Mikailalsys Journal of Mathematics and Statistics

e-ISSN : 3030-816X
p-ISSN : 3030-8399

Index : Harvard, Boston, Sydney University, Dimensions, Lens, Scilit, Semantic, Google, etc

https://doi.org/10.58578/MJMS.v2i2.2875

ENSEMBLE MACHINE LEARNING ALGORITHM FOR DIABETES PREDICTION IN MAIDUGURI, BORNO STATE

Emmanuel Gbenga Dada[1], Aishatu Ibrahim Birma[2], Abdulkarim Abbas Gora[3]
[1]University of Maiduguri, Maiduguri, Nigeria; [2,3]Borno State University, Maiduguri, Nigeria
gbengadada@unimaid.edu.ng; ayshabirma13@hotmail.com

Article Info:

Submitted:	Revised:	Accepted:	Published:
Mar 25, 2024	Apr 13, 2024	Apr 17, 2024	Apr 20, 2024

Abstract

Diabetes mellitus (DM) is a metabolic disease characterised by high levels of glucose in the blood, known as hyperglycemia, that can result in multiple problems within the body. The World Health Organisation (WHO) data for 2021 reveals a substantial increase in the prevalence of diabetes mellitus (DM), with the number of cases rising from 108 million in 1980 to 422 million in 2014. Between 2000 and 2019, there was a 3% increase in mortality rates associated with diabetes, categorised by age. In 2019, DM caused the deaths of more than 2 million people. These concerning figures clearly necessitate an immediate response. An alarming incidence of diabetes among the population of Maiduguri and Borno State inspired this investigation. This research proposed stacking ensemble learning approach to predict the rate of occurrence of diabetes cases in Maiduguri. The paper used different types of regression models to predict the occurrences of diabetes cases in Maiduguri over time. These models included adaptive boosting regression (Adaboost), gradient boosting regression (GBOOST), random forest regression (RFR), ordinary least square regression (OLS), least absolute shrinkage selection operator regression (LASSO), and ridge regression (RIDGE). The performance indicators studied in this work are root mean square (RMSE), mean absolute error (MAE), and mean square error (MSE). These metrics were used to assess the effectiveness of both the machine learning and proposed

MJMS Journal is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License

Stacking Ensemble Learning (SEL) approaches. Performance metrics considered in this study are root mean square (RMSE), mean absolute error (MAE), and mean square error (MSE), which were used to evaluate the performance of the machine learning and the proposed Stacking Ensemble Learning (SEL) technique. Experimental results revealed that SEL is a better predictor compared to other machine learning approaches considered in this work with an RMSE of 0.0493; a MSE of 0.0024; and a MAE of 0.0349. It is hoped that this research will help government officials understand the threat of diabetes and take the necessary mitigation actions.

Keywords: Ensemble learning, Diabetes, Stacking ensemble learning, Random forests, Gradient boost regressor

INTRODUCTION

Diabetes is a prevalent chronic disease that necessitates patients to continually adopt techniques for self-management for effective treatment (Azbeg et al., 2022; Diabetes, 2022). Diabetes is defined by the presence of hyperglycemia, which refers to an abnormally high concentration of sugar in the bloodstream (CDC, 2021). Diabetes, as defined by the World Health Organisation (WHO), is a chronic condition marked by inadequate release of insulin from the pancreas into the blood. In addition, it can occur when the body is unable to efficiently utilise the insulin it produces (Diabetes, 2022). Insulin is a hormone that controls the levels of glucose in the blood. Hyperglycemia, also referred to as high blood sugar levels, is a common outcome of uncontrolled diabetes and progressively leads to substantial damage to several biological systems, particularly the neurons and blood vessels. The main determinant of this syndrome is hereditary (Robinson et al., 2011); nevertheless, triggers from the environment can have some impact (Nahla et al., 2010).

There are four discrete types of diabetes: Roughly 5–10% of all verified cases of diabetes can be attributed to Type 1, commonly known as adolescent diabetes or insulin-dependent diabetes. Type 2 diabetes is defined by its onset in adulthood and its lack of need for insulin. Type 1 diabetes, often known as juvenile diabetes, usually begins before the age of 20 (Type 1 diabetes, 2022). An autoimmune disorder occurs when the immune system attacks the body's own tissues. In those suffering from type 1 diabetes, the pancreatic cells that are accountable for producing insulin suffer necrosis. However, type 2 diabetes usually manifests after the age of 30 and is sometimes called "old-age diabetes." However, young

individuals are prone to influence. Genetic factors, obesity, and inadequate cardiovascular exercise impact Type 2 diabetes (Type 2 diabetes, 2022).

The management and control of this condition are considered major public health concerns, and the number of cases is quickly rising. Diabetes is a life-threatening and harmful condition that is increasingly common, particularly in emerging countries and among economically disadvantaged populations. The worldwide incidence of diabetes has been steadily increasing and is expected to continue rising in the next few years (Tinajero & Malik, 2021; Saeedi, et al., 2019). The incidence of diabetes among Africans is increasing, and predictions indicate that the African continent will experience the highest growth rate (143%) in diabetes patients from 2019 to 2045.

The death rates linked to diabetes, classified by age, witnessed a 3% growth from 2000 to 2019. Approximately 2 million people succumbed to diabetes and diabetes-related kidney illnesses in 2019. It is imperative to consistently guarantee universal access to healthcare facilities for all individuals. However, there are certain individuals who have the right to use these facilities but live far away. Furthermore, the scarcity of resources necessitates a continuous need for proficient medical professionals to attend to pressing and significant concerns.

Adeleye (2021) describes a consistent increase in the prevalence of diabetes mellitus in Nigeria. The prevalence of DM in specific villages in Nigeria varies from 0.8% to 4.4%, as reported by Oladapo et al. (2010), Ejim et al. (2011), and Sabir et al. (2013). Sabir et al. (2011) and Enang et al. (2014) reported a frequency range of 4.6% to 7% in urban areas. Uloko et al. (2018) did an in-depth meta-analysis and review and found that the prevalence of diabetic complications among Nigerians is 5.77%. In 2019, Tinajero and Malik (2021) estimated that impaired sugar tolerance affected 8.2 million Nigerians, with a projected increase to 11.5 million by 2030. Gezawa et al. (2015) have demonstrated a notable prevalence of type 2 diabetes in the urban area of Maiduguri. Given the aforementioned knowledge, it is imperative to create a computerised system that may assist healthcare professionals in delivering medical services, particularly in the diagnosis of diabetes. This system would be especially advantageous in locations with restricted accessibility and limited availability of medical services, where there is a lack of skilled specialists, clinical decision support (CDS) systems, and electronic diabetes diagnosis systems. The contributions of this paper are summarised as follows:

i. The paper described an overview of ensemble machine learning methods that can be applied for diabetes prediction and management.
ii. The study identified the most accurate prediction model for diabetes cases in different areas of Maiduguri and its surrounding areas.
iii. A novel dataset of diabetic patients was created using data from the University of Maiduguri Teaching Hospital and Umaru Shehu Specialist Hospital in Maiduguri, Borno State, Nigeria, between 2018 and 2023. The collection contains information about 1030 patients.
iv. A variety of machine learning methods were utilised to construct a model for the prediction of diabetes in Maiduguri and its surrounding areas. Some of the models used to predict diabetes are adaptive boosting regression, gradient boosting regression, random forest regression, the least absolute shrinkage selection operator, ridge regression, conventional least squares regression, and new stacking ensemble learning approaches.
v. The study evaluated the effectiveness of the proposed stacking ensemble machine learning technique for predicting diabetes. The study provides insightful knowledge into the benefits and limitations of the proposed paradigm.

The paper is structured as follows in subsequent sections: Section 2 of the document contains the review of related work and a summary of contributions table. In Section 3, the discussion revolved around the machine learning models employed for prediction. The simulation and numerical results from the experiment analyses are reported in Section 4. Section 5 contains the paper's conclusion.

Summarily, this study introduces a novel approach for predicting diabetes through the analysis of patient data collected from individuals residing in Maiduguri and its surrounding regions in Nigeria. The findings suggest that the stacking ensemble machine learning technique is a better predictor compared to other ensemble approaches. The newly acquired dataset has the potential to be deployed in the future for training various algorithms aimed at detecting and predicting diabetes. This can serve as an effective tool for decision support system in diabetes detection and diagnosis.

Sarwar et al. (2020) proposed a hybrid ensemble model that employs machine learning approaches to accurately detect instances of type 2 diabetes. The authors utilised several machine learning classifiers. The authors constructed the models using the MATLAB and WEKA 3.6.13 software platforms. The study's database consists of around 400 individuals chosen from a wide geographical region, covering ten various physiological traits. The simulation results showed that the ensemble approach outperformed the other models used in the study, with an average accuracy of 98.60%. The drawback of this research is the relatively small size of the dataset utilised. Moreover, there is a need to improve the accuracy and dependability of the system. Alasaady et al. (2022) employed an adaptive neurofuzzy (ANFIS) method to diagnose diabetes. They trained the ANFIS model using the hybrid neural network algorithm. The statistical results demonstrate that the proposed model achieved a classification accuracy of 92.77%. An inherent limitation of the research is the relatively small size of the dataset utilised, which hinders the generalizability of the findings to a larger population. Moreover, the analysis primarily emphasised diagnostic techniques rather than predictive modelling. Moreover, it is necessary to improve the accuracy of the system. Abdulhadi and Al-Mousa (2021) utilised machine learning algorithms to predict the likelihood of diabetes, specifically emphasising early identification in females. The random forest model achieved an accuracy rate of 82%, making it the most desirable result among the several models being evaluated. The small size of the dataset used limits the study. Furthermore, the suggested approach exhibits a notable inadequacy in terms of precision.

Laila et al. (2022) employed ensemble machine learning models to carry out a scientific investigation into diabetes. Laila et al. (2022) obtained the dataset for this investigation from the UCI repository. The diabetes dataset consists of a total of 17 variables. The prediction was performed using machine learning methods, specifically AdaBoost, Bagging, and RF. Diverse metrics were employed to assess the efficacy of the framework. The simulation results demonstrated that the ensemble technique, namely the random forest model, outperformed the other models investigated in the study, attaining an accuracy rate of 97%. A possible limitation of this study is the relatively small size of the dataset used. Furthermore, enhancing the accuracy of the proposed system is crucial. Katarya and Jain (2020) utilised a variety of machine learning methods to diagnose diabetes in their research. The experiments used the Indian Pima dataset. The results of the study showed that the RF ensemble technique outperformed the other models. The dataset employed is

comparatively restricted in size, which is an inconvenience of the effort. Furthermore, it is important to acknowledge that the suggested approach demonstrates a rather low level of accuracy. The diabetes patient data acquired from the hospital was not efficiently utilised.

Researchers Rubaiat, Rahman, and Hasan (2018) conducted a comparative study to assess the efficacy of various machine learning models in extracting valuable features from a dataset related to diabetes. The algorithms were also used to predict the occurrence of diabetes in patients. The experimental research demonstrated that using a multilayer perceptron (MLP) in combination with a feature selection strategy produced better results compared to other strategies used in the study. The work is inadequate due to the relatively small size of the dataset used. Furthermore, the performance of the proposed system demonstrates subpar values. In addition, the study did not incorporate authentic patient data from individuals diagnosed with diabetes obtained from a medical facility.

Swapna, Vinayakumar, and Soman (2018) utilised deep learning methodologies to diagnose diabetes. The study utilised diverse deep learning methodologies to extract characteristics from heart rate variability (HRV) data. The support vector machine (SVM) for classification requires the retrieved features as input. The CNN model (0.03%) and the CNN-LSTM model (0.06%) did a little better than the deep learning models used in previous studies that did not use SVM, according to the researchers. The proposed research has the capacity to aid healthcare professionals in making well-informed judgements regarding patient care. The study of ECG signals can assist in the detection of diabetes, with an achieved accuracy rate of around 95.7%. The study is significantly limited by the extremely small size of the dataset used. Moreover, there is a requirement for additional enhancement in the efficiency of the suggested system. Azbeg et al. (2022) employed a probabilistic predictive model in a separate investigation to detect occurrences of diabetes-related emergencies. The authors proposed a system architecture utilising the Internet of Things (IoT) to ensure the collection of patient data for predicting key occurrences of diabetes. To ensure data security, the system has implemented blockchain and IPFS. Furthermore, in order to analyse the data, a statistical approach based on predictive modelling has been proposed. The method's performance was evaluated and compared to other existing prediction approaches. The experiment produced accuracy rates of 85.9%, 99.5%, and 99.8% for the three datasets, respectively. Hence, the suggested framework might be used to forecast diabetes and alert medical professionals or healthcare establishments about urgent situations that demand prompt

action. One drawback of the work is the lack of any innovative algorithmic advancement. Islam et al. (2020) employed data mining methodologies to forecast diabetes. 520 subjects completed questionnaires regarding potential characteristics that could influence the initial phases of diabetes. Following the data preparation procedure, the gathered data exhibited a total of 314 instances with positive outcomes and 186 instances with negative outcomes. Positive numerical numbers indicate the presence of diabetes in an individual, whereas negative values indicate the absence of the disease. The Random Forest (RF) algorithm exhibited exceptional performance, with a detection rate of 99%. Therefore, this approach demonstrates its efficacy when used on an existing dataset. However, one possible limitation of the study is the relatively small size of the dataset used. Shukla (2020) predicted the occurrence of diabetes by employing a linear regression model. The researchers employed a dataset that showed that characteristics such as glucose levels, body mass index (BMI), and pregnancy status would result in the greatest level of accuracy. Medical specialists have acknowledged that the sickness mostly depends on features that may appear insignificant to us but can increase vulnerability to future ailments. The logistic regression model achieved an accuracy rate of 82.92% by training it with the most impactful features. An inherent limitation of this work is the comparatively lower level of precision.

The motivation for using ensemble learning in the current research arises from the observation that machine learning models have typically been constructed based on the assumption that a model will achieve its highest level of accuracy when trained and tested on data that originates from the same feature space and distribution. However, if there are changes in the feature space or data distribution, it is necessary to create a new model. The costs related to creating a new model from scratch each time, together with obtaining new training data, are substantial. Ensemble learning enables the efficient acquisition of large amounts of training data by reducing the required effort and time. Ensemble learning possesses the potential to utilise pre-existing data in order to address new tasks or domains. The application of gained knowledge empowers individuals to tackle new challenges better and effectively.

METHODS

1. Adaptive Boosting Regression (AdaBoost)

In the field of machine learning, the AdaBoost algorithm, commonly referred to as adaptive boosting, employs a boosting approach as an ensemble method. AdaBoost reallocates the weights to each instance, assigning greater weights to examples that were incorrectly identified, hence earning the moniker "adaptive boosting." Supervised learning employs boosting to reduce both bias and variation. It functions based on the assumption that learners make gradual improvements. The adaptive boosting regression meta-estimator (Barrow & Crone, 2016) begins by training a regressor on the initial dataset. Subsequently, it applies more iterations of the regressor to the existing dataset. The algorithm alters the weights of the occurrences based on the error of the most recent forecast. Let us examine a dataset $S = (x_1, y_1), \ldots, (x_n, y_n)$ that is obtained through a time series. The dataset consists of n sets of observations, with each observation assigned a weight, w_i. For each observation i, we determine the probability of including it in the training set during iteration k based on its assigned weight. Calculating the weighted sum of the probabilities obtains the average loss(l_k) for the model k over all the observations i. The mathematical equations for the average loss (l_k) and probability p^k are as follows:

$$p_k = \frac{w_i}{\sum w_i} \quad (1)$$

$$l_k = \sum_{i=1}^{n} l_k p_k \quad (2)$$

$$w_i^{k+1} = w_i^k \beta_k (1 - l_k) \quad (3)$$

The equation defines the variables as follows: p_k represents the probability at iteration k, w_i^{k+1} denotes the average loss at iteration k, w_i^k represents the prior weight at iteration i, and β_k represents the model loss.

2. Gradient Boosting Regression (GBoost)

Gradient boosting employs a machine learning technique for tasks such as regression and classification problems. The system offers a predictive model that consists of a collection of weak prediction models resembling decision trees (Oyewola *et al.,* 2021). Gradient boosting iteratively chooses a function that opposes the gradient direction in order to optimise a cost function over the whole function space. Gradient boosting (Oluwagbemi *et al.,* 2010) is a widely used machine learning technique that constructs a final prediction

model by combining weak predictors in an ensemble. It is commonly used for regression and classification tasks. Gradient boosting commonly employs decision trees as weak predictors. Weakly learned models exhibit minimal variance and regularisation, a significant bias towards the training dataset, and outputs that only marginally outperform random guesses. Boosting strategies consist of three key components: an additive model, weak learners, and a loss function. Gradient-boosting machines use gradients to identify deficiencies in subpar models. This is achieved through the use of an iterative approach, with the objective of ultimately combining base learners to minimise prediction errors. In this process, decision trees are combined using an additive model, and the loss function is minimised through the utilisation of gradient descent. The gradient boosting tree (g) can be defined as follows:

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

$$argmin \sum_t L(y_t, g) + f_{n+1}\, x_t \tag{5}$$

Let g represent the gradient boosting tree, $L()$ denote the loss function, and $f_{n+1}x_t$ represent the newly generated decision tree.

3. Random Forest Regression

Random Forest Regression is a robust machine learning approach used for regression-related tasks that involve predicting continuous numerical data (Dada *et al.*, 2021). It falls under the domain of ensemble learning, where it combines numerous decision trees to create predictions. During the training phase, random forest regression constructs numerous decision trees (Lingjun *et al.*, 2019; Xue *et al.*, 2021). Each tree trains on a randomly selected portion of the data and selects a random collection of features to split at each node. The procedure utilises the method of bootstrap aggregating or bagging, wherein each decision tree is trained on a randomly selected subset of the dataset with replacement. This aids in mitigating overfitting and enhances generalisation. During the prediction phase, each decision tree within the ensemble generates a forecast by utilising the input information. Each individual tree in the forest makes predictions, and the ultimate forecast is derived by averaging these predictions (in the case of regression), resulting in a more resilient and precise forecast. The ensemble approach and bagging technique employed in random forest regression make it less susceptible to overfitting than standalone decision

trees. The model is capable of accurately capturing complicated and nonlinear patterns in datasets by successfully modelling the connections between input features and the target variable. Random Forest Regression is capable of efficiently handling large datasets with an enormous number of features, making it suitable for scaling to real-world applications (Gieseke & Igel, 2018).

4. Ordinary Least Square Regression (OLS)

Ordinary least squares (OLS) regression, commonly referred to as linear regression, is a fundamental statistical method employed to describe the association between a number of independent variables (predictors) and a dependent variable (outcome) (Gomila, 2021). The objective is to identify the most suitable linear equation that accurately represents the correlation between the variables. Ordinary least squares (OLS) regression aims to minimise the total sum of squared residuals, which are the differences between the observed and projected values of y. Maximum Likelihood and the expanded technique of moments estimator are two alternatives to OLS.

$$Y = \beta_o + \sum_{j=1} \beta_j x_j + \varepsilon \quad (6)$$

where Y is the dependent variable, β, is the intercept of the model, x_j corresponds to the jth explanatory variable of the model and ε is the random error.

The residuals are the discrepancies between the observed values of *y* and the expected values derived from the linear equation. OLS regression assumes that the independent and dependent variables are linear, that errors are distributed normally, and that the variance of errors stays the same (homoscedasticity). OLS regression yields coefficients that are readily interpretable and indicate the size and direction of the impact of each independent variable on the dependent variable (Keele, Stevenson & Elwert, 2020). Deviation from such presumptions can impact the precision and dependability of the regression findings. Its ease of implementation and interpretation contribute to the widespread usage of OLS regression in the fields of statistics and econometrics. OLS regression efficiently predicts the coefficients, particularly for large datasets, and offers normal errors and confidence ranges to measure the uncertainty of the results. OLS regression can be used with different types of data and is suitable for both straightforward and sophisticated regression models (Darlington & Hayes, 2016).

5. Least Absolute Shrinkage Selection Operator Regression (LASSO)

LASSO regression is a linear regression method that is employed for both the selection of features and normalisation (Shafiee *et al.*, 2021). The objective is to identify the subset of predictors that are most pertinent for forecasting the result while simultaneously minimising the intricacy of the model. LASSO regression employs optimisation methods such as coordinate descent or gradient descent to get the coefficients βj that minimise the objective function. LASSO regression employs automatic feature selection by assigning zero coefficients to certain variables, thereby identifying the most pertinent features. By prioritising the most significant predictors, this approach can enhance model interpretability and mitigate overfitting. LASSO regression incorporates regularisation techniques to mitigate overfitting and enhance the model's ability to generalise (Czajkowski, Jurczuk & Kretowski, 2023). LASSO regression generates models that have a reduced number of nonzero coefficients, resulting in a simpler and more interpretable model. This is particularly useful in datasets with many features. LASSO regression efficiently addresses multicollinearity, which refers to the presence of a significant correlation between independent variables. It achieves this by picking one of the correlated variables and reducing the influence of the others towards zero (Wang *et al.*, 2024). The following objective function is minimized to get the regression coefficient:

$$lasso = \left(\sum_{i=1}^{n} y_i - \beta_o - \sum_{j=1}^{p} \beta_j x_{ij}\right)^2 + \gamma \sum_{j=1}^{p} |\beta_j| \quad (7)$$

$$\sum_{j=1}^{p} |\beta_j| \leq \gamma \quad (8)$$

where β_j is the regression coefficient operating on the standardized covariate j, β_o is the intercept and γ is a penalty term which controls the value of shrinkage.

To summarise, LASSO regression is a valuable method for selecting features and applying regularisation, especially in situations involving highly dimensional data and multiple correlations. It enhances the interpretability of models and improves their capacity to generalise.

6. Ridge Regression (Ridge)

Ridge regression, sometimes referred to as Tikhonov regularisation or L2 regularisation, is a linear regression method employed to regularise and address multicollinearity in

predictive modelling (Rokem & Kay, 2020). Lasso regression is akin to ordinary least squares (OLS) regression, but it incorporates a penalty component into the objective function to constrain the coefficients and mitigate overfitting. Ridge Regression employs optimisation methods, such as gradient descent or closed-form solutions (e.g., singular value decomposition), to determine the coefficients βj that minimise the objective function. Ridge regression incorporates regularisation to mitigate overfitting and enhance the generalisation capability of the model, particularly in datasets with a large number of features (la Tour *et al.*, 2022).

Ridge regression mitigates the issue of multicollinearity, which refers to a high correlation between independent variables, by applying a technique that reduces the coefficients towards zero. This process aids in stabilising the estimations of the coefficients and mitigates their susceptibility to minor fluctuations in the data. Ridge regression generates models with reduced coefficients, resulting in a more streamlined and easily understandable representation, particularly in situations where there are numerous linked features.

The Ridge Regression approach [38] can be used to analyse data from multiple regressions that show multicollinearity. While least-squares approximations remain impartial in the presence of multicollinearity, their substantial variances increase the likelihood of a significant deviation from the true value. Ridge regression reduces the standard errors by including a certain amount of prejudice in the regression values. Ultimately, the expectation is that this will lead to more accurate forecasts. Consider a regression equation:

$$Y = X^{-1}\beta + \varepsilon \tag{9}$$

Regression coefficients of ordinary least square is:

$$\hat{\beta} = (X^{-1}X)^{-1}X'Y \tag{10}$$

The variance covariance matrix of the estimate is:

$$V(\hat{\beta}) = \sigma^2 R^{-1} \tag{11}$$

From the above, we find that:

$$V(\widehat{\beta_J}) = r^{ij} = \frac{1}{1-R^2} \tag{12}$$

The amount of bias in this estimator is given by:

$$E(\tilde{\beta} - \beta) = [X'X + kI)^{-1}X'X - I]\beta \tag{13}$$

The covariance matrix is given by:

$$V(\tilde{\beta}) = (X'X + kI)^{-1}\ X'X(X'X + kI) \quad (14)$$

Y represents the dependent variable, while X represents the independent variables. β refers to the regression coefficients that need to be projected, while ε denotes the errors or leftovers.

7. Proposed Model

Stacking Ensemble Learning (SEL)

Ensemble learning is a method in machine learning that merges the predictions of numerous separate predictive models (learners) to get a final prediction (Oyewola & Dada, 2022; Dada, Yakubu & Oyewola, 2021). Ensemble learning aims to enhance prediction performance, robustness, and generalisation ability by using the combined knowledge of various models. Ensemble learning commences by training numerous base models, sometimes referred to as base learners or weak learners, on the training dataset. These foundational models might vary in nature or be trained using diverse strategies. Ensuring diversity in the base models is crucial, as it allows for the inclusion of various elements of the data and enables predictions to be made based on various subgroups of features. The inclusion of diverse elements in an ensemble helps mitigate the problem of overfitting and enhances the overall performance of the ensemble. After training the base models, their predictions are merged or consolidated to provide the ultimate forecast. The process of aggregation may encompass the computation of the average of predictions, the determination of the majority vote (for classification problems), or the utilisation of more advanced methods such as weighted averaging or stacking. Combining the predictions of the base models determines the final forecast of the ensemble model. The ultimate forecast is frequently more precise and dependable compared to the forecasts made by individual base models, since it capitalises on the advantages of several models while minimising their limitations (Wang *et al.*, 2018; Moon *et al.*, 2020).

Stacking, often referred to as stacked generalisation or stacked ensemble learning, is an effective ensemble learning strategy utilised to enhance the prediction performance of machine learning models. The process involves the integration of several base models, or learners, to create a meta-model, also known as a meta-learner, which consolidates their

individual predictions. The process of stacking involves training multiple distinct base models using the training dataset. The base models can vary in their kinds, such as decision trees, support vector machines, and neural networks. Additionally, researchers may have trained them using various algorithms or with different hyperparameters. After training the base models, each of them generates predictions on the validation dataset, also known as out-of-fold predictions. The predictions are used as input features for the meta-model.

A meta-model, also known as a meta-learner, is trained by using the predictions made by the base models as input features and the true labels from the validation dataset as the target variable. The meta-model acquires the ability to amalgamate the forecasts generated by the basis models in order to formulate the ultimate prediction. When making predictions on new data, the base models generate distinct predictions, which the trained meta-model subsequently merges to provide the ultimate forecast. Stacking frequently leads to enhanced prediction performance in comparison to individual base models since it capitalises on the strengths of several models and mitigates their deficiencies. Stacking enhances model diversity by amalgamating various base models, hence potentially improving generalisation performance and fortifying resilience against diverse data and patterns. Stacking is a versatile technique that can adapt to many types of base models and meta-models, enabling practitioners to explore new methods and designs. Stacking is a technique that achieves a balance between bias and variance by merging models that possess distinct biases and variances. This frequently leads to a predictive model that is more stable and dependable (Wang *et al.,* 2023).

This work employed stacked ensemble learning (SEL) models to predict the occurrence of diabetes among the populace in Maiduguri, Borno State. Stacking is a widely recognised technique in machine learning that involves using ensemble modelling. The process involves integrating many weak learners simultaneously and then using meta-learners to improve future predictions (Mienye & Sun, 2022). The ensemble technique operates by amalgamating the predictions of multiple feeble learners and meta-learners to generate a higher-output prediction model. Stacking is a method in which an algorithm uses the outputs of sub-models as input and attempts to learn how to effectively merge the input predictions in order to achieve a superior output prediction. It is called "stacking" because it is positioned above the other models.

Architecture of Stacking

The stacking model architecture incorporates six base learner models: adaptive boosting regression (Adaboost), gradient boosting regression (GBOOST), random forest regression, ordinary least square regression (OLS), least absolute shrinkage selection operator regression (LASSO), and ridge regression (RIDGE). Additionally, it includes a random forest meta-model that aggregates the predictions from the base models. The level 0 models serve as the foundational models, whilst the level 1 model functions as the meta-model. The stacking ensemble approach comprises the initial training data, primary level models, primary level predictions, secondary level models, and the final prediction. The underlying framework of the stacking architecture is as follows:

- Original data: The dataset is partitioned into training data and test data.
- Base models: The Level-0 models consist of Adaboost, GBOOST, RFR, OLS, LASSO, and RIDGE regression techniques. These models utilise training data to generate aggregated predictions at level 0.
- Level-0 Predictions: Applying a base model to a set of training data generates several level-0 predictions.
- Meta Model: The stacking model's approach incorporates only one meta-model that uses random forest regression to efficiently combine the predictions of the base models. The level-1 model is also known as the meta-model.
- Level-1 Prediction: The meta-model acquires the ability to effectively amalgamate the predictions generated by the base models and is trained using the diverse predictions produced by each unique base model. For example, the meta-model is given data that was not used to train the base models. It makes predictions using this data, and these predictions, along with the predicted outputs, are used as input and output pairs to train the meta-model. Refer to Figure 1 for an illustration of the proposed system's architecture.

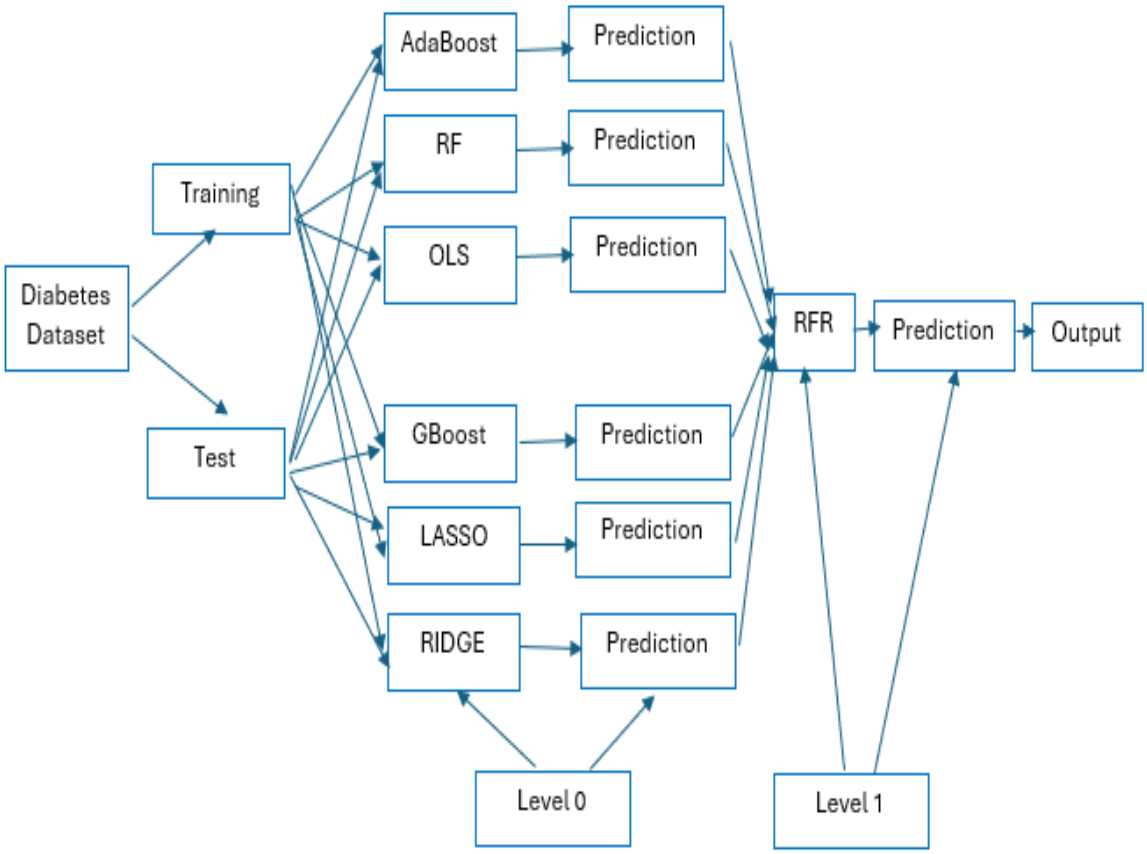


Figure 1: Proposed Architecture for Stacking Ensemble Learning

The proposed system makes use of stacking ensemble learning (SEL) techniques for time series forecasting. The diabetes prediction method under consideration uses ensemble learning approaches based on stacking for the purpose of prediction. The simulations use diabetes datasets acquired from various hospitals in Maiduguri. The dataset was partitioned into two sets, namely the training set, which included 80% of the data, and the test set, which included the remaining 20%. The overall structure of the proposed ensemble stacking system incorporated a total of six models. The main objective of this paper is to use the dataset to predict the risk of diabetes in patients using SEL. This is accomplished by arranging the six models (adaptive boosting regression, gradient boosting regression, random forest regression, regularised regression using the least absolute shrinkage selection operator, ridge regression, and ordinary least squares stacking ensemble learning methods) in a stacked manner. This approach is commonly used in both regression and classification tasks.

8. Dataset Description

The diabetes datasets used in this study were acquired from the University of Maiduguri teaching hospital and Umaru Shehu specialty hospital, Maiduguri. The datasets contain data on patients with diabetes collected from 2018 to 2023. The dataset was gathered from individuals aged 17 and above residing in Maiduguri, Borno State, Nigeria, and the surrounding areas, encompassing both males and females. The dataset comprises certain diagnostic measurements that serve as feature variables for the models. The dataset consists of 9 distinct features and 1030 individual cases. The included features encompass pregnancies, glucose levels, blood pressure, skin thickness, insulin levels, body mass index (BMI), diabetes pedigree function, and age.

9. Experimental Settings

The setup of the experiment and parameter settings for the study are illustrated in Table 1. The process of training a model entails the selection of appropriate values for each weight and bias parameter based on labelled samples. The setting of parameters is a crucial stage in the training process of machine learning models. The parameters utilised to regulate the diabetic datasets during the training and testing phases of the models are comprehensively depicted in Table 1. These factors play a crucial role in refining the effectiveness of the model. The models were trained with the scikit-learn tool for machine learning.

Table 1: Experimental settings and parameters tuning of Random Forest, Adaboost, GBOOST, LGBM, CatBoosting and WAEL

Model	**Hyperparameter**	**Values**
Adaboost	n_estimators	50
	learning_rate	0.2
	Loss	Exponential
Random forest	n_estimators	400
	Random_state	0.2
GBOOST	n_estimators	400
	max_depth	5
	Loss	Squared_error
	min_samples_split	2
	learning_rate	0.1
LASSO	Alpha	0.1
RIDGE	Alpha	0.1
Proposed model (SEL)	n_estimators	400
	Random_state	0

10. Performance Metrics

We evaluate all the models under comparison by computing the root mean square error (RMSE), mean square error (MSE), and mean absolute error (MAE) based on the predictions using the identical test set. The subsequent values represent the root mean square error (RMSE), mean square error (MSE), and mean absolute error (MAE). The equations (15), (16), and (17) depict the root mean square error (RMSE), mean square error (MSE), and mean absolute error (MAE), respectively.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(o-f)^2}{n}} \quad (15)$$

$$\text{MSE} = \frac{\sum_{i=1}^{n}(o-f)^2}{n} \quad (16)$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|o-f| \quad (17)$$

Where n is the number of observations, o is the observed values and f is forecasted values.

RESULTS AND DISCUSSION

This section provides an overview of the results and examines the significant discoveries derived from our experimental investigations. The experiments were conducted using the Jupiter notebook programming environment, namely the Python 3.9 version. The diabetes dataset was employed for the purpose of training and testing the ensemble machine learning models. This study examined the different features present in the dataset used for the purpose of training and evaluating the models. The paper presents an analysis of the computing times of six machine learning algorithms. The time intervals utilised in this paper for each machine learning algorithm commence at the initiation and termination of each machine learning method employed in this study. We calculated the time lapse by subtracting the initial time from the final time of the machine learning model. Figure 2 illustrates the duration of training for all the machine learning algorithms. The results indicate that ordinary least square has the shortest computational training time, followed by LASSO and RIDGE. Ensemble machine learning approaches, such as adaptive boost (Adaboost), gradient boost (GBOOST), and random forest, require more processing time compared to regularised regression algorithms like LASSO and RIDGE.

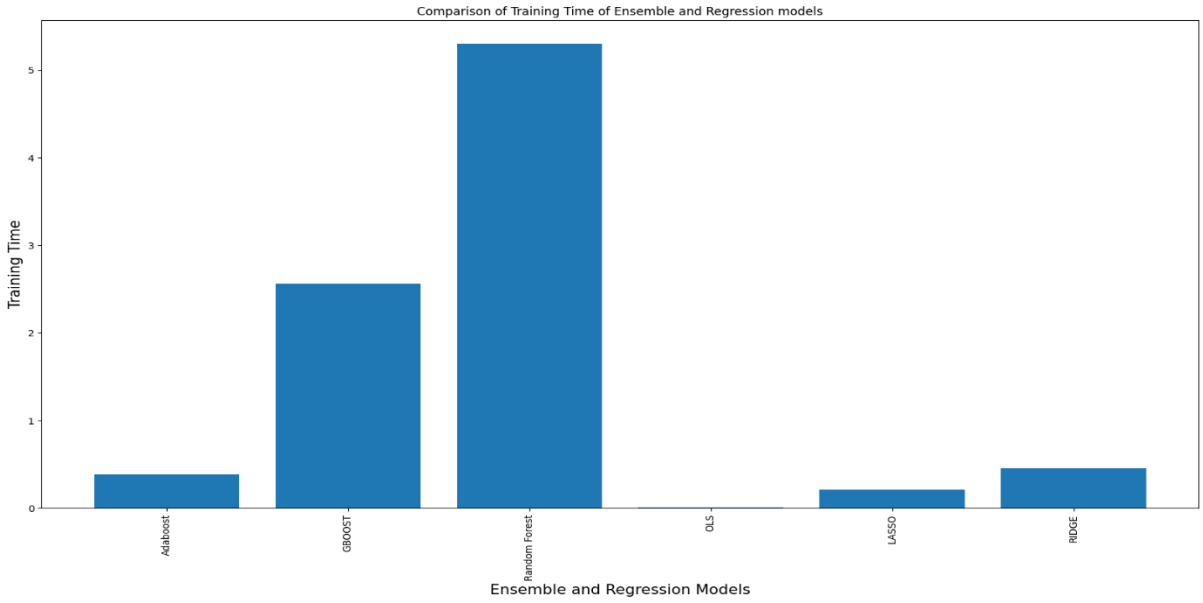


Figure 2: Comparison of Training time of ensemble and regression models

Figs. 3-9 is the time series of all the machine learning compared in this paper, which consist s of actual diabetes cases and predicted diabetes cases in Maiduguri, Borno State, Nigeria.

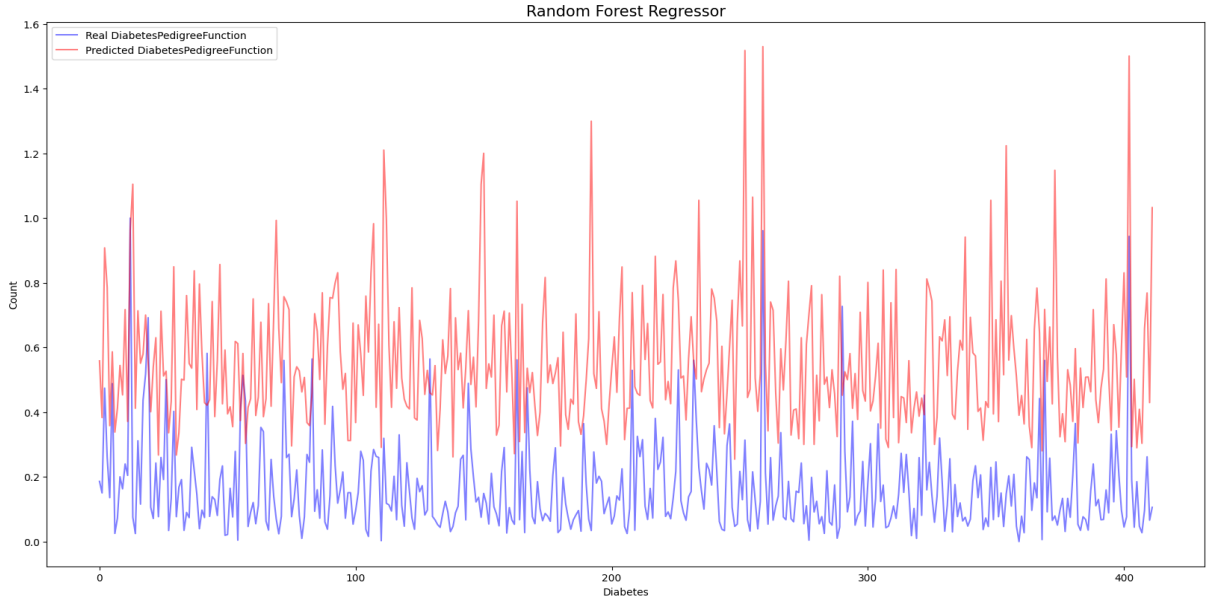


Figure 3: Time Series of Random forest regressor

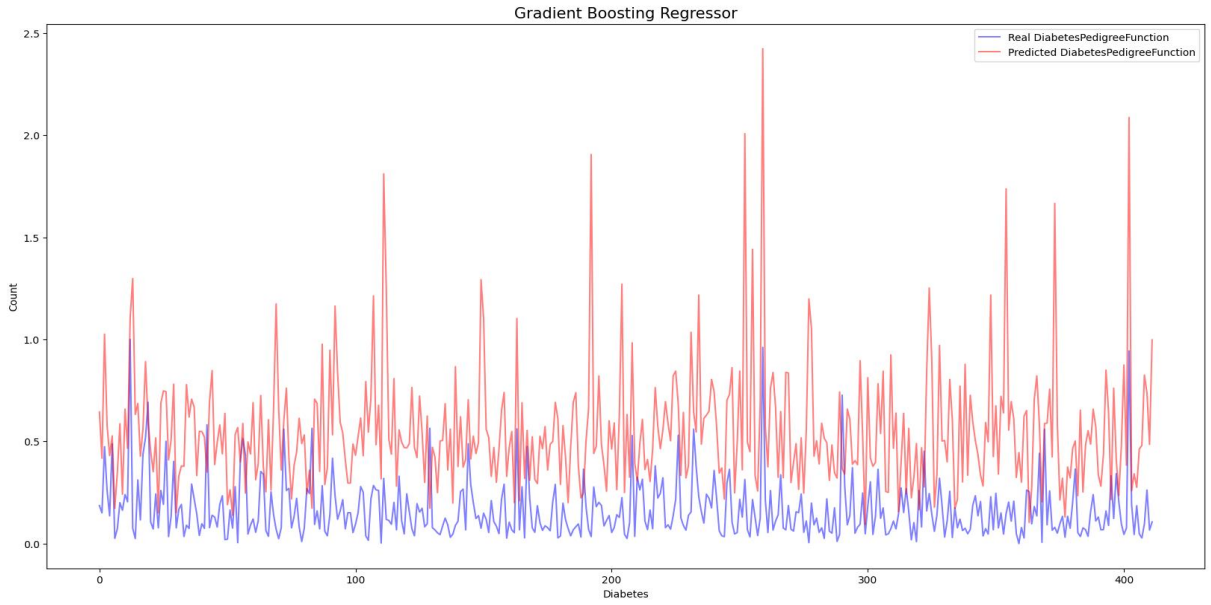


Figure 4: Time Series of Gradient boost regressor

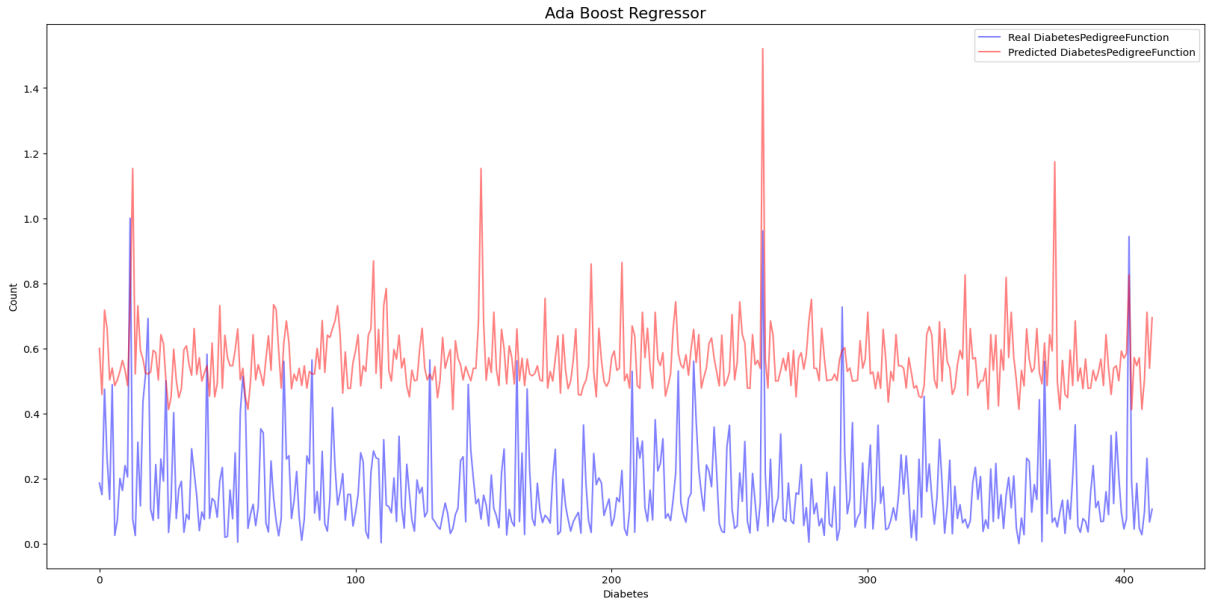


Figure 5: Time Series of Adaboost machine learning

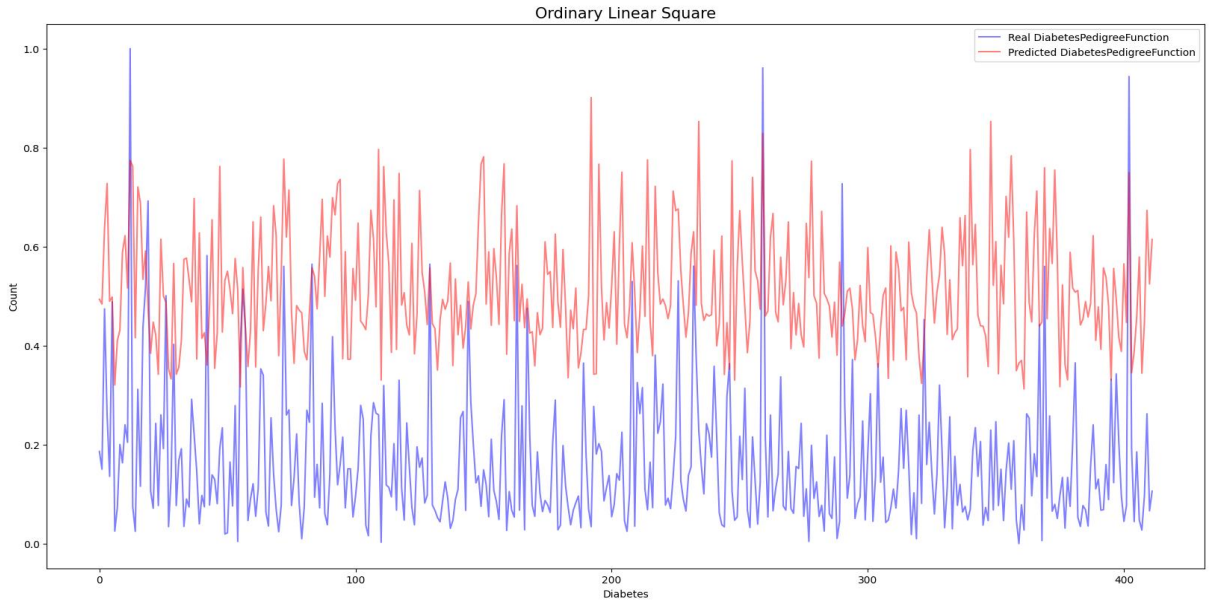


Figure 6: Time Series of Ordinary linear square

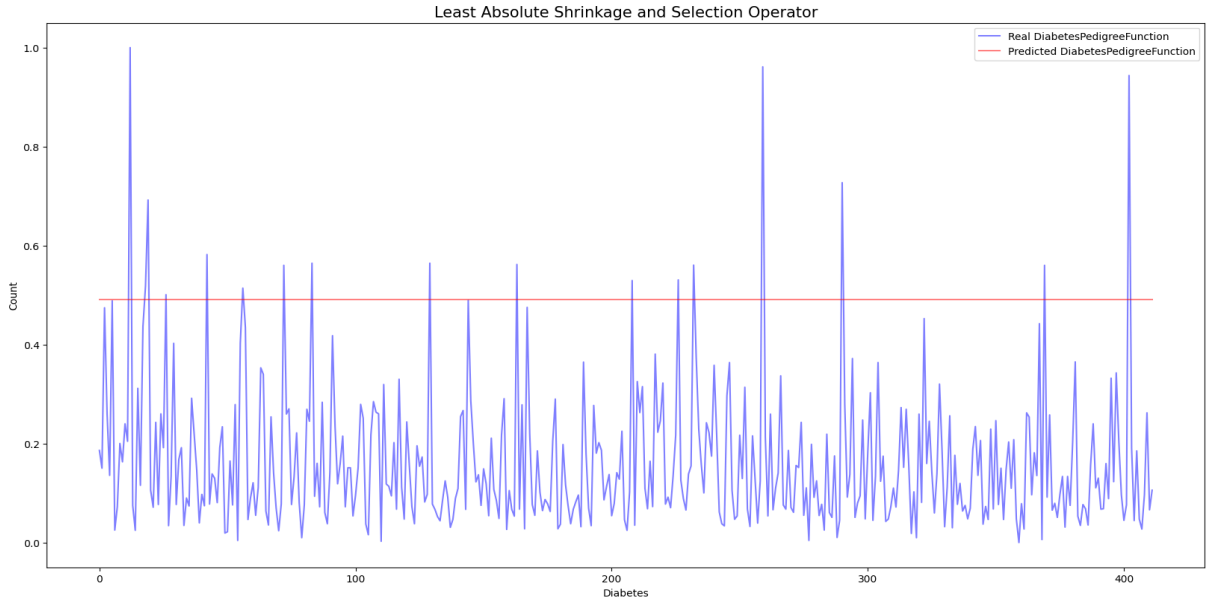


Figure 7: Time Series of Least absolute shrinkage and selection operator

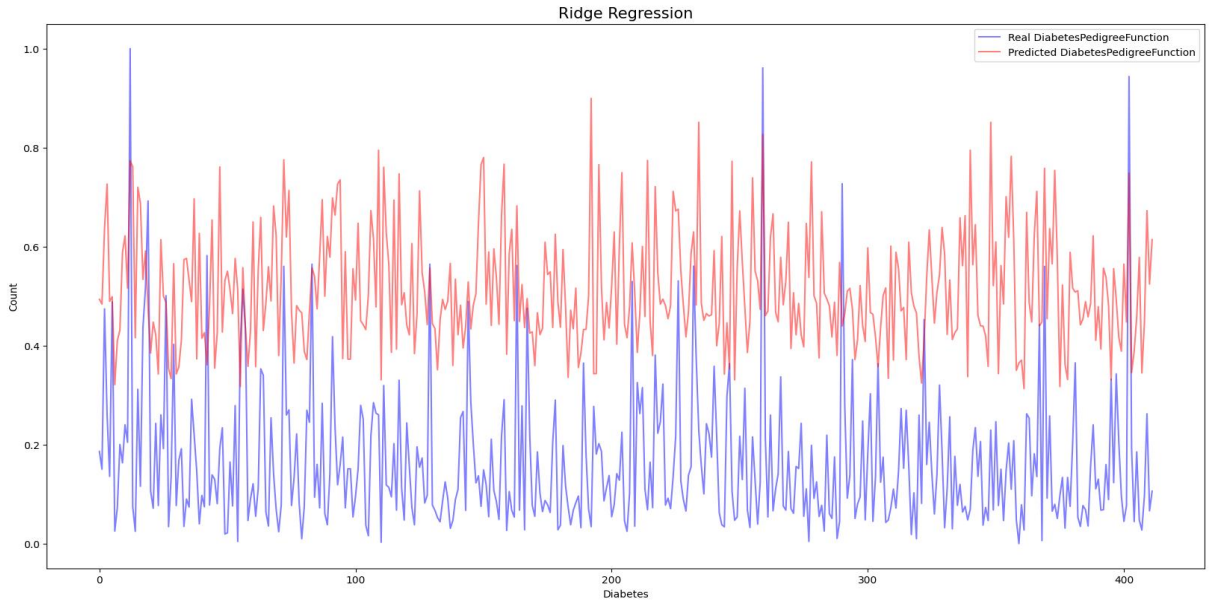


Figure 8: Time Series of Ridge regression

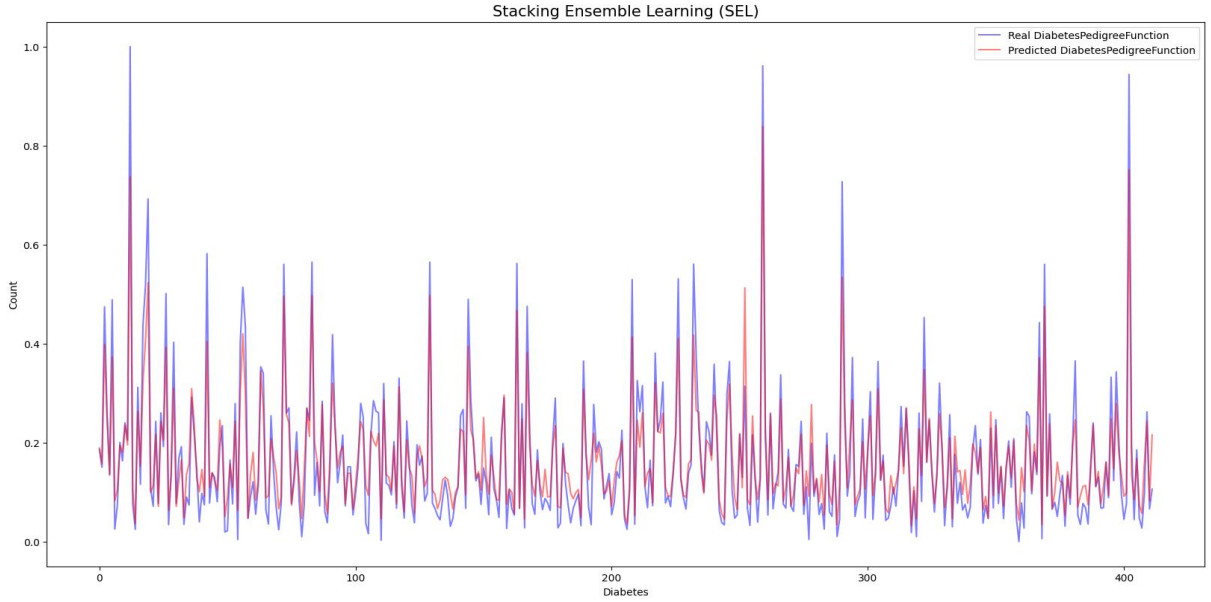


Figure 9: Time Series of Stacking ensemble learning model

About 80% of diabetes dataset was used for training the models while the remaining 20% of the dataset was used as test dataset. Three performance metrics were adopted in this paper, which were RMSE, MSE, and MAE. Fig. 9 shows the actual and predicted diabetes cases obtained from the Stacking Ensemble Learning (SEL) model proposed in this paper.

Table 2 displays the statistical results of all the machine learning methods considered in this paper. SEL performed excellently with a RMSE of 0.0493, followed by LASSO with 0.3573 when considering the RMSE.

Table 2: Performance comparison of machine learning using diabetes models dataset

Model	**RMSE**	**MSE**	**MAE**
Random Forest	0.4434	0.1966	0.3974
Gradient Boost	0.4833	0.2336	0.4003
Adaboost	0.4292	0.1842	0.4039
OLS	0.3828	0.1465	0.3551
LASSO	0.3573	0.1277	0.3390
Ridge	0.3826	0.1464	0.3550
SEL	0.0493	0.0024	0.0349

CONCLUSION

This research presents a technique that uses stacking ensemble learning for predicting diabetes cases in Maiduguri and its environs. We used the diabetes dataset to train and test the proposed method. Experiment results show that the proposed stacking ensemble learning outperformed other machine learning techniques considered in the paper. In addition, the three proposed training strategies - OLS, LASSO, and RIDGE - achieved the lowest computational training time. The proposed method can also train additional datasets. The stacking ensemble learning method was adopted for this paper because it saves us the trouble of having to train several machine learning models from scratch to fulfil similar tasks, therefore saving time and resources. SEL also serves as a cost-cutting measure in areas of machine learning that need a lot of resources, such as image classification or natural language processing. Moreover, it is very useful in compensating for a shortage of labelled training data maintained by an organisation, as pre-trained models are used. Stacking ensemble learning makes use of minimal computational resources and helps attain enhanced results using a smaller dataset. Furthermore, stacking ensemble learning models attains optimal performance quickly compared to conventional ML models. The reason for this is that the models leverage knowledge from base models and meta models.

Simulation results revealed that SEL exhibited the highest level of performance. The results from the evaluation metrics show that WAEL, as an ensemble algorithm, does a better job of accurately predicting diabetes in the dataset used in this study. The experimental findings indicate that the SEL algorithm promises to be a viable approach for achieving precise diabetes prediction. In conclusion, it is recommended that the government and healthcare practitioners dedicate additional resources and efforts towards the implementation of machine learning systems for the early detection and prediction of diabetes. Furthermore, it is imperative for hospitals and healthcare organisations to enhance their data collection efforts from patients. This would help researchers and academics engaged in the field of diabetes detection and prediction to construct highly precise prediction models. In the future, more work will be done of diabetes detection and prediction state-of-the-art algorithms such as Voting Classifier Ensemble method CatBoosting, blending ensemble method, extreme gradient boosting machines (XGBM), and other computational techniques to predict diabetes cases with high accuracy.

Declarations

Author contribution statement

All authors listed have significantly contributed to the development and the writing of this article.

Funding statement

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

Data availability statement

The data used for the research will be made public based on request.

Declaration of interests statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.

REFERENCES

Abdulhadi, N., & Al-Mousa, A. (2021, July). Diabetes detection using machine learning classification methods. In *2021 International Conference on Information Technology (ICIT)* (pp. 350-354). IEEE.

Adeleye, J. O. (2021). The hazardous terrain of diabetes mellitus in Nigeria: the time for action is now. *Research Journal of Health Sciences*, *9*(1), 69-76.

Alasaady, M. T., Aris, T. N. M., Sharef, N. M., & Hamdan, H. (2022). A proposed approach for diabetes diagnosis using neuro-fuzzy technique. *Bulletin of Electrical Engineering and Informatics*, *11*(6), 3590-3597.

Azbeg, K., Boudhane, M., Ouchetto, O., & Jai Andaloussi, S. (2022). Diabetes emergency cases identification based on a statistical predictive model. *Journal of Big Data*, *9*(1), 1-25.

Barakat, N., Bradley, A. P., & Barakat, M. N. H. (2010). Intelligible support vector machines for diagnosis of diabetes mellitus. *IEEE transactions on information technology in biomedicine*, *14*(4), 1114-1120.

CDC (2021). What is Diabetes? Available at https:// www.cdc.gov/diabetes/basics/diabetes.html Accessed 2022-01-27.

Diabetes (2022). Available at https://www.who.int/news-room/fact-sheets/detail/diabetes Accessed 27 Jan 2022.

Ejim, E. C., Okafor, C. I., Emehel, A., Mbah, A. U., Onyia, U., Egwuonwu, T., ... & Onwubere, B. J. (2011). Prevalence of cardiovascular risk factors in the middle-aged and elderly population of a Nigerian rural community. *Journal of tropical medicine*, *2011*.

Enang, O. E., Otu, A. A., Essien, O. E., Okpara, H., Fasanmade, O. A., Ohwovoriole, A. E., & Searle, J. (2014). Prevalence of dysglycemia in Calabar: a cross-sectional observational study among residents of Calabar, Nigeria. *BMJ Open Diabetes Research and Care*, *2*(1), e000032.

Gezawa, I. D., Puepet, F. H., Mubi, B. M., Uloko, A. E., Bakki, B., Talle, M. A., & Haliru, I. (2015). Socio-demographic and anthropometric risk factors for type 2 diabetes in Maiduguri, North-Eastern Nigeria. *Sahel Medical Journal*, *18*(5), 1.

Islam, M. M., Ferdousi, R., Rahman, S., & Bushra, H. Y. (2020). Likelihood prediction of diabetes at early stage using data mining techniques. In *Computer vision and machine intelligence in medical image analysis* (pp. 113-125). Springer, Singapore.

Katarya, R., & Jain, S. (2020, December). Comparison of different machine learning models for diabetes detection. In *2020 IEEE International Conference on Advances and Developments in Electrical and Electronics Engineering (ICADEE)* (pp. 1-5). IEEE.

Laila, U. E., Mahboob, K., Khan, A. W., Khan, F., & Taekeun, W. (2022). An ensemble approach to predict early-stage diabetes risk using machine learning: An empirical study. *Sensors*, *22*(14), 5247.

NCD Risk Factor Collaboration (NCD-RisC). (2017). Trends in obesity and diabetes across Africa from 1980 to 2014: an analysis of pooled population-based studies. *International journal of epidemiology*, *46*(5), 1421-1432.

Oladapo, O. O., Salako, L., Sodiq, O., Shoyinka, K., Adedapo, K., & Falase, A. O. (2010). 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*(1), 26-31.

Robinson, C. A., Agarwal, G., & Nerenberg, K. (2011). Validating the CANRISK prognostic model for assessing diabetes risk in Canada's multi-ethnic population. *Chronic Dis Inj Can*, *32*(1), 19-31.

Rubaiat, S. Y., Rahman, M. M., & Hasan, M. K. (2018, December). Important feature selection & accuracy comparisons of different machine learning models for early diabetes detection. In *2018 International Conference on Innovation in Engineering and Technology (ICIET)* (pp. 1-6). IEEE.

Sabir, A., Ohwovoriole, A., Isezuo, S., Fasanmade, O., Abubakar, S., & Iwuala, S. (2013). Type 2 diabetes mellitus and its risk factors among the rural Fulanis of Northern Nigeria. *Annals of African medicine*, *12*(4), 217.

Saeedi, P., Petersohn, I., Salpea, P., Malanda, B., Karuranga, S., Unwin, N., ... & IDF Diabetes Atlas Committee. (2019). 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*, 107843.

Sarwar, A., Ali, M., Manhas, J., & Sharma, V. (2020). Diagnosis of diabetes type-II using hybrid machine learning based ensemble model. *International Journal of Information Technology*, *12*, 419-428.

Shukla, A. K. (2020). Patient diabetes forecasting based on machine learning approach. In *Soft Computing: Theories and Applications: Proceedings of SoCTA 2019* (pp. 1017-1027). Springer Singapore.

Swapna, G., Vinayakumar, R., & Soman, K. P. (2018). Diabetes detection using deep learning algorithms. *ICT express*, *4*(4), 243-246.

Tinajero, M. G., & Malik, V. S. (2021). An update on the epidemiology of type 2 diabetes: a global perspective. *Endocrinology and Metabolism Clinics*, *50*(3), 337-355.

Type 1 diabetes (2022) - symptoms and causes. https:// www.mayoclinic.org/diseases-conditions/type1diabetes/symptoms-causes/syc-203530119 Accessed 27 Jun 2022.

Type 2 diabetes (2022) - symptoms and causes. https:// www.mayoclinic.org/diseases-conditions/type2diabetes/symptoms-causes/syc-20351193 Accessed 27 Jun 2022.

Uloko, A. E., Musa, B. M., Ramalan, M. A., Gezawa, I. D., Puepet, F. H., Uloko, A. T., ... & Sada, K. B. (2018). Prevalence and risk factors for diabetes mellitus in Nigeria: a systematic review and meta-analysis. *Diabetes Therapy*, *9*, 1307-1316.

Dada, E. G., Bassi, J. S., Chiroma, H., Adetunmbi, A. O., & Ajibuwa, O. E. (2019). Machine learning for email spam filtering: review, approaches and open research problems. *Heliyon*, *5*(6).

Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling. *Expert systems with applications*, *134*, 93-101.

Taherkhani, A., Cosma, G., & McGinnity, T. M. (2020). AdaBoost-CNN: An adaptive boosting algorithm for convolutional neural networks to classify multi-class imbalanced datasets using transfer learning. *Neurocomputing*, *404*, 351-366.

Chen, G., He, H., Zhao, L., Chen, K. B., Li, S., & Chen, C. Y. C. (2022). Adaptive boost approach for possible leads of triple-negative breast cancer. *Chemometrics and Intelligent Laboratory Systems*, *231*, 104690.

Sharma, A., & Singh, B. (2020). AE-LGBM: Sequence-based novel approach to detect interacting protein pairs via ensemble of autoencoder and LightGBM. *Computers in Biology and Medicine*, *125*, 103964.

Barrow, D. K., & Crone, S. F. (2016). A comparison of AdaBoost algorithms for time series forecast combination. *International Journal of Forecasting*, *32*(4), 1103-1119.

Oyewola, D. O., Dada, E. G., Misra, S., & Damaševičius, R. (2021). Predicting COVID-19 cases in South Korea with all K-edited nearest neighbors noise filter and machine learning techniques. *Information*, *12*(12), 528.

[91]Oyewola, D. O., Dada, E. G. Exploring machine learning: a scientometrics approach using bibliometrix and VOSviewer. *SN Applied Sciences*, **2022**, *4*(5), 1-18. DOI: https://doi.org/10.1007/s42452-022-05027-7.

[92] Dada, E. G., Yakubu, H. J., Oyewola, D. O. Artificial Neural Network Models for Rainfall Prediction. *European Journal of Electrical Engineering and Computer Science*, **2021**, *5*(2), 30-35.

[97] Oyewola, D.O., Ibrahim, A., Kwanamu, J.A. Dada, E.G. A new auditory algorithm in stock market prediction on oil and gas sector in Nigerian stock exchange. *Soft computing letters*, **2021,** *3*, p.100013, https://doi.org/10.1016/j.socl.2021.100013.

[98] Dada, E. G., Oyewola, D. O., Joseph, S. B., Duada, A. B. Ensemble Machine Learning Model for Software Defect Prediction. *Advances in Machine Learning & Artificial Intelligence*, **2021**, 2(1), 11-21. https://doi.org/10.33140/AMLAI.02.01.03

Lingjun, H., Levine, R. A., Fan, J., Beemer, J., & Stronach, J. (2019). Random forest as a predictive analytics alternative to regression in institutional research. *Practical Assessment, Research, and Evaluation*, *23*(1), 1.

Gieseke, F., & Igel, C. (2018, July). Training big random forests with little resources. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining* (pp. 1445-1454).

Xue, L., Liu, Y., Xiong, Y., Liu, Y., Cui, X., & Lei, G. (2021). A data-driven shale gas production forecasting method based on the multi-objective random forest regression. *Journal of Petroleum Science and Engineering*, *196*, 107801.

Darlington, R. B., & Hayes, A. F. (2016). *Regression analysis and linear models: Concepts, applications, and implementation*. Guilford Publications.

Gomila, R. (2021). Logistic or linear? Estimating causal effects of experimental treatments on binary outcomes using regression analysis. *Journal of Experimental Psychology: General*, *150*(4), 700.

Keele, L., Stevenson, R. T., & Elwert, F. (2020). The causal interpretation of estimated associations in regression models. *Political Science Research and Methods*, *8*(1), 1-13.

Shafiee, S., Lied, L. M., Burud, I., Dieseth, J. A., Alsheikh, M., & Lillemo, M. (2021). Sequential forward selection and support vector regression in comparison to LASSO regression for spring wheat yield prediction based on UAV imagery. *Computers and Electronics in Agriculture*, *183*, 106036.

Czajkowski, M., Jurczuk, K., & Kretowski, M. (2023). Steering the interpretability of decision trees using lasso regression-an evolutionary perspective. *Information Sciences*, *638*, 118944.

Wang, S., Chen, Y., Cui, Z., Lin, L., & Zong, Y. (2024). Diabetes Risk Analysis Based on Machine Learning LASSO Regression Model. *Journal of Theory and Practice of Engineering Science*, *4*(01), 58-64.

Rokem, A., & Kay, K. (2020). Fractional ridge regression: a fast, interpretable reparameterization of ridge regression. *GigaScience*, *9*(12), giaa133.

la Tour, T. D., Eickenberg, M., Nunez-Elizalde, A. O., & Gallant, J. L. (2022). Feature-space selection with banded ridge regression. *NeuroImage*, *264*, 119728.

Wang, L., Wang, Z., Qu, H., & Liu, S. (2018). Optimal forecast combination based on neural networks for time series forecasting. *Applied soft computing*, *66*, 1-17.

Moon, J., Jung, S., Rew, J., Rho, S., & Hwang, E. (2020). Combination of short-term load forecasting models based on a stacking ensemble approach. *Energy and Buildings*, *216*, 109921.

Mienye, I. D., & Sun, Y. (2022). A survey of ensemble learning: Concepts, algorithms, applications, and prospects. *IEEE Access*, *10*, 99129-99149.

Wang, X., Hyndman, R. J., Li, F., & Kang, Y. (2023). Forecast combinations: an over 50-year review. *International Journal of Forecasting*, *39*(4), 1518-1547.