Proceedings of the 1st International Conference on Statistical Sciences and Data Analytics

EVALUATING THE PREDICTIVE ACCURACY OF LOGISTIC REGRESSION AND MACHINE LEARNING ALGORITHMS FOR TYPE 2 DIABETES

Oguntade Emmanuel Segun[1] and Emmanuel Patrick Iwe*[2]

[1]Department of Statistics, Faculty of Science, University of Abuja, Abuja, Nigeria
[2]Department of Statistics, Faculty of Science, University of Abuja, Abuja, Nigeria
*Corresponding Autor**:** cejaelresearchconsultancy@gmail.com

ABSTRACT

This study evaluates the predictive accuracy of logistic regression and machine learning algorithms for Type 2 diabetes in Abuja, Nigeria. Type 2 diabetes is a growing global health condition with a very high rates especially in sub-Saharan Africa. Traditional predictive models often focus solely on clinical risk factors, neglecting socio-economic and lifestyle variables that are critical in diverse populations. To address this gap, this research aims to utilize hybrid models that combine logistic regression with machine learning techniques to enhance both predictive performance and interpretability. The study uses patient records from University of Abuja teaching hospital in FCT Abuja, incorporating clinical, socio-economic, and lifestyle data. Machine learning algorithms, including Random Forest, XGBoost, and Support Vector Machine (SVM), are employed, and their performances were compared with logistic regression and a hybrid ensemble model. Model accuracy was evaluated using metrics such as Accuracy, Precision, Recall, Area Under the Curve/Receiver Operating Characteristic, and Balanced Accuracy. Results indicated that the hybrid model outperforms individual models, achieving a Balanced Accuracy of 0.7951, compared to 0.784 for logistic regression, 0.788 for XGBoost, 0.787 for Random Forest, and 0.777 for the Support Vector Machine. Socioeconomic and lifestyle factors, in addition to clinical variables, significantly contribute to predictive accuracy. However, demographic factors such as education, gender and income do not significantly impact predictive accuracy. These findings highlight the importance of integrating diverse data sources and employing advanced analytical techniques to improve diabetes prediction. The results can inform the development of more effective interventions and public health strategies to address the growing burden of Type 2 diabetes in Nigeria.

Keywords: Type 2 diabetes, machine learning, hybrid models, socioeconomic factors, Nigeria.

1. INTRODUCTION

Type 2 diabetes (T2D) is a chronic metabolic disorder characterized by insulin resistance and impaired insulin secretion, resulting in persistent hyperglycemia (Nolan & Prentki, 2019). As insulin becomes less effective, the pancreas attempts to compensate by producing more insulin, but over time, this increased demand makes the pancreas exhausted, leading to a progressive drop in insulin secretion. Consequently, blood glucose levels remain high, resulting in hyperglycemia (Nolan & Prentki, 2019).

Globally, Type 2 diabetes poses a major public health challenge, affecting over 537 million individuals, with projections estimating 643 million cases by 2030 (International Diabetes Federation, 2021). In sub-Saharan Africa, the incidence of Type 2 diabetes has surged, largely due to rapid urbanization, changes

in dietary habits, increased sedentary behaviors, and limited access to quality healthcare. According to the IDF, over 19 million adults in sub-Saharan Africa were estimated to be living with diabetes in 2021, with the number projected to exceed 47 million by 2045 if current trends persist (IDF, 2021). Although once associated mainly with high-income countries, recent evidence indicates a substantial shift toward low- and middle-income regions, where healthcare systems are often under-equipped for managing chronic diseases (WHO, 2022). In Nigeria, the prevalence of Type 2 diabetes is rising rapidly, driven by urbanization, dietary shifts toward processed foods, sedentary lifestyles, and widening socioeconomic disparities (Gkrinia & Belančić, 2025). Carbone *et al.* (2019) also demonstrated that regular exercise enhances insulin sensitivity and glucose metabolism, thereby reducing the risk of diabetes. About 3.6 million Nigerians currently live with diabetes (International Diabetes Federation, 2021), though the true burden is likely higher due to limited diagnostic infrastructure. Gender disparities exist, with a slightly higher prevalence among females, attributed to hormonal changes, gestational diabetes, and socio-cultural constraints on physical activity (Angell *et al.,* 2022). These patterns align with the Social Determinants of Health (SDH) framework, identifying gender as a structural determinant influencing health outcomes (WHO, 2022).

Educational attainment has also emerged as a critical factor. With 78.7% of participants reporting low educational levels, poor health literacy, delayed diagnosis, and inadequate disease management are major concerns (Hill-Briggs *et al.,* 2020; Marciano *et al.,* 2019). Low education correlates with unhealthy behaviors that elevate diabetes risk. Income disparities further worsen these challenges; economic deprivation limits access to healthy foods and healthcare (Marmot, 2017; Ekure *et al.,* 2022). Nevertheless, in highly urbanized settings like Abuja, processed food consumption and sedentary behavior are widespread across income groups, weakening income's predictive strength.

Lifestyle factors, especially sedentary behavior and high-calorie diets, have emerged as dominant predictors of Type 2 diabetes risk (Popkin & Ng, 2022). Other indicators like body mass index (BMI), physical activity, and diet quality have been known to outperform others in predictive models. Although family history remains relevant, its influence appears to be weaker when compared to modifiable behaviors (Berumen *et al.,* 2023; Szczerbinski & Florez, 2023). Incorporating socioeconomic variables such as education, income, and gender helps to further improve the generalizability and robustness of diabetes prediction models (Ibitoye *et al.,* 2024).

Traditional predictive models have largely relied on logistic regression. Although studies like Olanrewaju and Ibikunle (2022) and Olusola *et al.* (2023) applied logistic regression to Nigerian datasets, they often omitted key socioeconomic and behavioral variables. They also did not address class imbalance or non-linear interactions. While logistic regression is great for simplicity and interpretability (Kost *et al.,* 2021), it struggles to capture the complex, multi-factorial nature of Type 2 diabetes risks (Edlitz & Segal, 2022).

As a result of this, attention has shifted to machine learning (ML) approaches, which excel with complex, high-dimensional data (Stiglic *et al.,* 2021; Jiang *et al.,* 2024). Hybrid models combining ML and traditional predictive models show improved predictive performance while retaining interpretability (Rahman *et al.,* 2023). However, variations in Type 2 diabetes prevalence over time, including a slight decline after 2023, require cautious interpretation without sequence modeling (Yoshioka et al., 2024).

In Nigeria, emerging studies by Shehu and Baha (2024) and Okwori *et al.* (2024) highlight growing ML applications for predicting chronic diseases. Yet, ML models face interpretability challenges ("black

box" problem) (Linardatos *et al.,* 2020). Class imbalance persists, although Synthetic Minority Over-sampling Technique (SMOTE) offers a solution (Chawla *et al.,* 2002), its application remains limited in Nigerian datasets.

Accordingly, this study aims to develop hybrid predictive models combining the interpretability of logistic regression with ML's predictive power. By incorporating clinical, socioeconomic, and lifestyle factors and applying SMOTE, the study seeks to enhance the accuracy, fairness, and applicability of T2D risk prediction within Abuja's population.

2. METHODOLOGY

2.1 Study Design

This study adopted a retrospective, cross-sectional design aimed at predicting Type 2 Diabetes Mellitus (T2DM) using a hybrid of statistical and machine learning techniques. A multi-level modeling strategy was implemented, combining traditional statistical inference, feature selection, and ensemble learning.

2.2 Data Source and Description of Model Variables

The dataset was obtained from the University of Abuja Teaching Hospital (UATH) and comprised 5,000 anonymized patient records collected between 2021 and 2024. Variables included demographic, clinical, socioeconomic, and lifestyle factors such as age, body mass index (BMI), hypertension status, education, income level, smoking status, physical activity, diet score, and family history of diabetes.

2.3 Class Imbalance Handling

To address the imbalance between diabetic and non-diabetic classes, the Synthetic Minority Over-sampling Technique (SMOTE) was applied. SMOTE generates synthetic minority class samples to balance the dataset, improving model training and prediction stability.

2.4 Model Development

2.4.1 Logistic Regression

Logistic regression models the probability $P(y = 1|X)$ using the logistic function (Hosmer, Lemeshow & Sturdivant, 2013):

$$P(y = 1|X) = \frac{e^{\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k}}{1 + e^{\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k}} \quad (1)$$

Where:
$P(y = 1 \mid X)$ = probability of having Type 2 diabetes
β_0 = intercept
β_j = coefficient for predictor X_j
X_j = predictor variables (e.g., BMI, age, diet score)

2.4.2 Random Forest (RF)

Random Forest aggregates predictions from multiple decision trees (Breiman, 2001):

(2)

$$h(X) = \frac{1}{N}\sum_{i=1}^{N} T_i(X)$$

Where:
$h(X)$ = final prediction of the Random Forest
N = number of trees
$T_i(X)$ = prediction from the i^{th}decision tree.

2.4.3 XGBoost

XGBoost fits additive regression trees to minimize a regularized loss function (Chen & Guestrin, 2016):

$$L = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(T_k) \quad 3$$

Where:
L = overall loss function
$l(y_i, \hat{y}_i)$ = differentiable loss function (e.g., log loss)
$\Omega(T_k)$ = regularization term penalizing model complexity
T_k = individual trees in the model.

2.4.4 Support Vector Machine (SVM)

For binary classification, SVM constructs a hyperplane (Cortes & Vapnik, 1995):

$$f(X) = sign(w^T X + b) \quad 4$$

Where:
w = weight vector
X = input features
b = bias term
$f(X)$ = classification output (diabetic or non-diabetic).

2.4.5 Weighted Soft Voting Ensemble

A Weighted Soft Voting Ensemble was developed, combining Logistic regression, Random Forest, XGBoost, and SVM based on validation scores (Rokach, 2010).
The final probability for each class c is:

$$w_i = \frac{A_i + BA_i + R_i + F1_i + MCC_i + P_i}{\sum_{j=1}^{n}(A_j + BA_j + R_i + F1_j + MCC_j + P_j)} \quad 5$$

Where:
w_i = weight assigned to model iii
A = accuracy, BA = balanced accuracy
R = recall, $F1$ = F1-score
MCC = Matthews Correlation Coefficient
P = precision.

The final prediction is obtained as the weighted average of class probabilities across models, enhancing robustness and interpretability.

2.5 Train–Test Split and Model Evaluation

To evaluate the models individually, the dataset was split into 80% training and 20% testing sets. Each model (Logistic Regression, Random Forest, XGBoost, and SVM) was trained using the training set and tested on the held-out test set. Performance was assessed using the following metrics:

i. Accuracy
ii. Precision
iii. Recall (Sensitivity)
iv. F1 Score
v. Matthews Correlation Coefficient (MCC)
vi. Area Under the ROC Curve (AUC-ROC)
vii. Cohen's Kappa
viii. Balanced Accuracy

Additionally, 10-fold cross-validation was performed during training to reduce overfitting and ensure model generalizability.

2.6 Ethical consideration

The data for this study were exclusive information about anonymous human subjects attending the University of Abuja Teaching Hospital (UATH). Approval was obtained from UATH before the commencement of this work.

3. RESULTS

3.1 Socio-Demographic and Clinical Characteristics of the Patients

The dataset of 5,000 individuals shows key differences between those with and without diabetes across various demographics. Among males, 32% have diabetes compared to 68% of females. Diabetes prevalence is higher in individuals without education (47%) compared to those with secondary education (36%) or university education (17%). Physical activity appears to be a protective factor: only 35.8% of those with diabetes reported being physically active, versus 53.8% without diabetes. A strong family history of diabetes is also notable, with 59.1% of diabetes patients having such a history compared to 49.8% without. Income levels show modest variation: low-income individuals make up the largest share of diabetes patients (47.9%), followed by middle (35.6%) and high-income (16.5%). Overall, diabetes is more common among females, those with lower education and income levels, physically inactive individuals, and those with a family history of the disease.

Table 1: Summary of Socio-Demographic and Lifestyle of Patients attending UATH

Categorical Variables	With Type 2 Diabetes	Without Type2 Diabetes	Count/Percentage
Gender			
Male	801 (32%)	1561 (62.5%)	2362 (47.2%)
Female	1701 (68%)	937(37.5%)	2638 (52.8%)
Total	**2502**	**2498**	**5000**
Education Level			
No Education	1176 (47%)	860 (34.4%)	2036 (40.7%)
Secondary School	901 (36%)	996 (39.9%)	1897 (39.9%)
University	425 (17%)	642 (25%)	1067 (19.4%)
Total	**2502**	**2498**	**5000**
Physical Activity			
Yes	895 (35.8%)	1345 (53.8)	2240 (44.8%)
No	1607 (64.2%)	1153 (46.2%)	2760 (55.2%)
Total	**2502**	**2498**	**5000**
Family History			**Count/Percentage**
Yes	1478 (59.1%)	1246 (49.8%)	2724 (54.5%)
No	1024 (40.9%)	1252 (50.2%)	2276 (45.5%)
Total	**2502**	**2498**	**5000**
Income			**Count/Percentage**
High Income	412 (16.5%)	382 (15.3%)	794 (15.9%)
Middle Income	890 (35.6%)	885 (35.4%)	1775 (35.5%)
Low Income	1200 (47.9%)	1231 (49.3%	2431 (48.6%)
Total	**2502**	**2498**	**5000**

Table 2: Summary of Clinical Variables of Patients attending UATH

	Diabetes	Age	BMI	SystolicBP	DietScore	Cholesterol
Count	5000	5000	5000	5000	5000	5000
Mean	0.5004	47.5568	28.95576	135.155	2.74432	1.1234
Std	0.50005	10.35169	4.629567	15.69186	0.904767074	0.828920902
Min	0	13	14.2	80	1	0
25th Percentile (Q1)	0	41	25.6	125	2	0
Median (50%)	1	48	28.5	135	2.7	1
75th Percentile (Q3)	1	55	32.025	146	3.5	2
Maximum	1	89	46.1	191	5	2

BMI: Body Mass Index; BP: Blood Pressure

Table 2 presents the descriptive statistics for six clinical variables among 5,000 patients: diabetes status, age, body mass index (BMI), systolic blood pressure, diet score, and cholesterol level. The prevalence of

diabetes is approximately 50%, as indicated by a mean of 0.5004 for the binary diabetes variable. The patient's ages ranges from 13 to 89 years, with a mean of 47.56 and a median of 48, suggesting a predominantly middle-aged population. BMI values range from 14.2 to 46.1, with a mean of 28.96 and a median of 28.5, indicating that many patients fall within the overweight or obese category. Systolic blood pressure ranges from 80 to 191 mmHg, with a mean of 135.16 and a median of 135 mmHg, showing that a considerable proportion may be hypertensive. The diet score, ranging from 1 to 5, has a mean of 2.74 and a median of 2.7, reflecting moderate dietary habits among respondents. Cholesterol levels range from 0 to 2, with a mean of 1.12 and a median of 1, suggesting a fairly balanced distribution across the cholesterol categories in the dataset.

3.2 Variable Ranking and Predictors of Type 2 diabetes using Random forest

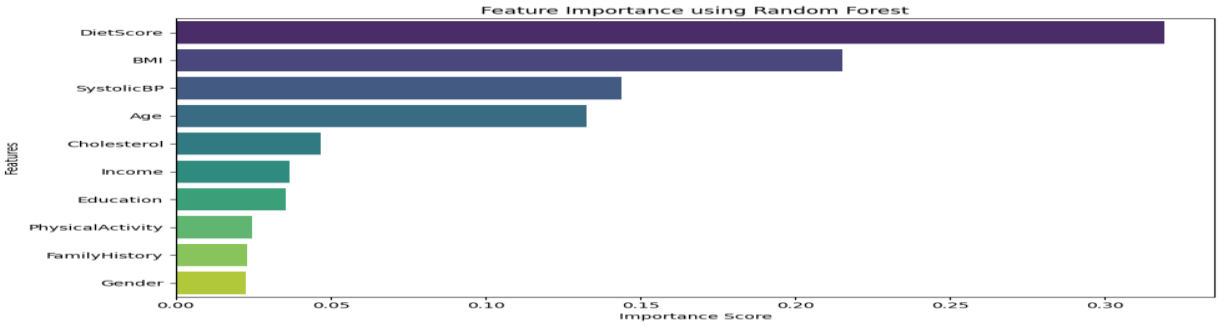

Figure 1: Feature Ranking Importance

Fig 1 shows the Feature Importance Analysis using the Random Forest algorithm reveals that Diet Score, BMI, Systolic Blood Pressure, and Age are the most influential predictors of diabetes. Diet Score has the highest importance, indicating that dietary habits significantly affect diabetes risk. BMI and Systolic Blood Pressure follow closely, showing the role of obesity and hypertension in diabetes prediction. Age also plays a critical role, as diabetes risk tends to increase with age. In contrast, features like Income, Education, Physical Activity, Family History, and Gender have lower importance, suggesting they contribute less to the model's predictive accuracy. This analysis highlights the need for lifestyle and clinical interventions in diabetes prevention.

3.3 Evaluating the Predictive Accuracy of Logistic Regression and Machine Learning Algorithms for Type 2 Diabetes Using Socioeconomic and Lifestyle Factors

The dataset of 5,000 patient records was randomly divided into a training set (80%, n = 4,000) and a testing set (20%, n = 1,000) using stratified sampling to maintain the original class distribution between diabetic and non-diabetic patients. All models were trained using the training set and evaluated exclusively on the testing set to ensure that the reported performance reflects true generalization ability rather than memorization of the training data.

Table 3 presents the results of the test set evaluation across different classification metrics. Among all models, XGBoost achieved the highest accuracy (0.788), precision (0.7905), recall (0.7905), and F1-Score (0.7905), showing the most balanced performance. Logistic Regression and Random Forest also performed well, with accuracies of 0.784 and 0.787, respectively. Logistic Regression recorded the highest specificity (0.7834) and negative predictive value (NPV = 0.7802). SVM had slightly lower performance, with an accuracy of 0.777 and a precision of 0.7758. These results demonstrate that all models achieved comparable predictive performance, with only modest differences between them.

Table 3: Models Predictive Performance

Model	Logistic Regression	XGBoost	Random Forest	SVM
Accuracy	0.784	0.788	0.787	0.777
Precision	0.787698	0.790514	0.78895	0.775828
Recall (Sensitivity)	0.784585	0.790514	0.790514	0.786561
F1-Score	0.786139	0.790514	0.789733	0.781158
MCC	0.57390	0.575939	0.573930	0.553915
Kappa	0.567959	0.575939	0.0573928	0.553861
Specificity	0.783401	0.785425	0.783401	0.767206
NPV	0.780242	0.778234	0.784990	0.778234
B. Accuracy	0.783993	0.787969	0.786957	0.776884

SVM: Support Vector Machine, Note: *All metrics were computed on the testing set after the train-test split (80/20). Stratified sampling ensured balanced class representation across splits.*

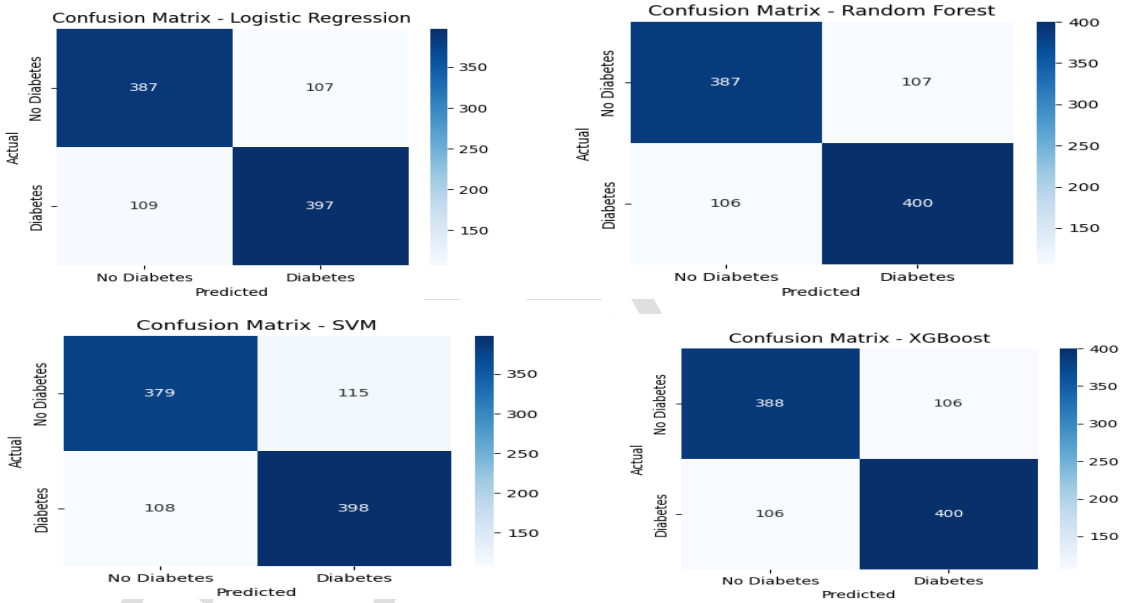

Figure 2: Confusion Matrix for Performance Assessment of Prediction Models

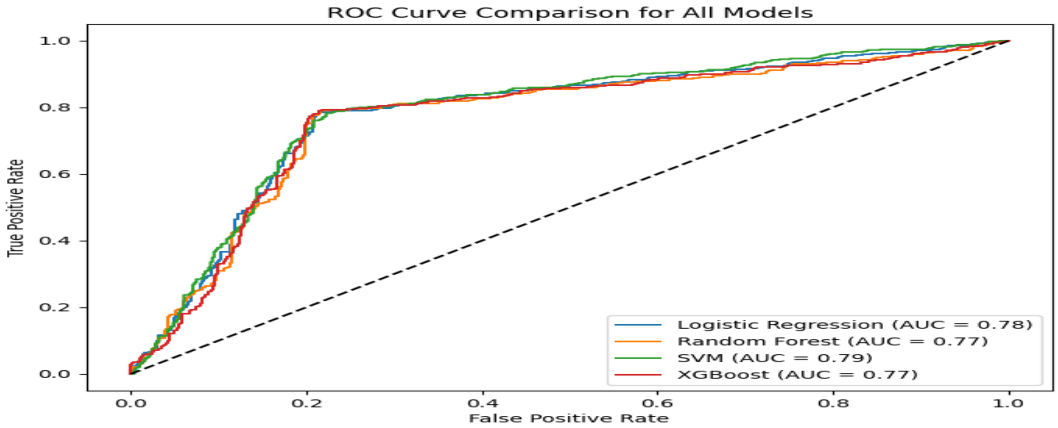

Figure 2: ROC Curve Comparison for all the Models

Figures 2 and 3 show the confusion matrix and Receiver Operating Characteristic (ROC). The results show that XGBoost achieved the best predictive performance for type 2 diabetes, with the highest accuracy and lowest misclassification rates. Random Forest and Logistic Regression followed, while SVM had the highest misclassification. However, based on the ROC curve, SVM achieved the highest AUC (0.79), slightly outperforming Logistic Regression (0.78), Random Forest, and XGBoost (both 0.77) in distinguishing diabetic from non-diabetic cases.

3.4 Development and Evaluation of Hybrid Models Combining Logistic Regression and Machine Learning Techniques for Enhanced Predictive Performance and Interpretability

Table 4: Performance comparison of the generated prediction models and Hybrid Model

Model	Logistic Regression	XGBoost	Random Forest	SVM	Hybrid Model
Accuracy	0.784	0.788	0.787	0.777	0.7874
Precision	0.787698	0.790514	0.78895	0.775828	0.7906
Recall (Sensitivity)	0.784585	0.790514	0.790514	0.786561	0.7895
F1-Score	0.786139	0.790514	0.789733	0.781158	0.7893
MCC	0.57390	0.575939	0.573930	0.553915	0.5760
Kappa	0.567959	0.575939	0.0573928	0.553861	0.5739
Specificity	0.783401	0.785425	0.783401	0.767206	0.7854
NPV	0.780242	0.778234	0.784990	0.778234	0.7870
B. Accuracy	0.783993	0.787969	0.786957	0.776884	0.7951

SVM: Support Vector Machine, Metrics are computed on the test set after an 80/20 stratified split.

Following the train–test split procedure (80% training set, n = 4,000; 20% testing set, n = 1,000), the predictive performance of four baseline models, Logistic Regression, XGBoost, Random Forest, and Support Vector Machine (SVM), was compared to a Hybrid Model built using weighted soft voting.

The split was carried out using stratified sampling to maintain the original class distribution between diabetic and non-diabetic patients. All reported results in Table 4 were obtained on the held-out testing set to evaluate the models' generalization performance. The Hybrid Model achieved the highest

balanced accuracy (0.7951), along with the best performance across nearly all metrics, including precision (0.7906), recall (0.7895), F1-score (0.7893), MCC (0.5760), and Kappa (0.5739). It also achieved the highest specificity (0.7854) and NPV (0.7870).

This result supports the well-documented advantage of ensemble learning in improving both predictive accuracy and robustness over single algorithms. By integrating multiple models through weighted soft voting, the Hybrid Model captures complementary decision boundaries, leading to superior performance on unseen data.

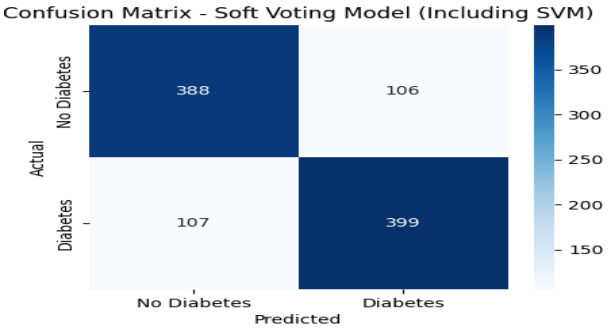

Fig 4. Confusion Matrix for the Hybrid Model. Developed Models, including the Hybrid Model Through Soft Voting Incorporating SVM

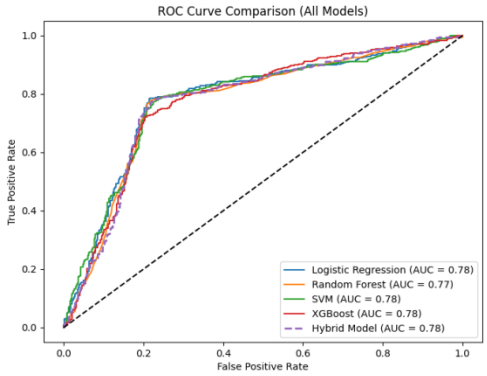

Figure 5: ROC Curve for all the

Figure 4 and Figure 5 present the confusion matrix and ROC curve for the Hybrid Model (Soft Voting Including SVM). Fig 3.4 shows the model correctly classified 388 non-diabetic and 399 diabetic cases, with 106 false positives and 107 false negatives, reflecting strong predictive performance but room for improving sensitivity. The incorporation of SVM, Logistic Regression, Random Forest, and XGBoost enhances both accuracy and interpretability. Fig 3.5 shows the ROC comparison of the Hybrid Model. It achieved an AUC of 0.78, matching Logistic Regression, SVM, and XGBoost, and outperforming Random Forest (0.77), confirming its competitive and stable performance.

3.5 Analysis of Temporal Aspects in Diabetes Prediction: Evaluating Model Performance across Different Time Horizons

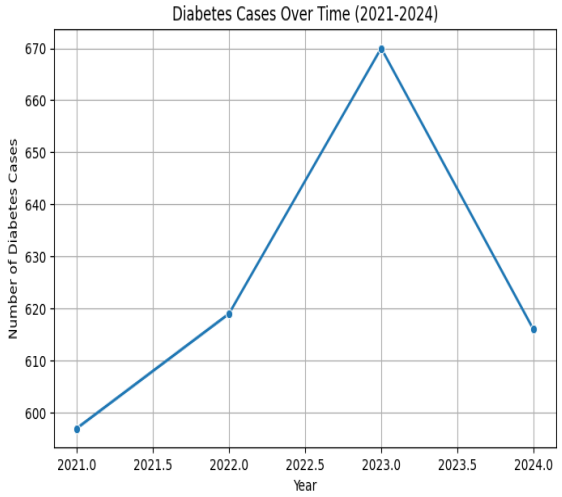

Fig 6 Plot of Diabetes Cases (2021-2024)

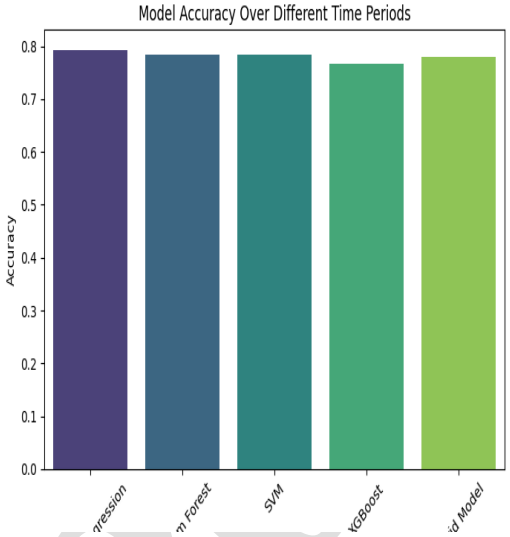

Fig 7: Model Accuracy over different Time periods

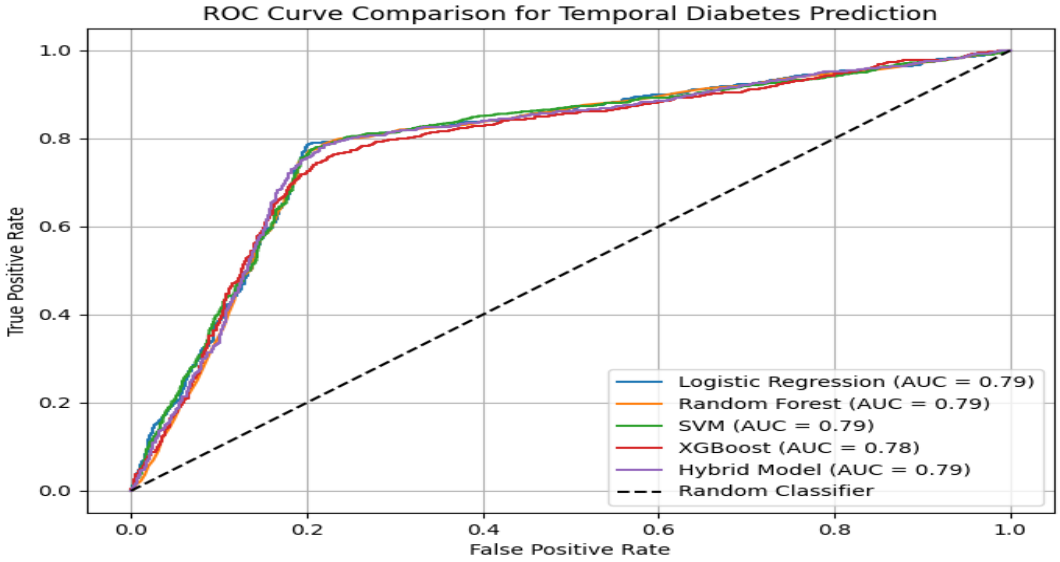

Figure 8: ROC Curve Comparison for Temporal Diabetes Prediction

Figures 6, 7, and 8 show trends and model performances over time. Diabetes cases increased gradually from 2021 to 2022, sharply rose in 2023, then declined in 2024, suggesting a possible intervention. Fig 3.7 compares model accuracies, showing that Logistic Regression, Random Forest, SVM, XGBoost, and the Hybrid Model performed similarly, with accuracies around 0.78–0.79, though the Hybrid Model was slightly better. Fig 3.8 presents the ROC curves for temporal diabetes prediction, where most models achieved an AUC of 0.79, except XGBoost at 0.78, indicating consistent and strong predictive performance across all models.

3.6 Assessment of Cross-Population Predictive Validity to Ensure Model Generalizability across Diverse Demographic Groups

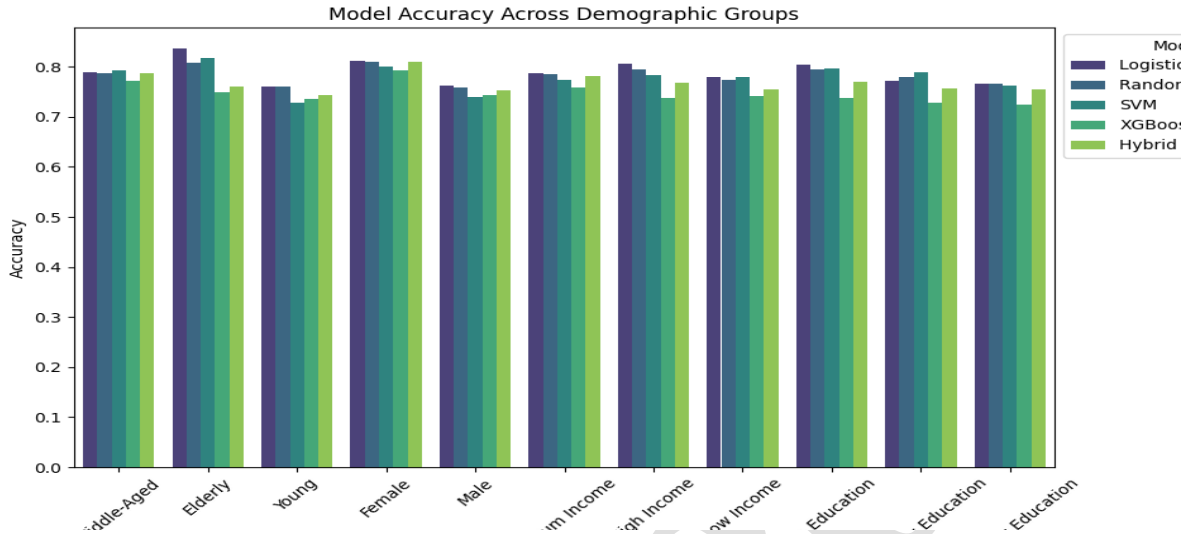

Figure 9: Model Accuracy across Demographic Groups

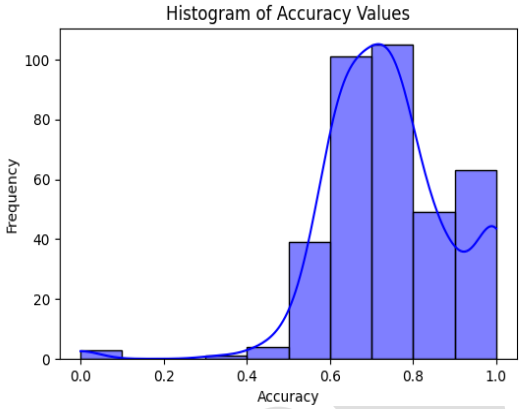

Figure 10: Histogram of Accuracy values

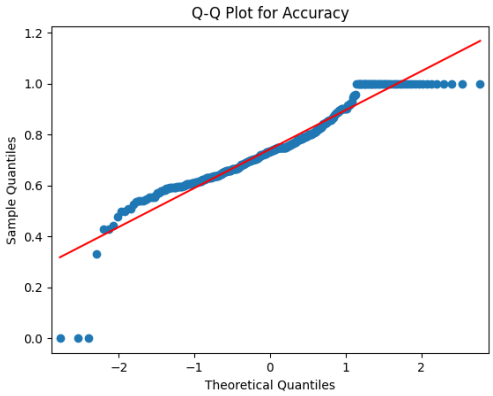

Figure 11: Q-Q Plot for Accuracy

Figures 9, 10, and 11 present model performance across demographics and distribution checks. Figure 9 shows that while complex models like XGBoost have higher accuracy and the Hybrid model has balanced accuracy, Logistic Regression offers more consistent performance across demographics, sometimes outperforming others in specific groups like the elderly. Figure 10's histogram reveals that accuracy values are right-skewed, peaking between 0.7 and 0.8, indicating a deviation from normality. Figure 11's Q-Q plot further confirms non-normality, as points deviate from the diagonal line, especially in the tails, suggesting skewness or possible outliers in the data.

4. DISCUSSION

This study developed and compared predictive models for Type 2 diabetes (T2D) using statistical and machine learning techniques, addressing challenges of class imbalance and model interpretability. The demographic distribution revealed a higher prevalence of diabetes among females (68%) than males (32%). This aligns with findings by Angell et al. (2022), who highlighted hormonal fluctuations,

gestational diabetes, and socio-cultural norms limiting female physical activity as underlying risk factors. It also reflects the Social Determinants of Health **(SDH)** framework (World Health Organization, 2022), which positions gender as a structural determinant influencing health outcomes. However, despite numerical dominance, gender was assigned low predictive weight in the models, suggesting that when lifestyle and clinical factors are accounted for, biological sex alone exerts a reduced independent influence.

Educational level, where 78.7% of participants had basic or no education (47% with no education and 36% with secondary education), theoretically suggests a risk pathway through poor health literacy (Hill-Briggs et al., 2020; Marciano et al., 2019). The SDH framework also links low education to unhealthy behaviors. Yet, education showed low feature importance in the models, possibly because its effect is indirectly mediated by lifestyle behaviors such as poor diet and physical inactivity, which were directly measured.

Similarly, income status showed 48.6% low-income participants, affirming SDH theory (Marmot, 2017; Ekure et al., 2022), where economic deprivation constrains access to healthy food and healthcare. Nonetheless, income had no strong predictive impact, possibly reflecting the urban Abuja context, where risk exposures such as processed food consumption and sedentary lifestyles are widespread across income groups, weakening the discriminatory role of income.

Lifestyle variables were more predictive: 64.2% of those with diabetes reported no physical activity, consistent with Popkin and Ng (2022) on urbanization-driven sedentary living. Higher feature importance was assigned to BMI and diet score, supporting Carbone et al. (2019), who showed direct links between these factors and glucose metabolism, highlighting that modifiable behaviors exert strong, immediate effects on T2D risk.

Family history, reported by 59.1% of those with diabetes, was consistent with findings by Berumen et al. (2023) and Szczerbinski & Florez (2023), emphasizing genetic predisposition. However, family history had lower predictive weight compared to modifiable factors, aligning with Edlitz & Segal (2022), who argued that machine learning models prioritize dynamic lifestyle features over static genetic markers.

Model evaluation showed that XGBoost was the strongest individual model, while the Hybrid Ensemble outperformed all other models overall. This supports findings by Stiglic et al. (2021) and Jiang et al. (2024) that ensemble methods handle complex, high-dimensional health data more effectively. The hybrid model retained interpretability through its logistic regression component (Rahman et al., 2023).

Temporal analysis revealed an increase in diabetes cases until 2023, followed by a decline in 2024. Without sequence modeling, it is difficult to determine whether this trend reflects genuine public health interventions or natural fluctuations (Yoshioka et al., 2024). Nonetheless, the stability of model performance over time underscores the robustness of the hybrid approach.

5. CONCLUSION

This study developed predictive models for Type 2 diabetes mellitus (T2DM) using logistic regression, machine learning algorithms, and ensemble techniques. By integrating clinical, socioeconomic, and lifestyle variables and applying SMOTE to address class imbalance, the models achieved high predictive

accuracy. The Weighted Soft Voting Ensemble performed best across all metrics, with an AUC-ROC of 0.931 and a Balanced Accuracy of 0.890, outperforming Random Forest, XGBoost, and SVM classifiers. Feature selection identified Age, BMI, Family history of diabetes, hypertension, and physical activity level as the most significant predictors.

REFERENCE

Angell, B., Sanuade, O., Adetifa, I. M., Okeke, I. N., Adamu, A. L., Aliyu, M. H. & Abubakar, I. (2022). Population health outcomes in Nigeria compared with other West African countries, 1998–2019: A systematic analysis for the Global Burden of Disease Study. *The Lancet*, *399*(10330), 1117–1129.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Berumen, J., Orozco, L., Gallardo-Rincón, H., Rivas, F., Barrera, E., Benuto, R. E., ... & Tapia-Conyer, R. (2023). Sex differences in the influence of type 2 diabetes (T2D)-related genes, parental history of T2D, and obesity on T2D development: A case–control study. *Biology of Sex Differences*, *14*(1), 39.

Carbone, S., Del Buono, M. G., Ozemek, C., & Lavie, C. J. (2019). Obesity, risk of diabetes and role of physical activity, exercise training and cardiorespiratory fitness. Progress in cardiovascular diseases, 62(4), 327-333.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273-297.

Edlitz, Y., & Segal, E. (2022). Prediction of type 2 diabetes mellitus onset using logistic regression-based scorecards. *Elife*, *11*, e71862.

Ekure, E., Ovenseri-Ogbomo, G., Osuagwu, U. L., Agho, K. E., Ekpenyong, B. N., Ogbuehi, K. C., ... & Naidoo, K. S. (2022). A systematic review of diabetes risk assessment tools in sub-Saharan Africa. *International Journal of Diabetes in Developing Countries*, *42*(3), 380–393.

Gkrinia, E. M. M., & Belančić, A. (2025). A narrative review on the risk factors and healthcare disparities of type 2 diabetes. *Diabetology*, *6*(4), 25.

Hill-Briggs, F., Adler, N. E., Berkowitz, S. A., Chin, M. H., Gary-Webb, T. L., Navas-Acien, A., ... & Haire-Joshu, D. (2020). Social determinants of health and diabetes: A scientific review. *Diabetes Care*, *44*(1), 258.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons.

Ibitoye, A. O., Akinyemi, J. D., & Onifade, O. F. (2024). Machine learning-based diabetes risk prediction using associated behavioral features. *Computing Open, 2*.

International Diabetes Federation. (2021). *IDF Diabetes Atlas* (10th ed.). Brussels, Belgium: International Diabetes Federation. Available at: https://www.diabetesatlas.org

Jiang, J., Xin, C., Wu, S., Chen, W., Li, H., & Ran, Z. (2024). Enhancing concrete workability prediction through ensemble learning models: Emphasis on slump and material factors. *Advances in Civil Engineering*, *2024*(1), 4616609.

Kost, S., Rheinbach, O., & Schaeben, H. (2021). Using logistic regression model selection towards interpretable machine learning in mineral prospectivity modeling. Geochemistry, 81(4), 125826.

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable AI: A review of machine learning interpretability methods. *Entropy*, *23*(1), 18.

Marciano, L., Camerini, A. L., & Schulz, P. J. (2019). The role of health literacy in diabetes knowledge, self-care, and glycemic control: A meta-analysis. *Journal of General Internal Medicine, 34*, 1007–1017.

Marmot, M. (2017). The health gap: The challenge of an unequal world: The argument. *International Journal of Epidemiology*, *46*(4), 1312–1318.

Matougui, Z., & Zouidi, M. (2025). A temporal perspective on the reliability of wildfire hazard assessment based on machine learning and remote sensing data. *Earth Science Informatics*, *18*(1), 1-20.

Nolan, C. J., & Prentki, M. (2019). Insulin resistance and insulin hypersecretion in the metabolic syndrome and type 2 diabetes: Time for a conceptual framework shift. *Diabetes and Vascular Disease Research*, *16*(2), 118–127.

Okwori, O. A., Agana, M. A., & Ofem, O. A. (2024). Application of support vector machine model for prediction of stroke vulnerability status. *Asian Journal of Pure and Applied Mathematics*, 174–181.

Olanrewaju, E. A. A. O., & Ibikunle, A. I. (2022). Socio-demographic profile, asymptomatic malaria parasitaemia and glycemic control among midled-aged and elderly type 2 diabetes mellitus patients in rural Southwestern Nigeria: A cross-sectional study. *African Journal of Diabetes Medicine*, *30*(3).

Olusola, A. A., Joel, O. O., Matthew, A. S., Toyin, E. O., Olusegun, O., Emmanuel, O. A. C., ... & Ibraheem, A. (2023). Socio-demographic, clinical profile and status of glycaemic control among diabetic patients in a rural tertiary hospital. *Acta Scientific Medical Sciences*, *7*(5).

Popkin, B. M., & Ng, S. W. (2022). The nutrition transition to a stage of high obesity and noncommunicable disease prevalence dominated by ultra- processed foods is not inevitable. *Obesity Reviews*, *23*(1), e13366.

Rahman, M. A., Islam, M. S., Miazee, A. J., & Rahman, M. S. (2024). Machine learning-based prediction of diabetes mellitus using clinical data: A comparative analysis. *International Journal of Intelligent Systems and Applications in Engineering*, *12*(12s), 332–343.

Rokach, L. (2010). Ensemble-based classifiers. *Artificial intelligence review*, *33*(1), 1-39.

Shehu, A., & Baha, B. Y. (2024). Artificial neural network prediction model for maternal health services quality in Nigeria. *International Journal of Development Mathematics (IJDM)*, *1*(1).

Stiglic, G., Wang, F., Sheikh, A., & Cilar, L. (2021). Development and validation of the type 2 diabetes mellitus 10-year risk score prediction models from survey data. *Primary Care Diabetes*, *15*(4), 699–705.

Szczerbinski, L., & Florez, J. C. (2023). Precision medicine of obesity as an integral part of type 2 diabetes management–past, present, and future. *The Lancet Diabetes & Endocrinology*, *11*(11), 861–878.

World Health Organization. (2022). *Global report on diabetes*. World Health Organization. https://apps.who.int/iris/handle/10665/204871

Yoshioka, H., Jin, R., Hisaka, A., & Suzuki, H. (2024). Disease progression modeling with temporal realignment: An emerging approach to deepen knowledge on chronic diseases. *Pharmacology & Therapeutics*, 108655.