A Framework for Optimized Diabetes Detection Model Based on Binary Butterfly and Machine Learning Algorithms

Yusuf Ayuba[1], Enesi Femi Aminu[2], MUHAMMAD Muhammad Kudu[3]
[1,2,3]Dept. of Computer Science, School of Information & Communication Technology,
Federal University of Technology, Minna.

ayubayusuf117@gmail.com[1], enesifa@futminna.edu.ng[2], muhammad_kudu@futminna.edu.ng[3]

Abstract

Diabetes has become a major cause of death in both developed and developing countries, affecting a large number of people globally. prompting significant investments in research to find a cure for this critical disease. Traditional approaches reliant on diabetes detection are time-consuming, this necessitates a paradigm shift towards more efficient methodologies. In response, this study introduces a conceptual framework for diabetes detection by leveraging the power of optimized machine learning algorithms. Addressing data preprocessing techniques and optimized feature selection algorithms, and machine learning algorithms, specifically Random forest, multilayer perceptron, and Gradient boosting model, the result shows that Random forest emerges as the potent model showcasing a remarkable performance metrics: accuracy score of 97.66%, F1-score of 97.56%, AUC-ROC of 98.54%, Multilayer perceptron achieved an accuracy of 96.10%, F1-score of 95.96%, AUC-ROC of 98.65% Gradient boost achieved and accuracy of 91.82%, F1-score of 91.49% and AUC-ROC of 98.01% respectively. These findings underscore the significant role of feature selection and machine learning in detecting diabetes offering transformative possibilities for global healthcare enhancement.

keywords : Diabetes melitus, Binary Butterfly, Random Forest, Multilayer Perceptron and Gradient Boost.

Introduction

Diabetes mellitus is a metabolic disease caused by an excessive amount of glucose in the blood due to the inadequate secretion of insulin or insulin resistance, the pancreas is the main source for producing insulin, the pancreas is a crucial hormone that is responsible for transferring the converted glucose through the bloodstream to different body parts, inappropriate secretion of insulin causes glucose to persist in the blood, which ultimately causes a surge in sugar level in the blood [1], new research indicates that about 50% of persons with type-2 diabetes are not aware of nor have a diagnosis for the illness [2] it is predicted that 537 million people worldwide have diabetes mellitus, a number that will have increased to 783 million By 2045, diabetes mellitus increases the risk of diabetic foot complications, such as ischaemia infection and peripheral neuropathy [3]. When blood sugar levels rise as a result of metabolic issues, diabetes develops. The heart, blood vessels, and eyes are just a few of the body systems and organs that may sustain harm from this exposure. It is noteworthy that hyperglycemia, or high blood sugar is the direct cause of these negative diseases [4]. It is also stated that diabetes can not be cured, but patients can lessen its impact and avoid more significant health issues down the road by being diagnosed early and adhering to a healthy diet [5] Diabetes detection uses a number of machine learning (ML) models, such as decision trees, k-nearest neighbors, artificial neural networks and SVMs etc. Although it is crucial to examine patterns in diabetes incidence and forecast future costs by utilizing risk factors in certain groups, not much has been done to implement ML classification techniques. A crucial step in the analysis of healthcare datasets, including those pertaining to diabetes, is feature selection. [6]. Machine learning algorithms are currently being used in a lot of wearable technology and smartphone apps to forecast blood sugar levels. But the requirement for large data volumes to improve accuracy limits their dependability. On the other hand, artificial neural networks are quite good at handling wild swings in blood sugar levels because they can represent the nonlinearities in the data very well [7].Therefore, this paper aims to design a framework to optimize diabetes detection model using binary butterfly optimization algorithm with RF, MLP, and Gboost machine learning techniques. The remaining sections of the research are organized as follows: section 2 presents the related works, and the proposed methodology is presented in section 3. Others are results, discussion and conclusion, which are presented by sections 4 and 5 respectively.

Related works

Based on the literatures, there are different approaches to detect diabetes, however machine learning approaches have been identified to have a significant role in mitigating the risk associated with diabetes detection based on different approaches. In view of this development, this review article has investigated systematic application of optimization algorithms to detect diabetes considering diabetes dataset 2019 and Pima indian dataset. Different techniques such as optimization algorithms to aid the detection of diabetes disease at an early stage have been reported in the following literatures. [5] proposed a ML-based diabetes detection model for false negative reduction, incorporating data preprocessing such as normalization, removing null values and encoding techniques and SMOTE technique for dataset balancing, and also integrated decision trees, k-nearest neighbors, logistic regression, RF, Naïve Bayes, and Support vector machine algorithms, these accentuate the importance of balanced datasets in reducing false negatives. where RF achieved an impressive accuracy of 97.54% on the Diabetes_dataset_2019 and 80% on the PIMA Indian dataset. In the same way [8] address a challenge on Predictive model and risk analysis for peripheral vascular disease in type-2 diabetes mellitus patients using MLand shapley additive explanation. They employed preprocessing techniques and FS recursive FS elimination and SMOTE for data balancing, integrating Six ML models, including decision tree, logistic regression, random forest, SVM, Extreme gradient boosting and Adaptive Boosting, grid search and 10-cross validation was used to optimize hyperparameters. The Extreme gradient Boost model achieved an accuracy of 89%. In Bangladesh, most of the people are not aware of the deadly clutch of diabetes which becomes rampant in course of time [9] solved a problem by investigating diabetes prediction and optimization of ML through FS and dimensionality reduction. They incorporate FS, dimensionality reduction techniques, and grid search optimization to preprocess data; they also employed five well-established ML algorithms: K-Nearest Neighbors, Support Vector Machines, RF, Extra Trees, and Gradient Boosting. Extra Trees achieved an optimal score of 92.5%. A review of the literature of the proposed diabetes techniques is good for understanding the significance of our suggested technique. Also [4] proposed an application of ML models for early detection and accurate classification of type-2 diabetes, they employed a robust exploratory data analysis for data preprocessing which include data encoding, data cleaning, outlier detection, data transformation, training, testing and validation, five ML models were trained including KNN, Bernoulli Naïve Bayes, decision tree, logistic regression, and SVM, k-nearest neighbor achieved an accuracy of 79.6%. [10] presents an Early detection of type-2 diabetes mellitus using ML-based prediction models, accentuating data preprocessing techniques, employing several ML-based prediction models such as Glmnet, RF, Extreme Gradient Boost, LightGBM were analyzed to evaluate models performance, Extreme Gradient Boost obtained an accuracy of 88.1%. In the same vein [11] proposed a model on diagnosis and classification of diabetes using ML algorithms, They integrated data preprocessing and data cleaning processes such as data imputation-mean technique removal of missing values and outliers from the dataset. They employed several ML models including K-Nearest Neighbors, Naive Bayes, Decision Tree, Extra Trees, Radial Basis Function, Multi-Layer Perceptron. The result obtained showed that the Multi-layer perceptron achieved an accuracy of 82.6%, ROC_AUC 86% respectively. Liu *et al.,* (2024) address a challenge on Predictive model and risk analysis for peripheral vascular disease in type-2 diabetes mellitus patients using MLand shapley additive explanation. They employed preprocessing techniques and FS recursive FS elimination and SMOTE for data balancing, integrating Six ML models, including decision tree, logistic regression, random forest, SVM, Extreme gradient boosting and Adaptive Boosting, grid search and 10-cross validation was used to optimize hyperparameters. The Extreme gradient Boost model achieved an accuracy of 89%. The section below presents the methodology of this research.

Methodology

This study is designed to develop diabetes detection machine learning models based on optimized machine learning algorithms. The methodology involves a step-by-step approach incorporating data collection and preprocessing, outlier detection, and use of optimization algorithms for the development of the detection model, the research subsumes three classification models for diabetes detection. Finally the model is evaluated to analyze the optimum classifier, the entire process is structured to ensure accuracy, efficiency, and reliability of diabetes detection. This study utilize the publicly available PIMA indian dataset and Diabetes_dataset_2019 obtained from Kaggle, the PIDD comprises 768 rows and 9 features, On the other hand Diabetes_dataset_2019 whose shape consists of 952 rows and 18 features, The data preprocessing stage is essential for transforming the dataset into a standard format suitable for model training. This study will focus on data normalization, standardization, removal of missing values, encoding categorical columns, and outlier detection. This study utilizes supervised learning classification models to efficiently detect diabetes, the models are Random Forest (RF), Gradient boost (GBoost) and multilayer perceptron

(MLP). The ML models are trained on the dataset based on the selected feature. In the evaluation phase the following metrics are considered which are: Accuracy, f1-score and Receiver Operating Characteristic (ROC) Area Under the Curve (AUC), which are essential metrics for evaluating the performance of models in detecting diabetes in patients. Also This study analyzes the power of three feature selection techniques. Which are the GOA, BBOA, and SA, and also evaluate the performance of different machine learning models in detecting diabetes. Finally the model is evaluated using the following metrics which include: Accuracy, F-score and Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC), which are essential metrics for evaluating the performance of machine learning models in detecting diabetes.

Accuracy is the ratio of the total number of input samples to the number of accurate predictions.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}. \qquad 1$$

F-measure is a machine learning metric that assesses a model's performance by combining precision and recall into a single score, the score helps to evaluate the Recall and Precision at the same time.

$$F1\text{-}Score = \frac{2 * Recall * Precision}{Recall + Precision}. \qquad 2$$

Receiver operating characteristic or ROC curve, is a visual plot that displays the performance of a binary classifier model at varying threshold values. The ROC curve is the plot of the True positive rate against the false positive rate at each threshold setting. Furthermore, Figure 1 depicts the framework for optimized machine-learning algorithms for diabetes detection.

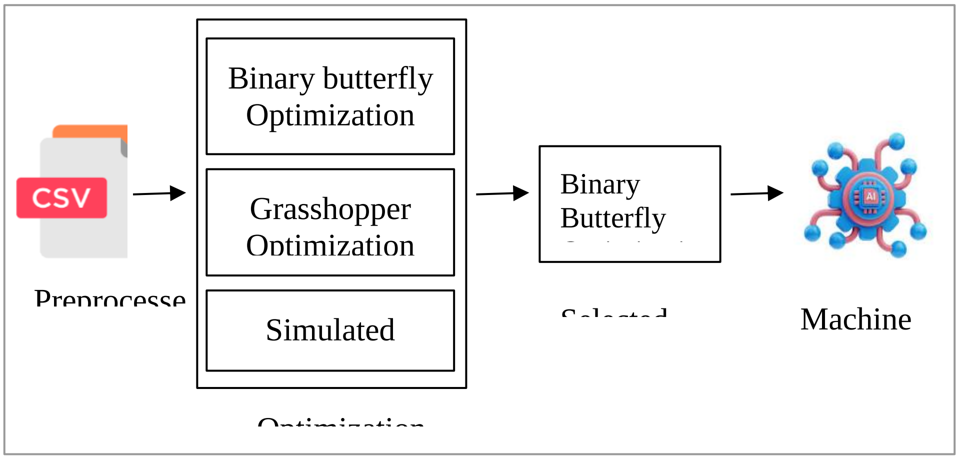


Figure 1 Framework for Optimize Machine-Learning Algorithms for Diabetes Detection.

Figure 1 shows the flow between preprocessed data which are the features, try all optimization algorithms and then select one optimization algorithm for feature selection, finally one algorithm is selected to evaluate the models. Similarly Figure 2 depicts Framework of the developed diabetes detection model using optimized machine learning algorithms.

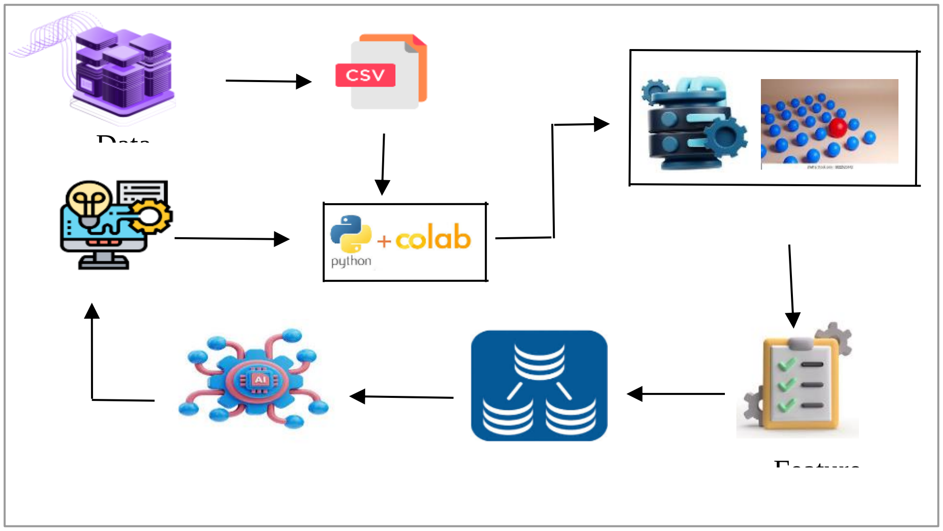


Figure 2 Framework Of The Developed Diabetes Detection Model Using Optimized Machine Learning Algorithms.

Figure 2 shows the conceptual framework of the proposed Developed diabetes detection model using optimized machine learning algorithms. The dataset utilized for this study was obtained from kaggle. Also, Google colab and python programming was used as our main IDE. Data preprocessing incorporates data normalization, removal of missing values, and outlier detection, BBOA was used for feature selection to select key features for analysis. Furthermore the data is splitted into training and testing sets, three machine learning models are used for training. Finally the diabetes detection models are evaluated using: Accuracy, F1-score, Roc_Auc. Table 1 shows the system algorithm of the developed diabetes detection model.

Table 1: System Algorithm

Algorithm

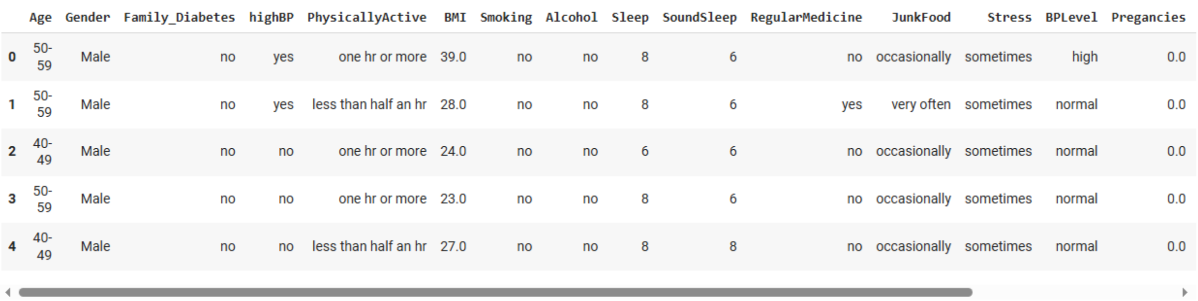
```
1. START
2. Get the Datasets (kaggle.com)
3. Load data {DD_2019}
4. D ← Preprocess (D, normalization, standardization, removing null_val, encoding
   categorical columns & outlier detection)
5. Features selection ←{BBOA, GOA, SA}
6. Split the Data into Training sets and Testing Sets
7. $D_{train}$, $D_{test}$←train_test_split ($D_{selected}$, 80%, 20%)
8. M ← train_model($D_{train}$)
9. Model evaluation: Accuracy, F1-score, AUC_ROC ($D_{test}$)
10. End
```

Table 1 shows the system algorithm starting from the data source, the DD_2019 dataset is loaded, then follows data preprocessing, these includes: normalization, standardization, removal of null values, encoding categorical columns and outlier detection. Also we employed three optimization techniques for feature selection which included BBOA, GOA, SA. The data is splitted into training and testing sets for model training, finally the model will be evaluated using three metrics which include: Accuracy, F1-score, AUC_ROC respectively.

Results and Discussion.

This section presents and discusses the results obtained in order to achieve research objectives. Data preprocessing : here we statistically analyzed the data for duplicates and Not-a-Number values, We converted categorical columns into numerical features, so it can be processed by the detection algorithms, also we normalized the datasets. Table 2. Shows the data sample for DD_2019 before preprocessing.

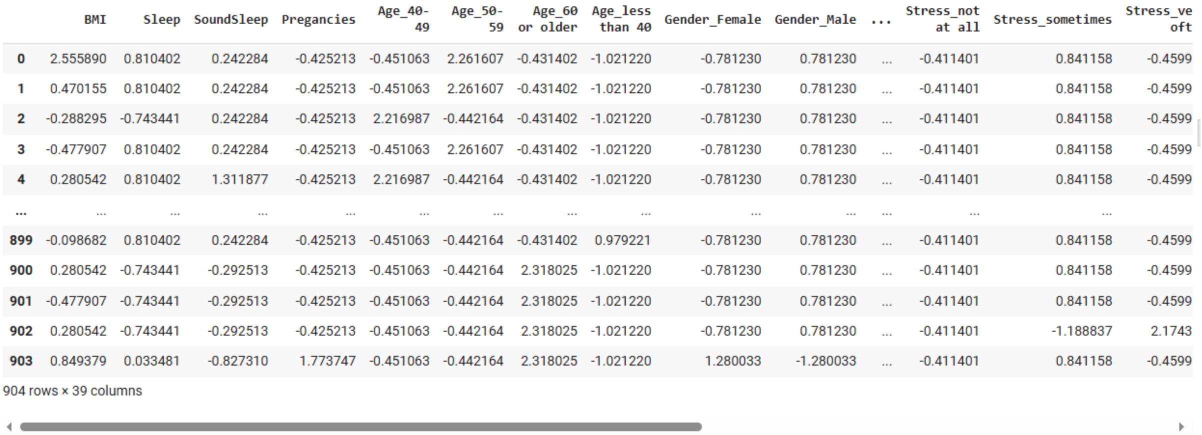
Table 2. Data sample for DD_2019.



	Age	Gender	Family_Diabetes	highBP	PhysicallyActive	BMI	Smoking	Alcohol	Sleep	SoundSleep	RegularMedicine	JunkFood	Stress	BPLevel	Pregancies
0	50-59	Male	no	yes	one hr or more	39.0	no	no	8	6	no	occasionally	sometimes	high	0.0
1	50-59	Male	no	yes	less than half an hr	28.0	no	no	8	6	yes	very often	sometimes	normal	0.0
2	40-49	Male	no	no	one hr or more	24.0	no	no	6	6	no	occasionally	sometimes	normal	0.0
3	50-59	Male	no	no	one hr or more	23.0	no	no	8	6	no	occasionally	sometimes	normal	0.0
4	40-49	Male	no	no	less than half an hr	27.0	no	no	8	8	no	occasionally	sometimes	normal	0.0

Table 2 depicts a visual representation of diabetes dataset 2019 before preprocessing containing 18 features and 952 data points, before applying feature selection algorithms and machine learning models for detection. Table 3 depicts the preprocessed data sample for DD_2019.

Table 3 preprocessed data sample for DD_2019.



	BMI	Sleep	SoundSleep	Pregancies	Age_40-49	Age_50-59	Age_60 or older	Age_less than 40	Gender_Female	Gender_Male	...	Stress_not at all	Stress_sometimes	Stress_ve oft
0	2.555890	0.810402	0.242284	-0.425213	-0.451063	2.261607	-0.431402	-1.021220	-0.781230	0.781230	...	-0.411401	0.841158	-0.4599
1	0.470155	0.810402	0.242284	-0.425213	-0.451063	2.261607	-0.431402	-1.021220	-0.781230	0.781230	...	-0.411401	0.841158	-0.4599
2	-0.288295	-0.743441	0.242284	-0.425213	2.216987	-0.442164	-0.431402	-1.021220	-0.781230	0.781230	...	-0.411401	0.841158	-0.4599
3	-0.477907	0.810402	0.242284	-0.425213	-0.451063	2.261607	-0.431402	-1.021220	-0.781230	0.781230	...	-0.411401	0.841158	-0.4599
4	0.280542	0.810402	1.311877	-0.425213	2.216987	-0.442164	-0.431402	-1.021220	-0.781230	0.781230	...	-0.411401	0.841158	-0.4599
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
899	-0.098682	0.810402	0.242284	-0.425213	-0.451063	-0.442164	-0.431402	0.979221	-0.781230	0.781230	...	-0.411401	0.841158	-0.4599
900	0.280542	-0.743441	-0.292513	-0.425213	-0.451063	-0.442164	2.318025	-1.021220	-0.781230	0.781230	...	-0.411401	0.841158	-0.4599
901	-0.477907	-0.743441	-0.292513	-0.425213	-0.451063	-0.442164	2.318025	-1.021220	-0.781230	0.781230	...	-0.411401	0.841158	-0.4599
902	0.280542	-0.743441	-0.292513	-0.425213	-0.451063	-0.442164	2.318025	-1.021220	-0.781230	0.781230	...	-0.411401	-1.188837	2.1743
903	0.849379	0.033481	-0.827310	1.773747	-0.451063	-0.442164	2.318025	-1.021220	1.280033	-1.280033	...	-0.411401	0.841158	-0.4599

904 rows × 39 columns

Table 3 shows the data was preprocessed by analyzing the data statistically where we removed data points containing duplicates and Not-a-Number values, also we converted categorical columns to numerical values, also we removed data points with missing values which are BMI, pregnancies, pdiabetes, and diabetic with a missing value of 4, 42, 1 and 1 respectively. Also inconsistent data were normalized to lowercase characters. And we also categorical columns to numerical features; also we used isolation forest for outlier detection, we use SMOTE to balance the dataset. We employed different optimization algorithms for feature selection. Table 4 shows the confusion matrix for DD_2019 using the grasshopper optimization algorithm.

Table 4 confusion matrix for DD_2019 using Grasshopper Optimization Algorithm.

Models	**True Positive(TP)**	**True Negative(TN)**	**False Positive(FP)**	**False negative(FN)**
RF	119	130	1	7
MLP	116	122	9	10
GBoost	119	128	3	7

Similarly Table 4. shows the confusion matrix of the random forest model which produced a TP of 119, TN of 130, FP of 1 and FN of 7 respectively. For the Multilayer Perceptron model, the TP, TN, FP and FN are 116, 122, 9 and 10 respectively. For the GBoosting model, the TP, TN, FP and FN are 119, 128, 3 and 7 respectively. Thus the Random forest model using the GOA was able to detect (TP) 119 diabetic patients and 130 (TN) non diabetic patients. Table 6 showcase the confusion matrix for DD_2019 using Simulated Annealing

Table 6 confusion matrix for DD_2019 using Simulated Annealing

Models	**True Positive(TP)**	**True Negative(TN)**	**False Positive(FP)**	**False negative(FN)**
RF	117	128	3	9
MLP	121	122	9	5
GBoost	115	128	8	11

Likewise Table 6 shows the confusion matrix of the random forest model which produced a TP of 117, TN of 128, FP of 3 and FN of 9 respectively. For the MLP model, the TP, TN, FP and FN are 121, 122, 9 and 5 respectively. For the GBoosting model, the TP, TN, FP and FN are 115, 128, 8 and 11 respectively. Table 7 confusion matrix for DD_2019 using binary butterfly optimization algorithm in an ordered manner.

Table 7 confusion matrix for DD_2019 using Binary Butterfly Optimization Algorithm

Models	**True Positive(TP)**	**True Negative(TN)**	**False Positive(FP)**	**False negative(FN)**
RF	120	131	0	6

MLP	119	128	3	7
GBoost	113	123	8	13

In like manner Table 7 shows the confusion matrix of the random forest model which produced a TP of 120, TN of 131, FP of 0 and FN of 6 respectively. For the MLP model, the TP, TN, FP and FN are 119, 128, 3 and 7 respectively. For the GBoosting model, the TP, TN, FP and FN are 113, 123, 8 and 13 respectively. Thus this shows the random forest model which was able to detect diabetic patients with a True negative of (TN) 131 and a true positive (TP) of 120 proving to be the most accurate model in respect to subsequent models. Table 8 shows the performance evaluation for Grasshopper Optimization Algorithm on DD_2019.

Table 8 performance evaluation for Grasshopper Optimization Algorithm on DD_2019

Models	**Accuracy**	**F1-score**	**Roc_AUC**
RF	96.88	96.74	98.43
MLP	94.16	93.92	98.73
GBoost	92.60	92.43	97.83

Table 9 performance evaluation for Simulated Annealing on DD_2019

Models	**Accuracy**	**F1-score**	**Roc_AUC**
RF	95.33	95.12	98.67
MLP	94.55	94.53	98.41

GBoost	92.60	92.36	97.71

Similarly from Table 8 the accuracy , F1-score and Roc_Auc for Grasshopper Optimization Algorithm for Random forest are 97.88%, 96.74%, and 98.43% respectively. For multilayer perceptron model, the Accuracy, F1-score and Roc_Auc are 94.60%, 93.92% and 98.73% respectively, for GBoosting model, the Accuracy, F1-score and Roc_Auc 92.60%, 92.43 and 97.83% respectively.

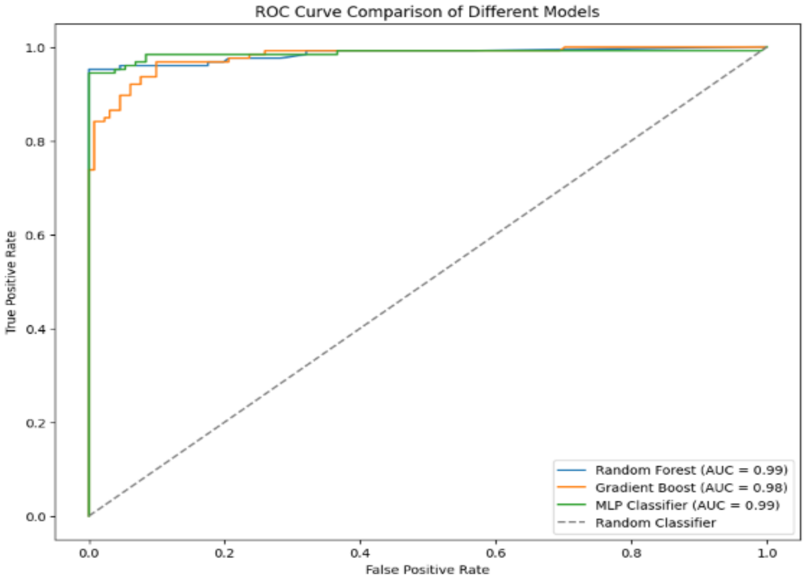
From Table 9 the accuracy, F1-score and Roc_Auc for Simulated Annealing, the random forest achieved 95.33%, 92.36% and 97.71% respectively. Also for the GBoost is 92.60%, 92.36% and 97.71% respectively. Similarly Table 11 shows the performance evaluation of Binary Butterfly Optimization algorithm on DD_2019.

Table 10 performance evaluation of Binary Butterfly Optimization algorithm on DD_2019

Models	**Accuracy**	**F1-score**	**Roc_AUC**
RF	97.66	97.56	98.54
MLP	96.10	95.96	98.65
GBoost	91.82	91.49	98.01

From Table 10 depicts the Random forest achieved an optimum result of 97.66% accuracy and F1-score of 97.56%, and Roc_Auc 98.54% respectively, for multilayer perceptron model, are 96.10%, 95.96% and 98.65% respectively. Also the GBoost model achieved an accuracy of 91.82%, and

98.01% respectively. Figure 1 shows the ROC Curve Comparison of different models.



ROC Curve Comparison of different models

Figure 1 shows the Roc curve for random forest, Gradient boost and multilayer perceptron are 99.00%, 98.00% and 99.00% respectively, for random classifiers on DD_2019.

Conclusion

The goal of this study is to develop a diabetes detection model based optimized machine learning algorithms for diabetes detection; the experiment was carried out using Diabetes Dataset 2019 This study assesses the models performance using these metrics, Accuracy, F1-score, and Roc_Auc. Thus this study proved an improvement in diabetes detection over the existing model by the previous research work done by uddin et al., 2023, they compared the improvements of diabetes detection, where Random forest classifier was considered the most. This study focused majorly on diabetes detection. From the result obtained, two different datasets were considered using machine learning algorithms. Some models perform better which increase performance of accuracy, F1-

score and Roc_Auc, but Random forest achieved an optimum result for this experiment. Based on the findings of this study, the following recommendations were made for future work, the results are helpful for implementation of detecting diabetes disease on clinical or medical records. Diabetes mellitus has become a global challenge nowadays. However individuals can lessen its impact by regularly going for medical checkup and adhering to a healthy diet. Since machine learning is now a major force in the medical industry and a powerful area in artificial intelligence which uses complex algorithms to analyze large datasets, it has opened a lot of unheard possibilities for diagnosing and treating disease. It has become a catalyst in diabetic sector able to improve accuracy for earlier detection of diabetes and optimize health care management. We look further to enhance the binary butterfly optimization algorithm.

References

[1] Z. Ullah *et al.*, "Detecting High-Risk factors and early diagnosis of diabetes using machine learning methods," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–10, Sep. 2022, doi: 10.1155/2022/2557795. Available: https://doi.org/10.1155/2022/2557795.

[2] E. Hadziabdic and K. Hjelm, "Beliefs about illness: comparing foreign- and native-born persons with type 2 diabetes living in Sweden in a cross-sectional survey," *Primary Health Care Research & Development*, vol. 24, Jan. 2023, doi: 10.1017/s1463423623000269. Available: https://doi.org/10.1017/s1463423623000269.

[3] V. Chuter *et al.*, "Effectiveness of bedside investigations to diagnose peripheral artery disease among people with diabetes mellitus: A systematic review," *Diabetes/Metabolism Research and Reviews*, vol. 40, no. 3, Jul. 2023, doi: 10.1002/dmrr.3683. Available: https://doi.org/10.1002/dmrr.3683

[4] O. Iparraguirre-Villanueva, K. Espinola-Linares, R. O. F. Castañeda, and M. Cabanillas-Carbonell, “Application of machine learning models for early detection and accurate classification of Type 2 diabetes,” *Diagnostics*, vol. 13, no. 14, p. 2383, Jul. 2023, doi: 10.3390/diagnostics13142383. Available: https://doi.org/10.3390/diagnostics13142383

[5] Md. A. Uddin *et al.*, “Machine learning based diabetes Detection Model for false negative reduction,” *Deleted Journal*, vol. 2, no. 1, pp. 427–443, Jun. 2023, doi: 10.1007/s44174-023-00104-w. Available: https://doi.org/10.1007/s44174-023-00104-w

[6] A. A. Alhussan *et al.*, “Classification of diabetes using feature selection and hybrid Al-Biruni Earth radius and dipper throated optimization,” *Diagnostics*, vol. 13, no. 12, p. 2038, Jun. 2023, doi: 10.3390/diagnostics13122038. Available: https://doi.org/10.3390/diagnostics13122038

[7] Y. Han, D.-Y. Kim, J. Woo, and J. Kim, “Glu-Ensemble: An ensemble deep learning framework for blood glucose forecasting in type 2 diabetes patients,” *Heliyon*, vol. 10, no. 8, p. e29030, Apr. 2024, doi: 10.1016/j.heliyon.2024.e29030. Available: https://doi.org/10.1016/j.heliyon.2024.e29030

[8] L. Liu, B. Bi, L. Cao, M. Gui, and F. Ju, “Predictive model and risk analysis for peripheral vascular disease in type 2 diabetes mellitus patients using machine learning and shapley additive explanation,” *Frontiers in Endocrinology*, vol. 15, Feb. 2024, doi: 10.3389/fendo.2024.1320335. Available: https://doi.org/10.3389/fendo.2024.1320335

[9] A. A. Aouragh, M. Bahaj, and F. Toufik, “Diabetes Prediction: Optimization of Machine Learning through Feature Selection and Dimensionality Reduction,”

International Journal of Online and Biomedical Engineering (iJOE), vol. 20, no. 08, pp. 100–114, May 2024, doi: 10.3991/ijoe.v20i08.47765. Available: https://doi.org/10.3991/ijoe.v20i08.47765

[10] L. Kopitar, P. Kocbek, L. Cilar, A. Sheikh, and G. Stiglic, "Early detection of type 2 diabetes mellitus using machine learning-based prediction models," *Scientific Reports*, vol. 10, no. 1, Jul. 2020, doi: 10.1038/s41598-020-68771-z. Available: https://doi.org/10.1038/s41598-020-68771-z

[11] P. Theerthagiri, A. U. Ruby, and J. Vidya, "Diagnosis and classification of the diabetes using machine learning algorithms," *SN Computer Science*, vol. 4, no. 1, Nov. 2022, doi: 10.1007/s42979-022-01485-3. Available: https://doi.org/10.1007/s42979-022-01485-3

[10] L. Liu, B. Bi, L. Cao, M. Gui, and F. Ju, "Predictive model and risk analysis for peripheral vascular disease in type 2 diabetes mellitus patients using machine learning and shapley additive explanation," *Frontiers in Endocrinology*, vol. 15, Feb. 2024, doi: 10.3389/fendo.2024.1320335. Available: https://doi.org/10.3389/fendo.2024.1320335