Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 9, September 2024

Diabetes Prevalence Prediction Among Academic Staff of Tertiary Institutions in South-Western Nigeria using Machine Learning Techniques

Olanegan Olayemi Ola
Department of General Studies
Federal Polytechnic, Ile Oluji
Ondo State, Nigeria

Aladesote Olomi Isaiah
Department of Computer Science
Federal Polytechnic, Ile Oluji
Ondo State, Nigeria

***Abstract*— Diabetes mellitus (DM) is a chronic health condition characterized by inadequate insulin production or ineffective utilization, which leads to elevated blood sugar levels. In Nigeria, the prevalence of DM has seen a sharp rise, particularly among individuals aged 20-79, with significant increases projected over the coming decades. Among academic staff in Southwestern Nigeria, the high demands of their professional duties have negatively impacted their health, leading to increased susceptibility to diabetes. This study seeks to address the limitations of existing diabetes prediction models, which primarily rely on secondary datasets, by utilizing primary data collected from academic staff in Southwestern Nigeria. A comprehensive diabetes prediction model is formulated using machine learning and ensemble methods such as K-Nearest Neighbors (KNN), Random Forest, and Logistic Regression. By employing feature selection techniques and model validation methods, the study offers novel insights into diabetes risk factors among academic staff. The results demonstrate that ensemble models, particularly Voting and AdaBoost, consistently outperformed individual machine learning algorithms, showcasing their potential for accurate diabetes prediction. This study provides a tailored and context-specific approach to diabetes prediction, with implications for public health interventions targeting tertiary institutions.**

***Keywords*—adaboost; diabetes mellitus; machine learning; mathematical model; voting.**

I. INTRODUCTION

Diabetes mellitus (DM) is a severe health condition that arises when the pancreas does not produce enough insulin and the body cannot effectively utilize the insulin produced responsible for regulating blood sugar levels [1]. It is a disease associated with microvascular and macrovascular complications, with serious effects on the quality of life [2]. The prevalence of diabetes among Nigerians aged 20-79 based on the International Diabetes Federation (IDF) data suggests a rapid increase over the years. From 2000 to 2011, the number of people with diabetes surged by a staggering 1358.69%. This upward trend continued, albeit at a slower pace, with an 18.63% increase from 2011 to 2021. Projections suggest that this rise will persist, with a 36.38% increase expected between 2021 and 2030, and a further 61.65% increase from 2030 to 2045 [3]. Notably, this age group includes academic staff in tertiary institutions, highlighting the growing public health challenge posed by diabetes in Nigeria.

Nigeria, situated in West Africa, is one of the most populous countries in the world. Its population has been growing rapidly and is projected to continue increasing in the coming decades [1], [4]. Southwestern Nigeria, the study area shown in Figure 1, is a region rich in cultural heritage and economic significance. It plays a crucial role in the nation's socio-economic landscape and offers a unique blend of ancient customs and contemporary advancements. The region is a hub of educational institutions, contributing to its dynamic and influential educational position within Nigeria.

Education is a significant driver of all socioeconomic, political, scientific, and technological development. As a result, higher education is an epicenter for knowledge and its applications. As such, it contributes to economic growth and development by encouraging invention and innovative ideas [4]. Achieving a higher level of productivity requires a healthy and sound academic staff. However, diabetes is a disease that can reduce the productivity level of any academic staff, suffering from this health challenge.

Research has shown that academic staff in Nigerian tertiary institutions sacrifice their well-being in favour of their professional duties (teaching, research, and community service), at the expense of their well-being. This imbalance not only jeopardizes their health but also significantly diminishes their overall productivity [5]–[7]. In addition, most existing work on diabetes prediction relies solely on secondary data (Pima Indian Diabetes Dataset) for detecting and predicting diabetes. This limitation underscores the need for a more comprehensive and context-specific investigation using primary data from academic staff in Southwestern Nigeria, to predict diabetes prevalence in tertiary institutions. The main contributions of the proposed study are to:

I. gather novel, context-specific datasets directly from academic staff at Southwestern tertiary institutions in Nigeria. This primary data collection addresses the gap in existing literature, which has largely relied on secondary sources.
II. formulate a mathematical model for diabetes prevalence and prediction
III. use machine learning and ensemble methods to predict diabetes prevalence among Nigerian academics. This new approach will provide valuable insights into diabetes risk and prevalence in tertiary institutions

IV. PRESENT FUTURE WORK

Fig. 1. Map of South Western Nigeria [8]

The rest of the paper is outlined as follows: Section 2 delves into the related study. Section 3 presents the methodology of the study. Section 4 presents the result and discussion of the study while Chapter 5 presents the study's conclusion.

II. REVIEW OF RELATED WORKS

This section reviews existing research on the prediction of Diabetes mellitus. The study addresses diabetes prediction using supervised learning by comparing the K-Nearest Neighbor (KNN) and Naive Bayes algorithms. Using the Pima Indians Diabetes Database from Kaggle and 10-fold cross-validation for model validation, the results showed that Naive Bayes outperformed KNN, achieving higher accuracy, precision, and recall. The research highlights the potential of machine learning in early diabetes detection, suggesting that Naive Bayes is a more reliable method for predicting diabetes [9].

The study addresses the problem of early detection and prediction of diabetes due to the lack of a permanent cure and the critical importance of early diagnosis. It utilizes various machine learning techniques, including Naive Bayes (NB), Support Vector Machine (SVM), Logistic Regression (LR), AdaBoost, Random Forest (RF), K Nearest Neighbor (KNN), Decision Tree (DT), and Neural Networks (NN) with different hidden layers and epochs, to accurately predict diabetes. Using the Pima Indian Diabetes (PID) dataset from the UCI Machine Learning Repository, the results show that Logistic Regression (LR) and Support Vector Machine (SVM) were particularly effective in predicting diabetes. Additionally, a Neural Network model with two hidden layers achieved an accuracy of 88.6% [10].

The study aims to improve diabetes prediction using various machine learning (ML) techniques, including K-Nearest Neighbors (KNN), Random Forest (RF), Support Vector Machine (SVM), Artificial Neural Network (ANN), and Decision Tree. Using the Pima Indian Diabetes Dataset and thorough data preprocessing, the results indicate that the Random Forest algorithm outperforms the others, achieving the highest accuracy at 88.31% [11].

To address accurate diabetes prediction and handle imbalanced datasets, the study employed a Support Vector Machine, Deep Learning, and Random Forest on the Pima Indian Diabetes Dataset. The experimental results show that Random Forest outperforms the others with the highest accuracy of 83.67%. Future work should explore more advanced machine-learning techniques on this dataset [12].

The study addresses the issue of predicting diabetes mellitus (DM) using machine learning algorithms to enhance early diagnosis and improve prediction accuracy. It employs various machine learning models (Support Vector Machine, Naïve Bayes, Decision Stump), the AdaBoostM1 ensemble method, and a proposed method on the Pima Indian Diabetes Dataset. The proposed method outperforms other models with an accuracy of 90.36% and a 9.64% error rate. However, the study does not address the potential impact of additional features on model performance, which could enhance the accuracy and reliability of the predictions [13].

The study developed a system to predict diabetes risk levels in patients with high accuracy using machine learning, the research employed the Pima Indian Diabetes Dataset on the following models: Decision Tree, Artificial Neural Network (ANN), Naive Bayes, and Support Vector Machine (SVM). This study demonstrates the potential of machine learning in predicting diabetes risk, with the Decision Tree model showing promising results, with an accuracy of 85% [14].

To solve the early prediction of diabetes to facilitate timely intervention and management of the disease, the researchers employed various machine learning classifiers (K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Decision Tree, Logistic Regression, Random Forest and Gradient Boosting) and ensemble techniques to predict diabetes mellitus with Pima Indians Diabetes Database from the UCI Machine Learning Repository. The article shows that the ensemble method outperforms other machine-learning methods [15].

Most existing studies on diabetes prediction rely heavily on secondary data, particularly the Pima Indian Diabetes Dataset, limiting the applicability of findings to other populations. While machine learning models like Naive Bayes and Random Forest have shown promise, their effectiveness is constrained by the relevance of the dataset used. To address this limitation, using primary data from academic staff in Southwestern Nigeria offers a more tailored approach, improving prediction accuracy by capturing specific characteristics of this population. This will enhance the reliability of machine learning models and lead to more effective early detection of diabetes.

III. METHODOLOGY

This section outlines a detailed approach to model formulation and diabetes prediction classification.

A. Model Formulation

The mathematical model developed, as depicted in Figure 2, involves five partitions: the Susceptible $S_P(t)$, the Diabetes $D_m(t)$, the Diabetes with Complication $D_mC_O(t)$, the Diabetes without Complication $D_mC(t)$, and the Hospitalised $H_P(t)$. The first partition (Susceptible) implies that the entire population is Susceptible based on family history (η_s) with diabetes and unhealthy lifestyle (γ_s) such as physical inactiveness, improper diet, unmanaged stress, obesity, and smoking. The second partition, Diabetes is divided into two compartments: Diabetes without complication (κs) and Diabetes with complication (λ_s). The Diabetes with complication compartment leads to the Hospitalised compartment (H_t) with Neuropathy, Retinopathy, and Nephropathy cases. The Neuropathy case can be managed, and the complication can be recovered from, while the other two cases lead to disability and mortality.

$$\frac{dS}{dt} = \frac{\lambda}{\alpha(1-\rho)D} - \left[(1-\varepsilon)\beta\frac{S_p}{N} + \mu S_p\right] \tag{1}$$

$$\frac{dD_m}{dt} = (1-\varepsilon)\beta\frac{S_p}{N} + \alpha\rho D_m - (mD_m + yD_m) \tag{2}$$

$$\frac{dD_m C_0}{dt} = yD_m + \mu N_p - (\lambda D_m C_0 + \sigma D_m C_0) \tag{3}$$

$$\frac{dD_{m_c}}{dt} = \lambda D_m C_0 + mD_m - \frac{\phi I}{1+KI} \tag{4}$$

$$\frac{dH_p}{dt} = \frac{\phi I}{1+KI} - \left(N_p + R_p - N_{e_p} - \mu_{N_p}\right) \tag{5}$$

B. Diabetes Prediction and Classification

The study developed a Diabetes Diagnosis model using a dataset collected from academic staff in southwestern Nigeria. The dataset was normalized with min-max scaling, ensuring all numerical features were adjusted to a range of 0 to 1 while preserving their original distribution. This scaling was applied before splitting the data to maintain consistency across training and testing sets, preventing data leakage. Significant features were selected using Gain Ratio and Information Gain, and SMOTE was applied to address data imbalance. The dataset was divided using three validation methods: 10-fold cross-validation, 80/20, and 70/30 splits. Two experiments were conducted using three machine learning algorithms—K-Nearest Neighbour (KNN), Support Vector Machine (SVM), and Logistic Regression (LR)—along with two ensemble models: Voting and AdaBoost. Weka, an open-source machine learning tool, was used for data analysis, offering features for preprocessing, classification, clustering, regression, visualization, and feature selection.

C. Statistical Analysis of the Dataset

The dataset for this experiment was gathered through Google Forms from academic staff at tertiary institutions in Southwestern Nigeria, with 149 male and 59 female respondents. It comprises 208 instances and 18 features. This dataset consists of 122 diabetes cases, with 43 classified as having Diabetes with Complications and 79 as having Diabetes without Complications. Additionally, there are 86 non-diabetes cases. Table 1 presents the attribute descriptions. The label is based on respondents' typical fasting blood sugar levels. Respondents with fasting blood sugar levels between 70 mg/dL (3.9 mmol/L) and 100 mg/dL (5.6 mmol/L) are labeled as Normal. Those with levels below 70 mg/dL (3.9 mmol/L) or equal to or above 126 mg/dL (7 mmol/L) are labeled as Diabetes with Complications. Respondents with fasting blood sugar levels between 100 mg/dL (5.6 mmol/L) and 125 mg/dL (6.9 mmol/L) are labeled as Diabetes without Complications. Additionally, for respondents who do not know their fasting blood sugar level, family history was used to determine the label; those with a family history of diabetes are labeled as Diabetes without Complications since every staff member is considered susceptible to diabetes at the initial stage.

TABLE 1 DATASET DESCRIPTION.

Feature	*Description*
Age	Age of the Academic Staff
Sex	Male or Female
S-Intake	Rate of Sugar Intake
FV-Intake	Rate of Fruit and Vegetable Intake
BD-Intake	Rate of Balanced Diet Intake
D-Info	Rate of Diabetes Reliable Information
DEBG-Info	Rate of Diet, Exercise, and Blood Glucose Control
BG-Monitoring	Blood Glucose Monitoring Rate
D-D	Have you ever been diagnosed with Diabetes?
PA-Barrier	Physical Active Barrier
SLW-Rate	Stress Level Work Rate
SLH-Rate	Stress Level at Home Rate
P-Fitness	Physical Fitness
HL-Style	Healthy Life-Style Choices
WL-Lifestyle	Workplace Healthy Lifestyle
EHW-Policy	Policies in Support of Employee Health and Wellness
RH-Checkups	Regular Health Check-ups

D. Statistical Analysis of the Dataset

Statistical analysis tools are essential for identifying important information for proper preprocessing before developing a model. Figure 2 illustrates the correlation between the variables and the target class. The heatmap reveals that HL Style and WH Lifestyle have the strongest positive correlation (0.67), while SLW Rate and SLH Rate show moderate positive correlations (0.33). On the negative side, SLW Rate and WH Lifestyle as well as SLH Rate and WH Lifestyle exhibit moderate negative correlations (-0.43). These correlations highlight important relationships within the dataset, offering benefits like feature reduction and improved predictive modeling by identifying potentially redundant features. Additionally, understanding these correlations provides deeper insights into the data, helping to inform more effective data-driven decisions.

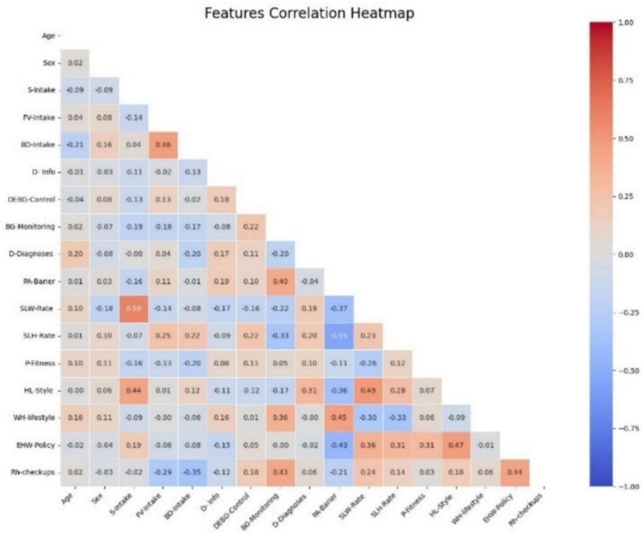

Fig. 2. Feature Correlation Heatmap

E. Preprocessing

Data preprocessing is critical in converting raw data into a format suitable for effective analysis. Our study utilized a range of preprocessing techniques, implemented through Python

Scikit-learn and Pandas libraries. This preparation was essential for machine learning models, which operate exclusively on numeric inputs and outputs. The preprocessing pipeline we developed ensured proper formatting, scaling, and encoding of all variables, thereby optimizing the performance of our subsequent modeling work. We implemented category reduction in our preprocessing pipeline to optimize model performance and reduce computational complexity. We applied this technique to categorical variables that had many unique values. For example, we combined the HIGH and VERY HIGH categories into a single HIGH category. Similarly, we merged the LOW and VERY LOW categories into one LOW category. This process simplified our feature space and helped prevent overfitting to rare categories. After reducing the categories, we used label encoding to transform them into numerical format, making the data suitable for our machine-learning algorithms. In addition, we employed label encoding to transform categorical data into numerical format, a necessary step before model training and evaluation. This approach allowed us to retain the original information while making it compatible with our chosen machine-learning algorithms.

F. Description of the proposed techniques

(a) K-nearest neighbor (KNN): KNN [16] identifies a group of k similar objects from the training set that are closest to the test object. The assigned label is based on the most frequent class within this group. Its straightforward nature makes it easy to understand and use.

(b) Random Forest: A random forest is a group of tree-based predictors, where each tree is built using a randomly selected set of features. This ensemble-supervised machine-learning method has recently gained significant attention [17].

(c) Multilayer Perception: Multilayer Perceptron (MLP) is a widely used supervised learning method in artificial neural networks. It is inspired by the human brain and nervous system and consists of three layers: input, hidden, and output. MLP is commonly applied to various predictive problems, as noted in numerous studies [18], [19].

(d) Voting Ensemble: A voting ensemble assigns classifiers to weighted categories based on training data. This study uses majority voting, combining the strengths of multiple machine learning classifiers.

(e) Adaboost: AdaBoost (Adaptive Boosting) improves weak learners by adjusting the weights of misclassified instances, focusing on difficult examples to create a stronger combined model. While effective in reducing errors, it can be sensitive to noisy data [20].

G Performance Metrics

Performance Metrics are used to assess how well a model performs. The following metrics are used: Accuracy(A), Precision (P), Recall (R), Specificity (S), and F1-Score (F).

$$A = \frac{\Omega + \Omega_0}{\kappa_T} \tag{6}$$

$$P = \frac{\Omega}{\Omega + \Omega'} \tag{7}$$

$$R = \frac{\Omega}{\Omega + \Omega_0'} \tag{8}$$

$$S = \frac{\Omega_0}{\bar{\Omega} + \Omega_0'} \tag{9}$$

$$F = 2 * \left(\frac{P * R}{P + R} \right) \tag{10}$$

Where: Ω= Number correctly classified as Diabetes, Ω_0 =Number correctly classified as non-diabetes, κ_T = Total Number of Prediction results, Ω' = Number incorrectly classified as Diabetes, Ω_0' = Number incorrectly classified as non-diabetes, $\bar{\Omega}$ = Number correctly classified as non-diabetes.

IV. RESULTS AND DISCUSSION

This section analyzes the results of the feature selection techniques and evaluates the model's performance on the dataset.

A. Feature Selection Results

The results of the feature selection techniques show that the Gain Ratio selected 8 features: BD-Intake, D-Info, FV-Intake, Age, SLH-Rate, D-Diagnoses, HL-Style, and SLW-Rate. Information Gain selected 9 features: D-Info, Age, SLH-Rate, S-Intake, FV-Intake, BD-Intake, HL-Style, SLW-Rate, and D-Diagnoses.

B. Classification Results

This section presents the classification results of 3 machine-learning techniques and 2 ensemble-based models on the diabetes dataset. The classification was performed on both the original dataset with duplicate records and a version without duplicates, using three dataset partitioning methods: 10-fold cross-validation, 80/20 split, and 70/30 split.

C. Classification Results on Selected Attributes Using Gain Ratio

Table II and Figures 4, 5 & 6 compare the performance of K-Nearest Neighbors (KNN), Random Forest, Multi-Layer Perceptron (MLP), Voting Ensemble, and AdaBoost using 10-fold cross-validation, 80/20 split, and 70/30 split validation techniques. AdaBoost and Voting Ensemble consistently achieve the highest accuracy and F1 score across all splits, indicating their robustness. While KNN performs well, it generally lags behind the ensemble methods. The consistency of metrics across different splits suggests the minimal impact of the validation method on model performance. Notably, the 80/20 split yields the best overall results, with each model achieving approximately 96.09% accuracy and an F1-Score of 96.50%, demonstrating well-balanced precision and recall.

The 70/30 split also performs well but shows slightly lower consistency. In contrast, 10-fold cross-validation provides robust metrics but with marginally lower accuracy and F1-Score, indicating that models generally perform better with the

80/20 split. Thus, the 80/20 split emerges as the most favorable validation method for achieving the highest and most consistent performance across models.

D. Classification Results on Selected Attributes Using Information Ratio

Table III and Figures 7, 8 & 9 compare the performance of K-Nearest Neighbors (KNN), Random Forest, Multi-Layer Perceptron (MLP), Voting Ensemble, and AdaBoost across three validation methods: 10-fold cross-validation, 80/20 split, and 70/30 split. Random Forest consistently shows the highest accuracy and F1-Score across all splits, particularly excelling in the 10-fold cross-validation with 98.28% accuracy and 98.35% F1-Score. Voting Ensemble and AdaBoost also perform strongly, with nearly identical results in the 80/20 split, both achieving 97.66% accuracy and 97.85% F1-Score. KNN is competitive but slightly lags behind the ensemble methods, while MLP has the lowest performance across all splits, indicating it may not generalize as well. Overall, the 10-fold cross-validation yields the highest metrics, especially for Random Forest and Voting Ensemble, highlighting their robustness and consistency.

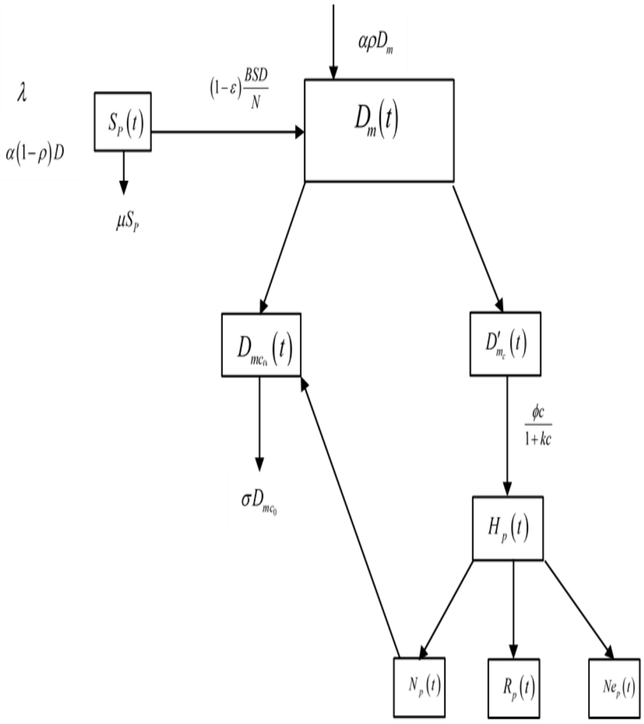

Fig. 3. Mathematical Model Chart for Diabetes Prediction

Table II: CLASSIFICATION RESULTS ON SELECTED ATTRIBUTES USING GAIN RATIO

Split	*Model*	*Accuracy (%)*	*Precision (%)*	*Recall (%)*	*F1-Score (%)*	*F-Measure (%)*
10-Fold Validation	KNN	95.14	95.7	95.1	95.40	95
	Random Forest	95.61	96.1	95.6	95.85	95.5
	MLP	93.26	93.5	93.3	93.40	93.1
	Voting Ensemble	95.45	96	95.5	95.75	95.4
	ADABOOST	95.77	96.1	95.8	95.95	95.7
80/20	KNN	96.09	96.9	96.1	95.40	95.7
	Random Forest	96.09	96.9	96.1	96.50	95.7
	MLP	96.08	96.9	96.1	96.50	95.7
	Voting Ensemble	96.09	96.9	96.1	96.50	95.7
	ADABOOST	96.09	96.9	96.1	96.50	95.7
70/30	KNN	96.34	96.6	96.3	96.45	96.2
	Random Forest	95.81	96.2	95.8	96.00	95.6
	MLP	96.09	96.9	96.1	96.50	95.7
	Voting Ensemble	96.34	96.6	96.3	96.45	96.2
	ADABOOST	95.81	96.2	95.8	96.00	95.6

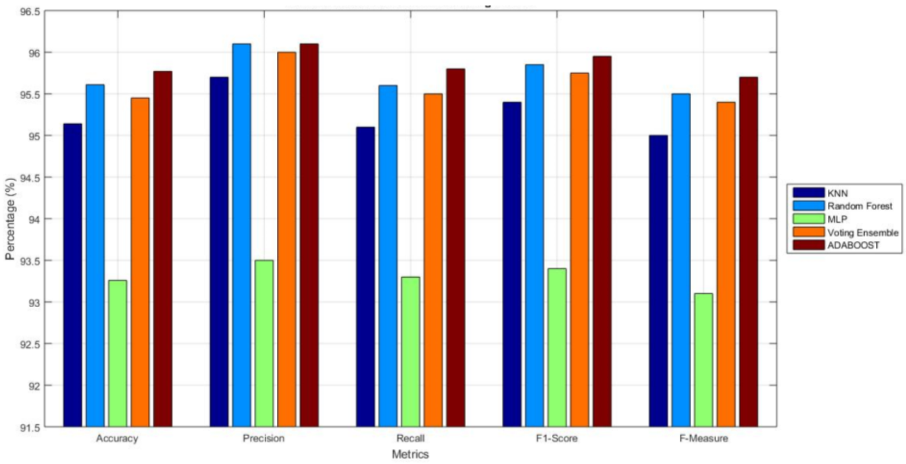

Fig. 4. Classification of selected attributes using Gain Ratio with 10-fold cross-validation on the algorithms

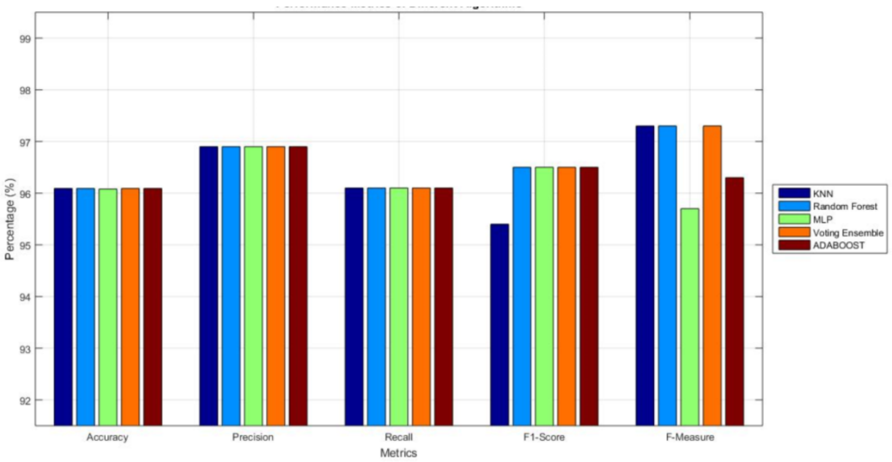

Fig. 5. Classification of selected attributes using Gain Ratio with 80/20 split on the algorithms.

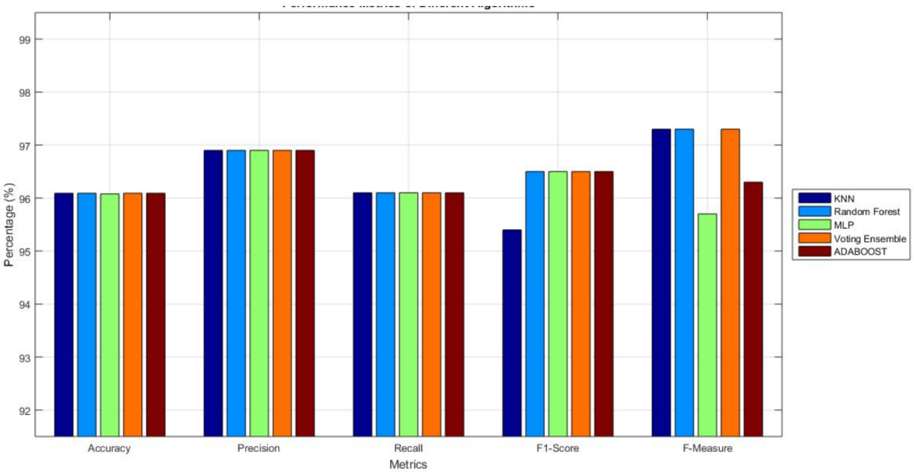

Fig. 6. Classification of selected attributes using Gain Ratio with 70/30 split on the algorithms.

Table III: CLASSIFICATION RESULTS ON SELECTED ATTRIBUTES USING INFORMATION GAIN

Split	Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	F-Measure (%)
10-Fold Validation	KNN	97.81	98	97.8	97.9	97.8
	Random Forest	98.28	98.4	98.3	98.35	98.3
	MLP	96.71	96.8	96.7	96.75	96.7
	Voting Ensemble	97.96	98.1	98	98.05	97.9
	ADABOOST	97.65	97.7	97.6	97.65	97.6
80/20	KNN	96.88	97.4	96.9	97.15	96.6
	Random Forest	97.66	98	97.7	97.85	97.5
	MLP	96.09	96.2	96.1	96.15	95.9
	Voting Ensemble	97.66	98	97.7	97.85	97.5
	ADABOOST	97.66	98	97.7	97.85	97.5
70/30	KNN	97.38	97.5	97.4	97.45	97.3
	Random Forest	97.38	97.5	97.4	97.45	97.3
	MLP	95.81	95.8	95.8	95.8	95.7
	Voting Ensemble	97.38	97.5	97.4	97.45	97.3
	ADABOOST	96.34	96.3	96.3	96.3	96.3

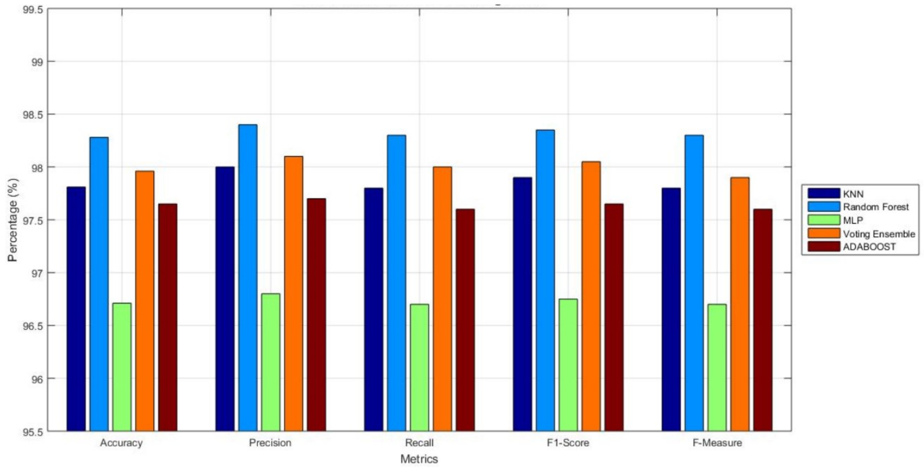

Fig. 7. Classification of selected attributes using Gain Ratio with 10-fold cross-validation on the algorithms.

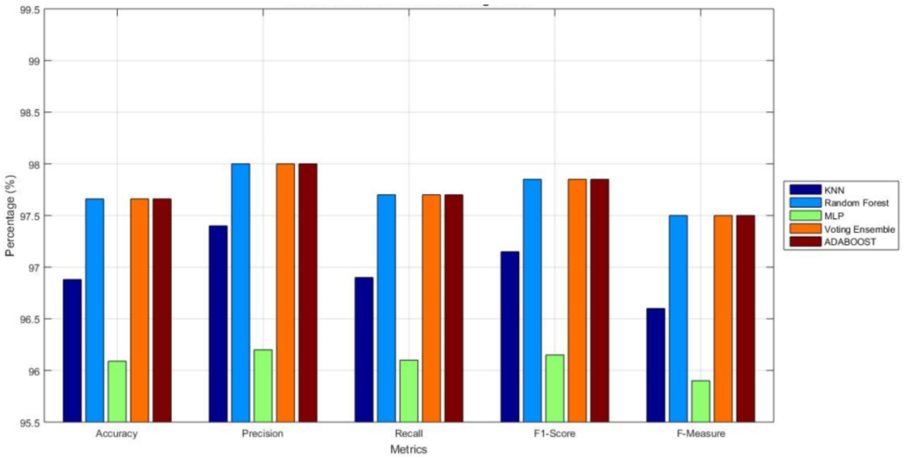

Fig. 8. Classification of selected attributes using Gain Ratio with 80/20 split on the algorithms

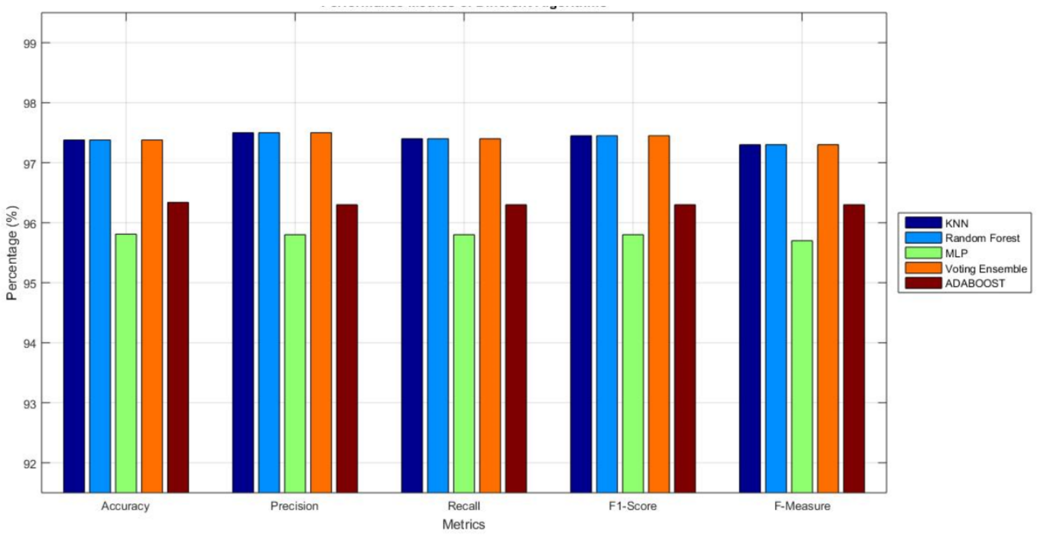

Fig. 9. Classification of selected attributes using Gain Ratio with 80/20 split on the algorithms.

V. CONCLUSION

The study successfully developed a predictive model for diabetes prevalence among academic staff in Southwestern Nigeria, addressing the gap in existing research that relies heavily on secondary datasets. By leveraging primary data and advanced machine learning techniques, this research offers a more precise and population-specific approach to diabetes prediction. The results show that ensemble methods, particularly Voting and AdaBoost, provide superior accuracy and robustness across various validation techniques. These findings underscore the importance of early detection and intervention strategies for diabetes in academic institutions, where the well-being of staff is critical to maintaining high productivity levels. Future research should explore the integration of additional health-related features and expand the dataset to include more regions within Nigeria, thereby improving model generalization and prediction accuracy.

ACKNOWLEDGMENT

Researchers of this present study acknowledge and sincerely appreciate the academic developer, (Tertiary Education Trust Fund (TETFund, Nigeria) for sponsoring this research via its Institution Based Research (IBR) Intervention. The researchers also acknowledge and appreciate the Management of the Federal Polytechnic, Ile-Oluji, Ondo State for its support.

REFERENCES

[1] M. U. Muhammad, R. Jiadong, N. S. Muhammad, and B. Nawaz, “Stratified diabetes mellitus prevalence for the Northwestern Nigerian States, a data mining approach,” Int. J. Environ. Res. Public Health, vol. 16, no. 21, 2019, doi: 10.3390/ijerph16214089.

[2] D. Adje, “Assessment of Knowledge of Self Care and Patient Satisfaction with Care in Patients with Type 2 Diabetes in Warri, Delta State, Nigeria,” Ann. Med. Health Sci. Res., pp. 333–337, 2022, [Online]. Available: https://www.amhsr.org/abstract/assessment-of-knowledge-of-self-care-and-patient-satisfaction-with-care-in-patients-with-type-2-diabetes-in-warri-delta--11475.html

[3] “IDF Diabtes Atlas,” Nigeria Diabetes report 2000 — 2045, 2021. https://diabetesatlas.org/data/en/country/145/ng.html (accessed Jul. 13, 2024).

[4] W. H, H. OM, E. RS, A. W. SM, and I. MM, “Impact of Diabetes Mellitus on Work Productivity in Construction Industry,” Egypt. J. Occup. Med., vol. 40, no. 1, pp. 129–143, 2016, doi: 10.21608/ejom.2016.836.

[5] N. O. Orunbon and M. Mohammed, “Effect of Occupational Stress on Academic Staff Productivity of Public Tertiary Educational Institutions in Lagos State, Nigeria,” J. Educ. Pract., no. August, 2023, doi: 10.7176/jep/14-11-02.

[6] L. U. Akah et al., “Occupational Stress and Academic Staff Job Performance in Two Nigerian Universities,” J. Curric. Teach., vol. 11, no. 5, pp. 64–78, 2022, doi: 10.5430/jct.v11n5p64.

[7] I. J. Iguoba, “STRESS MANAGEMENT AND ACADEMIC STAFF PERFORMANCE OF TERTIARY INSTITUTIONS IN EDO STATE,” vol. 2, no. 1, 2023.

[8] L. A. S. Agbetoye and O. A. Oyedele, “Investigations into some engineering properties of Gari produced in south–western Nigeria,” Int. J. AgriScience, vol. 3, no. 10, pp. 728–742, 2013.

[9] M. E. Febrian, F. X. Ferdinan, G. P. Sendani, K. M. Suryanigrum, and R. Yunanda, “Diabetes prediction using supervised machine learning,” Procedia Comput. Sci., vol. 216, no. 2022, pp. 21–30, 2022, doi: 10.1016/j.procs.2022.12.107.

[10] J. J. Khanam and S. Y. Foo, “A comparison of machine learning algorithms for diabetes prediction,” ICT Express, vol. 7, no. 4, pp. 432–439, 2021, doi: 10.1016/j.icte.2021.02.004.

[11] S. NAHZAT and M. YAĞANOĞLU, “Makine Öğrenimi Sınıflandırma Algoritmalarını Kullanarak Diyabet Tahmini,” Eur. J. Sci. Technol., no. 24, pp. 53–59, 2021, doi: 10.31590/ejosat.899716.

[12] A. Yahyaoui, A. Jamil, J. Rasheed, and M. Yesiltepe, “A Decision Support System for Diabetes Prediction Using Machine Learning and Deep Learning Techniques,” 1st Int. Informatics Softw. Eng. Conf. Innov. Technol. Digit. Transform. IISEC 2019 - Proc., pp. 1–4, 2019, doi: 10.1109/UBMYK48245.2019.8965556.

[13] M. Alehegn, R. Joshi, and P. Mulay, “Analysis and prediction of diabetes mellitus using machine learning algorithm,” Int. J. Pure Appl. Math., vol. 118, no. Special Issue 9, pp. 871–878, 2018.

[14] P. Sonar and K. Jaya Malini, “Diabetes prediction using different machine learning approaches,” Proc. 3rd Int. Conf. Comput. Methodol. Commun. ICCMC 2019, no. Iccmc, pp. 367–371, 2019, doi: 10.1109/ICCMC.2019.8819841.

[15] M. K. Hasan, M. A. Alam, D. Das, E. Hossain, and M. Hasan, “Diabetes prediction using ensembling of different machine learning classifiers,” IEEE Access, vol. 8, pp. 76516–76531, 2020, doi: 10.1109/ACCESS.2020.2989857.

[16] M. Steinbach and P. N. Tan, “kNN: k-Nearest Neighbors,” The Top Ten Algorithms in Data Mining. pp. 151–161, 2009. doi: 10.1201/9781420089653-15.

[17] V. Y. Kullarni and P. K. Sinha, "Random Forest Classifier: A Survey and Future Research Directions," Int. J. Adv. Comput., vol. 36, no. 1, pp. 1144–1156, 2013.

[18] A. Darvishan, H. Bakhshi, M. Madadkhani, M. Mir, and A. Bemani, "Application of MLP-ANN as a novel predictive method for prediction of the higher heating value of biomass in terms of ultimate analysis," Energy Sources, Part A Recover. Util. Environ. Eff., vol. 40, no. 24, pp. 2960–2966, 2018, doi: 10.1080/15567036.2018.1514437.

[19] W. Cao, X. Wang, Z. Ming, and J. Gao, "A review on neural networks with random weights," Neurocomputing, vol. 275, pp. 278–287, 2018, doi: 10.1016/j.neucom.2017.08.040.

[20] Z.-H. Zhou, "Ensemble Method Foundation." CRC Press, Taylor and Francis Group, LLC, pp. 1–232, 2012.