Diabetes Prediction Using Machine Learning with Feature Engineering and Hyperparameter Tuning

Hakim El Massari[1*], Noreddine Gherabi[2], Fatima Qanouni[3], Sajida Mhammedi[4]
LAMAI Laboratory, Faculty of Sciences and Techniques, Cadi Ayyad University, Marrakech, Morocco[1]
Lasti Laboratory, National School of Applied Sciences, Sultan Moulay Slimane University, Khouribga, Morocco[1, 2, 3, 4]
Higher School of Technology of El Kelâa des Sraghna, Cadi Ayyad University, El Kelâa des Sraghna, Morocco[1]

***Abstract*—Diabetes, a chronic illness, has seen an increase in prevalence over the years, posing several health challenges. This study aims to predict diabetes onset using the Pima Indians Diabetes dataset. We implemented several machine learning algorithms, namely Random Forest, Gradient Boosting, XGBoost, LightGBM, and CatBoost. To enhance model performance, we applied a variety of feature engineering techniques, including SelectKBest, Recursive Feature Elimination (RFE), Recursive Feature Elimination with Cross-Validation (RFECV), Forward Feature Selection, and Backward Feature Elimination. RFECV proved to be the most effective method, leading to the selection of the best feature set. In addition, hyperparameter tuning techniques are used to determine the optimal parameters for the models created. Upon training these models with the optimized parameters, XGBoost outperformed the others with an accuracy of 94%, while Random Forest and CatBoost both achieved 92.5%. These results highlight XGBoost's superior predictive power and the significance of thorough feature engineering and model tuning in diabetes prediction.**

***Keywords*—Machine learning; feature engineering; hyperparameter tuning; diabetes prediction; healthcare**

I. Introduction

The World Health Organization considers diabetes one of the world's leading causes of death. Diabetes mellitus is a metabolic disorder of the endocrine system in which the blood glucose levels remain high for longer than necessary, causing hyperglycemia. The majority of diabetes cases are type 2 diabetes. The symptoms of diabetes include frequent urination, excessive thirst, extreme fatigue, etc.

The proportion of individuals with diabetes has been increasing more rapidly than can be accounted for by a rapidly increasing population. The World Health Organization has predicted that diabetes will be the leading cause of disease burden in the world by 2030. Such reporting is important but still almost undoubtedly an underestimate of the total impact of diabetes since there are very many children in whom the diagnosis is not made, in whom there may be a very early onset of complications, and whose death is not reported as 'diabetic'.

Youth-onset type 2 diabetes will also provide a considerable burden to some populations, especially indigenous peoples. Many obese individuals already have the insulin resistance that promises eventual diabetes. A third of the American adult population is thought to have the insulin resistance syndrome. Individuals of Asian and African origin, as well as indigenous peoples, have an increased risk of diabetic complications, at least in part independent of the greater weight for height. Since even impaired fasting glucose has been reported to be associated with an increased independent risk of cardiovascular disease, such reports demonstrate the threat and the value of strategies to prevent or delay the onset of metabolic syndrome. With outcome metrics such as the development of retinopathy and cardiovascular events, a diagnosis may only come in time to prevent a diagnosis of type 2 diabetes

It is associated that diabetes is very long and gradually exerts its unwanted effects on all the body parts. It remains in the individual's body for a long time and then develops heart disease, chronic meningitis, hypertension, blindness, stroke, erectile dysfunction, nerve damage (neuropathy), among other things. The population growth, relocating from rural to urban areas, interacting with bestial food habits, lack of exercise, stress, and lifestyle changes in people of all ages also contribute to the development of diabetes.

Diabetes is now recognized as a global health problem. It creates a huge impact on people and countries around the world. The importance of diabetes lies in the fact that it increases a person's likelihood of having a stroke by 1.5 times. It is predicted that if the rising incidence of diabetes is not reversed, the overall death rate from diabetes and heart disease will also rise.

Machine learning (ML) in healthcare is used to diagnose diseases, create personalized treatment plans, and predict hospital readmissions [1], [2], [3]. It can also detect which patients are at high risk of developing diabetes, long before it occurs. There are also many other related problems in the medical field such as disease diagnosis, hospital readmission, personalized treatment, and patient hope. However, the main goal of this study is to establish how basic, everyday habits affect the early detection of diabetes [4].

At the moment, prediabetes and diabetes are diagnosed through massive blood tests (glucose, insulin, and so on) that only patients with symptoms undergo. The main advantages of predicting diabetes using machine learning are: once the algorithm is implemented, everybody can use it and the test can be done whenever one wants; it is cheap; it allows everyone to know if they are at risk of developing diabetes months/years in advance, and take action; it saves the time and resources of doctors and hospitals to spend on the real ill patients.

*Corresponding Author.

Machine learning (ML) in healthcare is used to diagnose diseases, create personalized treatment plans, and predict hospital readmissions. It can also detect which patients are at high risk of developing diabetes, long before it occurs. There are also many other related problems in the medical field such as disease diagnosis, hospital readmission, personalized treatment, and patient hope. However, the main goal of this study is to establish how basic, everyday habits affect the early detection of diabetes.

The remainder of this study is organized as follows: Section II covers previous research and related studies. Section III details the methodology, including the various algorithms employed, steps taken to prepare the dataset, techniques for creating and refining features, and the process of optimizing the parameters. Section IV presents the results and a thorough discussion of the findings. Finally, Section VI provides a conclusion summarizing the key insights and implications of this work.

II. Related Work

This section provides a detailed overview of related work in the field of diabetes prediction using machine learning techniques in particular.

A recent study [5] proposed an ensemble-based approach for predicting diabetes using the Pima Indian Diabetes Dataset. It evaluated LightGBM, XGBoost, AdaBoost, and Random Forest, finding that LightGBM alone achieved an accuracy of 94% and a ROC AUC of 95%. By introducing a Soft Voting classifier, the combined model's accuracy increased to 95% with a ROC AUC of 96%, demonstrating the potential of ensemble methods to improve prediction reliability.

TABLE I. Related Work Comparison

Reference	ML algorithms	Highest Accuracy
[3]	Logistic Regression, Decision Tree, Random Forest, k-Nearest Neighbors, Naive Bayes, Support Vector Machine, Gradient Boosting, and Neural Network	78,57%
[5]	LightGBM, XGBoost, AdaBoost and Random Forest	93%
[6]	decision tree (DT), logistic regression (LR), support vector machine (SVM), gradient boost (GB), extreme gradient boost (XGBoost), random forest (RF), and ensemble technique (ET)	93,27%
[7]	Ensemble learning, XGBoost, CatBoost, LightGBM, AdaBoost, gradient boost	92,85%
[8]	Random forest classifier (RF), logistic regression (LR), decision tree classifier (DT), support vector machine (SVM), Bayesian Classifier (BC) or Naive Bayes Classifier (NB), Bagging Classifier (BG), Stacking Classifier (ST), Moderated Ada-Boost(AB) Classifier, K Neighbors Classifier (KN) and Artificial Neural Network (ANN)	90,95%
[9]	ET, RF, SGB, AB	93.63%
[10]	decision tree, SVM, Random Forest, Logistic Regression, KNN, and various ensemble techniques.	81%
Our study	Random Forest, Gradient Boosting, XGBoost, LightGBM, and CatBoost.	**94%**

Numerous studies have explored the use of machine learning algorithms for predicting diabetes. The study in [6] employed a range of classification algorithms, including Logistic Regression, Decision Trees, and Support Vector Machines, to predict diabetes onset using the Pima Indians Diabetes dataset. Their research highlighted the effectiveness of Support Vector Machines in achieving high accuracy.

Further investigation [7] focused on the application of ensemble methods for diabetes prediction. They compared the performance of Bagging, Boosting, and Stacking techniques, demonstrating that ensemble methods generally outperformed single classifiers. Specifically, their results indicated that Boosting algorithms, particularly XGBoost, provided superior predictive performance.

Feature selection and engineering play a critical role in improving model accuracy. The study [8] implemented Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA) to enhance their machine learning models. They concluded that RFE, in combination with Gradient Boosting Machines, yielded the best results, emphasizing the importance of selecting relevant features.

The use of deep learning approaches has also been investigated [11], [12], [13], [14]. In study [15] authors proposed a deep neural network model for diabetes prediction, achieving remarkable accuracy. Their work demonstrated that deep learning models could capture complex patterns in the data, albeit at the cost of increased computational resources and the need for larger datasets.

Additionally, researches [9], [16] introduced the concept of hybrid models that combine multiple machine learning techniques to improve prediction accuracy. They developed a hybrid model integrating Random Forest and Neural Networks, which surpassed the performance of traditional models.

Recent advancements in explainable AI have also been applied to diabetes prediction. For instance, [10], [17] utilized SHapley Additive exPlanations (SHAP) to interpret the predictions of their machine learning models. This approach provided insights into the importance of different features, enhancing the transparency and trustworthiness of the predictive models.

In earlier studies, different feature selection and classification have been proposed to optimize the classifier model for 12 different classifiers over Pima Indians Diabetes Database from the UCI machine learning website [18], [19], [20]. The work was done on Pima Indians Diabetes Database in order to predict diabetes using different Data Mining algorithms. The main feature selection methods are Correlation, Wrappers, and Principal Components. Wrapper was the most successful feature selection method in obtaining a small data subset optimizing the classifier. Besides the feature

selection, a metaheuristic algorithm was used to optimize the classifier to fit the classifier model by using a small training data subset. Random Forest (RF) with 28 input features produced high accuracy (98.08%), sensitivity (94.6), and specificity (99.3). The study in [21], [22], dataset with six input attributes was tested using six different classifiers. Decision Trees (DT) and J48 classifiers gave the best results on both 10 × 10 cross-validation (CV) or independent test sets, resulting in 78.33% accuracy, 77.38% and 88.33% accuracy, 87.84%, respectively. These works have partially been compared with the work.

Research indicates that type 2 diabetes [23], [24], [25] is treatable and preventable through making lifestyle changes such as weight loss, improved diet, and increasing physical activity. Regular monitoring, at-home blood glucose testing, and A1C levels are pivotal for identifying high risk for diabetes and type 2 diabetes early on. However, individuals are experiencing continuous increases in weight and obesity because of the rising trends of high-calorie diets and sedentary lifestyles.

This trend could reach a breaking point for the healthcare system if our understanding of the factors leading to type 2 diabetes risk remains incomplete. Machine learning techniques, such as classification, regression, clustering, anomaly detection, and pattern recognition, are developed to make accurate predictions from data. Specifically, the current health status of pre-diabetic patients is used to assess potential for diabetes diagnoses.

By understanding how certain factors of lifestyle change can influence diabetes diagnoses, pre-diabetics may be able to take steps to avoid the implications of diabetes. With the implementation of machine learning algorithms and healthcare data, healthcare providers would benefit from a tool capable of early diagnosing patients who possess the highest risk of developing type 2 diabetes and cardiovascular diseases while developing personalized and cost-effective intervention strategies for pre-diabetics [26], [27], [28].

Overall, the related work in this field underscores the continuous evolution of machine learning techniques for diabetes prediction. The integration of advanced feature engineering, ensemble methods, and deep learning has significantly improved predictive accuracy, paving the way for more effective and reliable diabetes prediction models.

III. Methods and Evaluation

In this study, the methodology used is divided into several key components: data collection and preprocessing, data analysis techniques, feature engineering, hyperparameter tuning and performance evaluation metrics. Each component is discussed in detail to provide a comprehensive understanding of the research process. Fig. 1 depicts the overall process workflow for this experimental study.

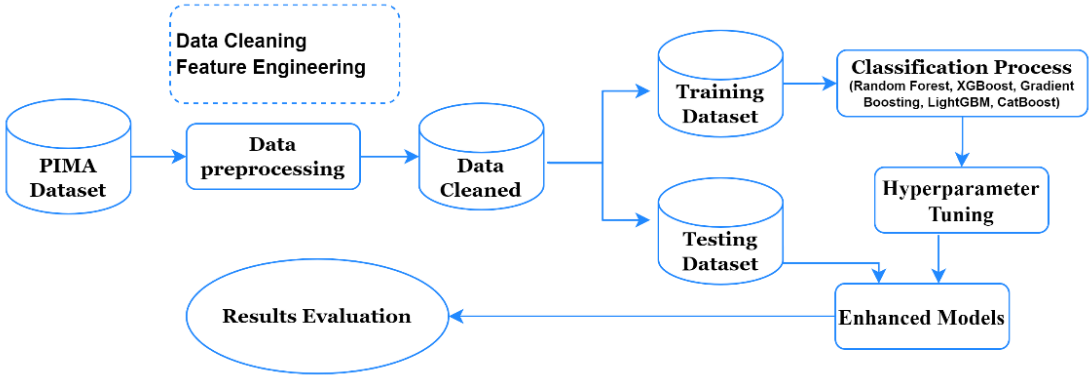

Fig. 1. Experimental workflow.

A. Data Collection and Preprocessing

To implement the predictive models in this study, an open-access diabetes dataset was utilized. This dataset, obtained from Kaggle, includes various medical predictor variables and a target variable. It comprises records of 768 patients. All patients are females of Pima Indian heritage, aged at least 21 years. Among them, 34.9% have diabetes, while 65.1% do not, as depicted in Fig. 2. Detailed attribute information is provided in Table II.

Data preprocessing is essential for all machine learning (ML) applications because the effectiveness of an ML algorithm depends significantly on how well the dataset is prepared and structured. This step ensures the data is tailored to meet the specific needs of the chosen algorithm. For the diabetes dataset, we employed several preprocessing techniques during this initial phase:

TABLE II. Dataset Feature's Information

Attribute	Description
1- Pregnancies	Count of pregnancies
2- Glucose	Plasma glucose levels
3- BloodPressure	Diastolic blood pressure (mm Hg)
4- SkinThickness	Triceps skin fold thickness (mm)
5- Insulin	2-Hour serum insulin (mu U/ml)
6- BMI	Body mass index
7- DiabetesPedigreeFunction	Diabetes pedigree function
8- Age	Age (years)
9- Outcome	1 = diabetic, 0 = non diabetic

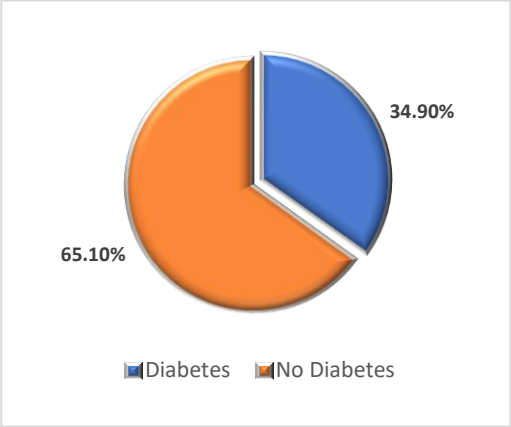

Fig. 2. Distribution of diabetes.

- Data Cleaning: We removed missing or null values, cleaned up noisy data, and detected and eliminated outliers.
- Outlier Handling: To enhance the robustness of our model, we used the "Replace with Thresholds IQR" method. This technique involves substituting extreme values with thresholds derived from the Interquartile Range (IQR), which helps create a more resilient and reliable model.
- Scaling and Normalization: We scaled and normalized the data to ensure that no single feature disproportionately influences the model due to differing scales.
- Handling Imbalanced Data: To prevent bias and ensure the model is not unduly influenced by the prevalence of a particular class, we applied the Synthetic Minority Oversampling Technique (SMOTE) as described in Table III. This technique generates a balanced dataset by synthesizing instances of the minority class, thereby improving the predictive accuracy for that class.

Throughout the study, we utilized various Python libraries, including NumPy, Pandas, Seaborn, Matplotlib, and Scikit-learn, for both exploratory data analysis (EDA) and data visualization. These tools helped us thoroughly analyze and prepare the data, setting a solid foundation for building effective predictive models.

TABLE III. RESULT OF THE SMOTE TECHNIQUE

	Diabetes	
	Yes	No
Before SMOTE	268	500
After SMOTE	500	500

B. Machine Learning Algorithms

Among the existing algorithms of machine learning [29], [30], we used in this study Random Forest, Gradient Boosting, XGBoost, LightGBM, and CatBoost. These algorithms were selected for their robustness and versatility in handling various types of data and predictive modeling tasks. Random Forest is known for its simplicity and effectiveness in reducing overfitting through ensemble learning. Gradient Boosting improves predictive accuracy by iteratively minimizing errors. XGBoost, an optimized version of Gradient Boosting, enhances performance and computational efficiency. LightGBM, designed for speed and scalability, handles large datasets and high-dimensional data efficiently. CatBoost is particularly effective with categorical data and requires minimal preprocessing. By leveraging the strengths of these algorithms, we aimed to achieve a comprehensive analysis and robust predictive performance for our study.

Random Forest is an ensemble learning algorithm that constructs multiple decision trees during training and outputs the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. It is robust to overfitting due to the averaging of multiple trees, handles large datasets effectively, and can manage missing data and maintain accuracy for large portions of the data.

Gradient Boosting is an iterative algorithm that builds a model in a stage-wise fashion from weak learners, typically decision trees. Each new model attempts to correct the errors of the previous one by minimizing a loss function. This approach results in high predictive accuracy, making it suitable for various machine learning tasks, although it can be computationally intensive and sensitive to overfitting without proper regularization.

XGBoost (Extreme Gradient Boosting) is an optimized version of Gradient Boosting designed for performance and speed. It incorporates advanced features like regularization to prevent overfitting, parallel processing, and efficient handling of missing data. XGBoost is known for its scalability and effectiveness in both regression and classification problems, making it a popular choice in competitive machine learning.

LightGBM (Light Gradient Boosting Machine) is a gradient boosting framework that uses tree-based learning algorithms. It is designed for efficiency and scalability, making it well-suited for large datasets with high-dimensional features. LightGBM achieves faster training speed and higher efficiency by using a histogram-based approach and leaf-wise tree growth, which leads to better accuracy.

CatBoost (Categorical Boosting) is a gradient boosting algorithm specifically designed to handle categorical data with minimal preprocessing. It automatically deals with categorical features and reduces overfitting through techniques like ordered boosting and efficient oblivious tree structures. CatBoost is robust, accurate, and user-friendly, making it ideal for applications where categorical data is prevalent.

C. Feature Engineering

Feature engineering plays a role, in the realm of machine learning. It involves the creation, adjustment and selection of features from data to boost the performance of predictive models. This process encompasses methods like normalization encoding variables and crafting features based on domain expertise. Skillful feature engineering can notably improve the precision and effectiveness of machine learning models by equipping them with valuable input data. It often necessitates testing and a profound comprehension of both the dataset and

the core issue to identify features that aptly capture the patterns and connections, for accurate forecasts.

In this study various techniques are used to enhance machine learning models such as SelectKBest, Recursive Feature Elimination (RFE), Recursive feature elimination with cross-validation (RFECV), Forward Feature Selection, Backward Feature Elimination. Fig. 3-7 represents the results of features used in this study.

SelectKBest is a feature selection method that selects the top k features from a dataset based on a scoring function. We used the chi2 function with SelectKBest which is based on the chi-squared statistical test, which measures the independence of each feature with respect to the target variable. Higher chi-squared values indicate a stronger relationship between the feature and the target. By using SelectKBest with chi2, we reduced the dimensionality of the dataset by keeping only the most relevant features, potentially improving the performance of the machine learning models employed. By using SelectKBest and chi2 function we find that all features give over 91% accuracy, and the highest score of 92.2% goes to Random Forest Classifier.

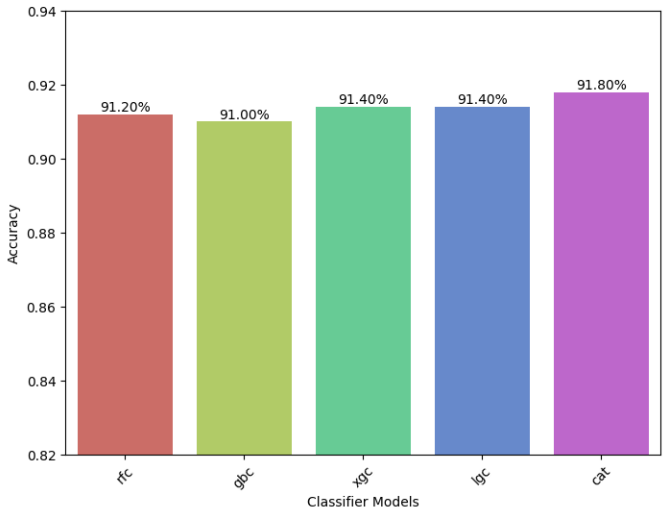

Fig. 3. Accuracy result using SelectKBest.

Recursive Feature Elimination (RFE) is a feature selection technique in machine learning that iteratively removes the least important features from the dataset. Starting with all features, RFE fits a model and evaluates the importance of each feature. The least important feature is then removed, and the model is re-fit on the remaining features. This process continues until the desired number of features is reached. By systematically eliminating features, RFE helps in identifying the most relevant subset, improving model performance and reducing overfitting by eliminating noise and irrelevant data.

Recursive Feature Elimination with Cross-Validation (RFECV) is an enhanced feature selection technique that combines Recursive Feature Elimination (RFE) with cross-validation to select the optimal number of features. RFECV iteratively removes the least important features while simultaneously evaluating model performance using cross-validation at each iteration. This approach ensures that the feature selection process is guided by model accuracy, helping to identify the subset of features that yields the best predictive performance. By incorporating cross-validation, RFECV provides a more robust and reliable method for feature selection, reducing the risk of overfitting and improving the generalizability of the model.

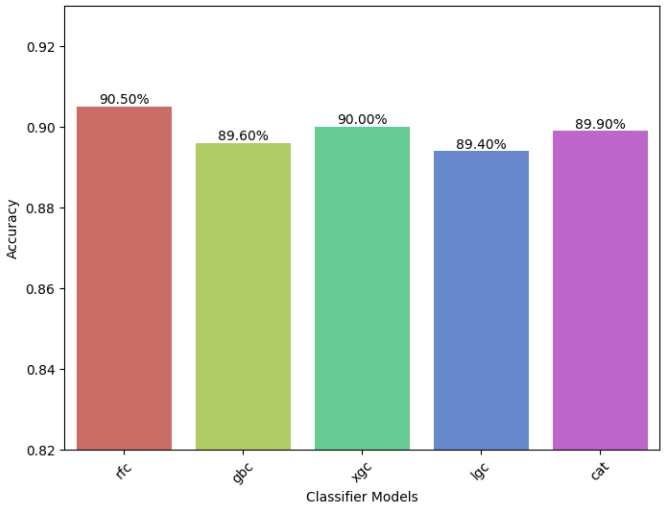

Fig. 4. Accuracy result using RFE.

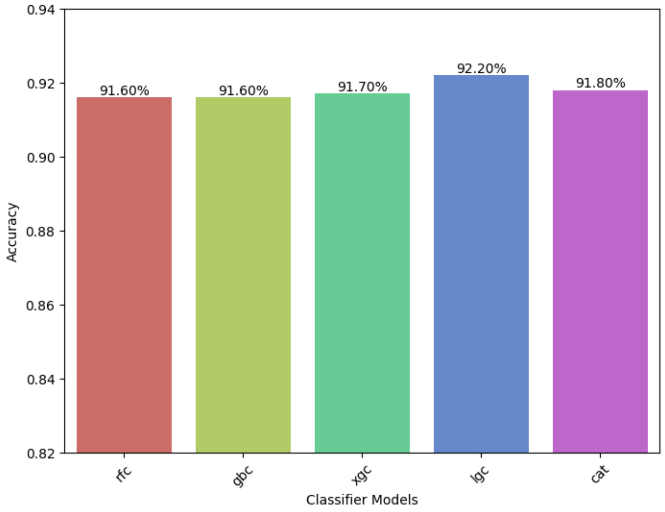

Fig. 5. Accuracy result using RFECV.

Forward Feature Selection is a feature selection technique in machine learning that starts with an empty model and iteratively adds the most significant features. At each step, the method evaluates all candidate features and adds the one that improves the model performance the most, based on a predefined criterion like accuracy or F1 score. This process continues until adding more features no longer significantly improves the model or a specified number of features is reached. Forward Feature Selection is effective for identifying a small, relevant subset of features, enhancing model interpretability and performance by including only the most impactful variables.

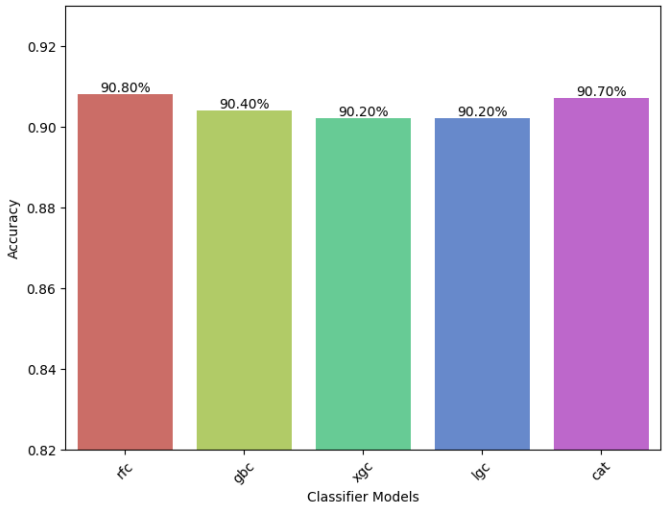

Fig. 6. Accuracy result using FORWARD.

Backward Feature Elimination is a feature selection technique in machine learning that starts with all available features and iteratively removes the least significant ones. At each step, the model is trained, and the importance of each feature is evaluated based on a predefined criterion such as p-values or model performance metrics. The least important feature is then removed, and the process is repeated until a specified number of features remains or further removal would degrade model performance. This method helps in simplifying the model by eliminating redundant or irrelevant features, improving interpretability and potentially enhancing predictive accuracy by reducing overfitting.

By comparing the results obtained from the five feature selection techniques used in this study, as shown in the Fig. 8, we conclude that the RFECV (Recursive Feature Elimination with Cross-Validation) feature selection method provides the best results. The top three algorithms identified are Random Forest, CatBoost, and XGBoost. For the remainder of this study, we will rely on these three algorithms.

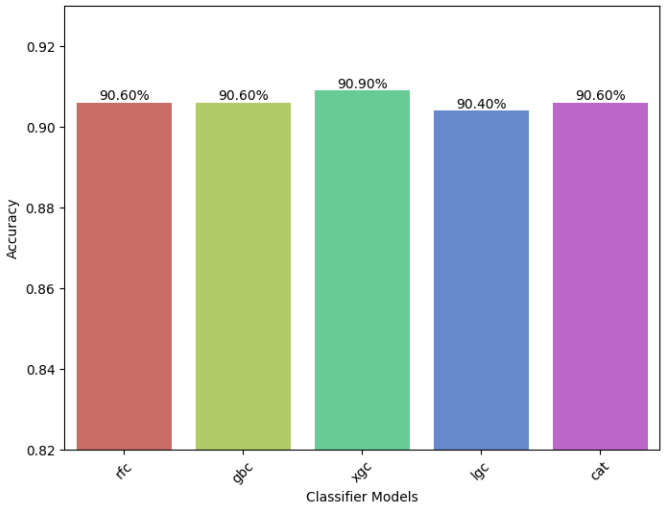

Fig. 7. Accuracy result using BACKWARD.

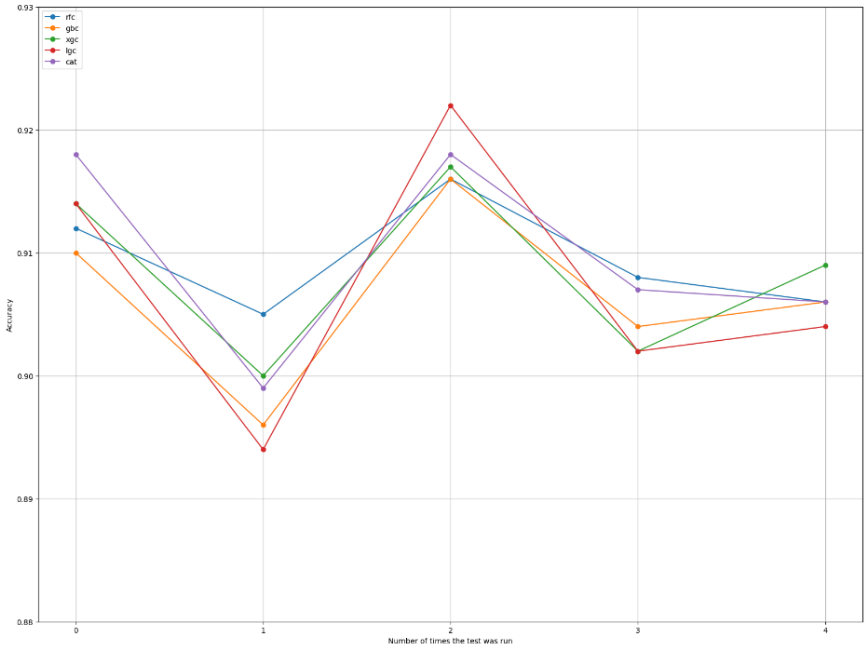

Fig. 8. Classifiers performance results.

D. Hyperparameter Tuning

The optimization of the hyperparameters of a machine learning model for optimizing the performance of a model is called hyperparameter tuning. Hyperparameters are different from model parameters because model parameters are learned during training while hyperparameters need to be set prior to the training process and influence different attributes of the learning process (learning rate, regularization strength, number

of layers in a neural network). Hyperparameter tuning is the process of finding the best selection of hyperparameters for a model which is typically done by systematically searching the space of hyperparameter values in a methodical manner. Some common methods of hyperparameter tuning for machine learning are grid search, random search, and Bayesian optimization. Model performance can be greatly improved by properly tuning the hyperparameters through achieving the ideal settings which enable the learning process to efficiently learn patterns in the data and avoid overfitting.

In this study, we utilize the GridSearchCV technique to determine the optimal parameters for our three models: Random Forest, CatBoost, and XGBoost. This method ensures that our models are fine-tuned for maximum accuracy and robustness. The results obtained from this parameter optimization process are presented in the accompanying Table IV, highlighting the best parameter settings for each model and their corresponding performance metrics.

TABLE IV. BEST PARAMETER SETTINGS FOR EACH MODEL

	Parameter
Random Forest	{bootstrap= False, ccp_alpha= 0, criterion= 'gini', max_depth= None, max_features= 'sqrt', n_estimators= 100, n_jobs= -1, verbose=0, random_state= 42}
XGBoost	{gamma= 0.1, learning_rate= 0.1, max_depth= 7, n_estimators= 200, reg_alpha= 0, reg_lambda= 0.001, verbose=0}
CatBoost	{bootstrap_type= 'Bernoulli', depth= 8, grow_policy= 'SymmetricTree', iterations= 200, l2_leaf_reg= 1, learning_rate= 0.1, verbose=0}

E. Performance Evaluation Metrics

Performance evaluation metrics are crucial tools in machine learning for assessing the effectiveness of predictive models. These metrics provide quantitative measures to evaluate how well a model performs on a given task. Common metrics include accuracy, precision, recall, F1-score, etc.

$$\text{ACC} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (1)$$

$$\text{PREC} = \frac{\text{TP}}{\text{TP+FP}} \quad (2)$$

$$REC = \frac{TP}{TP+FN} \quad (3)$$

$$\text{F-Measure} = 2 * \frac{PREC*REC}{PREC+REC} \quad (4)$$

IV. RESULTS AND DISCUSSION

The Pima Indians Diabetes dataset served as the foundation for this study, aiming to predict the onset of diabetes. This dataset includes several medical predictor variables and one target variable, which indicates whether or not the patient has diabetes. To enhance the predictive power of our machine learning models, we implemented a variety of feature engineering techniques.

We tested five machine learning algorithms: Random Forest, Gradient Boosting, XGBoost, LightGBM, and CatBoost. These algorithms were chosen for their robust performance in classification tasks. To optimize the input features for these models, we employed several feature engineering techniques, namely SelectKBest, Recursive Feature Elimination (RFE), Recursive Feature Elimination with Cross-Validation (RFECV), Forward Feature Selection, and Backward Feature Elimination.

Among these techniques, RFECV provided the most significant improvement in terms of accuracy and other performance metrics. RFECV methodically eliminates less important features while incorporating cross-validation to prevent overfitting. This approach identified the optimal set of features, which were then used to train our models.

Focusing on the top three algorithms identified by RFECV—Random Forest, CatBoost, and XGBoost—we used GridSearchCV to determine the optimal parameters for each model. This method involves an exhaustive search over specified parameter values to find the best combination that maximizes model performance.

Once the models were trained with these optimized parameters, we compared their performance Fig. 9 and Fig. 10. XGBoost emerged as the top-performing model, achieving an impressive accuracy score of 94%. In comparison, both Random Forest and CatBoost achieved accuracy scores of 92.5%. The superior performance of XGBoost can be attributed to its advanced tree boosting techniques and regularization methods, which help in managing data complexity and avoiding overfitting.

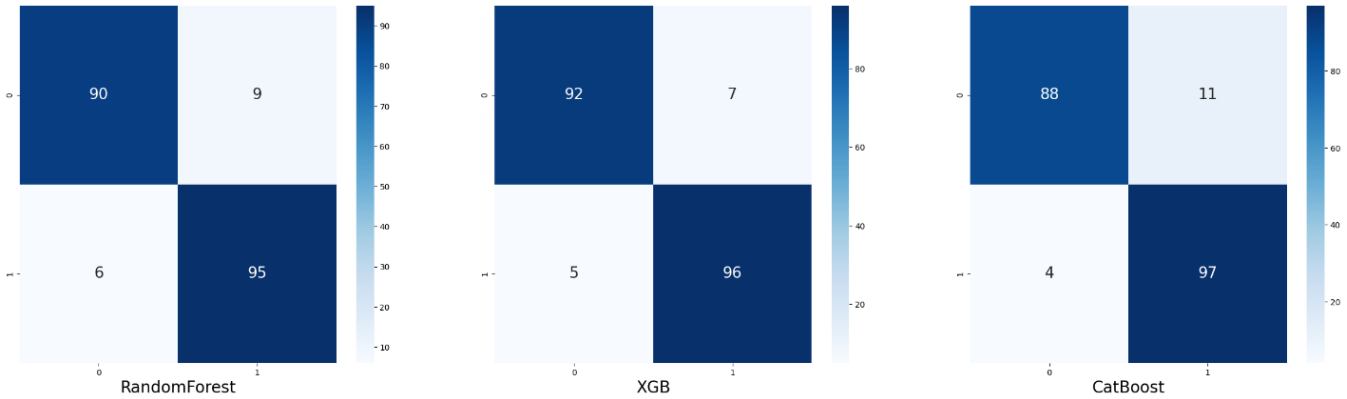

Fig. 9. Confusion matrix results.

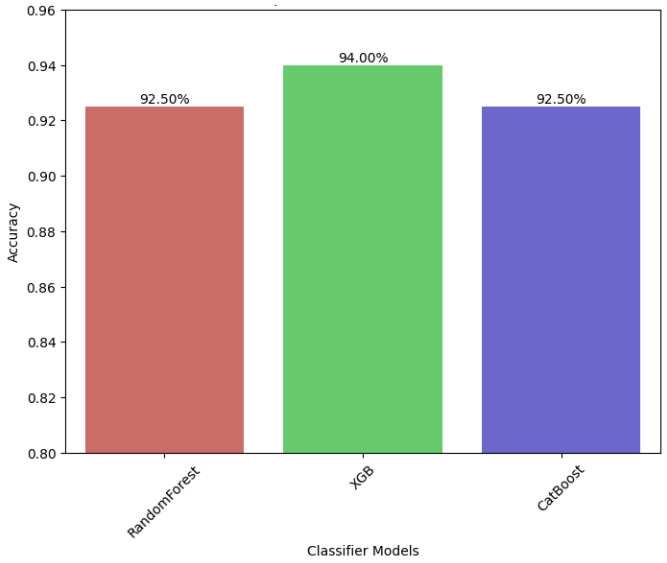

Fig. 10. Accuracy comparison of the models.

Comparing our findings with the most recent researches in the same field and used same the dataset, the feature engineering and hyperparameter tuning applied on our selective models, got the highest accuracy rate, Table I represent the comparison of different studies with ours.

In summary, the application of RFECV for feature selection and GridSearchCV for parameter optimization significantly enhanced the performance of our models. XGBoost, with its sophisticated boosting algorithms, proved to be the most effective model for predicting the onset of diabetes using the Pima Indians Diabetes dataset. Future work could explore additional data preprocessing steps and the inclusion of more complex models to further improve predictive accuracy.

V. Comparison with Other Studies

When examining the latest studies utilizing the Pima Indians Diabetes dataset, it's evident that our approach to feature engineering and model optimization has set a new benchmark in predictive accuracy.

Therefore, by leveraging RFECV for feature selection, we were able to identify the most relevant features and reduce noise in the dataset, leading to significant improvements in model performance. The subsequent application of GridSearchCV for hyperparameter tuning further optimized our models, ensuring that we achieved the best possible configuration for predictive accuracy. The integration of these techniques enabled our top-performing model, XGBoost, to reach an accuracy of 94%, surpassing the results of the aforementioned studies.

Our study not only highlights the importance of rigorous feature engineering and parameter optimization but also demonstrates the potential of advanced ensemble methods in predictive analytics. The substantial gains in accuracy underline the effectiveness of our approach compared to other contemporary methodologies. Table I provides a detailed comparison, showcasing the advancements our study brings to the field.

VI. Conclusion

This research focused on predicting the onset of diabetes using the Pima Indians Diabetes dataset. By applying various machine learning algorithms and feature engineering techniques, we aimed to identify the most effective model for this task. Among the algorithms tested — Random Forest, Gradient Boosting, XGBoost, LightGBM, and CatBoost — XGBoost demonstrated superior performance.

We employed several features engineering methods, including SelectKBest, Recursive Feature Elimination (RFE), Recursive Feature Elimination with Cross-Validation (RFECV), Forward Feature Selection, and Backward Feature Elimination, to refine our models. RFECV stood out as the most successful approach, yielding the best results in terms of accuracy and other metrics. Consequently, we concentrated on the top three algorithms identified by RFECV: Random Forest, CatBoost, and XGBoost. To further optimize these models, we utilized GridSearchCV to find the best parameter settings. After training the models with these optimized parameters, XGBoost achieved an accuracy score of 94%, outperforming Random Forest and CatBoost, which both scored 92.5%.

In summary, this study highlights the efficacy of XGBoost in predicting diabetes, this is due to its advanced boosting techniques and robust regularization methods. The importance of comprehensive feature engineering and parameter tuning was also underscored. Future research could explore additional preprocessing steps and incorporate more complex models to enhance predictive accuracy further.

References

[1] J. J. Khanam and S. Y. Foo, “A comparison of machine learning algorithms for diabetes prediction,” ICT Express, vol. 7, no. 4, pp. 432–439, Dec. 2021, doi: 10.1016/j.icte.2021.02.004.

[2] Z. Sabouri, N. Gherabi, M. Nasri, M. Amnai, H. El Massari, and I. Moustati, “Prediction of Depression via Supervised Learning Models: Performance Comparison and Analysis,” International Journal of Online and Biomedical Engineering (iJOE), vol. 19, no. 09, Art. no. 09, Jul. 2023, doi: 10.3991/ijoe.v19i09.39823.

[3] M. S. Alzboon, M. S. Al-Batah, M. Alqaraleh, A. Abuashour, and A. F. H. Bader, “Early Diagnosis of Diabetes: A Comparison of Machine Learning Methods,” International Journal of Online and Biomedical Engineering (iJOE), vol. 19, no. 15, Art. no. 15, Oct. 2023, doi: 10.3991/ijoe.v19i15.42417.

[4] H. El Massari, N. Gherabi, S. Mhammedi, H. Ghandi, M. Bahaj, and M. R. Naqvi, “The Impact of Ontology on the Prediction of Cardiovascular Disease Compared to Machine Learning Algorithms,” iJOE, vol. 18, no. 11, Art. no. 11, Aug. 2022, doi: 10.3991/ijoe.v18i11.32647.

[5] N. H. Taz, A. Islam, and I. Mahmud, “A Comparative Analysis of Ensemble Based Machine Learning Techniques for Diabetes Identification,” in 2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), Jan. 2021, pp. 1–6. doi: 10.1109/ICREST51555.2021.9331036.

[6] M. J. Uddin et al., “A Comparison of Machine Learning Techniques for the Detection of Type-2 Diabetes Mellitus: Experiences from Bangladesh,” Information, vol. 14, no. 7, Art. no. 7, Jul. 2023, doi: 10.3390/info14070376.

[7] S. M. Ganie, P. K. D. Pramanik, M. Bashir Malik, S. Mallik, and H. Qin, “An ensemble learning approach for diabetes prediction using boosting techniques,” Front Genet, vol. 14, p. 1252159, Oct. 2023, doi: 10.3389/fgene.2023.1252159.

[8] M. S. Alam, M. J. Ferdous, and N. S. Neera, "Enhancing Diabetes Prediction: An Improved Boosting Algorithm for Diabetes Prediction," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 15, no. 5, Art. no. 5, Jun. 2024, doi: 10.14569/IJACSA.2024.01505129.

[9] P. Talari et al., "Hybrid feature selection and classification technique for early prediction and severity of diabetes type 2," PLOS ONE, vol. 19, no. 1, p. e0292100, Jan. 2024, doi: 10.1371/journal.pone.0292100.

[10] I. Tasin, T. U. Nabil, S. Islam, and R. Khan, "Diabetes prediction using machine learning and explainable AI techniques," Healthcare Technology Letters, vol. 10, no. 1–2, pp. 1–10, 2023, doi: 10.1049/htl2.12039.

[11] R. Rajalakshmi, P. Sivakumar, L. K. Kumari, and M. C. Selvi, "A novel deep learning model for diabetes melliuts prediction in IoT-based healthcare environment with effective feature selection mechanism," J Supercomput, vol. 80, no. 1, pp. 271–291, Jan. 2024, doi: 10.1007/s11227-023-05496-6.

[12] M. A. Bülbül, "A novel hybrid deep learning model for early stage diabetes risk prediction," J Supercomput, May 2024, doi: 10.1007/s11227-024-06211-9.

[13] A. R. Mohamed Yousuff, M. Zainulabedin Hasan, R. Anand, and M. Rajasekhara Babu, "Leveraging deep learning models for continuous glucose monitoring and prediction in diabetes management: towards enhanced blood sugar control," Int J Syst Assur Eng Manag, vol. 15, no. 6, pp. 2077–2084, Jun. 2024, doi: 10.1007/s13198-023-02200-y.

[14] K. K. Patro et al., "An effective correlation-based data modeling framework for automatic diabetes prediction using machine and deep learning techniques," BMC Bioinformatics, vol. 24, no. 1, p. 372, Oct. 2023, doi: 10.1186/s12859-023-05488-6.

[15] M. F. Aslan and K. Sabanci, "A Novel Proposal for Deep Learning-Based Diabetes Prediction: Converting Clinical Data to Image Data," Diagnostics, vol. 13, no. 4, Feb. 2023, doi: 10.3390/diagnostics13040796.

[16] P. V and R. D. R, "A Hybrid Model for Prediction of Diabetes Using Machine Learning Classification Algorithms and Random Projection," Jun. 28, 2023. doi: 10.21203/rs.3.rs-3081331/v1.

[17] S. Mhammedi, H. El Massari, and N. Gherabi, "Composition of Large Modular Ontologies Based on Structure," in Advances in Information, Communication and Cybersecurity, Y. Maleh, M. Alazab, N. Gherabi, L. Tawalbeh, and A. A. Abd El-Latif, Eds., in Lecture Notes in Networks and Systems. Cham: Springer International Publishing, 2022, pp. 144–154. doi: 10.1007/978-3-030-91738-8_14.

[18] A. A. Alzubaidi, S. M. Halawani, and M. Jarrah, "Towards a Stacking Ensemble Model for Predicting Diabetes Mellitus using Combination of Machine Learning Techniques," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 14, no. 12, Art. no. 12, 47/29 2023, doi: 10.14569/IJACSA.2023.0141236.

[19] S. K. S. Modak and V. K. Jha, "Diabetes prediction model using machine learning techniques," Multimed Tools Appl, vol. 83, no. 13, pp. 38523–38549, Apr. 2024, doi: 10.1007/s11042-023-16745-4.

[20] K. Oliullah, M. H. Rasel, Md. M. Islam, Md. R. Islam, Md. A. H. Wadud, and Md. Whaiduzzaman, "A stacked ensemble machine learning approach for the prediction of diabetes," J Diabetes Metab Disord, vol. 23, no. 1, pp. 603–617, Jun. 2024, doi: 10.1007/s40200-023-01321-2.

[21] S. G. Choi et al., "Comparisons of the prediction models for undiagnosed diabetes between machine learning versus traditional statistical methods," Sci Rep, vol. 13, no. 1, p. 13101, Aug. 2023, doi: 10.1038/s41598-023-40170-0.

[22] A. A. Alzubaidi, S. M. Halawani, and M. Jarrah, "Integrated Ensemble Model for Diabetes Mellitus Detection," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 15, no. 4, Art. no. 4, 33/30 2024, doi: 10.14569/IJACSA.2024.0150423.

[23] M. Kawarkhe and P. Kaur, "Prediction of Diabetes Using Diverse Ensemble Learning Classifiers," Procedia Computer Science, vol. 235, pp. 403–413, Jan. 2024, doi: 10.1016/j.procs.2024.04.040.

[24] I. Nissar et al., "An Intelligent Healthcare System for Automated Diabetes Diagnosis and Prediction using Machine Learning," Procedia Computer Science, vol. 235, pp. 2476–2485, Jan. 2024, doi: 10.1016/j.procs.2024.04.233.

[25] A. Hennebelle, H. Materwala, and L. Ismail, "HealthEdge: A Machine Learning-Based Smart Healthcare Framework for Prediction of Type 2 Diabetes in an Integrated IoT, Edge, and Cloud Computing System," Procedia Computer Science, vol. 220, pp. 331–338, Jan. 2023, doi: 10.1016/j.procs.2023.03.043.

[26] S. Jangili, H. Vavilala, G. S. B. Boddeda, S. M. Upadhyayula, R. Adela, and S. R. Mutheneni, "Machine learning-driven early biomarker prediction for type 2 diabetes mellitus associated coronary artery diseases," Clinical Epidemiology and Global Health, vol. 24, p. 101433, Nov. 2023, doi: 10.1016/j.cegh.2023.101433.

[27] 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, Jan. 2023, doi: 10.1016/j.procs.2022.12.107.

[28] A. Nurdin, M. M. Tane, R. W. T. Tumewu, K. M. Suryaningrum, and H. A. Saputri, "Using Machine Learning for the Prediction of Diabetes with Emphasis on Blood Content," Procedia Computer Science, vol. 227, pp. 990–1001, Jan. 2023, doi: 10.1016/j.procs.2023.10.608.

[29] H. El Massari, N. Gherabi, S. Mhammedi, Z. Sabouri, H. Ghandi, and F. Qanouni, "Effectiveness of applying Machine Learning techniques and Ontologies in Breast Cancer detection," Procedia Computer Science, vol. 218, pp. 2392–2400, Jan. 2023, doi: 10.1016/j.procs.2023.01.214.

[30] H. El Massari, N. Gherabi, S. Mhammedi, H. Ghandi, F. Qanouni, and M. Bahaj, "Integration of ontology with machine learning to predict the presence of covid-19 based on symptoms," BEEI, vol. 11, no. 5, Art. no. 5, Oct. 2022, doi: 10.11591/eei.v11i5.4392.