

Prediction of Diabetes Mellitus using Ensemble Methods in Machine Learning

Dr. K. Malarvizhi¹, Dr. M. Sangeetha², V Deepak³, C Kavikumar⁴ and P Surya Prakash⁵

¹⁻⁵Coimbatore Institute of Technology, Coimbatore, India

Email: malarvizhi@cit.edu.in, {msangeetha@cit.edu.in, deepakvel82@gmail.com, ckavikumar2k@gmail.com, psurya521@gmail.com}

Abstract— Diabetes Mellitus is a chronic disease that affects millions of people worldwide and is known to cause an increase in blood sugar levels. The main objective of this study is to build a model that can accurately predict the early stages of diabetes and reduce the risk associated with it. To achieve this goal, several supervised machine learning algorithms such as Support Vector Machines, Decision Trees, K-Nearest Neighbors, and Logistic Regression have been employed. Additionally, ensemble techniques like Random Forest have been used to improve the accuracy of the machine learning models for the PIMA diabetes dataset, which is widely used for diabetes prediction. To further enhance the accuracy of the model, LGBM Feature Selector have been utilized. Moreover, Grasshopper Optimization, a metaheuristic optimization algorithm, has been employed to optimize the machine learning models. This study provides a comprehensive and effective method for diabetes prediction using machine learning and optimization techniques, which can be beneficial for healthcare professionals and researchers in the field of diabetes management.

Index Terms— Diabetes prediction, Random Forest, Grasshopper Optimization, LGBM Feature Selector.

I. INTRODUCTION

A. Diabetes Prediction

Diabetes is a chronic disease that affects millions of people worldwide. Early detection and accurate diagnosis of diabetes is crucial for effective treatment and management of the disease. Machine learning algorithms have been widely used in predicting diabetes in recent years, as they offer a powerful tool for analyzing large and complex datasets. In this paper, we focus on developing a robust and accurate diabetes prediction model using the Grasshopper Optimization Algorithm (GOA) and a feature selection technique called LightGBM (LGBM). In addition, we compare the performance of our model with other commonly used feature selection techniques, including SelectKBest, Recursive Feature Elimination (RFE), Mutual Information, and Lasso Regression. To evaluate the performance of our model, we also compare it with several commonly used machine learning algorithms, including Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Naive Bayes, Decision Tree, and Random Forest. The results of the paper will have important implications for the development of effective diabetes prediction models and can be used by healthcare professionals and researchers to improve the diagnosis and management of diabetes.

II. RELATED WORKS

The system presented in [1] uses few existing ML classification models for the prediction of diabetic patients have been discussed based on the accuracy. An expression of accuracy on the classification problem has been identified. ML technique was enforced on the PIDD data set. The results of the implementation method show how the LR performed better than other ML algorithms. The results show that glucose and BMI strongly correlate with diabetes using association rule mining. In this system it has been found that the ROC value of LR is 86%. RF, MLP, SVC, DT, GBC, and LR algorithms were used in [2]. The top-performing algorithm was Random Forest. Top-performing algorithms from each portion were selected and closely reviewed the outcomes. The splitting outcomes of MLA can produce accurate, reliable, and tested in some cases they may even outperform learnable ability.

It compares the Clinical dataset utilized in this study with the PIMA diabetes dataset in order to better diagnose Early Prediction of Diabetes based on the Machine Learning Algorithm as compared to the diabetes clinical data set has more accuracy. A smart continuous monitoring architecture was developed in [3] for monitoring the diabetic patients using the Internet of Things (IoT), which was based on machine learning categorization and the IoT. Patient blood sugar, temperature, and exercise levels were all monitored by portable sensors that were included in the design. Data collection and analysis were carried out using a number of different categorization strategies. The proposed technology also provided diabetes patients with the capability of predicting their blood sugar levels in the future, which was advantageous to them. The methodology proposed in [5] uses ML based system using LR-RF combination for feature selection technique and classifier gave the highest classification accuracy. The results demonstrated that the proposed combination reached an accuracy of **94.25%** for K10 protocol. Moreover, a comparative analysis was made using four classifiers, one feature selection technique, and three partition protocol (total 12) experiments.

III. PROPOSED SYSTEM

A. Data Collection

The PIMA dataset is a well-known and frequently used dataset in the field of diabetes prediction. It is named after the Pima Native American tribe, from which the dataset was originally collected. The dataset contains a total of 768 observations, where each observation represents an individual and consists of 8 attributes or features.

The features included in the PIMA dataset are:

- 1) *Pregnancies*: Number of times the person has been pregnant
- 2) *Glucose*: Plasma glucose concentration a 2 hours in an oral glucose tolerance test (OGTT)
- 3) *Blood Pressure*: Diastolic blood pressure (mm Hg)
- 4) *Skin Thickness*: Triceps skin fold thickness (mm)
- 5) *Insulin*: 2-Hour serum insulin (μ U/ml)
- 6) *BMI*: Body mass index ($\text{weight in kg}/(\text{height in m})^2$)
- 7) *Diabetes Pedigree Function*: A score that indicates the genetic predisposition for diabetes based on family history
- 8) *Age*: Age in years
- 9) The target variable in the PIMA dataset is a binary classification variable that indicates whether or not the individual has diabetes. It is represented by 1 if the individual has diabetes, and 0 if not.

B. Data Pre-processing

Data pre-processing is a crucial step in the machine learning pipeline, especially for real-world datasets that often contain noisy or missing values. Pre-processing techniques are used to clean and transform the data into a format that can be easily consumed by machine learning models, thereby improving their accuracy and efficiency.

One common issue in datasets is missing values, which can be handled in different ways. In our PIMA diabetes dataset, we have used two techniques to handle missing values:

Deleting rows: In this technique, rows that contain missing values are removed entirely from the dataset. However, this can result in a loss of information, and hence it should be used judiciously.

Mean value imputation: This technique involves calculating the mean value of a feature and replacing missing values with this mean value. It is mainly used for datasets with numeric features, like the PIMA dataset.

C. Feature Selection

Feature selection is a critical step in machine learning, as it helps to identify the most relevant features that are highly correlated with the target outcome. The selection of appropriate features not only enhances the accuracy of

machine learning models, but also reduces the computational complexity and improves the efficiency of the models. In our system, we have employed several feature selection techniques, including Select K Best, Recursive Feature Elimination, Mutual Information, Lasso Regression, and LGBM Feature Selector.

1. Recursive Feature Elimination

Recursive Feature Elimination (RFE) is a technique used in machine learning to select a subset of the most important features from a larger set of features. It works by recursively removing features from the original set, building a model using the remaining features, and then assessing the model's performance. The goal is to find the smallest set of features that still produces a model with high accuracy.

2. Mutual Information

Mutual Information estimates mutual information for fixed categories like in a classification problem or a continuous target variable in regression problems. Mutual Information works on the entropy of the variables. Let us take two random variables there mutual information between them will be zero if and only if the variables are completely independent otherwise the mutual information between them would be symmetric and non-negative.

3. Lasso Regression

Lasso regression (short for “Least Absolute Shrinkage and Selection Operator”) is a type of linear regression that is used for feature selection and regularization. It works by adding a penalty term to the standard linear regression cost function, which encourages the model to use fewer variables or features in the final regression equation.

4. Select K Best

Select K Best is a feature selection algorithm that selects the K best features from a given dataset. The algorithm is based on the chi-square statistical test, which measures the dependence between two variables. SelectKBest works by calculating the chi-square statistic for each feature in the dataset and selecting the K features with the highest chi-square values.

5. LGBM Feature Selector

LGBM Feature Selector is a feature selection algorithm that is based on the LightGBM algorithm, which is a gradient boosting framework that uses tree-based learning algorithms. This feature selection algorithm is specifically designed to work with high-dimensional datasets and provides a fast and efficient way to select the most important features for a given machine learning problem.

The LGBM Feature Selector works by training a LightGBM model on the dataset and then using the feature importance scores generated by the model to select the most important features. The feature importance scores generated by the LightGBM model are based on the split gain of each feature, which measures the contribution of each feature to the improvement of the model's performance.

D. Model Building

The process of model building is a crucial step in accurately predicting diabetes. In this study, we employed various supervised machine learning algorithms, including Logistic Regression, K-Nearest Neighbors (KNN), Naive Bayes, Support Vector Machines (SVM), Decision Tree, and Random Forest, to train and test the models for improved accuracy.

To optimize the performance of the machine learning models, we used the Grasshopper Optimization algorithm, which helped to refine the models and achieve better accuracy in predicting the diabetes data. By evaluating the performance of each algorithm, we identified the best model that provided the most accurate results. The selected model was then used to predict and classify the diabetes data more accurately

1. Logistic regression

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. In this, we use multinomial logistic regression. The logistic regression uses a more complex cost function which is the sigmoid function or logistic function. The hypothesis of logistic regression tends to limit the cost function between 0 and 1.

2. KNN

It is a Machine Learning algorithm based on the supervised learning technique. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for Classification problems. It is also called a lazy learner

algorithm because it does not learn from the training set. Instead, it stores the dataset and at the time of classification, it acts on the dataset.

3. Random Forest

Random Forest is one of the popular machine learning algorithms that belong to the supervised learning technique. The random forest can be used to solve both Classification and Regression problems in Machine Learning. Random forest is the ensemble classifier, which combines lots of classifiers to solve highly complex problems and it improves the model accuracy. The random forest contains the number of decision trees and is used for classifying the data correctly and improving the accuracy. In Random forest, each decision tree gives the prediction based on the majority votes of the prediction final output is decided.

4. Naïve Bayes

Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problem. It is mainly used in text classification that includes a high-dimensional training dataset. It is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.

5. Support Vector Machine

Support Vector Machine(SVM) is a supervised machine learning algorithm used for both classification and regression. The objective of the SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points. The dimension of the hyperplane depends upon the number of features. If the number of input features is two, then the hyperplane is just a line. If the number of input features is three, then the hyperplane becomes a 2-D plane. It becomes difficult to imagine when the number of features exceeds three.

6. Decision Tree

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

7. Grasshopper Optimization Algorithm

The GOA algorithm was proposed by Saremi et al. which is a recent and interesting swarm intelligence algorithm that mimics grasshoppers' natural foraging and swarming behaviors. Grasshoppers are insects well-known as a dangerous pests that affect and damage crop production and agriculture. Their life cycle includes two phases called nymph and adulthood.

The swarming behavior of grasshoppers is mathematically modeled as follows:

$$P_i = S_i + G_i + A_i \quad (1)$$

where P_i indicates the i -th grasshopper' position, S_i is the social interaction between grasshoppers, G_i denotes the gravity force on the i -th grasshopper, and A_i is the wind advection. To produce a random behavior of grasshoppers, Equation 1 can be rewritten as follows:

$$P_i = r_1 S_i + r_2 G_i + r_3 A_i \quad (2)$$

where r_1 , r_2 , and r_3 are random numbers in the range $[0, 1]$ The social interaction S_i is defined as follows:

$$S_i = \sum_{\substack{j=1 \\ j \neq i}}^N s(d_{ij}) \hat{d}_{ij} \quad (3)$$

where N denotes the number of grasshoppers,

$d_{ij} = |P_j - P_i|$ defines the Euclidean distance between the i -th and the j -th grasshopper, $\hat{d}_{ij} = (P_j - P_i) / d_{ij}$ is a unit vector from the i -th to the j -th grasshopper, and s represents the social forces designed by the following equation: where f and l are the attraction intensity and attraction length scale, respectively. The social interaction between grasshoppers can be defined as attraction and repulsion.

The gravity force G_i is given by the following equation:

$$s(r) = f \exp \frac{-r}{l} - \exp^{-r} \quad (4)$$

where f and l are the attraction intensity and attraction length scale, respectively. The social interaction between grasshoppers can be defined as attraction and repulsion. The gravity force G_i is given by the following equation:

$$G_i = -g \hat{e}_g \quad (5)$$

where g denotes the gravitational constant and $\hat{e}_g$ represents a unit vector toward the center of earth. The wind advection A_i is given by the following equation:

$$A_i = u \hat{e}_w \quad (6)$$

where u represents the drift constant and $\hat{e}_w$ is a unit vector in the wind direction.

$$P_i = \sum_{j=1, j \neq i}^N s(|P_j - P_i|) \frac{P_j - P_i}{d_{ij}} - g \hat{e}_g + u \hat{e}_w \quad (7)$$

Equation 7 cannot be used directly to solve optimization problems, as the grasshoppers reach quickly the comfort zone and the swarm system does not converge to a target location. An enhanced version of this equation is given as:

$$P_i^d = c \left(\sum_{j=1, j \neq i}^N c \frac{ub_d - lb_d}{2} s(|P_j^d - P_i^d|) \frac{P_j^d - P_i^d}{d_{ij}} \right) + \hat{T}_d \quad (8)$$

where ub_d and lb_d represent the upper and lower bounds in the d -th dimension, respectively. $\hat{T}_d$ denotes the best solution found so far in the d -th dimension space. Note that S is similar to S component in equation 1, G is equal to zero and A is always toward the best solution $\hat{T}_d$.

$$c = c_{max} - t \frac{c_{max} - c_{min}}{t_{max}} \quad (9)$$

where c_{max} and c_{min} represent the maximum and minimum values of c , respectively, t is the current iteration and t_{max} is the maximum number of iterations. The pseudo-code of the standard Grasshopper Optimization Algorithm is given in Algorithm 1. Its flowchart can be represented in below figure.

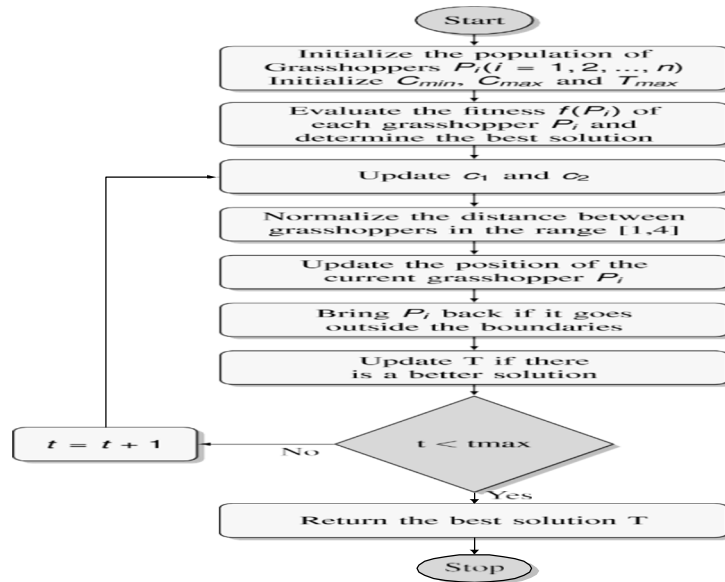


Fig 1 Flow chart of Grasshopper Optimization Algorithm

E. Performance Evaluation

In machine learning, it is important to evaluate the performance of a model using appropriate metrics. Sklearn is a popular Python library that provides various metrics for evaluating classification models. In this section, we will discuss three commonly used metrics in Sklearn, including accuracy score, confusion matrix, and classification report.

Accuracy Score = (Number of Correct Predictions) / (Total Number of Predictions)

Confusion matrix: The confusion matrix is a table that summarizes the performance of a classification model. It shows the number of true positive (TP), false positive (FP), true negative (TN), and false negative (FN) predictions for each class.

The accuracy of the confusion matrix is calculated as the sum of the diagonal elements divided by the total number of predictions

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN})$$

Classification report: The classification report provides a summary of the precision, recall, f1-score, and support for each class. The precision measures the proportion of true positive predictions among all positive predictions, while the recall measures the proportion of true positive predictions among all actual positive samples.

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$$

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$$

$$\text{F1-Score} = 2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$$

$$\text{Support} = \text{Number of Samples in Each Class}$$

IV. EXPERIMENTAL RESULTS

In this project, we compared the performance of six different machine learning models and evaluated their performance using the confusion matrix. We also used the grasshopper optimization technique to optimize the hyper parameters of a random forest classifier. Additionally, we conducted a comparative analysis of the models with and without using the Light GBM feature selection technique. The accuracies of the models were recorded and presented in a table.

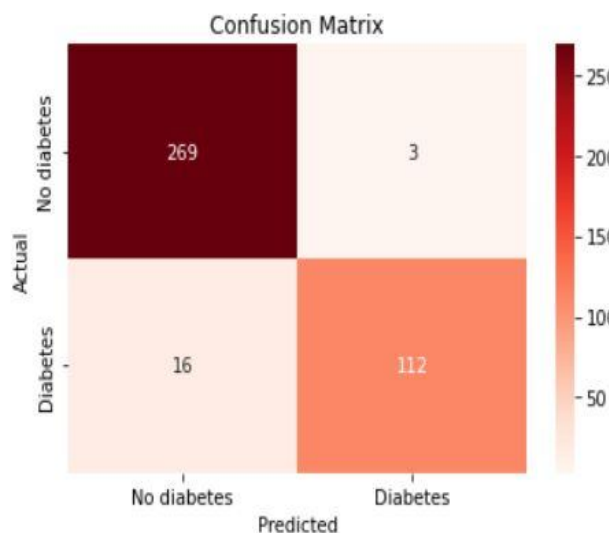


Fig 2. Grasshopper Optimization without LGBM

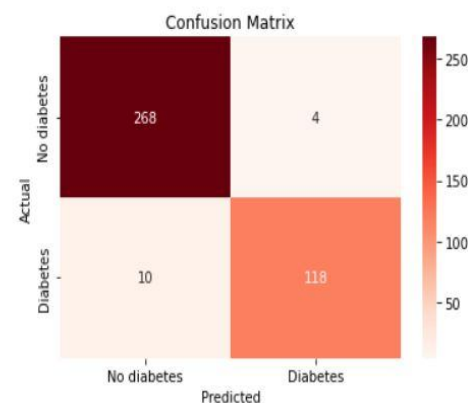


Fig 3 Random Forest with Grasshopper and LGBM

The above figure 2 shows the performance of the random forest with using grasshopper optimization technique and without using LGBM Feature selector.

The above figure 3 shows the performance of the random forest with using grasshopper optimization technique and LGBM Feature selector.

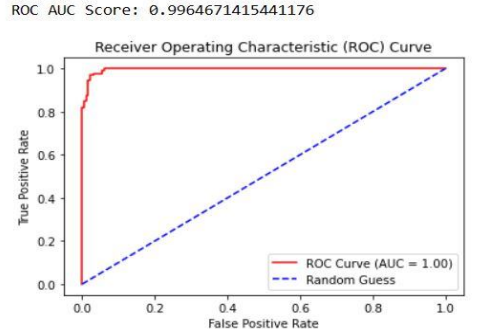


Fig 4 ROC Curve for Random Forest with using Grasshopper Optimization and LGBM Feature Selector

TABLE I: ACCURACIES OF ALGORITHMS

ALGORITHM	Without LGBM Feature Selector	With LGBM Feature Selector
Logistic Regression	78.57	79.25
KNN	72.07	81
Naïve Bayes	75.32	79
Support Vector Machine	79.87	83.75
Decision Tree	76.62	94
Random Forest	80.51	94.25
Random Forest with Grasshopper Optimization	95.5	97.5

V. CONCLUSION

The main purpose of the paper is to predict diabetes accurately using different machine learning algorithms and feature selection techniques. The PIMA dataset is used for training and testing with various models, including Logistic Regression, KNN, Naive Bayes, SVM, Decision Tree, and Random Forest Classifier. Our proposed grasshopper optimization algorithm achieved an accuracy of 94.75%, and by applying the LGBM feature selector to the grasshopper optimization technique, the accuracy increased to 97.5%. Furthermore, the results show that all other algorithms without feature selection and grasshopper optimization techniques yielded accuracy rates of only 70% to 80%. This study demonstrates the effectiveness of the proposed feature selection and grasshopper optimization techniques in improving the performance of machine learning models for diabetes prediction.

SCOPE FOR FUTURE

As future work, Intended to expand the research to include the detection and classification of diabetes into type 1 and type 2. Further enhancement is to developing a mobile application. This mobile application will enable easy and convenient diabetes detection, allowing users to assess their diabetes risk using the developed model. The application will provide an intuitive interface for users to input their relevant health information and receive an instant classification result.

REFERENCES

- [1] Raja Krishnamoorthi, Shubham Joshi, and Hatim Z. Almarzouki, "A Novel Diabetes Healthcare Disease Prediction Framework using Machine Learning Techniques," Journal of Healthcare Engineering, pp. 1-10 2022.
- [2] Salliah Shafi, Prof. Gufran Ahmad Ansari, "Early Prediction of Diabetes Disease & Classification of Algorithms Using Machine Learning Approach", International Conference on Smart Data Intelligence, 2021.
- [3] M. M. Bukhari, B. F. Alkhamees, S. Hussain, A. Gumaei, A. Assiri, and S. S. Ullah, "An improved artificial neural network model for effective diabetes prediction," Complexity, vol. 2021, Article ID 5525271, 10 pages, 2021.

- [4] KM Jyoti Rani, "Diabetes Prediction Using Machine Learning," International Journal of Scientific Research in Computer Science Engineering and Information Technology, volume. 6, pp. 294-305, 2020.
- [5] Md. Maniruzzaman, Md. Jahanur Rahman, B. Ahammed, and Md. Menhazul Abedin, "Classification and Prediction of Diabetes Disease Using Machine Learning Paradigm," Health Information Science and Systems, vol. 8, 2020.
- [6] 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, 2020.
- [7] S. Srivastava, L. Sharma, V. Sharma, and A. Kumar, "Prediction of diabetes using artificial neural network approach," in Engineering Vibration, Communication and Information Processing Vol. 29, Springer, Berlin/Heidelberg, Germany, 2020.
- [8] A. Mahabub, "A Robust Voting Approach for Diabetes Prediction Using Traditional Machine Learning Techniques," SN Applied Sciences, Springer, 2019.
- [9] Desmond Bala Bisandu, Dorcas Dachollom Datiri, Eva Onokpasa, Godwin Thomas, Musa Maaji Haruna, Aminu Aliyu and Jerry Zachariah Yakubu (2019), "Diabetes Prediction Using Data Mining Techniques", International Journal of Research and Innovation in Applied Science, ISSN 2454-6194, Volume-4, Issue-6, PP.103-111.
- [10] A. Mahabub, "A Robust Voting Approach for Diabetes Prediction Using Traditional Machine Learning Techniques," SN Applied Sciences, Springer, 2019.
- [11] Sneha, N., Gangil, T. Analysis of diabetes mellitus for early prediction using optimal features selection. J Big Data 6, 13 (2019).
- [12] S. Kapoor and K. Priya, "Optimizing hyper parameters for improved diabetes prediction," International Research Journal of Engineering and Technology, vol. 5, 2018.
- [13] Dutta, Debadri & Paul, Debpriyo & Ghosh, Parthajeet. (2018). Analysing FeatureImportances for Diabetes Prediction using Machine Learning. 924-928. 10.1109/IEMCON.2018.8614871
- [14] M. A. Sarwar, N. Kamal, W. Hamid, and M. A. Shah, "Prediction of Diabetes Using Machine Learning Algorithms in Healthcare," in Proceedings of the 2018 24th International Conference on Automation and Computing (ICAC), Newcastle upon Tyne, UK, September 2018.