes. For this purpose, ML classification algorithms were analyzed using different hyperparameters and feature selection

methods. The study demonstrates the availability of AI-based models as the basic building block for the model to be used in the early diagnosis of diabetes. Developing and testing the model provides a ML model that can be used for early diagnosis and treatment in the healthcare industry.

In this study, diabetes detection was carried out with hybrid models developed from classical algorithm parameters using different hybrid models.

II. Materials and Method

ML is a method within the field of AI that enables data analysis and the performance of specific tasks [8]. This technology is being used in various sectors and has gained significant importance in the healthcare sector in recent years [8]. Early diagnosis of chronic diseases, especially diseases like diabetes, holds critical importance in monitoring patients' health status and initiating necessary treatment promptly [2]. ML can be employed as a valuable tool to accelerate this process and enhance accuracy.

Within the scope of the study, 10-fold cross validation process was applied on the hybrid model in order to test the data more stablely in the test set. The main features that affect the classification in the dataset were determined with the XGBoost feature selection methods. The best parameter values were determined with Randomize Search, one of the optimization algorithms. KNN, SVM, LR, Extra Trees, AdaBoost and Gradient Boosting algorithms were chosen to be tested on a diabetes-related dataset. The methodology includes preprocessing steps for the dataset, such as data normalization, processing of missing data, and feature selection. After these steps are completed, the dataset is divided for train 70% and test 30%, and the selected ML algorithms are trained using these datasets. In Figure 1, the flow diagram of the applied hybrid model is given.

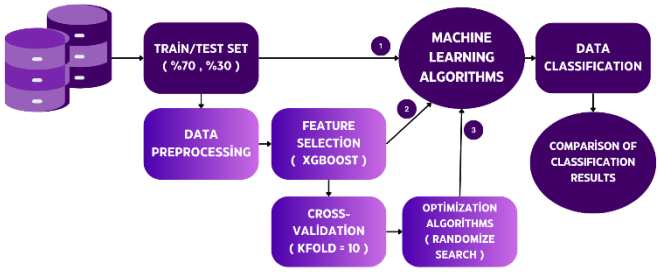


Fig. 1 Hybrid Model of Flow Diagram

After completing the training process, the performance of the algorithms has been evaluated using the test dataset, and the results have been compared using evaluation metrics such as prediction accuracy, precision, recall, f1 score and specificity. The obtained results have been interpreted to assess the performance of algorithms used in predicting diabetes and to determine the most effective algorithm.

A. *Dataset*

The open-source Pima Indians Diabetes dataset used in this study is approved by the National Institute of Diabetes and Digestive and Kidney Diseases [9-10]. This dataset contains information related to the diagnostic of diabetes in Pima Indian women. All patients included in the dataset are women of Pima Indian heritage and are at least 21 years old. Due to its inclusion of important data related to diabetes and its focus on Pima Indian women, this dataset serves as a valuable resource for constructing machine learning models for diabetes prediction.

The dataset consists of a total of 768 observations (rows) and 9 attributes (columns). Most of the attributes in the dataset have specific value ranges, but the Outcome (Class Variable) attribute has a categorical structure and contains values such as 0 or 1. Table 1 provides a detailed listing of these attributes. The dataset has been used as a fundamental resource in numerous studies that aim to predict diabetes.

Table 1. Attributes and value ranges in the dataset

Attributes	Value Ranges
Pregnancies	[0-17]
Glucose	[0-199]
BloodPressure	[0-122]
Skin Thickness	[0-99]
Insulin	[0-846]
BMI	[0-67]
Diabetes Pedigree Function	[0-2]
Age	[21-81]
Outcome	[0-1]

A Heatmap graph is a tool used to examine the correlation relationships between variables for dataset.

Correlation measures direction and strength of the relationship with variables. The Heatmap visualizes these correlations through a color scale.

Figure 2 shows the color scale in the Heatmap graph represents the correlation coefficients between variables. Positive correlation signifies a direct proportional relationship between variables, while negative correlation indicates an inverse proportional. Intense colors, such as dark purple or dark pink shades, represent high correlation, while lighter colors, such as light purple or light pink shades, represent low correlation. The heatmap graph allows us to understand the relationships by evaluating the correlations between variables. It's important to note that the heatmap graph provides information about correlations but does not establish causality or interactions, which may require more detailed analysis and research.

The Heatmap graph in the dataset displays correlations between variables, for example, dark-colored positive correlation among the 'Outcome' and 'Glucose,' suggesting a relationship between glucose levels and the likelihood of diabetes. Similarly, a positive correlation can be observed between 'BMI' and 'Skin Thickness.'

Another classification process performed on the data set was made by utilizing the feature selection capabilities of the XGBoost algorithm. It has been researched that the XGBoost algorithm is effective in feature selection in classification problems and this conclusion has been reached. It was concluded that only four attributes caused an increase in the results.

In this case, only four of the attributes 'Glucose', 'BMI', 'DiabetesPedigreeFunction' and 'Age' were classified.

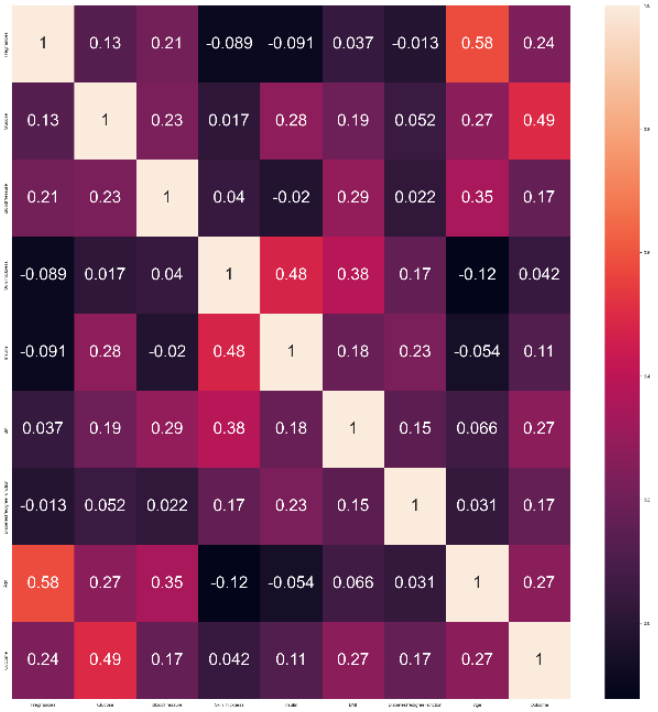


Fig. 2 Heatmap Chart Describing the Dataset

The Stripplot graph in the dataset gives the distribution of variables in relation to the 'Outcome' column. Stripplot is a type of graph used to visualize the distribution of a numerical variable associated with a categorical variable.

That's correct. The Stripplot graph provides a visual representation of how each variable is distributed based on the categorical variables in the 'Outcome' column. This graph is used to understand relationships between variables and analyze the structural characteristics of the dataset. By observing the distribution patterns of variables across different categories, insights can be gained regarding the potential influence of the categorical variable on the numerical variable.

When examining the graph in Figure 3, it is determined that the most concentrated distribution is observed in the "Age" and "Glucose" graphs. This finding highlights an interesting point to further investigate the impact of age and glucose levels on the "Outcome" column in more detail.

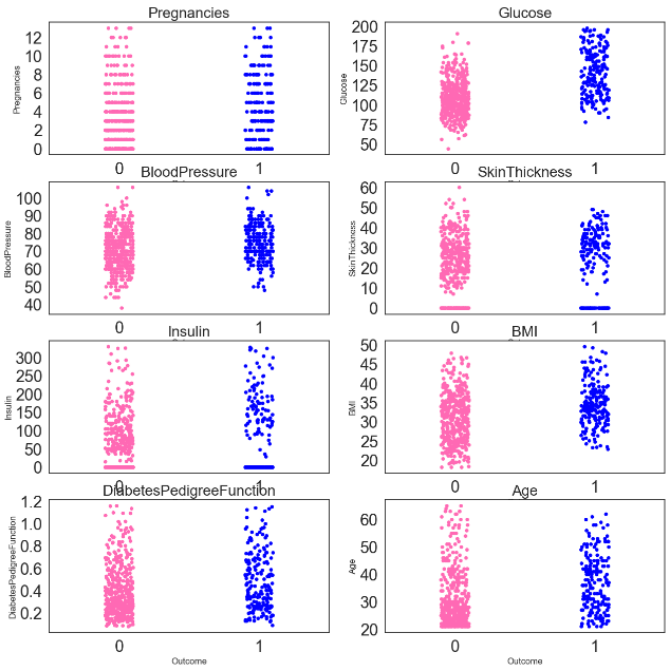


Fig. 3 Dataset Stripplot Graphics

The class distribution of the presented dataset has been examined in Figure 4. In the dataset represented by the "Class" column, there are 500 data class "0" (non-diabetic) and 268 data class "1" (diabetic).

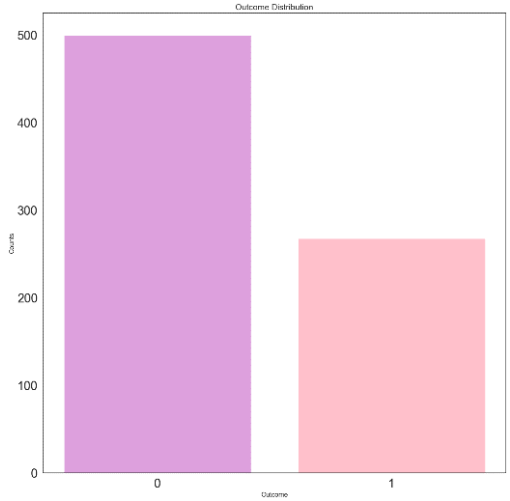


Fig. 4 Outcome Class Distribution

The observed class distribution exhibits an imbalance; class "0" has a larger number of data points, while class "1" has fewer data points. This imbalance is an important consideration in model training and evaluation processes [11]. Imbalanced class distribution suggests that the model may have a tendency to make biased predictions for the classes. To address this imbalance and achieve better results, preprocessing methods such as oversampling or undersampling have been employed. These preprocessing steps aim to stability the class distribution and enable the model to make more balanced and reliable predictions.

B. Construction of Models

ML models are commonly used in conjunction with various data preprocessing steps [12]. These preprocessing steps are necessary for accuracy of the model and ensure proper handling of the data. The data preprocessing process involves steps such as cleaning the dataset, scaling the data, handling missing values, and removing irrelevant features.

In this study, prediction was made using ML models such as SVM, KNN) LR, Extra Trees, AdaBoost and Gradient Boosting. These models are used to capture patterns in the data and classify new instances using different algorithms and learning approaches.

XGBoost Feature Selection using the features that can make the best prediction in the data set were selected.

The prediction process was performed using the Randomized Search method to optimize the model's predefined parameters and achieve the best performance [13]. Randomized Search is a parameter optimization method that aims to improve model performance by selecting random parameter combinations within a hyperparameter space.

In this study, the parameters specified in Table 2 were adjusted for each model using the Randomized Search method. This way, the prediction performance was optimized by obtaining the best parameters for the model. ML models are commonly used in conjunction with various data preprocessing steps [12]. These preprocessing steps are necessary to improve ensure proper handling of the data.

Randomized Search is a method that provides parameter optimization to enhance model performance [13]. In this work, the parameters listed in Table 2 were adjusted for each model using the Randomized Search method.

Cross-validation is a commonly used method to objectively evaluate the performance and measure models generalization ability [14]. For this study 10-fold cross-validation applied and data divided into equal parts; While some of it is for test data, and the other part is for train data [15]. This process creates ten different combinations where each part is used as both the train and test data.

The hybrid model is trained and tested for each combination, resulting in ten separate performance measurements

Table 2. Knowledge of the Best Performing Parameters for Each Algorithm

Algorithms	Parameters
KNN	n_neighbors = 3, 5, 11 weights = distance distance metric (p) = 1, 2, 3
SVM	fault tolerance = 0.1, 1 kernel = linear, rbf, sigmoid, poly gamma = scale
LR	penalty = l1, l2 penalty parameter(C) = 0.1, 10, 100 solver = liblinear, saga
ET	estimator = 100 criterion = entropy max_features = sqrt, log2
AB	Estimator = 100 learning_rate = 1.0, 10.0
GB	estimator = 50, 100, 200 learning_rate = 0.1, 0.5, 1.0 max_depth = 5, 7

The main purpose of using these models is to discover patterns and relationships in the dataset and perform classification. The prediction process was evaluated for each model using a specific performance metric.

The results demonstrate the performance of each model and the impact of specific parameter settings on prediction success. The parameter combinations obtained through the Randomized Search method were carefully selected to obtain the best prediction results.

III. Experimental Results And Discussion

The primary purpose of using these models is to explore patterns and relationships in the dataset and perform classification. The prediction process is evaluated for each model using a specific performance metric with the goal of achieving the best prediction results. Strategies such as K-fold cross-validation, data preprocessing steps feature selection and Randomized Search methods are employed to improve model performance and find the best parameter combinations. The parameter combinations obtained through the Randomized Search method are carefully selected to obtain the best prediction results.

interpreted to assess the performance of algorithms used in predicting diabetes and to determine the most effective algorithm.

A. *Evuluation*

In this section, the classification performances obtained from the conducted experiments were evaluated and compared with existing studies in the literature. The classification performances of the models were assessed using metrics which are sensitivity, precision, accuracy and f1-score. These metrics were calculated according to the formulas given in Figure 5.

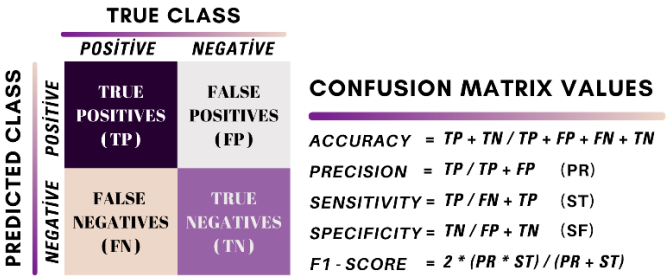


Fig. 5 Evaluation metrics

The classification performances obtained are presented in Table 3, 4 and 5. When classification is made without using any augmentation method, the highest result is obtained in the Extra Trees Algorithm. When classification is made using data preprocessing and feature selection, the highest result is obtained in the LR Algorithm. When classification is made using data Preprocessing, Feature Selection, Randomize search and Cross Validation, the highest result is obtained in the LR Algorithm. In Table 5 the highest classification performance with 77.06% accuracy was obtained by the LR algorithm with Hybrid Model. On the other hand, the KNN algorithm showed the lowest performance.

Table 3. Classic Classification Results

Performance Results (%)				
Models	**Accuracy (%)**	**Precision (%)**	**Sensitvy (%)**	**F-Score (%)**
KNN	68,83	65,69	65,87	65,78
SVM	73,59	71,13	67,75	68,61
LR	74,03	71,32	71,32	71,32
ET	76,19	73,87	72,09	72,76
AB	74,46	71,78	71,65	71,71
GB	74,46	71,93	72,53	72,19

Table 4. Feature Selection Classification Results

Performance Results (%)				
Models	**Accuracy (%)**	**Precision (%)**	**Sensitvy (%)**	**F-Score (%)**
KNN	71,86	68,92	67,83	68,24
SVM	74,03	72,14	68,07	69,00
LR	75,76	73,91	70,83	71,75
ET	74,03	71,59	69,35	70,07
AB	73,38	70,74	70,56	70,64
GB	73,16	70,46	68,83	69,40

Table 5. Hybrid Model Classification Results

Performance Results (%)				
Models	**Accuracy (%)**	**Precision (%)**	**Sensitvy (%)**	**F-Score (%)**
KNN	76,62	75,00	71,78	72,76
SVM	76,19	74,72	70,88	71,93
LR	77,06	75,64	72,11	73,16
ET	76,64	74,71	72,35	73,16
AB	76,16	74,13	72,01	72,76
GB	75,76	73,49	71,96	72,55

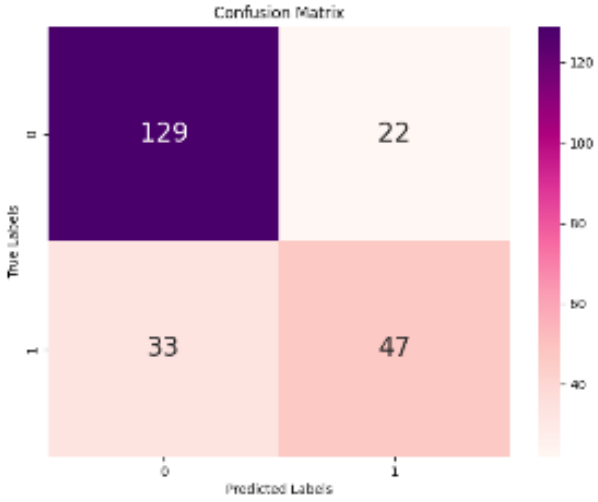
The performance results of the classification algorithms using all features are given in Table 3. According to these results, the ET algorithm obtained the best accuracy. In Table 4, normalization and standard scale were applied as data preprocessing. In addition, four features were used in classification by applying XGBoost feature selection to the data set. Accordingly, when the accuracy values obtained by the algorithms were examined, LR gave the best performance with 75.76%. In the hybrid model proposed in Table 5, data set normalization, standard scale, XGBoost Feature selection, Randomized Search and 10 KFold Cross Validation were used. As a result of these processes, the accuracy value of 77.06% was reached in the LR algorithm.

The issue of classifiers algorithms exhibiting imbalanced performance on test datasets is commonly known as the problem of overfitting. In the experiments of this study, the lower performance of the KNN classifier compared to other classifiers indicates that this algorithm does not exhibit a tendency for overfitting.

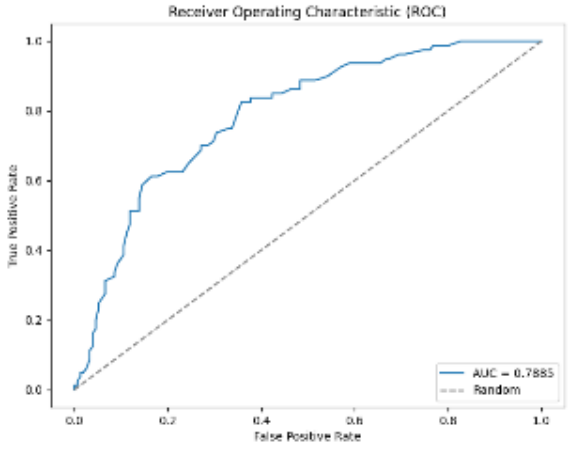
Based on the experimental results, the confusion matrices presented above allow us to evaluate the ability of each algorithm to correctly predict diabetic patients. The values in the confusion matrices demonstrate the agreement between the true class and the predicted class [17].

In conclusion, this evaluation demonstrates that the LR Classifier exhibits the highest success in correctly classifying non-diabetic individuals. Other algorithms show lower performance in accurately predicting diabetic patients. However, the performance of the algorithms can vary depending on the dataset and the hyperparameters used. Therefore, it is recommended to use different metrics and validation methods for a more comprehensive evaluation.

The performance evaluation of the classification models was conducted using the ROC curves presented in Figure 6,7,8. The Area Under the Curve (AUC) score of each model quantifies its discriminatory capability from area under the ROC curve [18].

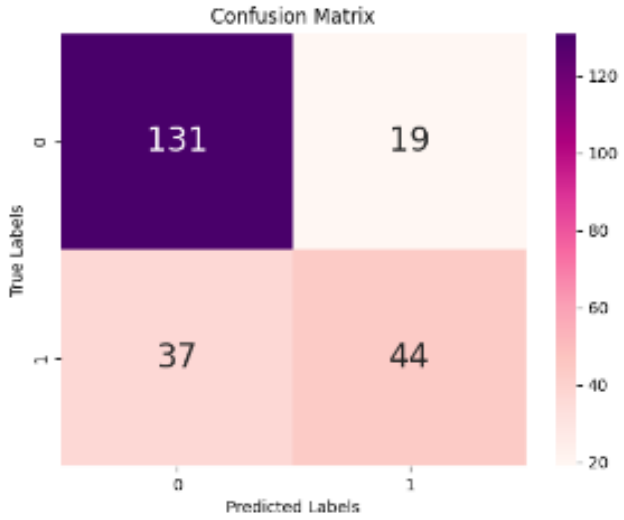


(a) confusion matrix

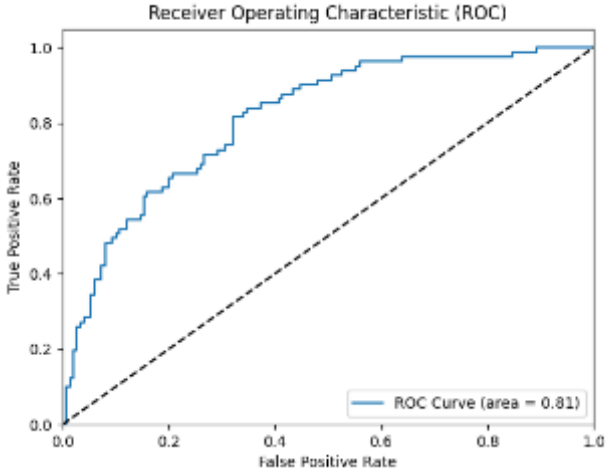


(b) ROC curve

Fig. 6 Evaluation metrics of algorithm Extra Trees

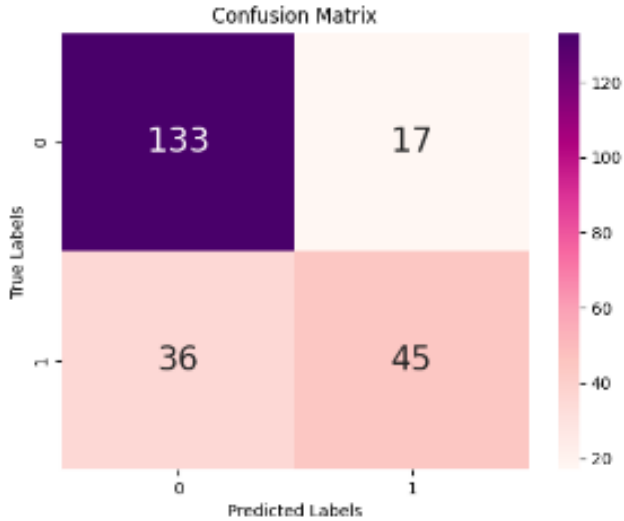


(a) confusion matrix

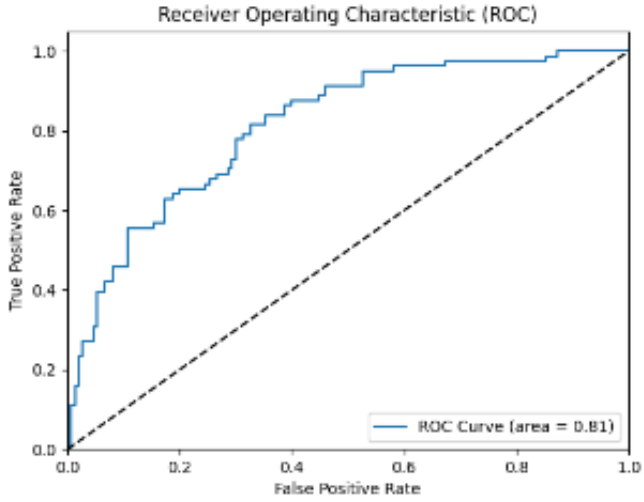


(b) ROC curve

Fig. 7 Evaluation metrics of Logistic Regression algorithm classified using feature selection



(a) confusion matrix



(b) ROC curve

Fig. 8 Evaluation metrics of Logistic Regression algorithm classified using feature selection, Randomized Search and Cross Validation

The AUC scores measure a model's ability to distinguish between classes, and higher scores represent better performance [18]. Upon examining the results, 'Logistic Regression' model has the highest 0.81 AUC score. The graphs of the results obtained in Figure 9 are given comparatively

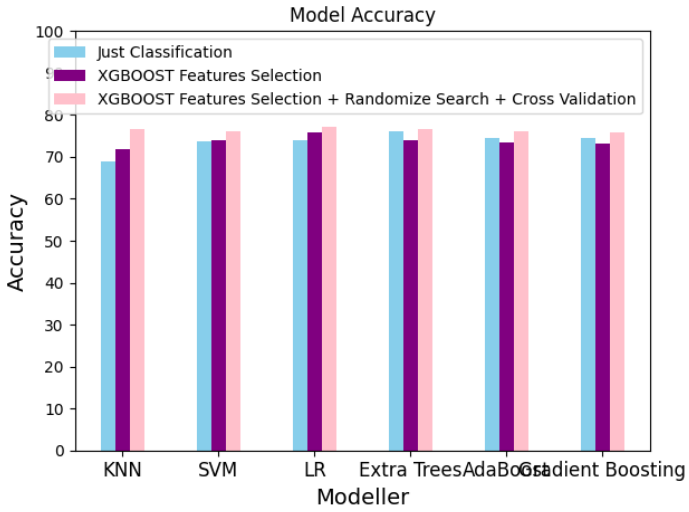


Fig. 9 Comparison Chart of Accuracy Values of Classification Results

Based on the provided accuracy scores, a performance comparison has been made among different classification models. When the models are examined, the LR model has the highest accuracy score of 77,06%. Possible reasons for this are the randomized search parameters used in the LR classification.

On the other hand, the 'KNN' model achieved a lower accuracy score of 71,86%. KNN performs classification based on a proximity measure between examples, but its performance can decrease as the complexity of the dataset increases [20].

Unlike the results obtained in some studies conducted with the Pima Indian diabetes dataset, in this study, both the feature selection and the models used directly increased the study performance results. Instead of studies that increase accuracy using this method, examples of different model applications made with this dataset. Evaluating the results obtained from these studies in comparison to previous research is important. Below, the methods, techniques used, and accuracy rates of some exemplary studies are summarized in Table 6.

Table 6. Dataset Models Literatur Review

Study	Method	Methodology	Accuracy
Lukmanta [21]	ML	DVM, fuzzy inference	%89,02
Vaishali [22]	ML	Genetic Algorithm Evolutionary Fuzzy Classifier	%83,04
Sehly [23]	ML	Feature Selection	%77,21
Bhalla [24]	ML	Cross Validaton	%72,9
This Study	ML	XGBoost Feature Selection, RandomizedSearchCV, KFold (10)	%77,06

In accordance with the findings of this study, our model demonstrates an accuracy rate of 77.06%. It is noteworthy that this accuracy rate appears notably lower when compared to the values reported in previous research endeavors conducted by Lukmanta [21] and Vaishali [22], with reported accuracies of 89,02% and 83,04%, respectively. The observed disparities in accuracy rates may be attributed to several potential factors that warrant consideration. Although the accuracy of our result was higher than classical classification models, it was lower than some literature studies. However, what should be taken into consideration here is the increase in accuracy achieved by the proposed hybrid model. In addition, various factors directly affect these accuracy values, including the size of the dataset, feature selection, training process of the model and evaluation measurements.

IV. CONCLUSION

Considering the model developed in the study, the performance of the algorithms used varies depending on factors such as data set properties and hyperparameters. XGBoost for feature selection is a critical consideration and it affects the accuracy of using model. The results obtained in the evaluation process have shown varying degrees of success in correctly classifying diabetic patients.

In this study, the performance of KNN, SVM, LR, Extra Trees, AdaBoost and Gradient Boosting based ensemble classifiers for early detection of this disease was investigated. The results showed that the Logistic Regression algorithm outperformed other algorithms in accurately identifying non-diabetic individuals.

These findings indicate that the overall performance of the algorithms can vary depending on the characteristics of the dataset. Therefore, it is recommended to use different metrics and validation methods for a more comprehensive evaluation. Additionally, it should be noted that hyperparameter tuning and model optimization are very important for model performance.

It is believed that the LR algorithm, which demonstrated the best performance, could support studies aiming to detect diabetes in the early stages through a real-time expert system.

V. ACKNOWLEDGEMENT

The authors thank The Data Catalog Platform [9] for the publicly available diabetes dataset.

REFERENCES

[1] P. Bala Manoj Kumar, R. Srinivasa Perumal, R. K. Nadesh, K. Arivuselvan, "Type 2: Diabetes mellitus prediction using Deep Neural Networks classifier," International Journal of Cognitive Computing in Engineering, vol. 1, pp. 55–61, 2020.

[2] R. D. Howsalya Devi, A. Bai, and N. Nagarajan, "A novel hybrid approach for diagnosing diabetes mellitus using farthest first and support vector machine algorithms," Obesity Medicine, vol. 17, p. 100152, 2020.

[3] D. Jashwanth Reddy, B. Mounika, S. Sindhu, T. Pranayteja Reddy, N. Sagar Reddy, G. Jyothsna Sri, et al., “Predictive machine learning model for early detection and analysis of diabetes,” Materials Today: Proceedings, 2020.

[4] 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.

[5] N. Masood, R. Ahmed, M. Tariq, Z. Ahmed, M. S. Masoud, I. Ali, R. Asghar, A. Andleeb, A. Hasan, “Silver nanoparticle impregnated chitosan-PEG hydrogel enhances wound healing in diabetes induced rabbits”, International Journal of Pharmaceutics, Vol 559, ISSN 0378-5173, 2019.

[6] H. Das, B. Naik, H. S. Behera, "Classification of Diabetes Mellitus Disease (DMD): A Data Mining (DM) Approach", "Progress in Computing, Analytics and Networking", 539-549, ISBN 978-981-10-7871-2, "Springer Singapore", 2018.

[7] F. D'Aiuto, N. Gkranias, D. Bhowruth, T. Khan, M. Orlandi, S. Masi, “Systemic effects of periodontitis treatment in patients with type 2 diabetes: a 12 month, single-centre, investigator-masked, randomised trial”, The Lancet Diabetes & Endocrinology, 954-965, Vol 6, ISSN 2213-8587, 2018.

[8] K. Avirneni, S. Prasad, “Predictive Modeling of Diabetes Risk Using Machine Learning Techniques” International Journal of Engineering and Technology (IJET), 9(1), 46-52, 2017.

[9] Data.world https://data.world/data-society/pima-indians-diabetes-database Irvine, CA: University of California, School of Information and Computer Science, data of access 20 June 2023.

[10] Gurney K., An introduction to neural networks, Taylor & Francis e-Library, UCL Press Limited, London, 2004.

[11] H. Zhang, Y. Zhao, “A Comparative Study of Naive Bayes Classifier and Decision Tree Algorithm”, Proceedings of the 2004 International Conference on Machine Learning and Cybernetics, 2, 1030–1034, 2004.

[12] R. Khan, N. Wajid, N. Ullah, M. A. Khan, “Using Machine Learning to Predict and Improve Gestational Diabetes Diagnosis”, International Journal of Advanced Computer Science and Applications, 9(10), 14-20, 2018.

[13] Asif, Md Asfi-Ar-Raihan, et al. "Performance Evaluation and Comparative Analysis of Different Machine Learning Algorithms in Predicting Cardiovascular Disease." Engineering Letters 29.2, 2021.

[14] Peco Chacon, Ana Maria, and Fausto Pedro García Márquez. "Support Vector Machine and K-fold Cross-validation to Detect False Alarms in Wind Turbines." Sustainability: Cases and Studies in Using Operations Research and Management Science Methods. Cham: Springer International Publishing, 2023. 81-97.

[15] Nguyen, May Huu, and Hai-Bang Ly. "Development of machine learning methods to predict the compressive strength of fiber-reinforced self-compacting concrete and sensitivity analysis." Construction and Building Materials 367 (2023): 130339.

[16] Terasawa, Teruhiko, et al. "Systematic review: computed tomography and ultrasonography to detect acute appendicitis in adults and adolescents." Annals of internal medicine 141.7, 537-546, 2004.

[17] D. M. Powers, “Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation”, Journal of machine learning technologies, 2(1), 37-63, 2011.

[18] Ahmad, N. Ghulab, et al., "Mixed machine learning approach for efficient prediction of human heart disease by identifying the numerical and categorical features", Applied Sciences, 12.15,7449, 2022.

[19] H. Saeed, M. Ahmed, "Diabetes type 2 classification using machine learning algorithms with up-sampling technique", Journal of Electrical Systems and Information Technology, 10.1, 1-10, 2023.

[20] Sowah, A. Robert, et al., "Design and development of diabetes management system using machine learning.", International journal of telemedicine and applications, 2020.

[21] R. B. Lukmanto, A. Nugroho, H. Akbar, “Early detection of diabetes mellitus using feature selection and fuzzy support vector machine”, Procedia Computer Science, 157, 46-54, 2019.

[22] R. Vaishali, R. Sasikala, S. Ramasubbareddy, S. Remya, S. Nalluri, “Genetic algorithm-based feature selection and MOE Fuzzy classification algorithm on Pima Indians Diabetes dataset”, International Conference on Computing Networking and Informatics (ICCNI), Lagos, Nijerya, 29-31 Ekim 2017.

[23] R. Sehly and M. Mezher, "Comparative Analysis of Classification Models for Pima Dataset," 2020 International Conference on Computing and Information Technology (ICCIT-1441), Tabuk, Saudi Arabia, 2020, pp. 1-5, doi: 10.1109/ICCIT-144147971.2020.9213821.

[24] Bhalla, Rajni & Bagga, Amandeep. (2019). RB-bayes algorithm for the prediction of diabetic in "PIMA Indian dataset". International Journal of Electrical and Computer Engineering (IJECE). 9. 10.11591/ijece.v9i6.pp4866-4872.