Franklin Open 8 (2024) 100153

Contents lists available at ScienceDirect

Franklin Open

journal homepage: www.elsevier.com/locate/fraope

Optimizing machine learning algorithms for diabetes data: A metaheuristic approach to balancing and tuning classifiers parameters

Hauwau Abdulrahman Aliyu [a], Ibrahim Olawale Muritala [b,*], Habeeb Bello-Salau [b], Salisu Mohammed [c], Adeiza James Onumanyi [d], Ore-Ofe Ajayi [b]

[a] Department of Biochemistry and Molecular Biology, Federal University Birnin-Kebbi, 1157, Kebbi, Nigeria
[b] Department of Computer Engineering, Ahmadu Bello University, Zaria, 810107, Nigeria
[c] Department of Maintenance Engineering, KRPC Ltd, Kaduna Nigerian National Petroleum Company, Kaduna, 800242, Nigeria
[d] AIoT, Next Generation Enterprises and Institutions, Council for Scientific and Industrial Research (CSIR), Pretoria, 0001, South Africa

ARTICLE INFO

Keywords:
Bioinformatics
Biotechnology
Computational genomics
Machine learning
Metaheuristic algorithm

ABSTRACT

Diabetes mellitus poses a global health concern, prompting the development of machine learning algorithms designed to construct a model for the accurate classification of patients, enabling precise diagnoses and early-stage treatment. However, the efficacy of classifying diabetes patients through machine learning relies on datasets, often plagued by imbalance, leading to biased classification and inaccurate diagnoses. Previous research attempts, employing techniques like random sampling (under-sampling and oversampling) and the Synthetic Minority Oversampling Technique (SMOTE), have struggled to achieve optimally balanced datasets. Additionally, setting the best parameters for machine learning classifiers remains a challenging task. To address these issues, this research focuses on devising a methodological metaheuristic optimization, a machine learning algorithm tailored for diabetes data balancing, and classifier hyperparameter tuning. Leveraging Particle Swarm Optimization (PSO) algorithm for diabetes data balancing and a genetic algorithm to select the optimal architecture for various machine learning classifiers. The study compares the performance of the proposed metaheuristic data balancer and classifier architecture parameter tuner using classification metrics (F1 score, Average Precision–Recall (APR), AUC, and accuracy). The PSO balanced dataset emerges as the most effective in classifying diabetes, with an Average Percentage Improvement (API) in classification performance metrics: 20.78% accuracy, 16.79% area under the curve for receiver operating characteristics, and a significant 32.78% enhancement in APR. Moreover, the XGBOOST classifier trained with a genetic algorithm demonstrates minimal computational training time for the Centre for Disease Control and Prevention (CDC) diabetes dataset compared to the artificial neural network and random forest classifier. Notably, the imbalanced CDC diabetes dataset exhibits the least APR compared to random under-sampling and the PSO data balancing technique.

1. Introduction

Elevated sugar levels in the body contribute to the onset of diabetes mellitus [1], a condition broadly categorized into three types: Type 1 (insulin-dependent), Type 2 (insulin-independent), and Type 3 (gestational diabetes) [2,3]. Insulin-dependent diabetes arises when the pancreas, a vital gland, fails to produce the necessary insulin for the body's cells. Conversely, insulin-independent diabetes occurs when the body's cells produce insulin but do not effectively utilize it. Gestational diabetes, on the other hand, manifests during pregnancy [4]. Insulin plays a crucial role in facilitating the absorption of glucose by cells, converting it into energy. When cells fail to absorb glucose, leading to its accumulation in the body, the pancreas struggles to produce insulin [1]. Consequently, elevated blood sugar levels pose a health risk. Normoglycemia, representing the normal glucose level ranging from 70–99 mg per deciliter, contrasts sharply with the diabetic threshold of 126 mg per deciliter [1]. Untreated diabetes can result in severe complications such as nerve damage, stroke, and kidney failure [5]. Early detection is paramount for effective disease management [3]. prompting researchers to explore the application of machine learning algorithms. These algorithms, adept at learning from empirical data without explicit programming, hold promise for enhancing the timely identification and treatment of diabetes [6].

* Corresponding author.
E-mail addresses: hauwaumi002@gmail.com (H.A. Aliyu), drolawalemi@gmail.com (I.O. Muritala), bellosalau@abu.edu.ng (H. Bello-Salau), salis2k10@gmail.com (S. Mohammed), aonumanyi@csir.co.za (A.J. Onumanyi), aoreofe@abu.edu.ng (O.-O. Ajayi).

https://doi.org/10.1016/j.fraope.2024.100153
Received 14 December 2023; Received in revised form 21 June 2024; Accepted 22 August 2024

Nomenclature

Abbreviations

ANN	Artificial Neural Networks
API	Average Percentage Improvement
APR	Average-Precision Recall
AUC	Area Under Curve
CDC	Center for Disease Control and Prevention
FN	False Negative
FP	False Positive
GA	Genetic Algorithm
MLP	Multilayer Perceptron
PCA	Principal Component Analysis
PSO	Particle Swarm Optimization
Ran	Random Forest
RNA	Ribonucleic Acid
SMOTE	Synthetic Minority Oversampling Technique
TN	True Negative
TP	True Positive
XGB	eXtreme Gradient Boosting

Constants and variables

ρ	Percentage average for PSO
η	Percentage average for random sampling
$f(x)$	Objective function
V_i^{t+1}	Future particle velocity
x_i^{t+1}	Future particle position

Numerous machine learning models have been proposed by researchers to predict diabetes using diverse datasets, given the substantial reliance of machine learning on data [7–9]. However, medical data sets for diabetes classification problems, available in repositories, often exhibit imbalance, creating a bias towards the major class and affecting the accuracy of the developed machine learning model [10]. Existing conventional data balancing methods, including the synthetic minority oversampling technique, oversampling, and under-sampling, lack heuristics or fitness functions [11,12]. In contrast, this study introduces metaheuristic algorithms, problem-independent algorithms that employ fitness functions to guide their search for optimal solutions, specifically optimal datasets. Notably, machine learning classifying algorithms lack a standard procedure for selecting the best parameters, hindering their performance improvement based on data patterns and observed evidence [13,14]. In addressing this gap, the proposed distributed metaheuristic optimization algorithm emerges as a potential solution, enhancing robust performance on unseen datasets in diabetes classification problems. While previous research has delved into machine-learning-related medical analysis classification issues, the exploration of diabetes data balancing and classifier parameter tuning using metaheuristic algorithms remains insufficient.

An imbalance in cancer medical data sets at different cancer stages or between malignant and benign cases affects predictive model performance. kabir et al. [15], analyze the impact of dimensionality reduction techniques on machine learning models for cancer prediction using RNA sequencing data. PCA, kernel PCA, and autoencoder were used, with neural network and support vector machine classifiers trained and tested on the original and reduced data. The study found that dimensionality reduction improves classifier performance, with the autoencoder outperforming PCA and kernel PCA, emphasizing its potential for high-dimensional data analysis. The prevalence of rare events like heart attacks can lead to data imbalance issues, thus [16] aims to enhance personalized treatments for cardiovascular diseases (CVD) using artificial intelligence and machine learning on RNA-sequencing gene-expression data. The study generated and processed RNA-sequence data from the serum of consented CVD patients, applying visualizing genes with disease-causing variants for gene-disease annotation and expression analysis. They developed a Findable, Accessible, Intelligent, and Reproducible (FAIR) approach based on the Random Forest algorithm for biostatistical evaluation, successfully predicting associations between significant genes and demographic variables, underscoring the model's potential for improving personalized CVD treatments. Also, an imbalance between different genetic markers can influence the outcome of association studies. As such, [17] applies machine learning techniques to identify and classify COVID-19 infections using lung computerized tomography scans and a computer-aided diagnosis system. Decision Tree, SVM, K-means clustering, and RBF were used with clinical samples, involving screening, pre-processing, segmentation, and classification phases. The significance lies in offering a more efficient, accurate, and less labor-intensive method for early COVID-19 diagnosis, potentially improving patient outcomes and reducing virus exposure.

This research aims to highlight the significance of employing a distributed metaheuristic optimization-based machine learning algorithm for diabetes data balancing and classifier parameter tuning, emphasizing improvements in model accuracy, training efficiency, and computational training time for imbalanced data sets. Mumjudar & Vaidehi [18], focus on predicting diabetes using machine learning classifiers, highlighting logistic regression as the top-performing classifier in terms of classification metrics without pipelining. Furthermore, they found that when employing a pipeline for the control and automation of workflow, the AdaBoost classifier outperformed others in predicting diabetes, specifically using the Pima Indian diabetes dataset, though details on other datasets were not disclosed. Nicolucci et al. [19], concentrate on constructing prognostic models for diabetes complications based on electronic medical records, specifically for insulin-independent diabetes. Their supervised learning machine learning algorithm, based on 148 patient data collected over 15 years from 23 centers, identifies diabetic patients at a higher risk of complications. Ganie & Malik [20]], explore the prediction of insulin-independent diabetes using lifestyle indicators, employing an ensemble machine learning method along with a synthetic minority oversampling technique to address imbalanced datasets consisting of 1939 records and 11 biological/lifestyle parameters. The study identifies urination as a crucial feature in predicting insulin-independent diabetes, with the bagged decision tree classifier demonstrating superior performance compared to other machine learning classifiers in predicting insulin-independent diabetes with lifestyle indicators.

Cheheltani et al. [21], delve into the identification of insulin-dependent patients erroneously diagnosed as insulin-independent, utilizing the XGBoost classifier and the IQVIA database of electronic medical records. Their findings underscore therapy history, body mass index, age, and blood glucose values as key predictors of misdiagnosis. Jangili et al. [22], focus on predicting biomarkers associated with insulin-independent diabetes and coronary artery diseases. Using the SMOTE data balancing method and various machine learning classifiers, they employ an imbalanced dataset from Mediciti Hospital, Hyderabad, consisting of 123 insulin-independent individuals aged 35 to 70 years. The study reveals that the random forest classifier outperforms support vector machine, K-nearest neighbor, logistic regression, and decision tree classifiers in predicting biomarkers, considering metrics such as precision, area under the curve, F1 score, and recall. Bhat et al. [23], explore diabetes prediction guidelines and risk analysis, utilizing the SMOTE balancing method for the PIMA Indian diabetes datasets with three machine-learning classifiers. They identify blood pressure, glucose level, and diabetes pedigree function as major contributors to diabetes, with weight having the least impact. Among the classifiers, the decision tree exhibits superior performance with precision (96%), accuracy rate (91%), recall (92%), and F1-score (94%).

The PIMA diabetes dataset, encompassing 768 patients and their respective features, is relatively small. Some feature values within these datasets were unavailable, necessitating the use of mean values as replacements. In contrast, the Centers for Disease Control and Prevention (CDC) diabetes health indicator dataset is larger than the PIMA dataset. While larger datasets generally enhance accuracy in machine learning classification problems, careful consideration of trade-offs is imperative. These trade-offs encompass factors such as extended training time and increased computational complexity. Moreover, the presence of highly correlated features can lead to overfitting, characterized by the memorization of the training dataset and resulting in low bias but high variation between the training and testing datasets. To address this issue, this research leverages the relationship between metaheuristic data balancing for diabetes datasets and distributed optimization which combines the ability to efficiently handle large-scale data balancing tasks.

By utilizing distributed optimization (optimizing imbalanced data set and hyperparameter tuning of machine learning classifiers), metaheuristic algorithms can scale to larger datasets, providing effective solutions in a computationally efficient manner [24–26]. This synergy is particularly valuable in medical data analysis, where large and imbalanced datasets are common. Also, the Pearson correlation was employed to scrutinize highly correlated features, mitigating problems associated with overfitting and ensuring the robustness of the machine learning classification process. Additionally, this research utilizes the imbalanced diabetes health indicator dataset from the Centers for Disease Control and Prevention (CDC). With a participant pool of 253,680 and 21 features, this dataset aims to unravel the relationship between lifestyle and diabetes in the United States, employing a fixed split train–test methodology. The study introduces a novel approach, proposing a diabetes Particle Swarm Optimization (PSO) data balancer and Genetic Algorithm (GA) for machine learning classifier parameter tuning. The key contributions of this research include:

1. The introduction of an under-sampled PSO CDC diabetes dataset to enhance training efficiency and model accuracy in diabetes classification [27].
2. The application of GA-tuned hyperparameters to machine learning classifiers, specifically the Artificial Neural Network (multilayer perceptron neural model), XGBOOST (eXtreme Gradient Boost) classifier, and the Random Forest classifier, for the under-sampled PSO CDC diabetes dataset.
3. The demonstration that the PSO balanced dataset exhibits an API over the random under-sampling technique, showcases 20.78% accuracy, 16.79% area under the curve for receiver operating characteristics, and 32.78% APR in terms of classification performance metrics.

Considering these significant contributions, the study aims to underscore the pivotal role of metaheuristic algorithms in addressing imbalances within datasets and optimizing hyperparameters for classifiers, thereby elevating performance scores through insights derived from data patterns and empirical evidence. The subsequent sections of the study are structured as follows: Section 2 delves into materials and methods, Section 3 offers results and discussion, and Section 4 serves as the concluding segment of the paper.

2. Material and methods

This section details the materials and methods employed in the current research. Table 1 provides an itemized list of materials along with their specifications. The analysis utilized various modules imported into the Jupyter Notebook server version 6.4.8, including sklearn, xgboost, matplotlib, scipy, pandas, random, and seaborn. The framework of the method is illustrated in Fig. 1. For metaheuristic optimization, the particle swarm optimization algorithm was employed to balance the majority-classified (negative diabetes) and minority-classified (positive diabetes) data sets.

Table 1
Materials and specification.

Materials	Specification
Laptop computer	HP EliteBook 830 G5 Intel(R) Core (TM) i5- 8350U CPU @ 1.70 GHz, 1.90 GHz
Anaconda Navigator	Jupyter Notebook (Anaconda 3)
Data set	[28]

Table 2
PSO parameters.

Parameter	Value
Swarm Size	10
Inertia weight	0.5
Cognitive coefficient	0.9
Social coefficient	0.5

2.1. Data set

The CDC diabetes dataset encompasses lifestyle information from individuals with and without diabetes in the United States. This multivariate dataset comprises 253,680 patients and encompasses 21 features utilized for classifying labels indicating the presence or absence of diabetes [28]. Funded by the CDC, the creation of this dataset was facilitated. Derived from mobile phone and landline data submitted in 2014 for all 50 states, the District of Columbia, Guam, and Puerto Rico, the Collective Behavioral Risk Factor Surveillance System dataset is notably imbalanced, with 86.07% representing cases without diabetes and 13.93% representing instances of diabetes.

2.2. Particle swarm optimization

The particle swarm optimization, a swarm-based algorithm [12,29–33], was used as the chosen method for data balancing in this study. In the CDC diabetes dataset, the majority class corresponds to the negative diabetes class, constituting 86.07% of the dataset. Leveraging its strong exploration capabilities, the PSO algorithm optimizes, systematically exploring the negative diabetes instances within the CDC dataset. Subsequently, it identifies the optimal participant profile using Euclidean distance as the performance metric. The PSO metaheuristic algorithm, outlined in Algorithm 1, employs particles to represent participants recorded in the historical CDC diabetes dataset. The under-sampled CDC dataset is partitioned into training, validation, and test subsets. The update process for the algorithm's velocity is calculated using (1) and (2) sequentially.

$$x_i^{t+1} = x_i^t + v_i^{t+1} \tag{1}$$

$$v_i^{t+1} = w * v_i^t + c_1 * rand * (P_{best_i} - x_i^t) + c_2 * rand * (g_{best} - x_i^t) \tag{2}$$

where x_i^t is the position of the particle, v_i^t is the velocity of the particle, $w * v_i^t$ is the exploration ability of PSO, c_j is a weighing factor, rand is a random number between 0 and 1. The P_{best_i} is the P_{best_i} of the ith agent, and g_{best} is the best solution.

The effective tuning of the key parameters of PSO in Table 2 such as swarm size, inertia weight, cognitive and social coefficients, and velocity limits in Algorithm 1 is crucial for optimizing training process of machine learning models. This optimization directly impacts the model's balance between exploration and exploitation, convergence speed, and ultimately its accuracy and generalization ability during the testing phase with unknown data. A comprehensive and precise adjustment of these parameters lead to a significant enhancement in the performance of machine learning models.

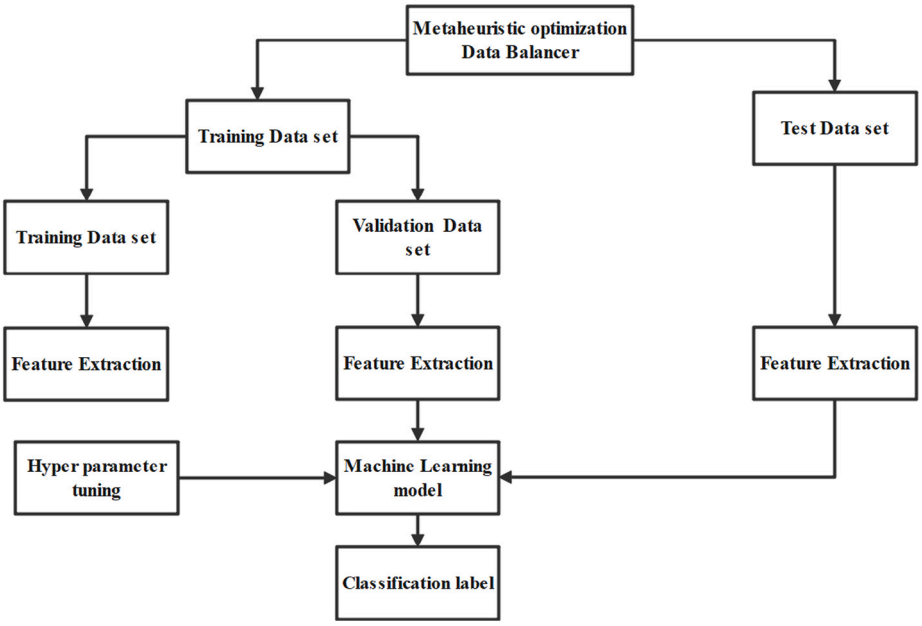

Fig. 1. The Framework of the Method.

Algorithm 1 PSO Algorithm.

1: **The participants' positions and velocities initialization**: The particles' positions, maximum and minimum velocities are initialized in n-dimensional search space.
2: **The evaluation of participant's fitness**: The objective function subject to constraints is evaluated.
3: **Assessment of the fitness of (personal best position):** Compare each of the particle's fitness with the fitness of the personal best position.
4: **if** the value at the current position < Best position **then**
5: Accept the current position
6: **else**
7: Use the Best position
8: **end if**
9: **Assessment of the fitness of (global best position):** Compare the fitness of the current position with the fitness of the population's overall best positions achieved before.
10: **if** the value at current space< the current position **then**
11: Reset the global best position as the current position and do the same for the fitness value
12: **end if**
13: **Updating each participant's velocity and position:** Calculate the update of the particle's position and velocity limit vector according to (1) and (2).
14: **Repeat the evolutionary cycle:** Return to Step 2 until a stopping criterion is satisfied.

2.3. Separation of the data sets

In this study, the balanced PSO CDC diabetes datasets underwent a fixed separation into training, test, and validation subsets. The training sets were employed to train the dataset parameters, while the test sets were utilized to examine the hyperparameters of the machine learning classifiers associated with the datasets. Following the completion of GA hyperparameter machine learning tuning, the validation sets were used to assess the model's performance. Specifically, the training, test, and validation subsets constituted 72%, 10%, and 18%, respectively, of the balanced 70,692 participants in the PSO CDC datasets. Given the susceptibility of balanced datasets to overfitting, especially in the presence of a high correlation between independent variables (features), steps were taken to mitigate this risk. Highly correlated features were excluded, and the remaining features were employed to address the classification problem. Fig. 2 illustrates the Pearson correlation map for the CDC diabetes dataset.

As observed in Fig. 2, it is evident that the differences in walk patterns, general health, and physical health features are highly correlated.[1] The physical health interpretation in the context of this study translates to physical illness and injury during the past month, general health means the state of general well-being on a scale of 1[2] to 5[3], and lastly, the differences in walk pattern translate to difficulty in walking or climbing stairs. This correlation results in prolonged training times, subsequently contributing to increased computational complexity. The forthcoming section elaborates on the proposed solution to address this issue.

2.4. Features extraction

A total of 21 features were employed for classifying positive or negative diabetes within the CDC diabetes dataset. Among these features, physical health, general health, and the difference in walking exhibited a high correlation. Consequently, two of these features were omitted to enhance computational efficiency and reduce training time. Additionally, this step led to a reduction in the dimensionality of the data. The resultant set of 19 features was then utilized to train the machine-learning model. Genetic Algorithm (GA) was employed to fine-tune the hyperparameters of the Artificial Neural Network (ANN), specifically the multilayer perceptron model, along with the XGBOOST classifier and the random forest classifier.

2.5. Genetic algorithm

The genetic algorithm, drawing inspiration from evolutionary principles [29,34] was employed to determine the optimal parameters (hyperparameters) for the classifiers, aiming to enhance performance

[1] Greater than or equal to 50%.
[2] Excellent.
[3] Poor.

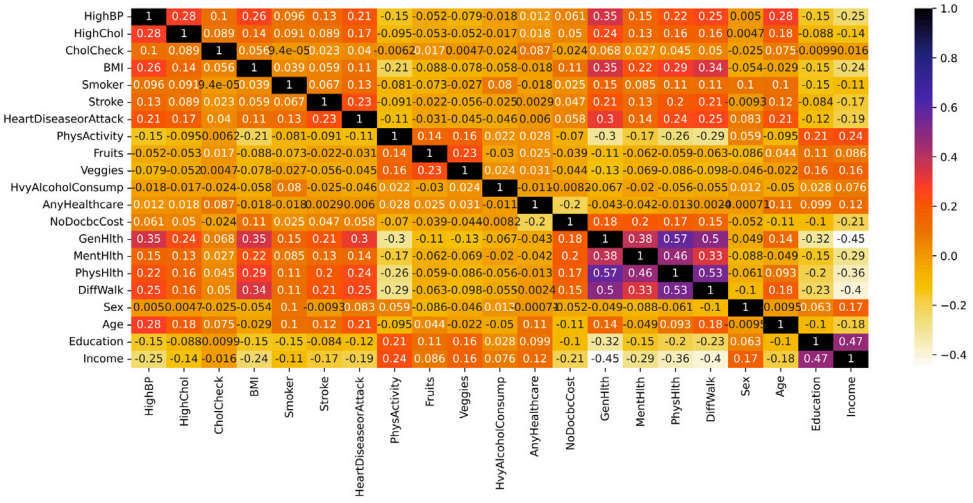

Fig. 2. Pearson Correlation Features Heatmap of the CDC Data Set.

scores based on discerned data patterns [35–37]. Due to the substantial size of the CDC diabetes dataset, traditional hyperparameter tuning techniques like grid and random search become time-consuming and intricate when selecting hyperparameters for classifiers. Moreover, the hyperparameter optimization problem is non-convex, meaning that local minima differ from the global optimum. Consequently, the GA algorithm, with its survival and adaptive features, is better positioned to persist and pass on these characteristics to future generations, facilitating the selection of the best parameters for machine learning classifiers. In the context of the GA hyperparameter problem, each chromosome symbolizes a hyperparameter and consists of multiple genes, each denoting either an active or inactive state. Crossover and mutation operations are then performed on these genes. The GA algorithm for hyperparameter tuning in machine learning classifiers is detailed in Algorithm 2.

Algorithm 2 GA Algorithm.

1: **The chromosome's initialization:** The chromosomes are initialized in n-dimensional search space.
2: **The evaluation of chromosome's fitness:** The objective function of the machine learning classifier is evaluated in (3), (4), and (5)
3: **Selection, crossover, and mutation:** The selection, crossover, and mutation of the genes from chromosomes that produce the next generation of the hyperparameters are evaluated.
4: **for** the number of epochs **do**
5: Repeat steps 2 and 3
6: **end for**
7: **Terminate**, and then output the best hyperparameters

2.6. Machine learning classifiers

This section introduces the GA-based machine learning classifiers employed for classifying the CDC diabetes datasets. The machine learning classifiers encompass the artificial neural network (multilayer perceptron neural model), eXtreme Gradient Boosting (XGBOOST), and Random Forest. Detailed hyperparameters for each classifier are outlined in Table 3.

2.6.1. Artificial neural network (multi-layer perceptron neural model)

The Artificial Neural Network (ANN) model draws inspiration from the functioning of the human brain and can utilize either the single-layer perceptron neural model or the multi-layer perceptron neural model. In the single-layer perceptron model, a single layer processes and learns data patterns, while the multi-layer perceptron employs multiple layers for more intricate data pattern processing and identification. This study chose the multi-layer perceptron neural model for its advantages in swift learning from extensive datasets compared to the single-layer perceptron. The ANN comprises three layers: the input layer, hidden layers, and output layer. The input layer receives the 20 features of the CDC diabetes dataset, which then activates the hidden layers through an activation function. The GA algorithm is responsible for selecting hyperparameters to optimize the activation (objective) function detailed in (3).

Table 3
Hyperparameters of the classifiers.

Machine learning classifiers	Hyperparameters
Artificial neural network (MLP)	alpha, hidden layers, learning rate, maximum iteration, number of iterations with no changes, and solver.
XGB	Number of estimators, maximum depth, learning rate, and colsample bytree.
Ran	Number of estimators, maximum depth, minimum samples split, and minimum samples leaf.

$$f(x) = g(\sum_{j=0}^{M} W_{ij}^{2} g(\sum_{i=0}^{d} W_{ji}^{1} x_i)) \tag{3}$$

where x_i are the 19 features, W_{ji}^{1} are the weights of the features, and W_{ij}^{2} are the weights of the hidden layers. The output is the classified label diabetes output.

2.6.2. XGBOOST

XGBOOST stands as an ensemble decision tree-based machine learning classifier. Conceived in March 2014 by Tianqi Chen [38], it was crafted in the $C++$ programming language to prioritize speed, parallel processing, and overall performance enhancement. In the XGBOOST classifier, the features of a new decision tree are interconnected with those of the preceding decision tree. The primary objective of the XGBOOST classifier is to optimize the objective function outlined in (4).

$$f(x) = \frac{-1}{2} \sum_{j=1}^{t} \frac{G_j^2}{H_j + \lambda} + \gamma t \tag{4}$$

where t is the number of leaves, γ, and λ are the penalty of coefficients, H, and G are first and second-order gradient statistical functions.

Table 4
Diabetes confusion matrix.

	Actual positive diabetes class	Actual negative diabetes class
Predicted positive diabetes class	True Positive (TP)	False Positive (FP)
Predicted negative diabetes class	False Negative (FN)	True Negative (TN)

2.6.3. Random forest

The random forest classifier, akin to XGBOOST, belongs to the category of ensemble tree-based machine learning classifier [39]. It employs multiple decision trees during the training phase, and the mode tree is then designated as the label output. Categorized under the bagging algorithm, this classifier's training process involves multiple models, with their combination enhancing the algorithm's overall performance and generalization. A distinctive characteristic of the random forest classifier is its absence of an explicitly formulated objective function. Instead, it relies on Gini impurity for classification problems, as defined in (5).

$$I(t) = 1 - \sum_{i=1}^{c} p(i|t)^2 \tag{5}$$

where $p(i|t)$ is the probability of training instances of class i at node t, c is the number of classes. A node's total impurity is the weighted sum of its impurities across all classes.

2.7. CDC diabetes classification metrics

This section presents the confusion matrix utilized for assessing the CDC diabetes classification. The specific details of the confusion matrix can be found in Table 4.

Table 4 defines the following classification metrics used in this research. The classification metrics are accuracy, F1-score, AUC, and the average-precision recall. Eq. (6) presents the accuracy,

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \tag{6}$$

From (6), the accuracy is defined as the ratio of the correctly classified diabetes samples to total diabetes samples [40,41]. Eq. (7) introduces the F1-score, which is calculated as the arithmetic mean of precision and recall,

$$F1 - score = \frac{2 * Precision * recall}{Precision + recall} \tag{7}$$

where

$$Precision = \frac{TP}{TP + FP}, \tag{8}$$

and

$$recall = \frac{TP}{TP + FN}, \tag{9}$$

The Area Under Curve (AUC) corresponds to the area under the receiver operating characteristics, providing a comprehensive assessment of the negative and positive diabetes classification performance [42]. A higher AUC value indicates superior performance of the machine learning classifier, signifying a high True Positive (TP) rate (recall) and a lower False Positive (FP) rate. Additionally, the APR metric is employed to evaluate the quality of the machine learning classifier output (labels), with particular significance in the context of imbalanced datasets such as the CDC dataset. Precision gauges the relevance of label classification by machine learning classifiers (negative or positive), while recall quantifies the number of relevant labels returned. Eq. (10) delineates the calculation for APR.

$$Averageprecision - -recall = \sum_{n} (R_n - R_{n-1}) P_n \tag{10}$$

where the precision, and recall nth are P_n, and R_n respectively. R_{n-1} is the $n-1th$ recall. Specificity is the actual negative diabetes class that is correctly classified. The equation of the specificity is in (11),

$$Specificity = \frac{TN}{TN + FP} \tag{11}$$

Fig. 3. The Imbalanced CDC Diabetes Data Set.

Table 5
Percentages of the negative and positive imbalanced and balanced CDC diabetes data set.

	Negative diabetes (%)	Positive diabetes (%)
Imbalanced	86.07	13.93
Random undersampling	50	50
PSO	50	50

3. Results and discussion

This section presents the findings and analysis of balancing the CDC diabetes data and tuning classifiers using a metaheuristic optimization-based machine learning algorithm.

3.1. The CDC's imbalanced and balanced data set

The CDC data set is imbalanced, Fig. 3 shows the pie chart representation of this data set.

From Fig. 3, it is evident that the CDC diabetes data set is highly imbalanced. The negative diabetes class has the majority data set with 86.07%, while the positive diabetes class has the remaining 13.93%. The imbalanced data set is biased towards the majority data set [10], consequently affecting the accuracy of the classification of diabetes. The machine learning classifiers (ANN, XGBOOST, and random forest) performance was examined on the imbalanced data set, random undersampling balanced data set, and PSO balanced data set. Table 5 shows the percentage of the negative and positive imbalanced and balanced CDC diabetes data set.

3.2. Performance evaluation of the balanced and imbalanced test data sets

The classification performance metrics were examined using accuracy, specificity, AUC, and the average-precision recall on the balanced and imbalanced test data sets. The balanced techniques were random undersampling and PSO balancing techniques. Figs. 4, 5, and 6 show the classification metrics of the imbalanced, and balanced (random undersampling and PSO) respectively.

In Fig. 4, the APR of the three machine learning classifiers has the lowest performance classification score, with MLP at 20.62%, XGB at 20.48%, and Ran 19.75%. On the other hand, specificity has the highest performance score, MLP has 97.99%, XGB has 97.89%, and

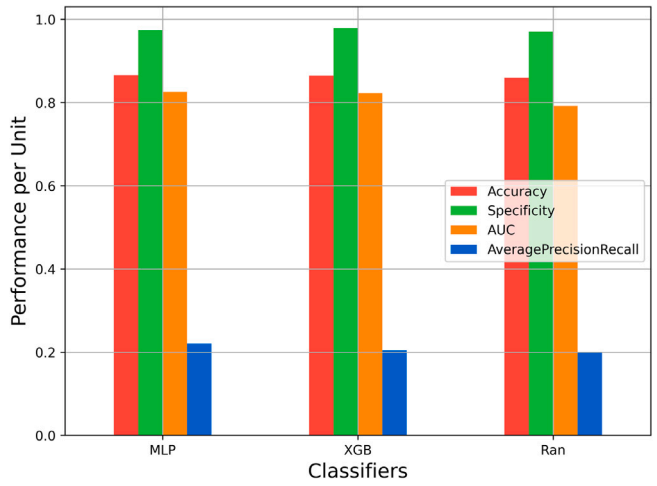

Fig. 4. Imbalanced CDC Diabetes Test Data Set.

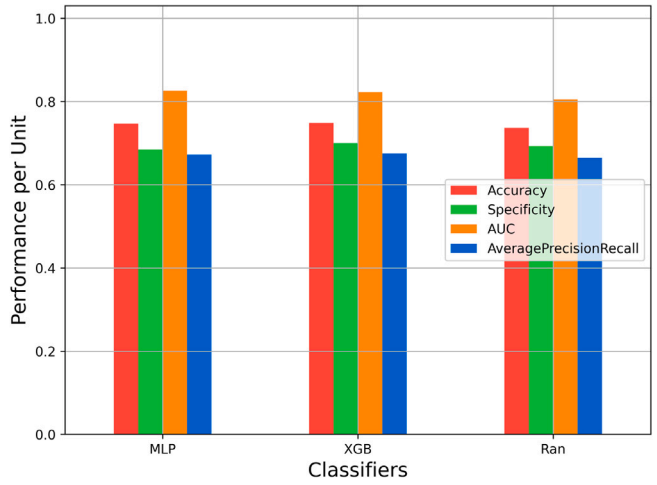

Fig. 5. Random Under-sampling of CDC Diabetes Test Data Set.

Ran has 97.02%. This is because the CDC's imbalanced data set is biased to the majority negative class. From Fig. 5, the random under-sampling enhances the APR compared to the imbalanced data set with MLP having 67.08%, XGB with 67.55%, and Ran with 61.17%. This is because the random under-sampling reduces the data set of the majority data to be equivalent to that of the minority class. Consequently, the area under curve receiver operating characteristics has the highest performance score with MLP having 82.10%, XGB has 82.30%, and Ran has 80.60%. It is worth noting that the random under-sampling technique samples the majority class of data sets without a fitness function. The PSO under-sampling technique used Euclidean distance as the fitness function in selecting the optimal data sets from the majority class. From Fig. 6, the PSO under-sampling technique improves the classification performance score compared to the random under-sampling technique. Out of the four performance classification scores, AUC has the best scores. MLP has 95.80%, XGB has 95.90%, and Ran has 94.90%. Figs. 7–9 show the area under the curve receiver operating characteristics of the CDC test data sets for the imbalanced, balanced data sets.

From Fig. 7, it can be noted that the MLP classifier has the best predictive performance for the imbalanced diabetes data set with 82.60% compared to the XGB classifier and Random Forest classifier which are 82.30% and 79.20% respectively.

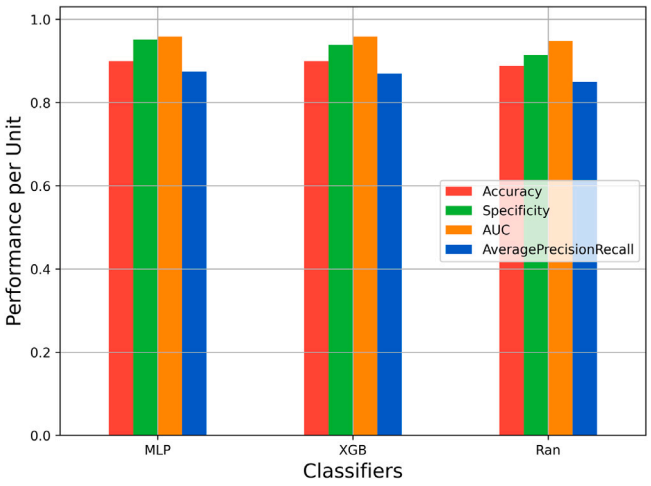

Fig. 6. PSO CDC Diabetes Test Data Set.

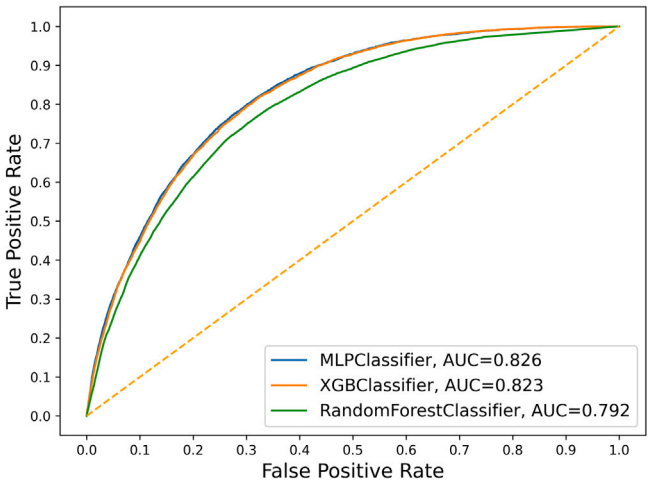

Fig. 7. Imbalanced CDC Diabetes Test Data Set AUC.

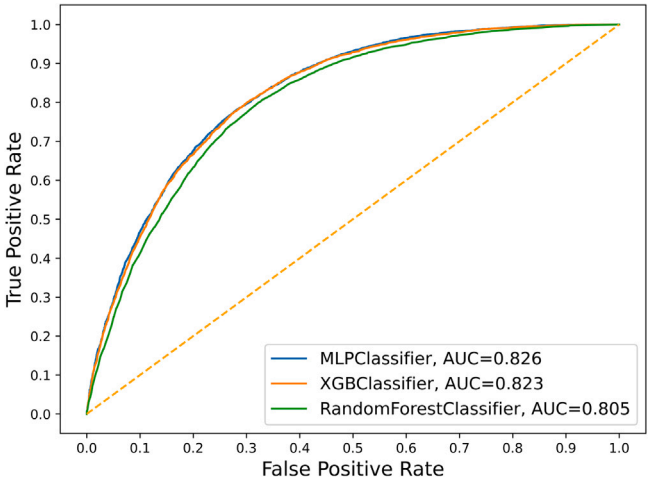

Fig. 8. Random Under-sampling of CDC Diabetes Test Data Set AUC.

Table 6
Results of the imbalanced and balanced CDC test data set classification metrics.

	Accuracy (%)			Specificity (%)			AUC (%)			APR (%)		
	MLP	XGB	Ran	MLP	XGB	Ran	MLP	XGB	Ran	MLP	XGB	Ran
Imbalanced data set	86.56	86.48	85.89	97.99	97.89	97.02	82.60	82.30	79.20	20.62	20.48	19.75
Random undersampling	74.60	74.86	73.35	67.27	70.02	69.11	82.60	82.30	80.50	67.08	67.55	61.17
PSO	89.95	90.05	89.11	95.07	93.96	91.79	95.90	95.90	94.80	87.47	87.13	85.40

Table 7
Average performance improvement.

	Accuracy (%)			API for accuracy	AUC (%)			API for AUC	APR (%)			API for APR
	MLP	XGB	Ran		MLP	XGB	Ran		MLP	XGB	Ran	
Random undersampling	74.60	74.86	73.35	20.78	82.60	82.30	80.50	16.79	67.08	67.55	61.17	32.78
PSO	89.95	90.05	89.11		95.90	95.90	94.80		87.94	87.13	85.40	

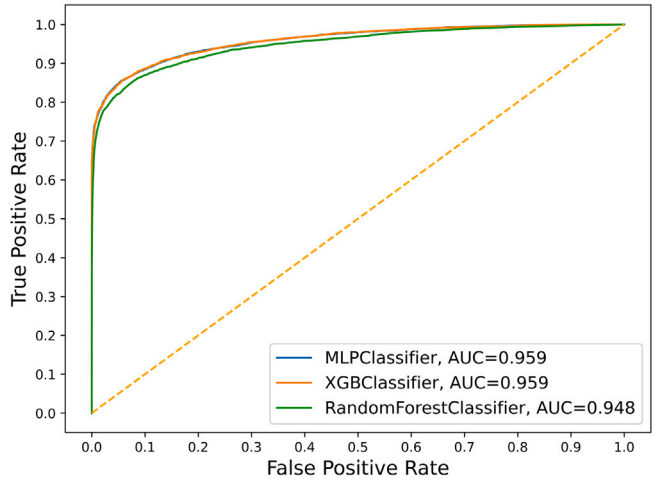

Fig. 9. PSO CDC Diabetes Test Data Set AUC.

Table 8
GA-trained machine learning tuned hyperparameter.

Machine learning classifiers	Hyperparameters
ANN(MLP)	Alpha = 1e−05, hidden layers (7,4), learning rate 0.09, maximum iteration = 1100, number of iterations with no changes = 80, and solver = stochastic gradient descent.
XGB	Number of estimators = 100, maximum depth =6, learning rate 0.1, and colsample bytree =0.7.
Ran	Number of estimators =200, maximum depth =100, minimum samples split =10, and minimum samples leaf =4,

In Fig. 8, the MLP classifier returned a higher AUC of 82.60% compared to 82.30% and 80.50% of the XGB, and Ran classifiers respectively for the random under-sampling balancing technique.

For the PSO balancing technique, the XGB and MLP classifier returned 95.90% and Ran returned 94.80%. Thus, the MLP performed better for both data balancing techniques. This is because of its enhanced processing capabilities. Table 6 gives comprehensive results for the imbalanced, and balanced data sets classification metrics from Figs. 4–9.

In Table 6, Particle Swarm Optimization (PSO) exhibits the best performance in terms of accuracy, achieving 89.95%, 90.05%, and 89.11% for MLP, XGB, and Random Forest (RF), respectively. This superiority stems from PSO's use of Euclidean distance in sampling CDC datasets, in contrast to random undersampling, which selects datasets randomly. The imbalanced datasets boast the highest specificity values of 97.99%, 97.89%, and 97.02% for MLP, XGB, and RF, respectively. This is

Table 9
GA parameters.

Parameter	Value
Number of iterations	10
Population Size	10
Mutation Probability	0.8

attributed to the prevalence of the negative class in the imbalanced CDC dataset. In terms of Area Under Curve Receiver Operating Characteristics (AUC–ROC), the PSO balanced dataset outperforms, achieving 95.90%, 95.90%, and 94.80% for MLP, XGB, and Ran, respectively. This can be linked to the creation of the best-undersampled dataset using Euclidean distance. APR for the imbalanced CDC dataset lags behind that of the random undersampling and PSO data balancing methods, recording 20.62% for MLP, 20.48% for XGB, and 19.75% for Ran. This disparity arises from the predominance of the negative diabetes class in the imbalanced CDC dataset, impacting the quality of the classified output. The PSO-balanced dataset exhibits higher APR than the random undersampling-balanced dataset. This is attributed to the naïveté of the random undersampling technique, while the PSO data balancing technique employs Euclidean distance as its fitness function for selecting optimal datasets. For the PSO balanced dataset, APR stands at 87.47% for MLP, 87.13% for XGB, and 85.40% for Ran (Table 7 showcases the Average Performance Improvement (API) between random undersampling and PSO data balancing techniques, derived from Table 6).

From Table 7, the API for the classification performance metric of the classifiers is calculated as follows in (12), and (13),

$$percentageAverage = \frac{MLP + XGB + Ran}{3} \tag{12}$$

where the average is the average of the 3-classification metric in Table 7, and the classification power metric in (12) is for both the random under sampling, and PSO technique. The percentage average from (12) is further processed to obtain (13),

$$API = \frac{\rho - \eta}{\eta} \tag{13}$$

where ρ is the percentage average for PSO, and η is the percentage average for random under-sampling.

3.3. PSO balanced data GA-trained machine learning classifiers

The genetically optimized hyperparameters for each classifier were selected to train the balanced PSO dataset. Table 8 showcases the optimal parameters selected by the genetic algorithm for training the classifiers, leading to improved classification scores as depicted in Fig. 10. The effective tuning of the key parameters of GA in Table 9 such as the number of iterations, population size, and mutation probability in Algorithm 2 is important in selecting the best architecture

Table 10
Results of the PSO balanced data GA-tuned classifiers classification metrics.

	Accuracy (%)	F1-score (%)	AUC (%)	APR (%)	Computational training time (hh:mm: ss. ms)
ANN(MLP)	89.37	89.35	95.71	86.98	1:50:41.601415
XGB	90.22	90.20	96.20	87.88	0:06:17.415970
Ran	91.77	91.76	97.70	89.91	2:32:38.268341

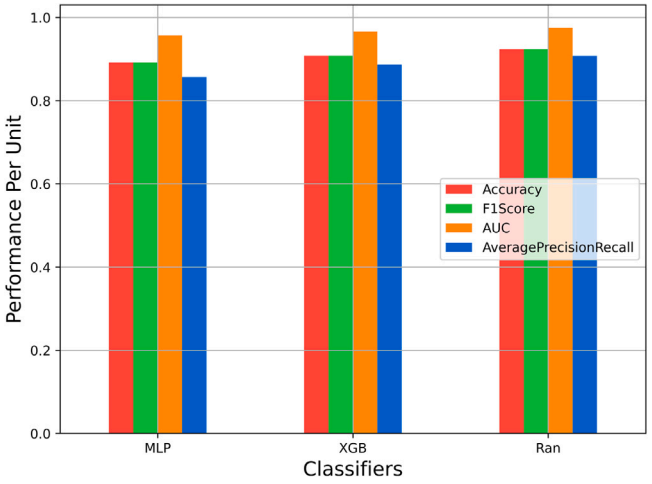

Fig. 10. PSO-GA trained CDC diabetes data set.

of the hyperparameters for the machine learning classifiers in Table 8. The mutation probability directly impacts the mutation diversity and prevents premature convergence of the classifier architecture.

The PSO-GA trained CDC diabetes data set in Fig. 10 shows an increase in classification performance metrics for the random forest classifier compared to MLP and XGB classifiers respectively. This is because the GA adapts better to the hyperparameters of random forest classifiers for the PSO CDC balanced data than MLP and XGB. The GA-trained random forest classifier has 91.77% accuracy, 91.76% for F1-score, 97.70% for AUC, and 89.91% for APR. For the MLP GA-trained classifier, accuracy was 89.37%, 89.35% for F1-score, 95.71% for AUC, and 86.98% for APR. The accuracy of 90.22%, 90.20% for F1-score, 96.20%, and 87.58% was returned for the XGB GA trained classifier. The comprehensive results of the CDC datasets trained with the PSO-GA, including the computational time for each classifier, are presented in Table 10. In Table 10, it is observed that the GA-tuned XGBOOST classifier on the PSO-balanced dataset exhibits significantly lower computational time compared to MLP and the random forest classifier. The GA completed the training of the PSO balanced dataset in 6 min, contrasting with MLP and RF, which took 1 h 50 min and 2 h 32 min, respectively. This outcome aligns with expectations, as XGBOOST is designed for speed, parallel processing, and enhanced performance. Additionally, the GA-tuned classifiers on the PSO-balanced datasets demonstrate no overfitting to the test data, as evidenced by the minimal variation between values in Fig. 10.

4. Conclusion

In conclusion, this study delved into the realm of diabetes data balancing and classifier parameter tuning through the utilization of a metaheuristic optimization-based machine learning algorithm. Metaheuristic algorithms are designed to avoid local optima, but they do not guarantee to find the global optimum, especially in highly complex or multimodal datasets. Effective optimization also relies on the assumption that the initial parameter settings of these algorithms are chosen appropriately. The outcomes of this investigation underscore the effectiveness of the PSO data balancing GA-tuned machine learning classifier hyperparameters technique, showcasing notable enhancements in both training efficiency and model accuracy. A comparative analysis against the random under-sampling data balancing technique revealed the superiority of the PSO data balancing approach, exhibiting an API of 20.78% for accuracy, 16.79% for area under curve receiver operating characteristics, and 32.78% for APR in CDC diabetes classification. Furthermore, during the data training phase, the XGBOOST classifier demonstrated superior efficiency by training the balanced dataset in significantly less time than its counterparts, namely ANN (MLP) and the random forest classifier. Notably, the imbalanced CDC dataset exhibited the least APR. This research underscores the significance of leveraging metaheuristic data balancing and classifier parameter tuning optimization in machine learning algorithms. It is recommended that future research endeavors explore hybrid metaheuristic data balancing techniques and extend the application to domains such as Natural Language Processing, particularly in areas like Nigerian local language translation.

Funding

The authors would like to thank Tertiary Education Trust Fund (TETFUND), Nigeria, for funding this research as part of the 2021 National Research Fund (NRF) grant cycle, under the project titled “A Novel Artificial Intelligence of Things (AIoT) Based Assistive Smart Glass Tracking System for the Visually Impaired” (TETF/DR&D/CE/NRF2021/SETI/ICT/00140/VOL.1).

CRediT authorship contribution statement

Hauwau Abdulrahman Aliyu: Conceptualization, Methodology, Software, Writing – original draft. **Ibrahim Olawale Muritala:** Conceptualization, Methodology, Software, Writing – original draft, Writing – review & editing. **Habeeb Bello-Salau:** Validation, Writing – review & editing. **Salisu Mohammed:** Writing – review & editing. **Adeiza James Onumanyi:** Methodology, Writing – review & editing. **Ore-Ofe Ajayi:** Methodology, Writing – original draft.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that support the findings of this study are available in Open Science Framework HOME at https://osf.io/9jkqm.

References

[1] J.J. Khanam, S.Y. Foo, A comparison of machine learning algorithms for diabetes prediction, ICT Express 7 (2021) 432–439, http://dx.doi.org/10.1016/J.ICTE.2021.02.004.

[2] D. Parkhi, N. Periyathambi, Y.G. Weldeselassie, V. Patel, N. Sukumar, R. Siddharthan, L. Narlikar, P. Saravanan, Prediction of postpartum prediabetes by machine learning methods in women with gestational diabetes mellitus, iScience 26 (2023) 107846, http://dx.doi.org/10.1016/J.ISCI.2023.107846.

[3] M.R. Islam, S. Banik, K.N. Rahman, M.M. Rahman, A comparative approach to alleviating the prevalence of diabetes mellitus using machine learning, Comput. Methods Programs Biomed. Update 4 (2023) 100113, http://dx.doi.org/10.1016/J.CMPBUP.2023.100113.

[4] Y. Belsti, L. Moran, L. Du, A. Mousa, K.D. Silva, J. Enticott, H. Teede, Comparison of machine learning and conventional logistic regression-based prediction models for gestational diabetes in an ethnically diverse population; the Monash GDM Machine learning model, Int. J. Med. Inform. 179 (2023) 105228, http://dx.doi.org/10.1016/J.IJMEDINF.2023.105228.

[5] M. khan, B.K. Singh, N. Nirala, Expert diagnostic system for detection of hypertension and diabetes mellitus using discrete wavelet decomposition of photoplethysmogram signal and machine learning technique, Med. Nov. Technol. Devices 19 (2023) 100251, http://dx.doi.org/10.1016/J.MEDNTD.2023.100251.

[6] C.B. Giorda, A. Rossi, F. Baccetti, R. Zilich, F. Romeo, N. Besmir, G. Di Cianni, G. Guaita, L. Morviducci, M. Muselli, A. Ozzello, F. Pisani, P. Ponzani, P. Santin, D. Verda, N. Musacchio, Achieving good metabolic control without weight gain with the systematic use of GLP-1-RAs and SGLT-2 inhibitors in type 2 diabetes: A machine-learning projection using data from clinical practice, Clin. Ther. 45 (2023) 754–761, http://dx.doi.org/10.1016/J.CLINTHERA.2023.06.006.

[7] M. Allwright, J.F. Karrasch, J.A. O'Brien, B. Guennewig, P.J. Austin, Machine learning analysis of the UK Biobank reveals prognostic and diagnostic immune biomarkers for polyneuropathy and neuropathic pain in diabetes, Diabetes Res. Clin. Pract. 201 (2023) 110725, http://dx.doi.org/10.1016/J.DIABRES.2023.110725.

[8] B. Li, S. Riaz, Y. Zhao, Experimental validation of iterative learning control for DC/DC power converters, Energies 16 (18) (2023) 6555.

[9] S. Riaz, H. Lin, M.P. Akhter, Design and implementation of an accelerated error convergence criterion for norm optimal iterative learning controller, Electronics 9 (11) (2020) 1766.

[10] A.S. Desuky, S. Hussain, An improved hybrid approach for handling class imbalance problem, Arab. J. Sci. Eng. 46 (2021) 3853–3864, http://dx.doi.org/10.1007/s13369-021-05347-7.

[11] F. Asnicar, A.M. Thomas, A. Passerini, L. Waldron, N. Segata, Machine learning for microbiologists, Nat. Rev. Microbiol. (2023) http://dx.doi.org/10.1038/s41579-023-00984-1, URL: https://www.nature.com/articles/s41579-023-00984-1.

[12] Z. Mustaffa, M.H. Sulaiman, Enhancing battery state of charge estimation through hybrid integration of barnacles mating optimizer with deep learning, Frankl. Open (2023) 100053, http://dx.doi.org/10.1016/j.fraope.2023.100053, URL: https://linkinghub.elsevier.com/retrieve/pii/S2773186323000476.

[13] I.O. Muritala, M.B. Mu'azu, A.T. Salawudeen, I.J. Umoh, H. Bello-Salau, Z. Haruna, A moving-horizon estimator of the unmeasured states for fuel sloshing dynamics in close proximity operation, in: 2023 3rd International Conference on Computing and Information Technology, ICCIT, IEEE, 2023, pp. 78–83.

[14] I.O. Muritala, M.B. Mu'azu, A.T. Salawudeen, I.J. Umoh, H. Bello-Salau, Z. Haruna, S. Mohammed, Moving horizon estimator for space vehicle dynamics with measurement noise in close propinquity operation, Frankl. Open 6 (2024) 100070.

[15] M.F. Kabir, T. Chen, S.A. Ludwig, A performance analysis of dimensionality reduction algorithms in machine learning models for cancer prediction, Healthc. Anal. 3 (2023) 100125.

[16] V. Venkat, H. Abdelhalim, W. DeGroat, S. Zeeshan, Z. Ahmed, Investigating genes associated with heart failure, atrial fibrillation, and other cardiovascular diseases, and predicting disease using machine learning techniques for translational research and precision medicine, Genomics 115 (2) (2023) 110584.

[17] M.T. Ahemad, M.A. Hameed, R. Vankdothu, COVID-19 detection and classification for machine learning methods using human genomic data, Meas.: Sens. 24 (2022) 100537.

[18] A. Mujumdar, V. Vaidehi, Diabetes prediction using machine learning algorithms, Procedia Comput. Sci. 165 (2019) 292–299, http://dx.doi.org/10.1016/J.PROCS.2020.01.047.

[19] A. Nicolucci, L. Romeo, M. Bernardini, M. Vespasiani, M.C. Rossi, M. Petrelli, A. Ceriello, P. Di Bartolo, E. Frontoni, G. Vespasiani, Prediction of complications of type 2 Diabetes: A machine learning approach, Diabetes Res. Clin. Pract. 190 (2022) 110013, http://dx.doi.org/10.1016/J.DIABRES.2022.110013.

[20] S.M. Ganie, M.B. Malik, An ensemble machine learning approach for predicting Type-II diabetes mellitus based on lifestyle indicators, Healthc. Anal. 2 (2022) 100092, http://dx.doi.org/10.1016/J.HEALTH.2022.100092.

[21] R. Cheheltani, N. King, S. Lee, B. North, D. Kovarik, C. Evans-Molina, N. Leavitt, S. Dutta, Predicting misdiagnosed adult-onset type 1 diabetes using machine learning, Diabetes Res. Clin. Pract. 191 (2022) 110029, http://dx.doi.org/10.1016/J.DIABRES.2022.110029.

[22] S. Jangili, H. Vavilala, G.S.B. Boddeda, S.M. Upadhyayula, R. Adela, S.R. Mutheneni, Machine learning-driven early biomarker prediction for type 2 diabetes mellitus associated coronary artery diseases, Clin. Epidemiol. Glob. Health 24 (2023) 101433, http://dx.doi.org/10.1016/J.CEGH.2023.101433.

[23] S.S. Bhat, M. Banu, G.A. Ansari, V. Selvam, A risk assessment and prediction framework for diabetes mellitus using machine learning algorithms, Healthc. Anal. 4 (2023) 100273, http://dx.doi.org/10.1016/J.HEALTH.2023.100273, URL: https://linkinghub.elsevier.com/retrieve/pii/S2772442523001405.

[24] Q. Hou, J. Dong, Distributed dynamic event-triggered consensus control for multiagent systems with guaranteed L_2 performance and positive inter-event times, IEEE Trans. Autom. Sci. Eng. (2022).

[25] Q. Hou, J. Dong, Robust adaptive event-triggered fault-tolerant consensus control of multiagent systems with a positive minimum interevent time, IEEE Trans. Syst. Man Cybern.: Syst. (2023).

[26] I.O. Muritala, H. Bello-Salau, A.T. Salawudeen, S. Mohammed, The effect of an event-triggered controller on non-strict feedback system based on the single input interval type-2 fuzzy method, Frankl. Open 4 (2023) 100035, http://dx.doi.org/10.1016/j.fraope.2023.100035, URL: https://www.sciencedirect.com/science/article/pii/S2773186323000294.

[27] H.A. Abdulrahman, I.M. Olawale, S. Habeeb-Bello, S. Mohammed, A.J. Onumanyi, O.-O. Ajayi, Centers for disease control and prevention particle swarm optimization data set for diabetes classification, Open Sci. Framew. (2023) http://dx.doi.org/10.17605/OSF.IO/9JKQM, URL: https://osf.io/9jkqm/.

[28] CDC, Diabetes health indicators dataset, 2015, Kaggle, URL: https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset.

[29] J. Kennedy, R. Eberhart, Particle swarm optimization, in: Proceedings of ICNN'95 - International Conference on Neural Networks, Vol. 4, IEEE, 1995, pp. 1942–1948 vol.4, http://dx.doi.org/10.1109/ICNN.1995.488968.

[30] O.O. Akinola, A.E. Ezugwu, J.O. Agushaka, R.A. Zitar, L. Abualigah, Multiclass feature selection with metaheuristic optimization algorithms: a review, Neural Comput. Appl. 34 (2022) 19751–19790, http://dx.doi.org/10.1007/s00521-022-07705-4.

[31] Z. Sun, G. An, Y. Yang, Y. Liu, Optimized machine learning enabled intrusion detection 2 system for internet of medical things, Frankl. Open (2023) 100056, http://dx.doi.org/10.1016/j.fraope.2023.100056, URL: https://www.sciencedirect.com/science/article/pii/S2773186323000506.

[32] S. Mohammed, A.S. Yusuf, U. Ime, S. Ahmed-Tijjani, O.M. Ibrahim, A study of hybridized smell agent symbiotic organism search in congress on evolutionary computation functions, SLU J. Sci. Technol. (2023) 44–58, http://dx.doi.org/10.56471/slujst.v6i.350.

[33] A.T. Salawudeen, M.B. Mu'azu, A. Yusuf, A.E. Adedokun, A Novel Smell Agent Optimization (SAO): An extensive CEC study and engineering application, Knowl.-Based Syst. 232 (2021) 107486.

[34] D.J. Hemanth, J. Anitha, Modified genetic algorithm approaches for classification of abnormal magnetic resonance brain tumour images, Appl. Soft Comput. 75 (2019) 21–28, http://dx.doi.org/10.1016/j.asoc.2018.10.054.

[35] H. Bello-Salau, A.M. Aibinu, Z. Wang, A.J. Onumanyi, E.N. Onwuka, J.J. Dukiya, An optimized routing algorithm for vehicle ad-hoc networks, Eng. Sci. Technol., Int. J. 22 (2019) 754–766, http://dx.doi.org/10.1016/J.JESTCH.2019.01.016.

[36] I.D. Raji, H. Bello-Salau, I.J. Umoh, A.J. Onumanyi, M.A. Adegboye, A.T. Salawudeen, Simple deterministic selection-based genetic algorithm for hyperparameter tuning of machine learning models, Appl. Sci. 12 (2022) 1186, http://dx.doi.org/10.3390/app12031186.

[37] A.T. Sulaiman, H. Bello-Salau, A.J. Onumanyi, M.B. Mu'azu, E.A. Adedokun, A.T. Salawudeen, A.D. Adekale, A particle swarm and smell agent-based hybrid algorithm for enhanced optimization, Algorithms 17 (2) (2024) 53.

[38] T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, in: ACM SIGKDD, ACM, New York, NY, USA, 2016, pp. 785–794, http://dx.doi.org/10.1145/2939672.2939785,

[39] O.R. Olaniran, M.A.A. Abdullah, Bayesian weighted random forest for classification of high-dimensional genomics data, Kuwait J. Sci. 50 (2023) 477–484, http://dx.doi.org/10.1016/J.KJS.2023.06.008.

[40] H. Bello-Salau, A.J. Onumanyi, R.F. Adebiyi, E.A. Adedokun, G.P. Hancke, Performance analysis of machine learning classifiers for pothole road anomaly segmentation, in: IEEE 30th International Symposium, IEEE, 2021, pp. 1–6, http://dx.doi.org/10.1109/ISIE45552.2021.9576214,

[41] T.O. Omotehinwa, D.O. Oyewola, E.G. Dada, A light gradient-boosting machine algorithm with tree-structured parzen estimator for breast cancer diagnosis, Healthc. Anal. 4 (2023) 100218, http://dx.doi.org/10.1016/j.health.2023.100218.

[42] N.A. Azit, S. Sahran, V.M. Leow, M. Subramaniam, S. Mokhtar, A.M. Nawi, Prediction of hepatocellular carcinoma risk in patients with type-2 diabetes using supervised machine learning classification model, Heliyon 8 (2022) e10772, http://dx.doi.org/10.1016/J.HELIYON.2022.E10772.