Published by
NIGERIAN SOCIETY OF PHYSICAL SCIENCES
Available online @ https://journal.nsps.org.ng/index.php/jnsps

J. Nig. Soc. Phys. Sci. 6 (2024) 2175

Journal of the Nigerian Society of Physical Sciences

Ensemble machine learning algorithm for cost-effective and timely detection of diabetes in Maiduguri, Borno State

Emmanuel Gbenga Dada [a,b,*], Aishatu Ibrahim Birma [a], Abdulkarim Abbas Gora [a]

[a] Department of Mathematics and Computer Science, Faculty of Science, Borno State University, Maiduguri
[b] Department of Computer Science, Faculty of Physical Sciences, University of Maiduguri, Maiduguri, Nigeria

Abstract

Diabetes is a serious medical condition that severely hinders the body's ability to produce or properly regulate insulin, leading to detrimental carbohydrate metabolism and dangerously high blood sugar levels. This ultimately causes inadequate carbohydrate metabolism and heightened blood glucose levels. Alarmingly, from 2000 to 2019, diabetes-related mortality rates rose by 3%. In the year 2019 alone, diabetes was tragically responsible for nearly 2 million deaths. This groundbreaking research introduces the improved weighted average ensemble learning (WAEL) model as an innovative solution for detecting diabetes. The enhanced WAEL model effectively addresses the overfitting challenge by integrating multiple models that have gained unique insights from the data. The proposed WAEL model ingeniously combines five feature spaces through the grey wolf optimisation (GWO) algorithm to uncover the optimal weight combination. GWO plays a vital role in weight optimization, enabling the reduction of weights in models that are particularly sensitive to noise. The results demonstrated that the improved WAEL achieved an astounding level of accuracy, soaring to 98.90%. The LGBM algorithm followed closely, achieving an impressive accuracy of 85.00%. The RF method recorded an accuracy of 81.00%. When it comes to accurately identifying diabetes, the improved WAEL ensemble model significantly outperformed the other five individual models, as evidenced by metrics such as accuracy, precision, recall, and F1-score. Therefore, the proposed model stands as a compelling alternative tool for healthcare professionals in the early detection of diabetes.

DOI:10.46481/jnsps.2024.2175

Keywords: Ensemble learning, Diabetes, Weighted average ensemble, Random forests, Light gradient boosting machine

Article History :
Received: 4 June 2024
Received in revised form: 07 September 2024
Accepted for publication: 10 September 2024
Published: 24 September 2024

Communicated by: Oluwatobi Akande

1. Introduction

Diabetes is an alarmingly frequent chronic condition around the world, necessitating patients to actively involve themselves in self-management strategies to effectively battle this serious health issue [1, 2]. This condition is marked by hyperglycemia, which signifies dangerously increased sugar levels in the bloodstream [3]. The World Health Organization (WHO) defines diabetes as a chronic disorder stemming from inadequate insulin production by the pancreas [2]. Furthermore, it can occur when the body fails to utilize the insulin it produces efficiently. Insulin plays a crucial role as a hormone that regulates blood glucose levels. Uncontrolled diabetes frequently results in hyperglycemia, commonly known as high blood sugar, which can inflict severe damage on various bodily systems, particularly the nerves and arteries [4]. While genetic factors are the pri-

*Corresponding author: Tel.: +234-706-577-0567.
Email address: gbengadada@unimaidedu.ng (Emmanuel Gbenga Dada)

mary contributors to this condition, environmental influences also significantly impact its development [5].

Diabetes is divided into four distinct categories: Type 1 diabetes, accounting for approximately 5-10% of all diagnosed cases, is often referred to as adolescent diabetes or insulin-dependent diabetes. Type 2 diabetes becomes recognizable by its symptoms and the lack of insulin dependence, typically manifesting later in life. Type 1 diabetes, commonly known as juvenile diabetes, usually presents before the age of 20 [6]. This condition is driven by an autoimmune disorder where the immune system attacks its own tissues, leading to the destruction of pancreatic cells that produce insulin. In contrast, type 2 diabetes generally arises after age 30, and is sometimes dubbed ”old-age diabetes”; however, younger individuals can also be at risk. The onset of Type 2 diabetes is significantly influenced by genetic predispositions, obesity, and inadequate cardiovascular activity [7].

Managing diabetes is an urgent public health challenge, and its prevalence is escalating at an alarming rate. This potentially life-threatening and debilitating disease is increasingly pervasive, particularly in developing countries and among impoverished communities with low socioeconomic status [8, 9]. The global incidence of diabetes has been on a steady rise and is anticipated to continue its upward trajectory in the coming years. The rate of diabetes among Africans is surging, with projections indicating that the African continent will experience the highest growth rate of 143% in diabetes cases from 2019 to 2045.

In 2019, nearly 2 million lives were tragically cut short due to diabetes and the renal diseases it brings. This staggering loss is largely attributed to the lack of accessible healthcare facilities for these vulnerable individuals. Many of those who desperately need these services are confined to remote villages, far from the help they require. Moreover, the limited number of healthcare facilities creates an urgent and ongoing need for highly skilled medical professionals to tackle critical health issues effectively. Adeleye [10] underscores a troubling and persistent rise in diabetes mellitus (DM) incidences throughout Nigeria. Research has shown that the prevalence of DM in certain villages in Nigeria varies between 0.8% and 4.4%, as reported by Refs. [11–13]. In urban centers, the prevalence is even higher, ranging from 4.6% to 7%, according to findings by Refs. [13, 14]. A systematic review and meta-analysis by Uloko *et al.* [15] has determined that the diabetes mellitus frequency among Nigerians stands at 5.77%. Alarmingly, Tinajero and Malik [8] projected that the number of Nigerians with impaired sugar tolerance was 8.2 million in 2019 and could surge to 11.5 million by 2030. Additionally, Gezawa *et al.* [16] have demonstrated a significant presence of type 2 diabetes in the Maiduguri metropolitan area. Given this critical information, it is essential to develop a computerized system that assists doctors in providing medical care, especially for diagnosing diabetes. This innovative approach would be particularly beneficial in regions where healthcare services and facilities are scarce. Furthermore, it would serve as a vital resource in areas lacking skilled medical professionals or clinical decision support (CDS) systems, as well as in electronic diabetes diagnosis systems. This work presents the following contributions:

1. Development of an innovative weighted average ensemble learning (WAEL) model for diabetes detection has established an unparalleled standard in performance on an extensive dataset of diabetes patients. This groundbreaking model has not only surpassed but has also redefined the capabilities of other ensemble learning models, including random forest (RF), adaptive boosting regression (Adaboost), gradient boosting regression (GBOOST), light gradient boosting machine (LGBM), and CatBoosting ensemble.
2. A new dataset of diabetes patients was collected from 2018 to 2023 from University of Maiduguri teaching hospital, and Umaru Shehu specialist hospital in Maiduguri, Borno State Nigeria. The dataset has a record of 1030 patients.
3. The grey wolf optimisation (GWO) algorithm and weighted average ensemble learning technique was used to diagnose diabetes. The improved WAEL that is proposed in this paper have overcome the limitation associated with the traditional WAEL model as the grey wolf optimisation (GWO) that was used to dynamically optimise the weights assigned to each model thereby finding the best combination that minimizes detection error.
4. The present study involved a comparative analysis of various ensemble techniques, namely RF, Adaboost, GBOOST, LGBM, and CatBoosting, in terms of their performance. The improved WAEL model demonstrates better results compared to all other techniques in terms of accuracy, precision, recall, and F1-score.
5. A comprehensive examination was conducted on the outputs. The information presented in this paper offers valuable insights on the advantages and drawbacks of the proposed model. This work also provides valuable insights into some underlining causes of diabetes.

The current research presents a new method for promptly detecting diabetes by analysing patient data from people living in Maiduguri and its neighbouring regions in Nigeria. The findings suggest that the improved weighted average ensemble learning (WAEL) model exhibits superior performance compared to alternative ensemble approaches while also offering vital insights into the significant elements contributing to the development of diabetes. We anticipate using the recently obtained dataset over time to train various algorithms aimed at diagnosing and predicting diabetes. This can serve as a powerful tool for decision-making in the identification and treatment of diabetes.

2. Related works

In their study, Sarwar *et al.* [17] developed a hybrid ensemble model that uses machine learning techniques to effectively identify cases of type 2 diabetes. The authors employed a number of machine learning classifiers. The study's database comprises approximately 400 individuals selected from a diverse geographical area, encompassing ten distinct physiological characteristics. The outcomes of the simulation demonstrated that the ensemble approach had superior performance, with an average accuracy of 98.60%, surpassing the performance of the other models employed in the paper. The downside of this research lies in the comparatively limited size of the dataset used. Furthermore, there is a need to enhance the performance of the system. Alasaady *et al.* [18] used an adaptive neurofuzzy (ANFIS) technique for the purpose of diagnosing diabetes. Statistical findings indicate that the model's performance is relatively good. A potential weakness of the research lies in the comparatively limited size of the dataset used, thus impeding the extent to which the findings of the study can be applied to a broader population. Additionally, the focus of the analysis was mostly on diagnostic procedures rather than predictive modelling. Another limitation is that the performance of the system is poor. In their study, Abdulhadi and Al-Mousa [19] employed machine learning algorithms to forecast the probable occurrence of diabetes, with a particular focus on early detection among female individuals. Out of all the models under consideration, the random forest model yielded the most favourable outcome, achieving an accuracy rate of 82%. The drawback of the work is that the size of the dataset used is small. Moreover, the proposed system has a significant deficiency in terms of accuracy.

Laila *et al.* [20] applied ensemble machine learning models to conduct a scientific investigation on diabetes. The dataset used for the study was acquired from the UCI repository. The diabetes dataset comprises a total of 17 variables. Machine learning algorithms, namely AdaBoost, Bagging, and RF, were used for the prediction. Various performance indicators were used to evaluate the effectiveness of the framework. One potential shortcoming of the study is the relatively limited size of the dataset employed. Also, the accuracy of the proposed system is relatively low. In their study, Katarya and Jain [21] applied several machine learning approached for diagnosing diabetes. The authors used the Indian Pima dataset for their experiments. The findings of the work demonstrated that the RF ensemble technique had superior performance compared to the other models. A drawback of the work is that the dataset used is small in size. Moreover, the proposed system exhibits a rather low level of accuracy.

Rubaiat, Rahman, and Hasan [22], did a comparable study to see how well different machine learning models could extract useful features from a diabetes dataset. The models were later used for the purpose of predicting diabetes in patients. Experimental analysis of the work showed that the utilisation of multilayer perceptron (MLP) in conjunction with a feature selection technique yielded superior results when compared to alternative approaches used in the paper. The study's inadequacy lies in the extremely small dataset it used. Moreover, the efficiency of the suggested system is subpar. Furthermore, the study did not use authentic patient data acquired from hospitalized individuals confirmed to have diabetes.

Swapna, Vinayakumar, and Soman [23] applied deep learning techniques for diabetes diagnosis. The research applied a range of advanced deep learning algorithms to extract features from heart rate variability (HRV) data. The obtained features serve as input for the support vector machine (SVM) for classification. The suggested approach has the capacity to aid medical personnel in making informed choices about providing treatment to patients. It can aid in the identification of diabetes through the analysis of ECG signals, having achieved an accuracy rate of approximately 95.7%. One potential constraint of the study is the very limited quantity of the dataset employed. Furthermore, there is a need for further improvement in the performance of the proposed system.
In another study, Azbeg *et al.* [1] used a probabilistic predictive model to identify instances of diabetes-related emergencies. The authors introduced a system architecture based on the Internet of Things (IoT) that guarantees the acquisition of patient data for the purpose of forecasting critical instances of diabetes. To ensure data security, the use of blockchain and IPFS has been employed. Additionally, for the purpose of data analysis, a statistical-based approach for predictive modelling has been suggested. The evaluation and comparison of the method's performance were conducted in relation to other contemporary prediction methods. Therefore, the proposed framework can be used for predicting diabetes, and notifying medical practitioners or healthcare facilities about critical instances that require immediate attention. A limitation of the work is the absence of any novel algorithmic development. Islam *et al.* [24] applied data mining approaches to predict diabetes. The dataset had a total of 520 participants who received questionnaires with respect to possible variables that may contribute to the beginning stages of diabetes. This method proves to be very effective when applied to a previously created dataset. One potential constraint of the study is the relatively limited size of the dataset employed.

Shukla [25] forecasted diabetes using a linear regression model. The authors demonstrated that diabetes can be because of some factors that may seem inconsequential to us but have been recognised by medical professionals as possible contributors to an enhanced susceptibility to this disease in the future. The logistic regression model, trained using the most influential features, had an accuracy rate of 82.92%. One drawback associated with this task is the relatively low level of accuracy.

Using ensemble learning in this study is because machine learning models have been built in the past with the idea that they will work best when they are trained and tested on data that comes from a similar feature space and distribution. Nevertheless, in the event of changes in the feature space or data distribution, it becomes imperative to generate a new model. The cost associated with developing a novel model from scratch on each occasion, in addition to the acquisition of fresh training data, is significant. Ensemble learning facilitates the efficient extraction of extensive training data by minimising the amount of effort and time needed. Ensemble learning has the capability to leverage already-existing data to tackle novel tasks or domains. The application of acquired knowledge enables the individual to address novel issues with greater speed and efficiency.

3. Materials and methods

3.1. Random Forests (RF)

The random forest algorithm is a versatile ensemble methodology that entails the construction of several decision

Table 1. Experimental settings and parameters tuning of Random Forest, Adaboost, GBOOST, LGBM, CatBoosting and Improved WAEL by GWO.

Model	Hyperparameter obtained from GWO	Values
Random forest	n_estimators	87
	max_depth	29
	min_samples_split	3
	min_samples_leaf	1
AdaBoost	n_estimators	93
	learning_rate	0.8356
GBOOST	n_estimators	400
	max_depth	5
	loss	Squared_error
	min_samples_split	2
	learning_rate	0.1
LGBM	n_estimators	101
	max_depth	5
	loss	Squared_error
	min_samples_split	2
	learning_rate	0.1
CatBoosting	n_estimators	101
	max_depth	5
	loss	Squared_error
	min_samples_split	2
	learning_rate	0.1
Improved WAEL	n_estimators	1
	Weight threshold	>=0.5

trees (DT). A random forest (RF) algorithm demonstrates the ability to efficiently handle large datasets and displays a reduced vulnerability to overfitting in comparison to individual decision trees. The decision tree methodology outlined by Ref. [26] is well acknowledged for its ability to generate various decision trees using a given dataset. The methodology involves randomly partitioning the dataset into various segments prior to constructing distinct decision trees for each sub-part. The projected result of each decision tree is later combined to obtain a prediction that exhibits increasing levels of accuracy and precision. In the context of random forest regression, the outcome value associated with each input or subset is computed by calculating the mean of the predicted values obtained from many decision trees. The process of generating a bootstrapping population of an n-tree in random forest regression involves utilising the actual input dataset [27]. The next step in the process entails the construction of an unpruned regression tree utilising distinct bootstrap sets. Nevertheless, by combining the results generated by the decision trees, a novel data value is computed. The computation of the frequency of error involves the utilisation of the average result that was generated for a data point inside the initial dataset by base learners that did not undergo training from the training data.

3.2. Adaptive Boosting Regression (Adaboost)

The AdaBoost approach, commonly called adaptive boosting, is a boosting method employed as an ensemble approach in the domain of machine learning. This feature emphasises the data points that have been wrongly classified by the preceding models within the ensemble. The deployment of this algorithm is straightforward and has the potential to significantly improve the learning outcomes of learners with limited capabilities. The process of adaptive boosting involves the reassignment of weights across every instance, with a greater emphasis placed on instances that were mistakenly classified. This approach is sometimes referred to as "adaptive boosting" due to its dynamic nature. Boosting is a commonly used technique in supervised learning that aims to mitigate both bias and variance. According to Refs. [28–30], the underlying premise of this approach is that learners exhibit incremental advancement. Afterwards, more iterations of the regressor are applied to the current dataset. Nevertheless, the weights of the incidences are just adjusted according to the most recent results.

In this analysis, we will consider a dataset $D=(x_1, y_1), \ldots, (x_n, y_n)$ which has been constructed by aggregating accurate data points obtained through ongoing observations conducted over a specific time frame. The dataset comprises a collection of n possible combinations of measurements, where each observation is assigned, a weight denoted as w_i. The calculation of the likelihood for an observation to be included in the training set during iteration c is determined by the allocated weight of each measurement i. The subsequent step involves calculating the average loss ($loss_c$) for the model c by utilising the weighted sum of the probability for each measurement. The mathematical expressions representing the average loss ($loss_c$) and probability (pr_c) are shown in equations (1)–(3).

$$pr_c = \frac{w_i}{\sum w_i}, \tag{1}$$

$$loss_c = \sum_{i=1}^{n} loss_c pr_c, \tag{2}$$

$$w_i^{c+1} = w_i^c \beta_c (1 - loss_c), \tag{3}$$

assuming p_k represents the probability at iteration k, average loss at iteration c, w_i^{c+1} is the updated weight at iteration i, w_i^c refers to the prior weight and β_c represents the model loss.

3.3. Gradient Boosting Regression (GBOOST)

Gradient boosting is a widely employed ensemble machine learning technique used for addressing many tasks, including regression, classification, and other problems. Refs. [31, 32] proposed a prediction model that takes the form of an ensemble, including many weak prediction models resembling decision trees. The gradient boosting algorithm iteratively chooses a function that is oriented in the reverse direction of the gradient, with the aim of maximising a given cost function throughout the whole function space. The GBOOST algorithm constructs decision trees in a sequential manner, where each subsequent tree is designed to rectify any flaws caused by the preceding tree. The algorithm frequently exhibits superior predicted accuracy compared to alternative methods. Decision trees are commonly

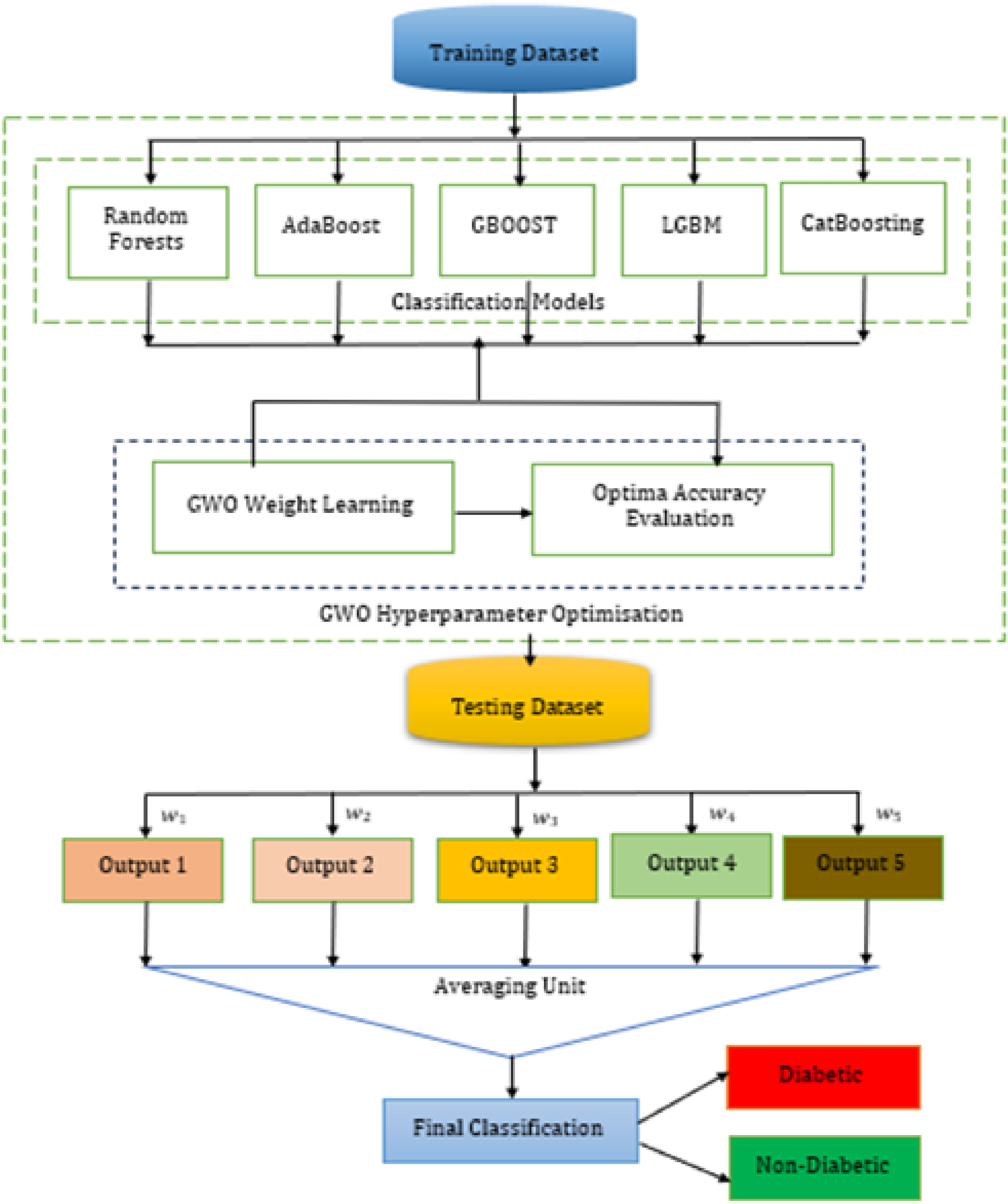

Figure 1. Architecture of the proposed diabetes detection system.

employed as suboptimal predictors in the context of gradient boosting. Weakly learned models can be characterised by low levels of variance and regularisation, as well as a substantial bias towards the training dataset. According to Ref. [33], these models provide outputs that only exhibit minor improvements over random predictions. The three primary constituents of

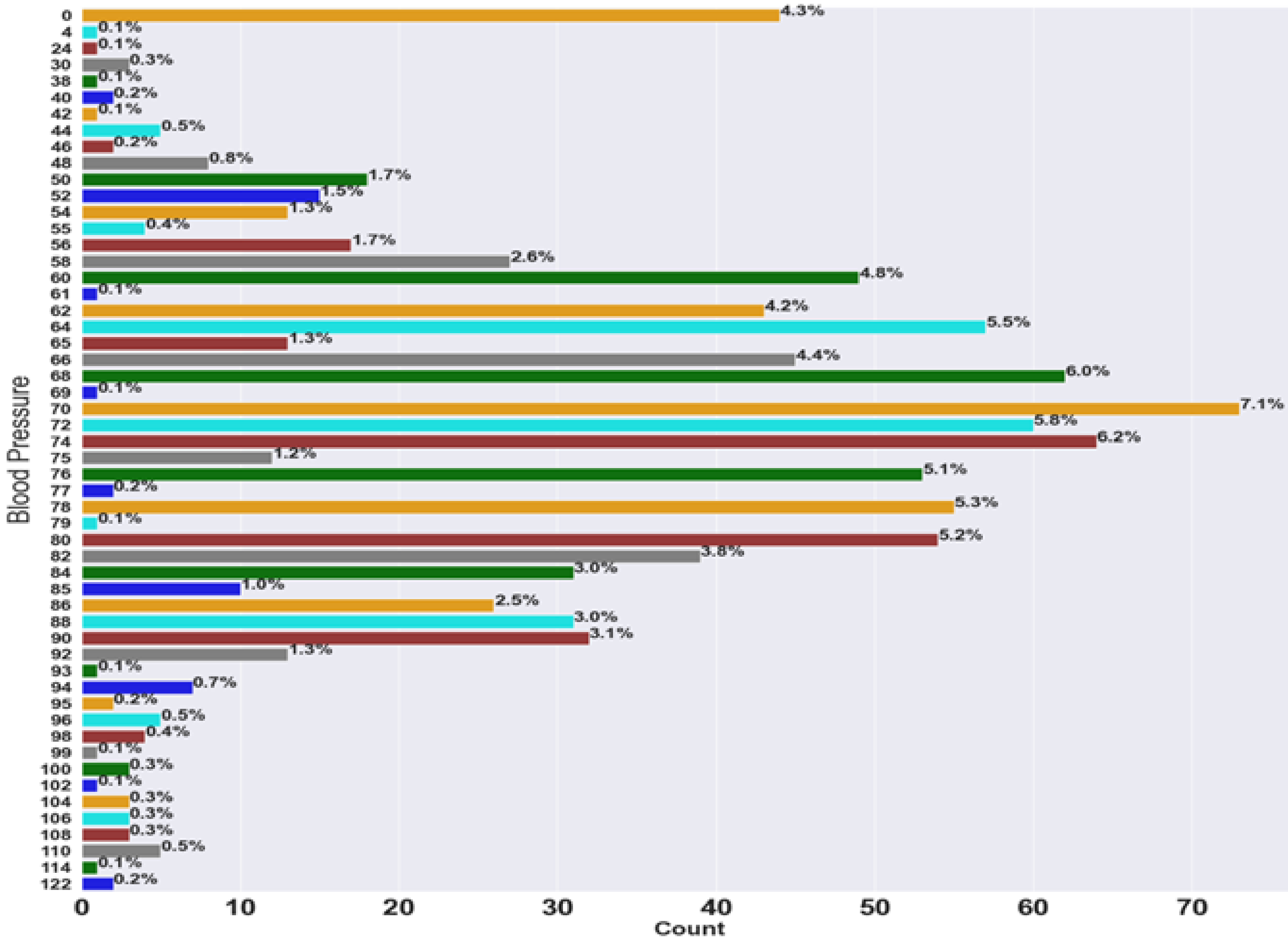

Figure 2. Blood pressure of diabetic patients.

boosting techniques encompass an additive model, weak learners, and a loss function. Gradient-boosting machines operate by leveraging gradients to identify the deficiencies present in suboptimal models. The process involves employing an iterative approach with the objective of ultimately combining base learners to minimise detection errors. This is achieved by merging decision trees through an additive model, while the reduction of the loss function is accomplished through the utilisation of gradient descent [34]. The mathematical representation of the gradient boosting tree (g) is illustrated by equation (4) and equation (5).

$$g = \sum_{i=1}^{n} f_i x_t, \tag{4}$$

$$arg_{min} = \sum_{t} L\left(y_t, g\right) + f_{n+1}\ x_t, \tag{5}$$

where "g" designate the gradient boosting tree, "$L()$" represent the loss function, and "$f_{n+1}x_t$" signify the newly produced decision tree.

3.4. Light Gradient Boosting Machine (LGBM)

The LGBM ensemble approach is a gradient-boosting technique that is widely accessible and known for its exceptional computational capabilities and effectiveness in the domain of machine learning. The acronym LGBM stands for the gradient-boosting framework that Microsoft Inc. developed. The purpose of its development was to enable the decentralised and efficient training of machine learning models on a large scale, as noted by Refs. [35–37]. The algorithm in question is a member of the gradient-boosting family of machine learning algorithms. These algorithms function by combining the results of multiple weak learners, typically in the form of decision trees, to create a reliable predictive model [38]. The Light Gradient Boosting Machine (LGBM) algorithm has been specifically developed to prioritise the improvement of computing speed as well as effectiveness. The system has been acknowledged for its outstanding performance in efficiently handling and analysing large volumes of data. The LGBM approach is based on the gradient boosting framework, wherein weak learners (particularly decision trees) are trained sequentially to correct any errors made by the preceding models. The algorithm being discussed utilises a

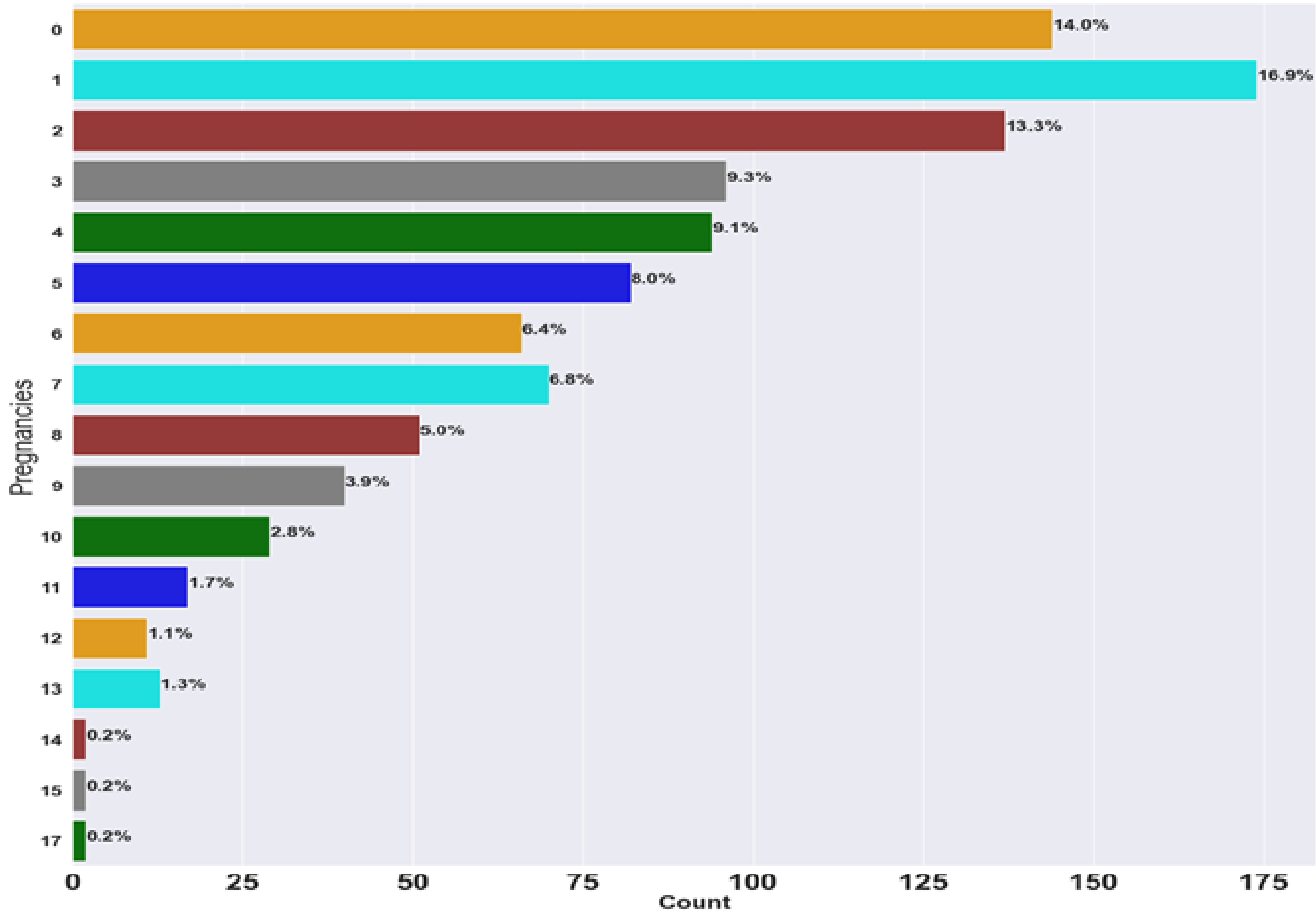

Figure 3. Pregnancy distribution of diabetic patients.

leaf-wise expansion strategy as opposed to the level-wise technique employed by various alternative gradient-boosting algorithms. Dev and Eden [39] suggested that the utilisation of this particular methodology has the potential to lead to shorter training durations and reduced memory usage. The LGBM algorithm utilises a learning strategy based on histograms, where continuous features are discretised into bins. The method of discretization facilitates the acceleration of the training process. Chen *et al.* [40] assert that the application of gradient-based optimisation techniques is utilised to create models in a highly efficient manner. The LightGBM (LGBM) algorithm provides built-in support for categorical features, eliminating the need for preprocessing or the use of one-hot encoding methods for categorical variables. Within the realm of LGBM, the optimisation of hyperparameters, the proficient management of categorical features, and the understanding of the impact of different parameters on model performance are frequently emphasised by researchers and practitioners [36]. The use of this technique extends to a wide array of machine learning applications, among others, categorisation, statistical modelling and rating.

3.5. *CatBoosting ensemble method*

The CatBoost algorithm, known as categorical boosting, belongs to the gradient boosting family within the domain of machine learning [41]. The method was specifically designed to effectively handle categorised aspects, making it well-suited for datasets that include both quantitative and qualitative variables. Zeng *et al.* [42] have emphasised the exceptional effectiveness, easy-to-use user interface, and extraordinary, unorthodox abilities of CatBoost in the domain. These achievements are notable as they were attained without the requirement for significant hyperparameter tuning. CatBoost is a machine learning algorithm that has been designed with the specific purpose of efficiently handling categorical information. This eliminates the need for complex preprocessing methods such as one-hot encoding, as demonstrated by Refs. [41, 43]. The algorithm utilises various methodologies, including target encoding and ordered boosting, to proficiently manage categorical variables within its internal processes. Like other gradient boosting strategies, CatBoost builds an ensemble of decision trees in a sequential way, aiming to minimise the loss function. CatBoost has been purposefully developed with a focus on enhancing performance and optimising memory usage. The acceleration of the training process is

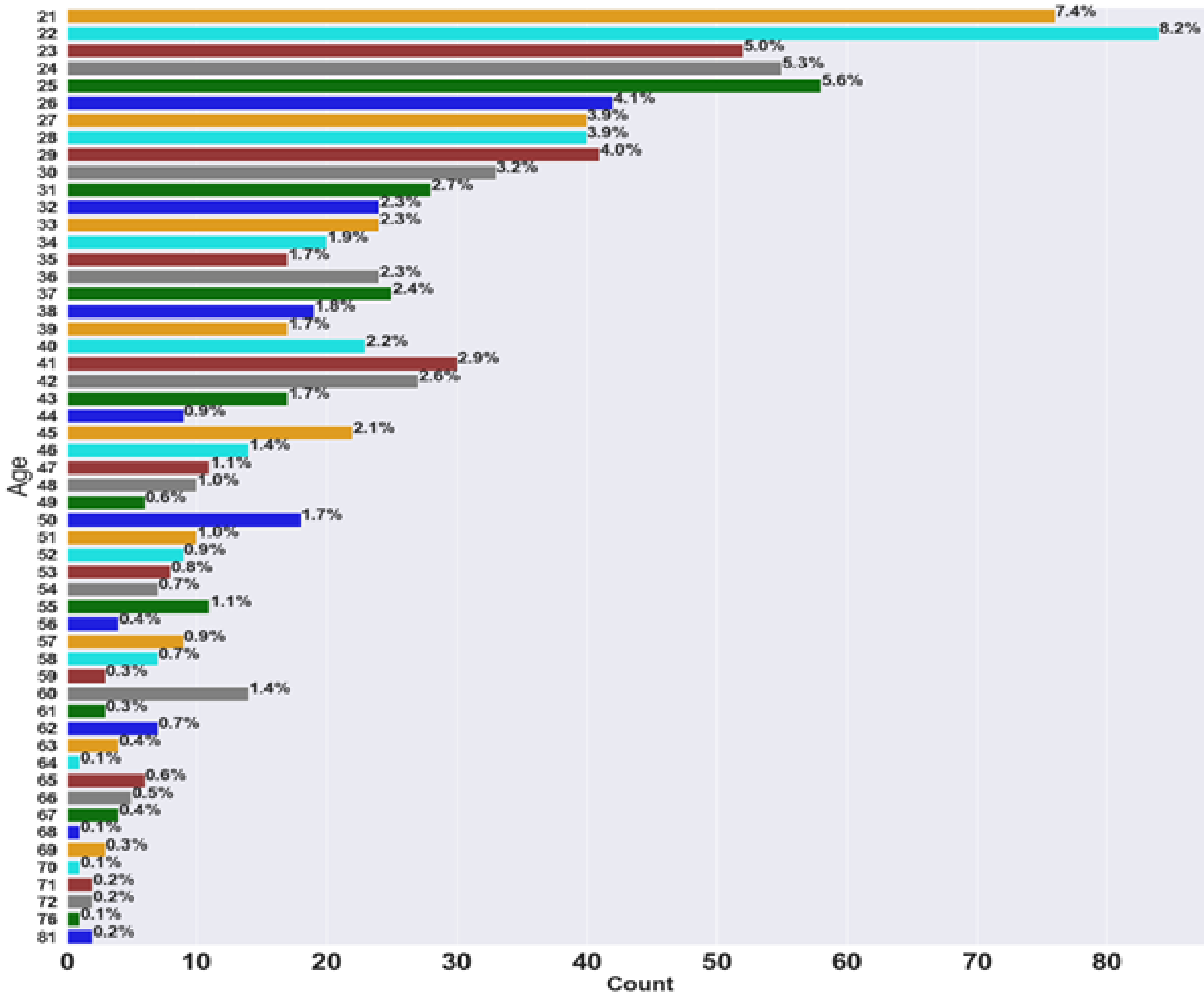

Figure 4. Age distribution of diabetic patients.

achieved by employing methodologies such as oblivious trees and implementing matrix operations [44, 45]. CatBoost integrates a variety of built-in regularisation methods to address the problem of overfitting. The model integrates both L1 and L2 regularisation techniques to effectively handle the intricacy of the system. Olsson and Acharya [46] assert that CatBoost provides a variety of tools that speed up cross-validation, thereby enabling users to efficiently assess the resilience of model performance. While CatBoost often exhibits robust performance using its default configurations, it provides users with a wide range of hyperparameters that may be customised to accommodate the distinct attributes of their datasets and goals. Commonly employed attributes include the pace of learning, the depth of trees, and the number of trees. Lazar, Sim, and Wu [47] assert that CatBoost provides GPU acceleration, thereby enabling expedited training times and improved efficiency, especially when dealing with extensive datasets.

3.6. *Grey Wolf Optimisation algorithm (GWO)*

GWO algorithm is a metaheuristic optimisation approach motivated by nature that imitates the predatory and social behaviours of grey wolves in their natural habitat. GWO was created in 2014 by Ref. [48] and is mostly used to address challenging optimization problems, especially those that involve continuous spaces. It is renowned for being easy to use, adaptable, and efficient in locating global optima. The core idea behind grey wolf optimisation algorithm is as follows:

1. Grey wolf social hierarchy: The four primary types of grey wolf packs, which are separated into groups according to rank, are modelled by the GWO algorithm as the leadership hierarchy:
   Alpha (α): The highest-ranking wolf and the best possible candidate answer thus far discovered. Alphas make all the decisions, including movements and hunting tactics.

Beta (β): The second tier, supporting the alpha and exerting a great deal of control over the pack. Betas stand for the runner-up options.

Delta (δ): The third tier, below both beta and alpha. When an alpha or beta dies or is replaced, deltas step forward to take the leadership role and assist in upholding order within the pack. These stand for the third-best options.

Omega (ω): The lowest-ranking wolves, who assist the higher-ranking wolves and have little say in matters of policy. They stand in for the remaining search space solutions.

2. Hunting behaviour: Encircling, hunting, and attacking prey are the three primary stages of the hunting method used by grey wolves. To discover and take advantage of the search space, the GWO algorithm imitates these stages: **Encircling Prey:** Estimating the distance that exists between themselves and the victim, grey wolves encircle their prey. This is reflected in the GWO method by updating the wolves' position according to the prey's position (i.e., the best solution thus far).

   Hunting: Alpha, beta, and delta wolves drive the hunting process, and the positions of the other wolves are regularly updated in relation to these three leaders. This cooperative strategy aids in striking an equilibrium between exploitation (focusing the search on favourable locations) and exploration (exploring unfamiliar territories).

   Attacking Prey: The wolves take fewer steps when they get closer to their victim. This is simulated by gradually lowering the values of a few control parameters, which enables the algorithm to converge on the best solution.

3.7. The proposed model

3.7.1. Improved Weighted Average Ensemble Learning (WAEL)

The improved weighted average ensemble learning is a technique used in ensemble learning where the results produced by many base models are combined using a weighted average technique [49]. The proposed approach entails the allocation of a distinct weight to the result generated by each separate base model. The ensembled classification is subsequently obtained through the aggregation of these weighted classifications [50, 51]. The assigned weights to each model are representative of their perceived importance or reliability in the ensemble. The core principle of the average ensemble, also known as the weighted average ensemble, is to mitigate overall errors by amalgamation of results derived from a diverse set of learners. The fundamental approach is to first propose that each classifier would manifest separate errors throughout the training and classification stages [52, 53]. Following this, a systematic process is implemented to produce a group of classifiers that demonstrate a diverse array of variances, which are then combined to give a consolidated result. It is anticipated that the combined process will result in the reduction of the general rate of incorrect classifications. The application of weighted average ensemble learning offers a flexible and understandable method, allowing for the incorporation of the beneficial characteristics of multiple models into a combined detection result [54]. The distribution of weights among separate models has a significant influence on the entire efficacy of the ensemble. The determination of the most appropriate weights often requires a procedure involving the testing and validation of an independent dataset. This approach is often used for both regression and classification tasks.

3.7.2. Architecture of weighted average ensemble learning

In weighted average ensemble learning, classifications made by many base models are combined using a weighted average as the building block. Although the notion is very straightforward, the architecture can be comprehended by following a sequential procedure. The proposed architecture of the weighted average ensemble learning model incorporates five base learner models, namely RF, Adaboost, GBOOST, LGBM, and CatBoosting ensemble. Additionally, a random forest meta-model is employed to aggregate the detection results generated by the base models. Five distinct feature spaces are derived from the five models and subsequently combined to create an optimised feature space. To determine the best-optimised classification model, a grid search is employed to assign five different weights, namely weight 1, weight 2, weight 3, weight 4, and weight 5, to five distinct models. The use of an ensemble model, which combines the outputs of multiple models, can enhance the accuracy of classification. The ensemble approach exhibits more robustness compared to standalone models since it leverages the collective intelligence of multiple models. In cases where one model within the ensemble produces an erroneous classification, the remaining models can effectively compensate for this error and yield the right classification. The underlying structure of the ensemble learning architecture is based on the concept of an elementary weighted average.

1. Original data: The dataset is segregated to two subgroups - the training data and the test data.
2. Base models: Several base models are trained individualistically using various techniques, hyperparameters, or subsets of the data. Let $M_1, M_2, M_3, \ldots M_n$ depicts the group of base models.
3. Generate detection results: Apply each trained base model to provide detection results on a consistent dataset of either test or validation data. Find detection result $P_1, P_2, P_3, \ldots P_n$ from each base model.
4. Assign weights: Ascribe a weight $W_1, W_2, W_3, \ldots W_n$ to each base model based on factors such as its performance, accuracy, or any other pertinent statistic. The determination of weights can be achieved either through empirical methods or by employing optimisation techniques.
5. Compute weighted average: The detection results of each model are multiplied by their respective given weights. To derive the ultimate ensemble detection E, it is necessary to consolidate the weighted detection. The mathematical representation of the ensemble detection results

E is given by equation (6).

$$E = W_1.P_1 + W_2.P_2 + W_3.P_3 + \ldots W_n.P_n \quad (6)$$

6. Normalize weights (optional): It is possible to consider normalising the weights in order to guarantee that their total sum equals 1. The process of normalisation can prove advantageous in terms of enhancing understanding and ensuring uniformity. The equation is depicted in equation (7).

$$Normalised\ weight = \frac{Original\ weight}{Sum\ of\ weights} \quad (7)$$

7. Final ensemble detection result: The ultimate ensemble detection, denoted as E, is derived by aggregating the detection result generated by each individual base model using appropriate weighting factors.

The architectural design of this system is characterised by its simplicity and modularity, which enable adaptability in selecting fundamental models, weights, and the overall configuration of the ensemble. The architectural design can be utilised for both regression and classification tasks. The visualisation of the architecture may take the form of a flowchart or diagram that depicts the sequential processes entailed in the training, detection, and amalgamation of the basic models to derive the ensemble detection result. It is imperative to acknowledge that the efficacy of weighted average ensemble learning is contingent upon the broad range and effectiveness of the underlying models, as well as the meticulous allocation of weights. The process of experimenting and evaluating different datasets is frequently essential in ascertaining the best weights for each model. Refer to Figure 1, which shows the design of the proposed system.

The diabetes detection system being considered employs ensemble learning techniques that rely on weighted averages to diagnose the condition. Experiments were conducted using diabetes data from various hospitals in Maiduguri. The dataset was divided into two sets: the training set, containing 80% of the data, and the test set, containing the remaining 20%. The proposed ensemble learning system consisted of a total of five models, forming its entire framework. This paper's primary goal is to categorise the dataset as either diabetic or non-diabetic using WAEL. This is realised by averaging the weights and classifications the five models (AdaBoost, random forest, GBOOST, LGBM, and CatBoosting ensemble learning methods). Weighted average ensemble learning is a flexible and interpretable method that allows the incorporation of the strengths of multiple models into a single detection. The weights assigned to each model influence the overall performance of the ensemble, and finding the optimal weights often involves experimentation and validation on a separate dataset.

The improved weighted average ensemble model is an exceptional strategy in machine learning that synergises various models to significantly boost detection performance. This research leverages the proposed model for diabetes detection by utilizing key features such as pregnancies, glucose levels, blood pressure, skin thickness, insulin, body mass index (BMI), diabetes pedigree function, and age; the ensemble model delivers a more robust and precise classifier than any standalone model, as demonstrated by the compelling results. The improved WAEL model amalgamates detection from multiple models to achieve detection accuracy and reliability that surpasses any individual model.

The proposed model employs GWO to fine-tune the weight of each model, ensuring that each model's contribution is valued based on its performance, resulting in better-performing models exerting greater influence on the final detection. In contrast to the traditional WAEL model, the improved weighted average ensemble further refines this methodology by optimizing the weights assigned to each model through the GWO technique, thereby identifying the optimal combination that minimizes detection error. GWO facilitates the optimization of weights assigned to each model within the ensemble. This crucial step entails assigning a weight to each model's detection based on its remarkable performance. The improved WAEL model employs the GWO algorithm to discover the most effective combination of weights that improves the ensemble's performance on the test dataset. By integrating multiple models, the ensemble reliably attains enhanced accuracy in comparison to any single model, leveraging the advantages of various algorithms.

3.7.3. Dataset description

The technique proposed in this paper was evaluated on the diabetes datasets obtained from the University of Maiduguri teaching hospital, and the one gotten from Umaru Shehu specialist hospital, Maiduguri. The datasets consist data of diabetic patients collected from 2018 to 2023.

The dataset was collected for male and female of 17 years and above living in Maiduguri, Borno State, Nigeria, and its environs. This dataset is based on certain diagnostic measurements which are used as models' feature variables. It is made up of 9 features and 1030 instances. The features include pregnancies, glucose, blood pressure, skin thickness, insulin, body mass index (BMI), diabetes pedigree function, insulin, and age.

3.7.4. Experimental settings

Table 1 presents the experimental setting and parameter configurations for the research. Training a model involves determining optimal values for each weight and bias variable using annotated samples. Parameter tuning is a critical step in the development of machine learning models. Table 1 fully illustrates the parameters used to tune the models during the training and evaluation stages using the diabetic dataset. These parameters are essential to enhancing the efficiency of the model.

3.7.5. Performance metrics

The effectiveness of the proposed WAEL model was evaluated using the following metrics:

1. **Accuracy:** The degree of accuracy in the field of machine learning serves as an evaluation term that quantifies the proportion of accurate detections generated by a given model relative to the overall number of detections generated. The calculation is performed by splitting the

Table 2. Performance comparison of the ensemble models.

Model	Accuracy	Precision	Recall	F1-Score
RF	0.81	0.79	0.79	0.79
AdaBoost	0.76	0.73	0.71	0.72
GBOOST	0.76	0.73	0.70	0.71
LGBM	0.85	0.85	0.85	0.85
CatBoosting	0.77	0.74	0.72	0.73
Improved WAEL	0.98	0.97	0.97	0.95

total number of accurate detections by the overall number of detections as seen in equation (8).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{8}$$

2. **Precision:** Precision is a primary measure employed to assess the efficiency of a machine-learning model. The metric measures the precision of the system's positive forecasts as seen in equation (9).

$$Precision = \frac{TP}{TP + FP}. \tag{9}$$

3. **Recall:** It is the metric used to measure the correctness of a model in properly detecting instances classified as true positives. It is mathematically presented in equation (10).

$$Recall = \frac{TP}{TP + FN}. \tag{10}$$

4. **F1-score:** This is a system of measurement that computes the accuracy of a model when applied to a given dataset. Binary classification methods are employed for the purpose of assessing and evaluating samples by categorizing them into either a 'positive' or 'negative' class. It is depicted in equation (11).

$$F1\ score = \frac{2\ x\ Precision\ x\ Recal}{Precision +\ Recall}. \tag{11}$$

5. **Confusion Matrix:** A confusion matrix is a potent instrument for comprehending the complex nature of a classification model's performance. It assists in recognising not only the accuracy but also the kinds of errors made by the model.

4. Descriptive analysis, results and discussions

This section presents an introduction to the findings and analyses the important breakthroughs obtained from the experiments we conducted. We conducted the investigations using the Python programming language on a Jupiter notebook. The ensemble machine learning models were trained and tested using the diabetes dataset. This study looked at all of the features in the dataset used for training and assessing the models.

4.1. Descriptive analysis of the diabetes dataset

Figure 2 illustrates the blood pressure readings of the diabetes patient as recorded in the dataset. The prevalence of individuals with a blood pressure reading of 70 mmHg is 7.1%. Approximately 6.2% of the patient population has a blood pressure reading of 74 mmHg. A mere 27% of the patient population exhibits blood pressure readings falling within the range of 80 mmHg to 122 mmHg. The typical blood pressure range for individuals is commonly observed to fall within the range of 90/60 mmHg and 120/80 mmHg. Hypertension is characterised by a blood pressure measurement below 90/60 mmHg. Hypertension and diabetes mellitus are frequently comorbid conditions within the framework of the metabolic syndrome. There are several potential explanations for this phenomenon, including the presence of shared risk factors among individuals with high blood sugar levels as well as the detrimental impact of elevated blood glucose on cells within the cardiovascular system. Hypertension and diabetes exhibit some shared etiological variables and risk elements. Individuals with a particular medical disease are more susceptible to the development of another related condition. Similarly, individuals who experience both diseases may observe a reciprocal exacerbation of each condition [55, 56]. Although diabetes is commonly associated with high blood pressure, it can also be correlated with low blood pressure. According to Chokshi, Grossman, and Messerli [57], individuals with diabetic neurological disorders may experience hypotension, especially after assuming a standing position or following meals, because of their sympathetic nervous system dysfunction.

Figure 3 depicts the number of pregnancies among individuals diagnosed with diabetes. The proportion of patients who have never experienced pregnancy is 14%. Approximately 16.9% of the patient population has experienced a single pregnancy. The proportion of individuals who have experienced two pregnancies is 13.3%. Furthermore, it is worth noting that a total of 9.3% of the patients included in the study have experienced three pregnancies. Approximately 9.1% of the population has experienced four pregnancies. The prevalence of patients with a history of five pregnancies is 8%. Additional examination reveals that a total of 1.3% of the patient population has experienced a remarkable 13 pregnancies. A total of 0.2% of the patients experienced pregnancy on 14, 15, or 17 occasions, respectively. It is imperative to acknowledge that although the quantity of pregnancies may contribute, it is not the exclusive predictor of the risk of developing diabetes. Additional variables, such as age, familial background, race, and individual lifestyle decisions, also exert notable influences. The association between the frequency of pregnancies and the occurrence of diabetes is frequently investigated within the framework of gestational diabetes mellitus (GDM) and type 2 diabetes. GDM is a form of diabetes that manifests specifically during the gestational period of pregnancy. The frequency of pregnancies may influence the probability of developing GDM. There is a potential correlation between repeated pregnancies in women and an elevated likelihood of acquiring GDM, particularly if they have previously experienced GDM in prior pregnancies.

Table 3. Performance comparison with other techniques.

Author(s)	Technique	Dataset	Classifier(s)	Accuracy (%)
Dutta *et al.* [58]	WAEL	Bangladesh Diabetes dataset	Naive Bayes, RF, DT, XGBoost, and LGB.	73.5
Albadri *et al.* [59]	Hybrid Machine Learning	Pima Indian Diabetes dataset	SVM, DT, and RF.	76.80
Karthikeyan *et al.* [60]	Weighted Average	UCI Diabetes dataset	Adaboost, RF or Randomization, Bagging or Bootstrap Aggregation, Voting, and Stacking.	86.79
Li, Fu and Li [61]	Ensemble	Pima Indian Diabetes dataset	XGBoost, XGBoost + logistic regression, data feature stitching + XGBoost.	80.20
Mushtaq *et al.* [62]		Do	kNN, RF, naive Bayes, SVM, GBoost, logistic regression, and voting classifier.	81.30
Atif, Answer & Talib [63]	Ensemble	Pima Indians Diabetes dataset and the Early Stage Diabetes Risk Prediction Dataset,	Hard voting classifier	81.7
Aurpa, Jeba & Rasel [64]	WAEL	Bangladesh Diabetes Dataset	LGBM, XGB, SVC, RF, and Gradient Boosting Classifier.	96.10
Saihood & Sonuç [65]	Stacking	Pima Indian Diabetes dataset	RF+SVM	97.50
This paper	Improved WAEL	Diabetes dataset from University of Maiduguri teaching hospital, and Umaru Shehu specialist hospital, Maiduguri	RF, AdaBoost, GBOOST, LGBM, and CatBoosting.	98.90

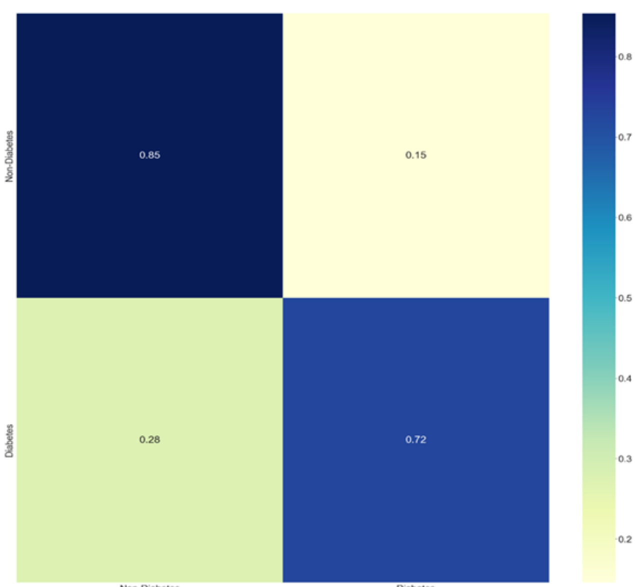

Figure 5. Confusion matrix of random forest (RF).

Figure 4 presents the age of the diabetic patients. The number of patients that are 21 years old is 7.4%. About 8.2% of the patients are 22 years old. The percentage that are 23 years old are 5.0%. Also, 5.6% of the patients are 25 years old. About 2.2% are 40 years old. The percentage of patients that are 50 years old is 1.7%. About 1.4% are 60 years old. Also, 0.1% of the patients are 70 years old. While 0.2% of the diabetic patients are 81 years old. The relationship between age and diabetes is well-established, and age is considered one of the significant risk factors for the development of diabetes. Type 1 diabetes is commonly diagnosed in people during their formative years, adolescence, or early adult years. Although it has the potential to manifest at any stage of life, it is not commonly linked to the process of ageing. Type 2 diabetes is frequently correlated with the process of ageing. The likelihood of having type 2 diabetes becomes more intense with advancing age, and many individuals confirmed to have type 2 diabetes are of adult age. The risk tends to rise significantly after the age of 45, and the prevalence increases with each subsequent decade of life. Aging can affect the function of pancreatic beta cells, which are responsible for insulin production. The decline in beta cell function can contribute to impaired glucose metabolism and an increased risk of developing diabetes.

4.2. Results and discussion

This sub-section gives a detailed description of the results of the experiments performance in this study. Figure 5 is confusion matrix of the RF model.

Figure 5 presents the confusion matrix of the RF algorithm, which offers significant information regarding the model's performance across several classes. The confusion matrix displays the expected and observed class labels for a binary-

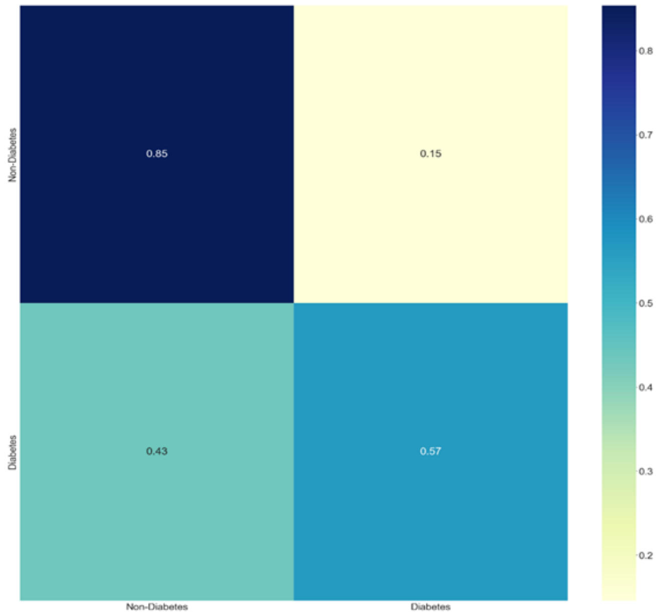

Figure 6. Confusion matrix of Adaboost.

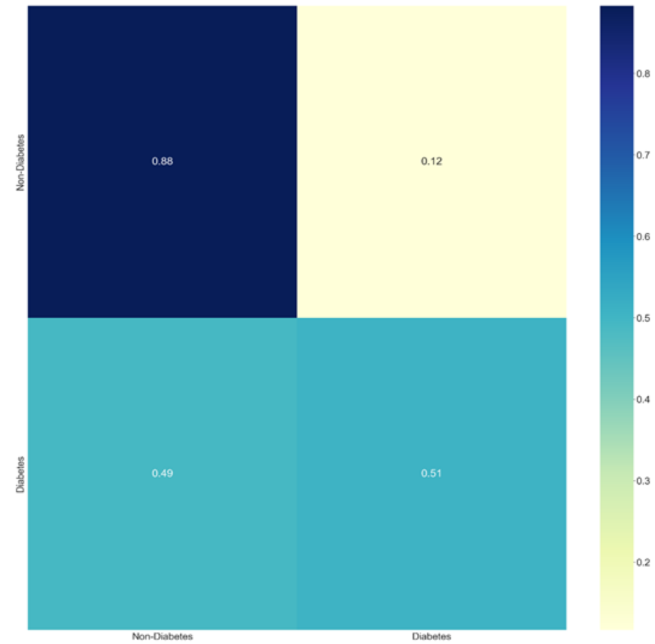

Figure 7. Confusion matrix of GBOOST.

classification task encompassing two distinct categories: individuals with diabetes and individuals without diabetes. The RF model accurately classified 117 cases of diabetes as diabetes, representing real positive detections. There were 20 instances of misclassification of diabetic as non-diabetic (false positives). The model correctly classified 50 instances of non-diabetes as non-diabetes (true negative). While RF model incorrectly classified 19 instances of non-diabetes as diabetes (false negative). Depicted in Figure 6 is the confusion matrix of Adaboost. The confusion matrix for AdaBoost is displayed in Figure 6. The confusion matrix presents the anticipated and actual class labels for a binary-classification task encompassing two distinct categories: people who have diabetes and those without diabetes. The AdaBoost algorithm accurately classified 117 occurrences as diabetes, which were indeed diabetes (true positives). There were 20 instances of misclassification of diabetic as non-diabetic (false positives). The model correctly classified 39 instances of non-diabetes as non-diabetes (true negative). While AdaBoost model incorrectly classified 30 instances of non-diabetes as diabetes (false negative). Presented in Figure 7 is the confusion matrix of GBOOST.

Figure 7 represent the confusion matrix of GBOOST. The GBOOST model correctly predicted 121 instances of diabetes as diabetes (true positives). There were 16 instances of misclassification of diabetic as non-diabetic (false positives). The model correctly classified.

5 instances of non-diabetes as non-diabetes (true negative). While GBOOST model incorrectly classified 34 instances of non-diabetes as diabetes (false negative). The confusion matrix of LGBM is presented in Figure 8.

Figure 8 represents the confusion matrix of LGBM. The LGBM model accurately predicted 121 instances of diabetes as diabetes (true positives). There were 16 instances of misclassification of diabetic as non-diabetic (false positives). The model correctly classified 55 instances of non-diabetes as non-diabetes (true negative). While LGBM model incorrectly classified 14 instances of non-diabetes as diabetes (false negative). The confusion matrix of CatBoosting is portrayed in Figure 9. The confusion matrix of CatBoosting can be seen in Figure 9. The confusion matrix displays the anticipated and observed class labels for a classification task where the dataset is classified either as diabetic or non-diabetic. The CatBoosting

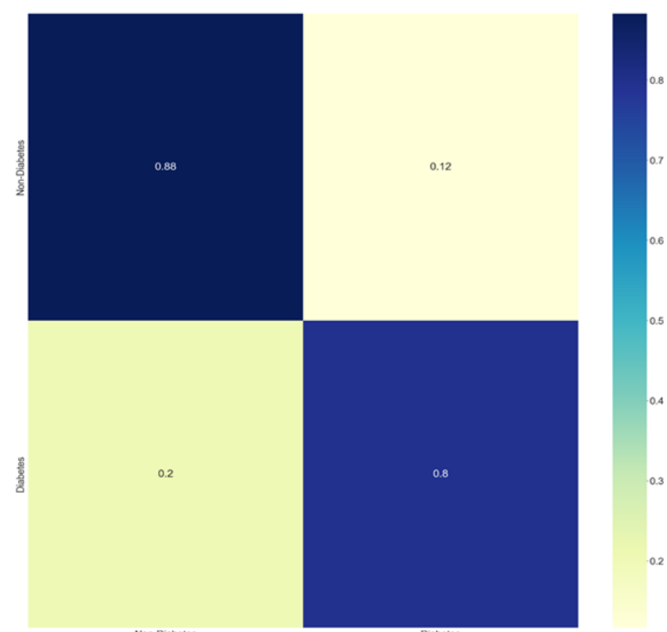

Figure 8. Confusion matrix of LGBM.

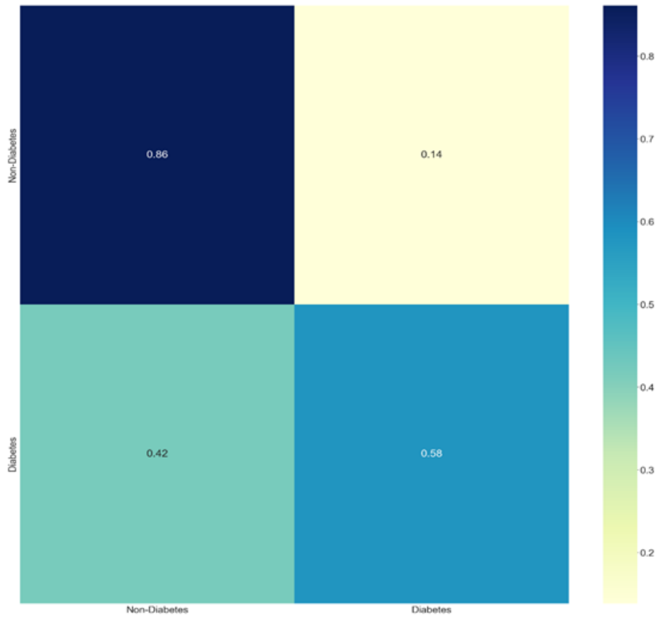

Figure 9. Confusion matrix of CatBoosting.

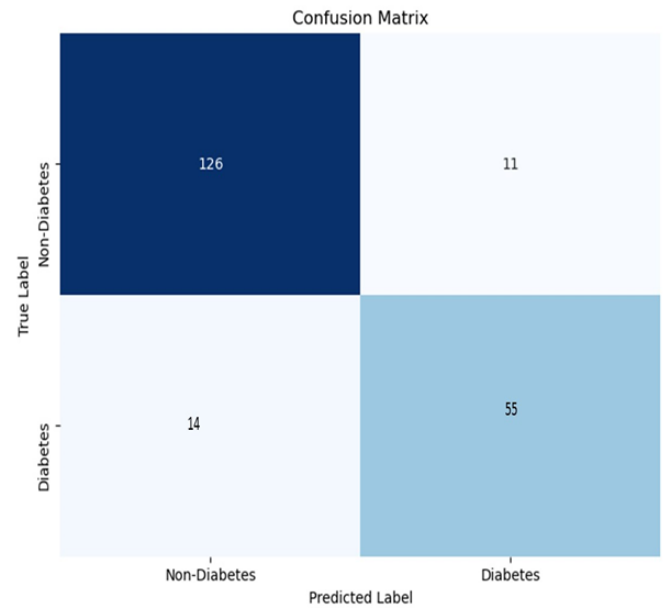

Figure 10. Confusion matrix of improved Weighted Average Ensemble Learning (WAEL).

model correctly predicted 118 instances of diabetes as diabetes (true positives). There were 19 instances of misclassification of diabetic as non-diabetic (false positives). The model correctly classified 40 instances of non-diabetes as non-diabetes (true negative). While CatBoosting model incorrectly classified 28 instances of non-diabetes as diabetes (false negative). The confusion matrix of WAEL model is depicted in Figure 10.

Figure 10 illustrates the confusion matrix of improved WAEL model. The proposed model correctly predicted 126 instances of diabetes as diabetes (true positives). There were 11 instances of misclassification of diabetic as non-diabetic (false positives). The model correctly classified 55 instances of non-diabetes as non-diabetes (true negative). While improved WAEL model incorrectly classified 14 instances of non-diabetes as diabetes (false negative).

Table 2 provides a comprehensive performance evaluation of all the models on the diabetes dataset used for in this paper. Depicted in the table is the average accuracy of the models considered in this study.

The proposed model used a weighted average ensemble model on 1030 data of patients and achieved an accuracy of 98.90%, precision of 97.00%, recall of 97.00%, and F1-Score of 95.00%. After giving the proposed improved WAEL model a close look, the results show that the performance is generally good, correctly putting most of the instances into the right classes. Nevertheless, there are some instances in which incorrect classifications have occurred. The improved WAEL model exhibits a significant degree of accuracy, showing its efficacy in accurately detecting and classifying instances of diabetes and non-diabetes within the dataset used in this paper. The experimental findings indicate that improved WAEL outperforms the remaining models utilised in this work, as indicated by several performance metrics such as accuracy, recall, precision, and F1 score. The utilisation of the improved WAEL model is justified by its substantial computing power. Hence, it can be inferred that the proposed improved WAEL model has remarkable performance in comparison to any of the other models presently under consideration. The outcomes of the research suggest that the improved WAEL model outperforms all other classifiers. There was a substantial improvement in accuracy, with an increase of 22.45% seen from the lowest-performing models, namely AdaBoost and GBOOST. The improved WAEL technique exhibited superior performance compared to other top-performing individual models, the LGBM Classifier, by a margin of 13.27%. This is a notable improvement.

5. Conclusion

This study proposed improved weighted average ensemble learning (WAEL) algorithm for effective and efficient diabetes classification. And evaluated the performance of the improved WAEL algorithm on diabetes dataset obtained from University of Maiduguri teaching hospital, and Umaru Shehu specialist hospital, Maiduguri using various measures to determine the effectiveness of the algorithm. The development and study of ensemble machine learning models are of utmost importance as they can aid healthcare professionals in the early identification and management of diabetes. Initially, a substantial proportion of the research included data analysis methodologies to comprehend, process, and present the data. Subsequently, the models undergo training using the training dataset employing six ensemble learning methods, including Random Forest, AdaBoost, GBOOST, LGBM, CatBoost, and weighted average ensemble classification models. Afterwards, the ensemble models' weights were adjusted by grid search to fine-tune the hyperparameters. Additionally, a 10-fold cross-validation was performed. For the purpose of accuracy analysis, the evaluation metrics employed across all models included the confusion matrix, precision, recall, and F1-score.

The simulation results demonstrated that improved WAEL achieved the highest degree of effectiveness. The classification process utilized key variables from the dataset, including blood pressure, age, glucose levels, number of pregnancies, BMI, and insulin levels. The analysis revealed that improved WAEL demonstrated superior performance in correctly diagnosing diabetes within the dataset used in this work. Experimental results suggest that the improved WAEL algorithm has potential as a reliable method for accurately classifying diabetes. Finally, we advise healthcare professionals to allocate more resources towards the exploitation of machine learning algorithms for timely detection of diabetes. Furthermore, it is critical for hospitals and health care organisations in Borno State to improve their data acquisition activities with patients. This has the potential to help scholars and scientists working to develop more effective diabetes detection and diagnosis methods.

Acknowledgement

This research received TETFUND institutional based grant from Borno State University, Nigeria.

References

[1] K. Azbeg, M. Boudhane, O. Ouchetto & S. Jai Andaloussi, "Diabetes emergency cases identification based on a statistical predictive model", Journal of Big Data **9** (2022) 1. https://doi.org/10.1186/s40537-022-00582-7.

[2] D. Hunt, K. Lamb, J. Elliott, B. Hemmingsen, S. Slama, R. Scibilia & B. Mikkelsen, "A WHO key informant language survey of people with lived experiences of diabetes: media misconceptions, values-based messaging, stigma, framings and communications considerations", Diabetes Research and Clinical Practice **193** (2022) 110109. https://doi.org/10.1016/j.diabres.2022.110109.

[3] A. M. Egan & S. F. Dinneen, "What is diabetes?", Medicine **47** (2019) 1. https://doi.org/10.1016/j.mpmed.2018.10.002.

[4] C. A. Robinson, G. Agarwal & K. Nerenberg, "Validating the CANRISK prognostic model for assessing diabetes risk in Canada's multi-ethnic population", Chronic Dis. Inj. Can. **32** (2011) 19. https://api.semanticscholar.org/CorpusID:12644541.

[5] L. M. Villeneuve & R. Natarajan, "The role of epigenetics in the pathology of diabetic complications", American Journal of Physiology-Renal Physiology **299** (2010) F14. https://doi.org/10.1152/ajprenal.00200.2010.

[6] M. R. Chetan, S. L. Thrower & P. Narendran, "What is type 1 diabetes?", Medicine **47** (2019) 5. https://doi.org/10.1016/j.mpmed.2018.10.006.

[7] M. G. Tinajero & V. S. Malik, "An update on the epidemiology of type 2 diabetes: a global perspective", Endocrinology and Meolism Clinics **50** (2021) 337. https://doi.org/10.1016/j.ecl.2021.05.013.

[8] V. Bellou, L. Belbasis, I. Tzoulaki & E. Evangelou, "Risk factors for type 2 diabetes mellitus: an exposure-wide umbrella review of meta-analyses", PloS one **13** (2018) e0194127. https://doi.org/10.1371/journal.pone.0194127.

[9] P. Saeedi, I. Petersohn, P. Salpea, B. Malanda, S. Karuranga & N. Unwin, "Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: results from the International Diabetes Federation Diabetes Atlas", Diabetes research and clinical practice **157** (2019) 107843. https://doi.org/10.1016/j.diabres.2019.107843.

[10] J. O. Adeleye, "The hazardous terrain of diabetes mellitus in Nigeria: the time for action is now", Research Journal of Health Sciences **9** (2021) 69. https://doi.org/10.4314/rejhs.v9i1.8.

[11] O. O. Oladapo, L. Salako, O. Sodiq, K. Shoyinka, K. Adedapo & A. O. Falase, "A prevalence of cardiometabolic risk factors among a rural Yoruba south-western Nigerian population: a population-based survey: cardiovascular topics", Cardiovascular Journal of Africa **21** (2010) 26. https://www.researchgate.net/publication/41911100_A_prevalence_of_cardiometabolic_risk_factors_among_a_rural_Yoruba_south-western_Nigerian_population_A_population-based_survey

[12] E. C. Ejim, C. I. Okafor, A. Emehel, A.U. Mbah, U. Onyia, T. Egwuonwu & B. J. Onwubere, "Prevalence of cardiovascular risk factors in the middle-aged and elderly population of a Nigerian rural community", Journal of tropical medicine **2011** (2011) 308687. https://doi.org/10.1155/2011/308687.

[13] A. Sabir, A. Ohwovoriole, S. Isezuo, O. Fasanmade, S. Abubakar & S. Iwuala, "Type 2 diabetes mellitus and its risk factors among the rural Fulanis of Northern Nigeria", Annals of African medicine **12** (2013) 217. https://doi.org/10.4103/1596-3519.122689.

[14] O. E. Enang, A. A. Otu, O. E. Essien, H. Okpara, O.A. Fasanmade, A. E. Ohwovoriole & J. Searle, "Prevalence of dysglycemia in Calabar: a cross-sectional observational study among residents of Calabar, Nigeria", BMJ Open Diabetes Research and Care **2** (2014) e000032. https://doi.org/10.1136/bmjdrc-2014-000032.

[15] A. E. Uloko, B. M. Musa, M. A. Ramalan, I. D. Gezawa, F. H. Puepet, A. T. Uloko & K.B. Sada, "Prevalence and risk factors for diabetes mellitus in Nigeria: a systematic review and meta-analysis", Diabetes Therapy **9** (2018) 1307. https://doi.org/10.1007/s13300-018-0441-1.

[16] I. D. Gezawa, F. H. Puepet, B. M. Mubi, A. E. Uloko, B. Bakki, M. A. Talle & I. Haliru, "Socio-demographic and anthropometric risk factors for type 2 diabetes in Maiduguri, North-Eastern Nigeria", Sahel Medical Journal **18** (2015) 1. https://doi.org/10.4103/1118-8561.149495.

[17] A. Sarwar, M. Ali, J. Manhas & V. Sharma, "Diagnosis of diabetes type-II using hybrid machine learning based ensemble model", International Journal of Information Technology **12** (2020) 419. https://doi.org/10.1007/s41870-018-0270-5.

[18] M. T. Alasaady, T. N. M. Aris, N. M. Sharef & H. Hamdan, "A proposed approach for diabetes diagnosis using neuro-fuzzy technique", Bulletin of Electrical Engineering and Informatics **11** (2022) 3590. https://doi.org/10.11591/eei.v11i6.4269.

[19] N. Abdulhadi & A. Al-Mousa, "Diabetes detection using machine learning classification methods", International Conference on Information Technology (ICIT), Amman, Jordan, 2021, pp. 350-354. https://doi.org/10.1109/ICIT52682.2021.9491788.

[20] U. E. Laila, K. Mahboob, A. W. Khan, F. Khan & W. Taekeun, "An ensemble approach to predict early-stage diabetes risk using machine learning: An empirical study", Sensors **22** (2022) 5247. https://doi.org/10.3390/s22145247.

[21] R. Katarya & S. Jain, "Comparison of different machine learning models for diabetes detection", IEEE International Conference on Advances and Developments in Electrical and Electronics Engineering (ICADEE), Coimbatore, India, 2020, pp. 1-5. https://doi.org/10.1109/ICADEE51157.2020.9368899.

[22] S. Y. Rubaiat, M. M. Rahman & M. K. Hasan, "Important feature selection and accuracy comparisons of different machine learning models for early diabetes detection", International Conference on Innovation in Engineering and Technology (ICIET), Dhaka, Bangladesh, 2018, pp. 1-6. https://doi.org/10.1109/CIET.2018.8660831.

[23] G. Swapna, R. Vinayakumar & K. P. Soman, "Diabetes detection using deep learning algorithms", ICT express **4** (2018) 243. https://doi.org/10.1016/j.icte.2018.10.005.

[24] M. M. Islam, R. Ferdousi, S. Rahman & H. Y. Bushra, "Likelihood prediction of diabetes at early stage using data mining techniques", Computer vision and machine intelligence in medical image analysis, Springer, Singapore, 2020, pp. 113-125. http://dx.doi.org/10.1007/978-981-13-8798-2_12.

[25] A. K. Shukla, "Patient diabetes forecasting based on machine learning approach", In Soft Computing: Theories and Applications: Proceedings of SoCTA **1154** (2020) 1017. https://doi.org/10.1007/978-981-15-4032-5_91.

[26] E. G. Dada, J. S. Bassi, H. Chiroma, A. O. Adetunmbi & O. E. Ajibuwa, "Machine learning for email spam filtering: review, approaches and open research problems", Heliyon **5** (2019). https://doi.org/10.1016/j.heliyon.2019.e01802.

[27] J. L. Speiser, M. E. Miller, J. Tooze & E. Ip, "A comparison of random forest variable selection methods for classification prediction modeling", Expert systems with applications **134** (2019) 93. https://doi.org/10.1016/j.eswa.2019.05.028.

[28] A. Taherkhani, G. Cosma & T. M. McGinnity, "AdaBoost-CNN: An adaptive boosting algorithm for convolutional neural networks to classify multi-class imbalanced datasets using transfer learning", Neurocomputing **404** (2020) 351.https://doi.org/10.1016/j.neucom.2020.03.064.

[29] G. Chen, H. He, L. Zhao, K. B. Chen, S. Li & C. Y. C. Chen, "Adaptive boost approach for possible leads of triple-negative breast cancer", Chemometrics and Intelligent Laboratory Systems **231** (2022) 104690. https://doi.org/10.1016/j.chemolab.2022.104690.

[30] E. G. Dada, D.O. Oyewola, S. B. Joseph, O. Emebo & O. O. Oluwagbemi, "Ensemble machine learning for monkeypox transmission time series forecasting", Applied Sciences **12** (2022) 12128. https://doi.org/10.3390/app122312128.

[31] C. N. Obiora, A. Ali & A. N. Hasan, "Implementing extreme gradient boosting (XGBoost) algorithm in predicting solar irradiance", IEEE PES/IAS PowerAfrica conference, Nairobi, Kenya, 2021, pp. 1-5. https://doi.org/10.1109/PowerAfrica52236.2021.9543159.

[32] A. Banerjee, A. Gajewicz-Skretna & K. Roy, "A machine learning q-RASPR approach for efficient predictions of the specific surface area of perovskites", Molecular Informatics **42** (2023) 2200261. https://doi.org/10.1002/minf.202200261.

[33] E. K. Sahin, "Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using XGBoost, gradient boosting machine, and random forest", SN Applied Sciences **2** (2020) 1308. https://doi.org/10.1007/s42452-020-3060-1.

[34] J. Dong, Y. Chen, B. Yao, X. Zhang & N. Zeng, "A neural network boosting regression model based on XGBoost", Applied Soft Computing **125** (2022) 109067. https://doi.org/10.1016/j.asoc.2022.109067.

[35] A. Sharma & B. Singh, "AE-LGBM: Sequence-based novel approach to detect interacting protein pairs via ensemble of autoencoder and LightGBM", Computers in Biology and Medicine **125** (2020) 103964. https://doi.org/10.1101/2020.07.03.186866.

[36] M. Massaoudi, S. S. Refaat, I. Chihi, M. Trabelsi, F. S. Oueslati & H. Abu-Rub, "A novel stacked generalization ensemble-based hybrid LGBM-XGB-MLP model for Short-Term Load Forecasting", Energy **214** (2021) 118874. https://doi.org/10.1016/j.energy.2020.118874.

[37] T. D. Pham, N. Yokoya, T. T. T. Nguyen, N. N. Le, N. T. Ha, J. Xia & T. D. Pham, "Improvement of mangrove soil carbon stocks estimation in North Vietnam using Sentinel-2 data and machine learning approach", GIScience & Remote Sensing **58** (2021) 68. https://doi.org/10.1080/15481603.2020.1857623.

[38] S. González, S. García, J. Del Ser, L. Rokach & F. Herrera, "A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities", Information Fusion **64** (2020) 205. https://doi.org/10.1016/j.inffus.2020.07.007.

[39] V. A. Dev & M. R. Eden, "Formation lithology classification using scalable gradient boosted decision trees", Computers and chemical engineering **128** (2019) 392. https://doi.org/10.1016/j.compchemeng.2019.06.001.

[40] P. Chen, Y. Deng, X. Zhang, L. Ma, Y. Yan, Y. Wu & C. Li, "Degradation trend prediction of pumped storage unit based on MIC-LGBM and VMD-GRU combined model", Energies **15** (2022) 605. https://doi.org/10.3390/en15020605.

[41] J. T. Hancock & T. M. Khoshgoftaar, "CatBoost for big data: an interdisciplinary review", Journal of big data **7** (2020) 1. https://doi.org/10.1186/s40537-020-00369-8.

[42] H. Zeng, B. Shao, H. Dai, Y. Yan & N. Tian, "Prediction of fluctuation loads based on GARCH family-CatBoost-CNNLSTM", Energy **263** (2023) 126125. https://doi.org/10.1016/j.energy.2022.126125.

[43] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush & A. Gulin, "CatBoost: unbiased boosting with categorical features", Advances in neural information processing systems **31** (2018) 6639. https://dl.acm.org/doi/abs/10.5555/3327757.3327770.

[44] H. Luo, F. Cheng, H. Yu & Y. Yi, "SDTR: Soft decision tree regressor for tabular data", IEEE Access, **9** (2021) 55999. https://doi.org/10.1109/ACCESS.2021.3070575.

[45] R. Yang, P. Wang & J. Qi, "A novel SSA-CatBoost machine learning model for credit rating", Journal of Intelligent and Fuzzy Systems **44** (2023) 1. http://doi.org/10.3233/JIFS-221652.

[46] R. Olsson & S. Acharya, "Using automatic programming to improve gradient boosting for classification", International Conference of Artificial Intelligence and Soft Computing (ICAISC), Zakopane, Poland, 2022, pp. 242–253. https://doi.org/10.1007/978-3-031-23492-7_21.

[47] S. Mehta, P. Rana, S. Singh, A. Sharma & P. Agarwal, "Ensemble learning approach for enhanced stock prediction", twelfth international conference on contemporary computing (IC3), Noida, India, 2019, pp. 1-5. http://dx.doi.org/10.1109/IC3.2019.8844891.

[48] M. A. Ganaie, M. Hu, A. K. Malik, M. Tanveer & P.N. Suganthan, "Ensemble deep learning: A review", Engineering Applications of Artificial Intelligence **115** (2022) 105151. https://doi.org/10.1016/j.engappai.2022.105151.

[49] M. A. I. Neloy, N. Nahar, M. S. Hossain & K. Andersson, "A weighted average ensemble technique to predict heart disease", International Conference on Trends in Computational and Cognitive Engineering (TCCE) , Tangail, Bangladesh, 2022, pp. 17–29. https://doi.org/10.1007/978-981-16-7597-3_2.

[50] H. Bonab & F. Can, "Less is more: A comprehensive framework for the number of components of ensemble classifiers", IEEE Transactions on neural networks and learning systems **30** (2019) 2735. https://doi.org/10.48550/arXiv.1709.02925.

[51] Z. H. Zhou, *Ensemble learning*, In: Li, S.Z., Jain, A. (eds) Encyclopedia of Biometrics, Springer, Boston, MA, 2009, pp. 270–273. https://doi.org/10.1007/978-0-387-73003-5_293.

[52] C. Molnar, "Interpretable machine learning", Metamorphosis: A Journal of Management Research **23** (2020) 318. https://doi.org/10.1177/09726225241252009

[53] B. M. Cheung & C. Li, "Diabetes and hypertension: is there a common metabolic pathway?", Current atherosclerosis reports **14** (2012) 160. https://doi.org/10.1007/s11883-012-0227-2.

[54] M. J. Cryer, T. Horani & D. J. DiPette, "Diabetes and hypertension: a comparative review of current guidelines", The Journal of Clinical Hypertension **18** (2016) 95. https://doi.org/10.1111/jch.12638.

[55] N. P. Chokshi, E. Grossman & F. H. Messerli, "Blood pressure and diabetes: vicious twins", Heart **99** (2013) 577. https://doi.org/10.1136/heartjnl-2012-302029.

[56] A. Lazar, A. Sim & K. Wu, "GPU-based classification for wireless intrusion detection", Systems and Network Telemetry and Analytics, Virtual, 2020, pp. 27–31. https://doi.org/10.1145/3452411.3464445

[57] A. Dutta, M. K. Hasan, M. Ahmad, M. A. Awal, M. A. Islam, M. Masud & H. Meshref, "Early prediction of diabetes using an ensemble of machine learning models", International Journal of Environmental Research and Public Health **19** (2022) 12378. https://doi.org/10.3390/ijerph191912378.

[58] R. F. Albadri, S. M. Awad, A.S. Hameed, T. H. Mandeel & R. A. Jabbar, "A diabetes prediction model using hybrid machine learning algorithm", Mathematical Modelling of Engineering Problems **11** (2024) 2119. https://doi.org/10.18280/mmep.110813.

[59] R. Karthikeyan, P. Geetha & E. Ramaraj, "Prediction of diabetes and cholesterol diseases based on ensemble learning techniques", International Journal of Scientific and Technology Research **9** (2020) 491. https://api.semanticscholar.org/CorpusID:214635142.

[60] M. Li, X. Fu & D. Li, "Diabetes prediction based on XGBoost algorithm", IOP Conf. Ser. Mater. Sci. Eng. **768** (2020) 072093. https://doi.org/10.1088/1757-899x/768/7/072093.

[61] Z. Mushtaq, M. F. Ramzan, S. Ali, S. Baseer, A. Samad & M. Husnain, "Voting classification-based diabetes mellitus prediction using hypertuned machine-learning techniques", Mob. Inf. Syst. **18** (2022) 1. https://doi.org/10.1155/2022/6521532.

[62] M. Atif, F. Anwer & F. Talib, "An ensemble learning approach for effective prediction of diabetes mellitus using hard voting classifier", Indian Journal of Science and Technology **15** (2022) 1978. https://doi.org/10.17485/IJST/v15i39.1520.

[63] T. T. Aurpa, S. M. Jeba & S. U. Rasel, "Ensemble Methods of Machine Learning Algorithms for Early Diabetic Detection in Comparison", International Conference on Circuits, Power and Intelligent Systems (CCPIS), Bhubaneswar, India, 2023, pp. 1-6. https://doi.org/10.1109/CCPIS59145.2023.10291566.

[64] Q. Saihood & E. Sonuç, "A practical framework for early detection of diabetes using ensemble machine learning models", Turkish Journal of

Electrical Engineering and Computer Sciences **31** (2023) 722. https://doi.org/10.55730/13000632.4013.

[65] S. Mirjalili, S.M. Mirjalili & A. Lewis, "Grey wolf optimizer", Adv Eng Softw **69** (2014) 46. https://doi.org/10.1016/j.advengsoft.2013.12.007.