RESEARCH PAPER

Machine Learning-Driven Insights into Socio-Economic, Demographic, and Health Determinants of Type-II Diabetes: A Comparative Analysis of Tuned Hyperparameter Random Forest and XGBoost Models Using NFHS-5 Data

Meghna Athwani[1], Hitesh Athwani[2], Kuhu Awasthi[3*], Deepankshi Agnihotry[4]

[1]Assistant Professor cum Statistician, Department of Community Medicine, Autonomous State Medical College, Amethi – 229309

[2]Ph.D. Scholar, Department of Statistics, University of Lucknow, Lucknow, India - 226007

[3*]Assistant Professor, Department of Management, Bennett University, Greater Noida, India – 201310 | Email: kuhu.awasthi@bennett.edu.in | +91-6393685095 (Corresponding Author)

[4]Programme Officer, Scientist B, Vikram A Sarabhai Community Science Centre, Ahmedabad, Gujarat 380009

ABSTRACT

Type-II diabetes represents a growing public health challenge in India, necessitating accurate predictive models for early detection. This study presents a comparative analysis of tuned hyperparameter Random Forest and XGBoost machine learning models using comprehensive data from the National Family Health Survey-5 (NFHS-5). The dataset comprised 825,954 cases with 27 independent variables encompassing socio-economic, demographic, lifestyle, comorbidity, and physiological factors. Synthetic Minority Over-sampling Technique (SMOTE) addressed significant class imbalance (2% diabetic cases). Chi-Square tests confirmed statistically significant associations between all explanatory variables and diabetes status ($p < 0.05$). The XGBoost model achieved superior overall performance with 79.66% accuracy, 68.70% recall, 48.00% log loss, and 80.29% ROC-AUC, while the Random Forest model attained 73.48% accuracy, 74.40% recall, 54.17% log loss, and 81.17% ROC-AUC. Feature importance analysis identified glucose levels, age group, and hypertension as the most significant predictors. While XGBoost demonstrated better generalization, Random Forest excelled in recall and interpretability, making it particularly suitable for screening scenarios. This study addresses limitations of prior research by utilizing nationally representative NFHS-5 data, enabling scalable and clinically viable diabetes prediction models tailored for personalized healthcare in the Indian context.

Keywords: Not provided.

How to cite this article: Athwani M, Athwani H, Awasthi K, Agnihotry D. Machine Learning-Driven Insights into Socio-Economic, Demographic, and Health Determinants of Type-II Diabetes: A Comparative Analysis of Tuned Hyperparameter Random Forest and XGBoost Models Using NFHS-5 Data. Int J Drug Deliv Technol. 2026;16(54s): 1575-1585. DOI: 10.25258/ijddt.16.54s.145

Source of support: Nil.

Conflict of interest: None.

Introduction:

Diabetes is a significant public health concern, with its incidence rising significantly worldwide. The number of people living with diabetes has doubled since 1980, driven by factors such as insufficient diets, sedentary lifestyles, and an aging population. The growing prevalence of diabetes is connected to serious health concerns such as heart disease, kidney failure, vision loss, and limb amputations.[1] The 10th edition of the Diabetes Atlas by International Diabetes Federation projected that an estimated 53.7 crore adults worldwide have diabetes, making up approximately 10.5% of the global adult demographic. Estimates indicate that it will increase to 64.3 crore by 2030 and reach 78.3 crore by 2045.[2] Alarmingly, over 240 million adults, or around 44.7% of instances, remain untreated, increasing their risk of disastrous effects.[3]

In India, identifying diabetes frequently happens too late because of inadequate medical services and low levels of understanding, particularly in rural locations.[4] This delay is crucial considering the growing incidence of diabetes, the lifelong burden of disease management, and the escalating healthcare expenditures associated with its complications.[2] Postponing diagnosis ramps up the chances of severe or even deadly problems, which really stresses why we need reliable methods for catching it early. A successful approach to addressing disparities in a country's healthcare system depends on eliminating knowledge gaps and improving screening and diagnostic programs, particularly among younger populations. These efforts are essential for controlling the disease, reducing complications, and enhancing the overall quality of life for diabetic patients.[5]

The prediction and diagnosis of diabetes through machine learning have shown impressive prospects, various algorithms such as K-Nearest Neighbors (KNN), Random Forest Support Vector Machines (SVM), Decision Trees, Naive Bayes, Logistic Regression are employed.[6,7] Among the various approaches, Random Forest proved to be effective because of its capability to handle high-dimensional datasets, manage outliers effectively, and identify intricate non-linear patterns.[8] It has been proven to beat linear models in dealing with medical data and understand complex interactions between features in predicting Type 2 diabetes as it reduces the risk of overfitting.[9] Moreover, researchers have shown that

successful Random Forest hyperparameters can strongly stimulate model performance.[10]

Alongside Random Forest, XGBoost has emerged as a powerful machine learning technique, particularly suitable for classification problems like diabetes prediction due to its gradient boosting framework offering strong predictive accuracy and robustness to overfitting.[11] Comparative analyses between Random Forest and XGBoost demonstrate that while Random Forest provides robust, interpretable models well-suited for initial screening, XGBoost's boosting approach frequently achieves superior overall accuracy and precision in predictions. [12] Combining insights from both models and selecting the best-fitting model based on dataset characteristics often yields optimal results in clinical prognostic tasks.

Therefore, this study conducts a comparative evaluation of Tuned Hyperparameter Random Forest and XGBoost models using the comprehensive and representative National Family Health Survey dataset. This enables the development of scalable, accurate, and clinically viable diabetes prediction models tailored for personalized healthcare. On top of that, a bunch of earlier works depend on narrow or made-up data collections, like the Pima Indian Diabetes one, which gets used a ton but struggles with its small scale and how well it applies broadly.[13] Although synthetic data allows flexibility and control, it usually skips over the tricky parts of actual patient information, causing disproportionate performance evaluations and limiting the practical application of these models in clinical settings.[14] To overcome these challenges, this research employs the National Family Health Survey, a comprehensive and diverse dataset encompassing demographic, socio-economic, and health-related variables. In contrast to smaller datasets, this survey provides data that truly reflects the population, enabling the development of precise, scalable, and clinically viable diabetes prediction models tailored for personalized healthcare.

1. Literature Review

In recent years, researchers have paid a lot of attention to using machine learning (ML) methods for predicting diabetes, mainly because Type-II diabetes is becoming more common worldwide, and we need ways to catch it early and handle it better. Plenty of studies have investigated different algorithms like decision trees, support vector machines (SVM), and ensemble approaches to figure out diabetes risks from things like medical records, personal backgrounds, and economic details. This overview pulls together some important findings, especially around Random Forest and XGBoost models, how they perform against each other, and how they've been used with various kinds of data. Some of the earlier research pointed out how ensemble methods can really help with tricky, detailed health information. Take Breiman (2001)[15], for example—he came up with Random Forest to create a bunch of decision trees and then combine their results through voting, which helps cut down on overfitting and makes classifications more accurate. When it comes to predicting diabetes, Random Forest gets a lot of credit for being tough on uneven data sets and spotting patterns that aren't straightforward. In one case, Kandhasamy et al. (2015) [16] tried out J48 (a decision tree variant), Random Forest, k-nearest neighbors, and Support Vector Machine (SVM) on Pima Indians Diabetes Dataset for diabetes prediction, and with preprocessing (noise removal), Random Forest hit 100% accuracy, doing better than the rest because of its robustness in handling complex features. Along similar lines, Xu et al. (2017) [17] used Random Forest to predict risks for Type-II diabetes, getting good precision and showing why it beats out naive Bayes algorithm, ID3 algorithm and AdaBoost algorithm.

Then there's XGBoost, which is a step up in gradient boosting and has become a strong option, giving better accuracy by building trees one after another and adding rules to avoid overfitting, as Chen and Guestrin (2016)[11] explained how its design scales well, includes L1 and L2 regularization, and deals with missing info smoothly, making it great for big health datasets. For diabetes particularly, Wang et al. (2020)[18] developed an XGBoost model for predicting type 2 diabetes risk from questionnaire data of middle-aged and elderly Chinese patients, and it performed exceptionally with 89.09% accuracy and 91.82% AUC, outperforming models like SVM, Random Forest, and K-NN across metrics such as accuracy, sensitivity, specificity, precision, MCC, and AUC. Zhang et al. (2024)[19] backed this up in a comprehensive diabetes dataset, where XGBoost model, reached 94.82% accuracy for predicting cases, outperforming models like AdaBoost, LightGBM, Decision Tree, and Logistic Regression.

When people compare Random Forest and XGBoost directly, they usually find that each has its own advantages—XGBoost tends to win on accuracy and precision, but Random Forest is easier to interpret. Zhao et al. (2020) did a straight comparison on diabetes data and saw XGBoost get to 85% accuracy compared to 82% for Random Forest, mostly because of how it fixes mistakes from earlier trees. Gürdüler and Özkan (2020) tested them on the Pima set and noted XGBoost's higher AUC of 0.89 and quicker results, though Random Forest handled noisy data a bit better. Overall, these works suggest that blending ideas from both—like using Random Forest to pick out key features and XGBoost for the final call—often leads to the best outcomes in medical predictions. Even with all this progress, a lot of studies stick to small datasets like Pima Indian,

which don't capture the full range of socioeconomic and demographic factors needed for predictions across whole populations, especially in places like India. The NFHS-5 data helps fill that hole by offering thorough, country-wide info that lets models factor in real-life influences. Our research adds to these comparisons by using tuned versions of Random Forest and XGBoost on NFHS-5, with the goal of boosting prediction accuracy and spotlighting important socioeconomic elements

2. Methodology

2.1 Data

Between 2019 and 2021, the National Family Health Survey's fifth round was conducted (NFHS-5). The survey was implemented by 17 field agencies under India's Ministry of Health and Family Welfare, while the International Institute for Population Sciences supervised the operation. Information on health, family welfare, and socio-economic measures was gathered by the survey. It captured data from 636,699 households, as well as 724,115 women aged from 15 to 49 and 101,839 men aged from 15 to 54, with impressive participation rates of 98% for households, 97% for women, and 92% for men. For this study, data were extracted from the NFHS-5 dataset using the Statistical Package for Social Sciences (SPSS) tool, converting the .sav file into a Comma-Separated Values (CSV) format.

2.2 Independent variables

The input (explanatory) variables encompassed a range of health indicators across multiple dimensions (Table 1). Demographic factors included sex, age brackets (15–64 years), residence type (urban/rural), and state groups (northern, southern, eastern, western etc.). Socioeconomic variables covered educational attainment, literacy levels, religious affiliation, and wealth index. Lifestyle and health behaviour indicators accounted for tobacco and alcohol consumption, dietary indices, and healthcare facility visits. Health conditions such as cancer, hypertension, thyroid disease, respiratory disease, kidney disease, and heart disease were also considered. Additionally, biometric measures included Body Mass Index (BMI) group, glucose level group, arm circumference, waist-to-hip ratio, anaemia level, haemoglobin level, high blood pressure, and high glucose status, providing a comprehensive assessment of individual health status.

Table 1: Feature variable list

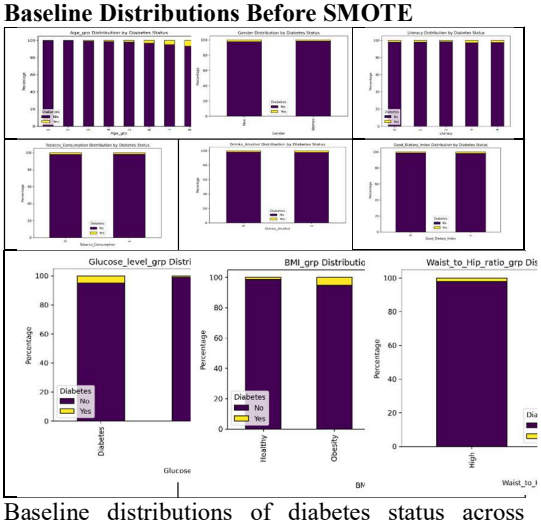
S.N.	Variable Name	Description	Levels
1.	Gender	Gender of the individual	Men, Women
2.	Age_grp	Age group of the individual	1:15-19, 2:20-24, 3:25-29, 4:30-34, 5:35-39, 6:40-44, 7:45-49, 8:50-54, 9:55-59, 10:60-64
3.	Residence_Type	Type of residence	1: Urban areas, 2: Rural areas
4.	Educ_Level	Level of education	0: No education, 1: Primary, 2: Secondary, 3: Higher
5.	Religion	Religion of the individual	1: Hinduism, 2: Islam, 3: Christianity, 4: Sikhism, 5: Buddhism/Neo-Buddhism, 6: Jainism, 7: Judaism, 8: Parsi/Zoroastrianism, 9: No religion, 96: Other
6.	Literacy	Literacy of the individual	0: Cannot read at all, 1: Able to read only parts of sentence, 2: Able to read whole sentence, 3: No card with required language, 4: Blind/visually impaired
7.	Wealth_Index_Combined	Wealth_Index group of the individual	1: Poorest, 2: Poorer, 3: Middle, 4: Richer,5: Richest
8.	Visited_HealthP	Visited health care facility	0: No, 1: Yes
9.	Drinks_Alcohol	Whether individual drinks alcohol	0: No, 1: Yes
10.	State_regrp	State group of the individual	Central, Southern, Northern, Eastern, North-Eastern, Western
11.	Check_Cancer_grp	Whether individual has cancer	0: No, 1: Yes
12.	Check_Hypertension_grp	Whether individual has hypertension	0: No, 1: Yes
13.	Check_ThyroidD_grp	Whether individual has thyroid	0: No, 1: Yes
14.	Check_RespD_grp	Whether individuals have	0: No, 1: Yes

		respiratory disease	
15.	Check_KidneyD_grp	Whether individuals have kidney disease	0: No, 1: Yes
16.	Check_HeartD_grp	Whether individuals have heart disease	0: No, 1: Yes
17.	Tobacco_Consumption	Whether individual consumes tobacco	0: No, 1: Yes
18.	Good_Dietary_Index	Good_Dietary group index of the individual	0: No, 1: Yes
19.	Bad_Dietary_Index	Bad_Dietary group index of the individual	0: No, 1: Yes
20.	Check_High_BP_regrp	Whether individual has High BP	0: No, 1: Yes
21.	Check_High_Glucose_regrp	Whether individuals have High Glucose	0: No, 1: Yes
22.	BMI_grp	BMI group of the individual	Healthy, Overweight, Underweight, Obesity
23.	Glucose_level_grp	Glucose_level group of the individual	Prediabetes, Diabetes, Normal
24.	Arm_Circum_grp	Arm Circumference (AC) group of the individual	Normal_AC, High_AC, Low_AC
25.	Waist_to_Hip_ratio_grp	Waist to Hip Ratio group of the individual	High, Normal
26.	Anaemia_Level_grp	Anameia level group of the individual	1: Severe,2: Moderate, 3: Mild, 4: Not anaemic
27.	Haem_Alt_Smok_grp	Haemoglobin level and smoking grp	Normal_Hb, Low_Hb, High_Hb

2.3 Outcome variable

The outcome variable of interest was whether an individual had diabetes. Respondents aged 15 and above were asked if they had diabetes. Responses were categorized as "Yes" for those who had diabetes, "No" for those who did not, and "Don't know" for uncertain responses.

Baseline Distributions Before SMOTE

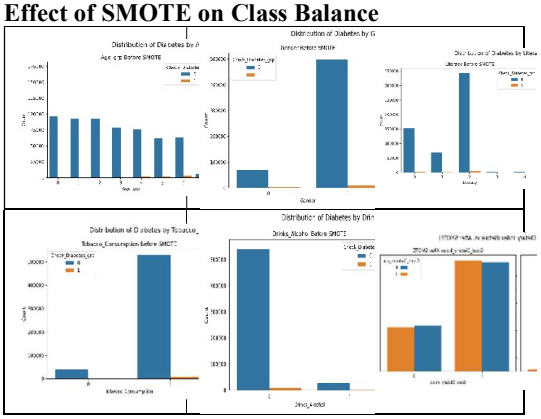


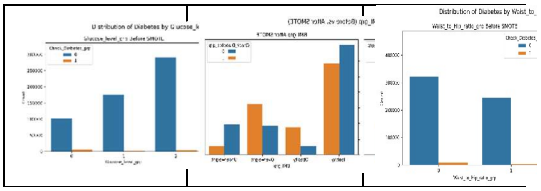
Baseline distributions of diabetes status across demographic, lifestyle, and physiological features before oversampling

Data Preprocessing

The data was prepared by cleaning, and preprocessing it, which included filling in any missing entries, converting category types to numbers, undertaking scaling of features, and dividing the whole set into parts for training and testing. To fix the uneven spread in class sizes during our sorting work, we turned to a machine learning approach called Synthetic Minority Over-sampling Technique (SMOTE), which sorts out imbalances by making up new examples for the smaller group in a skewed dataset—this produced fake samples for that group and helped create a more even balance overall. On top of that, we handled category variables with one-hot encoding, turning each type into a binary setup where '1' shows if it's there and '0' means it's not.

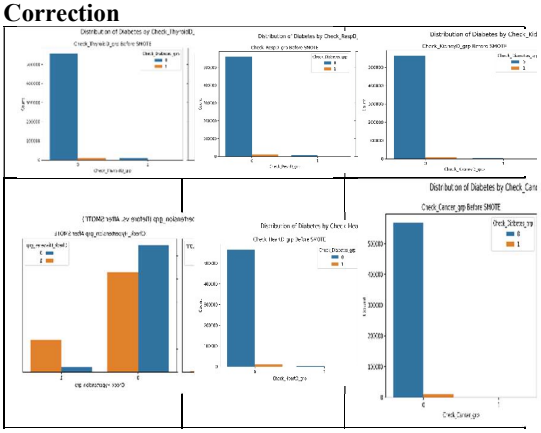
Effect of SMOTE on Class Balance





Comparison of diabetes-positive vs. diabetes-negative class proportions before and after SMOTE oversampling

Health Condition Imbalances and SMOTE Correction



Distributions of comorbidity indicators by diabetes status, before and after SMOTE

To assess the link between two categorical variables, the Chi-Square test of independence was utilized. The null hypothesis (H_0) posited that there was no significant relationship, whereas the alternative hypothesis (H_1) proposed the presence of an association.[14] The Chi-Square statistic compared observed and expected frequencies under independence, calculated as $\chi^2 = \sum((O_i - E_i)^2 / E_i)$. Expected frequencies were determined using (Row Total * Column Total) / Grand Total. The chi2_contingency function from the scipy.stats library was utilized for analysis. A p-value of less than 0.05 was interpreted as evidence of a statistically significant association, with results considered reliable at the 95% confidence level.

Machine Learning Models

We implemented two ensemble models—Random Forest and XGBoost—for comparative diabetes detection, both tuned via tuned hyperparameters to optimize performance on the preprocessed NFHS-5 data.

Random Forest Model

For diagnosis, the best fit was detected using an ensemble learning technique in which multiple decision trees are generated throughout the training process, and the most frequent result is selected through majority voting to predict the output class, known as Random Forest.[15] The algorithm utilizes the bagging technique, where each decision tree is trained on a random subset of features. This process reduces the correlation between trees and enhances the model's generalization capability.[16,17,18]

2.6 Mathematical Formulation

Each decision tree in Random Forest learns hierarchical splitting rules using metrics like Gini impurity or entropy. Gini impurity is calculated as $Gini(p) = 1 - \sum_{i=1}^{C} p_i^2$, where $p\,i$ denotes the share of samples in class i, and C is the total number of classes.[19,20] Entropy is computed as $Entropy(p) = -\sum_{i=1}^{C} \log p_i * p_i$(Shannon, 1948). The model aggregates predictions from multiple trees, where for a given input χ, each tree produces a prediction Ti(χ), and the final output is determined by majority voting as $\hat{y} = majority\ vote\{T_1(\chi), T_2(\chi), \ldots . T_k(\chi)\}$.[21]

2.7 Hyperparameters of Random Forest

To optimize performance, key hyperparameters were tuned. The number of trees (n_estimators) influences variance reduction, as variance is inversely proportional to n_estimators. The depth of trees (max_depth) balances bias-variance trade-off, with the optimal depth approximately proportional to log(n_samples)/log(m_features). Parameters such as min_samples_split and min_samples_leaf prevent overfitting by restricting tree growth. max_features is the number of features being considered for splitting, while bootstrap enables sampling with replacement. oob_score estimates generalization error using out-of-bag samples. n_jobs determines computational parallelism, criterion sets the splitting metric, max_leaf_nodes limits complexity, and random_state ensures reproducibility.[15]

XGBoost Model

XGBoost (eXtreme Gradient Boosting) is an optimized gradient boosting algorithm based on the principle of boosting, where multiple weak learners are combined sequentially to create a strong predictive model. XGBoost refines predictions in each iteration by focusing more on samples with higher errors, leading to better overall performance.

3.3.1 Mathematical Formulation

a) Objective Function

XGBoost minimizes a regularized objective function which includes both the loss and regularization terms**.**

For a given training set with n samples and m features, let y_i be the target for sample i, and $\hat{y}_i^{(t)}$ be the prediction at iteration t, $l(y_i, \hat{y}_i^{(t)})$ is the loss function, $\Omega(f_k)$ is a regularization term on the k^{th} tree.

The objective function at iteration t is defined as:

$$Obj^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^{t} \Omega\,(f_k)$$

The regularization term for each tree f_k is $\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$ where T is the number of leaves in the tree, w_j is the weight of

leaf j, and γ and λ are regularization hyperparameters.

b) Additive Learning

At each iteration, XGBoost adds a new tree $f_t(x)$ to correct the errors of the previous model:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i)$$

Here, $f_t(x)$ is chosen to minimize the objective function

c) Second-Order Taylor Approximation

XGBoost optimizes the objective function using a second-order Taylor expansion, making the process efficient. The approximate objective function for a tree $f_t(x)$ at iteration t is:

$$Obj^{(t)} \approx \sum_{i=1}^{n} \left[g_i f_t(x_i) + , \hat{y}_i^{(t)}) + \frac{1}{2} h_i f_t(x_i)^2 \right] + \Omega(f_t)$$

where $gi = \partial_{\hat{y}} l(y_i, \hat{y})$ and $h_i = \partial^2{}_{\hat{y}} l(y_i, \hat{y})$ are the first and second derivatives of the loss function with respect to $\hat{y}$.

d) Hyperparameters of XG Boost

XG Boost has several hyperparameters that need to be tuned to optimize the model's performance. Below are the key hyperparameters along with their mathematical impact:

max_depth: Controls the maximum depth of each tree. Higher values increase model complexity and risk of overfitting.

min_child_weight: Minimum sum of instance weights needed in a child node. Higher values make the algorithm more conservative by requiring more samples per leaf node.

gamma: Minimum loss reduction required to make a further partition. Higher values make the algorithm more conservative.

subsample: The fraction of samples used to grow each tree. Lower values prevent overfitting by adding randomness.

colsample_bytree: The fraction of features considered when building each tree, adding randomness and reducing overfitting.

learning_rate (or eta): Shrinks the weight of each added tree. Smaller values require more trees but can improve generalization.

n_estimators: Number of trees (iterations) in the model. Higher values often improve performance but can lead to overfitting if too high.

scale_pos_weight: Controls balance between positive and negative classes, useful for imbalanced datasets.

lambda (L2 regularization): Controls the regularization on leaf weights. Higher values make the model more conservative.

alpha (L1 regularization): Adds regularization on leaf weights, encouraging sparsity.

early_stopping_rounds: Stops training when validation score does not improve for a specified number of rounds.

objective: Defines the learning task and the corresponding loss function. Common values include reg:squarederror for regression and binary:logistic for binary classification.

eval_metric: Specifies the evaluation metric used during training, such as rmse for regression or logloss for binary classification.

3.3.2 Fitting the model and Model Diagnostics for both the models

Fitting the Model

The model was trained by minimizing a loss function and optimizing hyperparameters using a random search approach, which randomly selects combinations from a defined range.[22] Cross-validation (CV) was used to evaluate the model. The dataset was split into k subsets, with the model trained on k−1 subsets, validated on the remaining one, and the process repeated iteratively to ensure robust performance.[23]

Data Preprocessing and Model Training

The dataset comprised a total of 825,954 cases, with 811,660 which is 98% of the cases classified as "without diabetes" and 14,294 which is 2% of the cases classified as "with diabetes," highlighting a significant class imbalance. The data was divided into training and testing sets to formulate it for machine learning at a 70%-30% split, resulting in 578,167 cases which is 70% of cases for training and 247,787 cases which is 30% of cases for testing, and to address class imbalance, the training set was processed using SMOTE. Before applying SMOTE, the training set contained 568,189 cases without diabetes and 9,978 cases with diabetes. Categorical variables were converted into numerical values using one-hot encoding, and after SMOTE was applied, the minority class (diabetic cases) was oversampled to match the majority class, resulting in a balanced training set of 1,136,378 cases (568,189 per class), while the target variable (diabetes status) was label-encoded. The Chi-Square test confirmed that all explanatory variables had a statistically significant association with diabetes, ensuring their relevance. For diagnosis, the Random Forest and XGBoost models were employed, with hyperparameter tuning conducted using random search and k-fold cross-validation. Training was conducted on k-1 folds and evaluated on the remaining fold, with outcomes aggregated to ensure reliability. The optimal hyperparameter combination, selected based on the highest cross-

validation score, ensured the model's generalizability to unseen data.

Model Diagnostics

Performance for both the models was evaluated using diagnostic tools, with a confusion matrix detailing the counts of true negatives and positives (TN, TP) and false negatives and positives (FN, FP). Key classification metrics were computed:

Accuracy = ((TP+TN) /(TP+TN+FP+FN)), Precision = (TP/(TP+FP)), Recall = (TP/(TP+FN)), and F1-score=(2×(Precision×Recall)/(Precision+Recall)).

Log loss, a probability-based evaluation metric, is defined as $Log\ loss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log(p_{ij})$ where y_{ij} is a binary indicator, p_{ij} is the predicted probability, N and M is the sample size and number of classes respectively. The relationship between the True Positive Rate (TPR) and the False Positive Rate (FPR) at various thresholds was graphically represented by the Receiver Operating Characteristic (ROC) curve, while the overall efficacy of the model was measured by the Area Under the Curve (AUC), with higher values (near 1) indicating superior classification capability.[24,25]

Table 2: Frequency table of Target variable

Label	Frequency	Percent of cases
Cases without Diabetes	811,660	98%
Cases with Diabetes	14,294	2%
Total cases	**825,954**	**100%**

The data was divided training and testing sets at 70%-30% respectively to train the machine learning model.

SMOTE technique was applied to the dataset to address class imbalance by generating synthetic samples for the minority class, ensuring a more balanced distribution of classes.

Table 2: Frequency table of data for training and testing

Dataset	Frequency	Percent of cases
Training set	578,167	70%
Testing set	247,787	30%
Total cases	**825,954**	**100%**

Table 3: Frequency table of target variable in the training set

Label	Frequency before applying SMOTE	Frequency before applying SMOTE
Cases without Diabetes	568,189	568,189
Cases with Diabetes	9,978	568,189
Total cases	**578,167**	**1,136,378**

All the categorical variables in the data were converted into numerical values using one-hot encoding. The Target variable was converted into numerical values using label encoding. We checked each factor's significant relationship with the target variable, with the help the Chi-Square test of independence. All variables had a statistically significant relationship with the Target variable.

6. Results

Results for random forest

(Table 2) presents a Random Forest-based model for diabetes prediction, achieving an accuracy of 73.48%, recall of 74.40%, log loss of 54.17%, and ROC-AUC of 81.17%, demonstrating strong predictive performance. These metrics highlight the model's ability to identify at-risk individuals, particularly in correctly classifying positive cases (high recall) and distinguishing between diabetic and non-diabetic cases (high ROC-AUC). The dataset, sourced from NFHS-5, initially had significant class imbalance, which is 2% of diabetic cases, which were addressed using SMOTE to balance the classes, enhancing model performance.

Table 2.
Performance Metrics of the Random Forest Model on Training and Testing Datasets

Dataset	Accuracy	Recall	Logloss	ROC-AUC
Training set	74.69%	76.07%	54.88%	81.76%
Testing set	73.48%	74.40%	54.17%	81.17%

This approach ensured that the minority class (diabetic cases) were adequately represented, enabling the model to learn effectively from both classes and improving its overall performance.

A Chi-Square test of independence confirmed that all selected explanatory variables had a statistically significant relationship with diabetes ($p < 0.05$). These variables included demographics (age, gender, education level, literacy, and residence type), lifestyle factors (diet, tobacco consumption, alcohol use), socioeconomic status (wealth index, occupation), comorbidities (hypertension, heart disease, kidney disease, thyroid disease, respiratory disease, cancer), and anthropometric and physiological measures (BMI, waist-to-hip ratio, arm circumference, anemia levels, hemoglobin levels, and blood glucose levels). The inclusion of these diverse factors ensured a comprehensive analysis of diabetes risk, capturing both medical and socio-economic determinants in the Indian context.

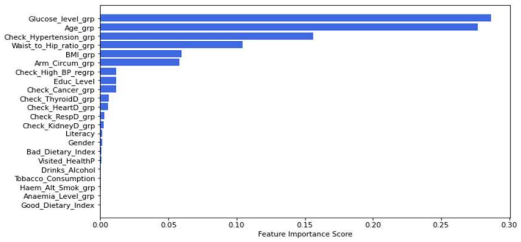
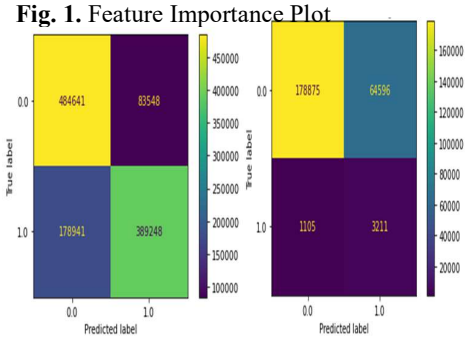


Fig. 1. Feature Importance Plot

Fig. 2(a). Confusion Matrix of Training Set

Fig. 2(b). Confusion Matrix of Testing Set



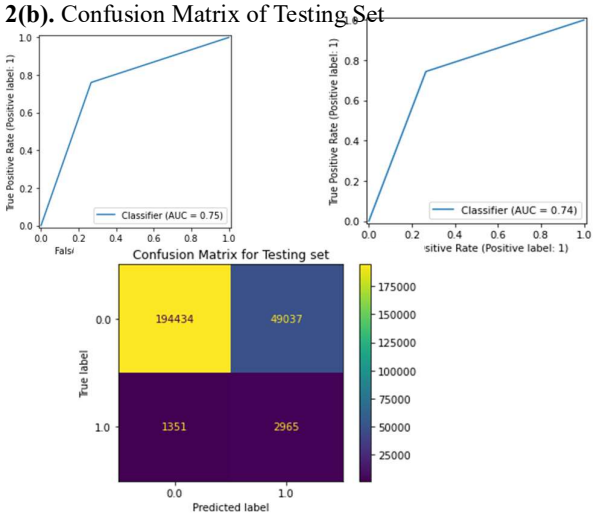


Fig. 3(a). ROC Curve Plot of Training Set

Fig. 3(b). ROC Curve Plot of Testing Set

Feature importance analysis (Fig. 1) highlighted key predictors of diabetes, with glucose levels, age group, and hypertension status emerging as the most significant contributors. This aligns with established medical knowledge and further validates the model's reliability. The confusion matrices for the training and testing sets (Fig. 2(a) and Fig. 2(b)) provided detailed insights into the model's classification performance, displaying the distribution of true negatives and positives, as well as false negatives and positives. The ROC curves (Fig. 3(a) and Fig. 3(b)) illustrated the model's ability to balance sensitivity and specificity across different thresholds, with the high AUC score confirming its effectiveness in distinguishing between diabetic and non-diabetic cases.

Results for XG Boost Model

The model diagnostic results after fitting XG Boost model is as follows:

The model achieved an accuracy of 79.66%, a recall of 68.7%, log loss of 48% and ROC-AUC value as 80.29%. The overall good performance on all model diagnostic metrics suggests the model is reliable, generalizes well and performs consistently across training and testing sets.

Thus, the construction of the diabetes prediction model provides early warning signals for effective diagnosis and efficient prognosis of the patient.

Table 3: Model Diagnostic Results

Dataset	Accuracy	Recall	Logloss	ROC_AUC
Training set	76.78%	73.98 %	49.00 %	84.36%
Testing set	79.66%	68.70 %	48.00 %	80.29%

Fig. 1: Confusion Matrix of Training set

Fig. 2: Confusion Matrix of Testing set

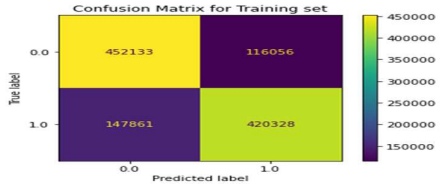


Fig. 3: Feature importance plot

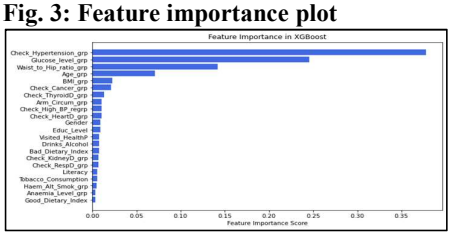
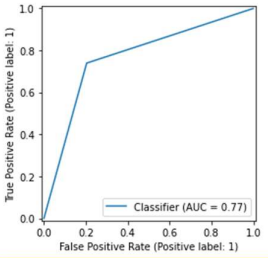
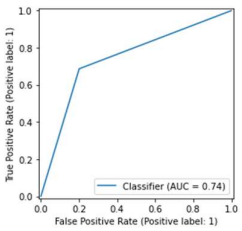


Fig. 4: ROC Curve plot of Training set

Fig. 5: ROC Curve plot of Testing set





Discussion

(Table 2) highlights the performance of our matched Random Forest model for diabetes prediction using the National Family Health Survey (NFHS-5) dataset, achieving an accuracy of 73.48% and a recall of 74.40%, demonstrating its proficiency in detecting individuals susceptible to diabetes, particularly in correctly classifying positive cases (high recall). This is consistent with studies such as Alam et al. (2020), which achieved similar results using Random Forest and highlighted the accuracy of diabetes classification.[26] Although Maniruzzaman et al. (2020) achieved slightly higher accuracy, the different data source and methodology make direct comparison challenging. Still, their reported effectiveness of Random Forest for diabetes prediction furthermore supports our results.[27]

The ROC AUC of our model of 81.17% reflects strong discriminatory power, in accordance with studies such as Esmaily et al. (2018), highlighting Random Forest's ability to distinguish between diabetic and non-diabetic cases. The similarity between training and test metrics indicates that the model establishes a good generalization without overfitting the training data. Furthermore, addressing class imbalance (2% diabetic cases) with SMOTE ensured balanced representation, enhancing performance, as supported by Alghamdi et al. (2017). Explanatory variables were selected using a Chi-Square test ($p < 0.05$), incorporating demographics, lifestyle, socioeconomic factors, comorbidities, and physiological measures for a comprehensive risk analysis, aligning with Razavian et al. (2015). Confusion matrices and ROC curves are powerful tools for evaluating a model's predictive accuracy. They help measure how well the model balances true positive rates and false positive rates. The high AUC scores observed in the analysis further confirm the model's strength in distinguishing between different classes or cases. These findings are consistent with the work of Hasan et al. (2020), who also demonstrated the effectiveness of these metrics in assessing predictive performance.

Conclusion

XGBoost generally performed better overall, particularly in accuracy and log loss, which suggest stronger generalization and lower error rates on unseen data. However, Random Forest excels in recall and ROC-AUC, making it potentially preferable if prioritizing the detection of diabetic cases (e.g., in screening scenarios where missing positives is costly). The conclusion emphasizes Random Forest's reliability and interpretability for real-world diabetes screening in India, while noting both models' strengths in handling the dataset's complexities. The choice ultimately depends on the specific priorities, such as balancing false negatives versus overall accuracy.

This study's Random Forest model for diabetes prediction is dependable and interpretable, with good performance metrics that might enhance diabetes screening and care in India. The model is trustworthy and useful for real-world applications since it identifies diabetes risk factors and uses feature significance analysis. To enhance diabetes management, future research should validate the model across demographic groups and implement it in healthcare. Addressing class inequality and using extensive information to construct accurate, scalable public health forecasting models are highlighted in the paper. Similar to the UN Sustainable Development Goals (SDGs) and the WHO's decade of nutrition (2016–2025), focused treatments and policies are needed to battle India's increasing diabetes pandemic. Advanced prediction models like Random Forest can enhance public health outcomes by ensuring prompt diagnosis and successful treatment in national health initiatives.

Addressing class imbalance using SMOTE improves relevance and accuracy for real-world datasets with underrepresented minority classes. For converting predictive models into practical health solutions, coordination between data scientists, healthcare professionals, and policymakers is vital. Increasing multidisciplinary interactions can help healthcare systems become proactive and preventative. This reduced diabetes and its consequences and improved patient outcomes. Therefore, to promote fair access to healthcare and combat the war against the expanding diabetes pandemic, the adoption of data-driven, evidence-based solutions using machine learning models will play a vital role.

Data availability statement

The NFHS 2019-21 that supports our findings in this study is available at https://dhsprogram.com.

Funding sources

This study did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declarations of interest

The authors do not have any conflict of interest with other entities or researchers.

References

Ahmad, A., Mustapha, A., Zahadi, E., Masah, N., & Yahaya, N. (2011). Comparison between Neural Networks against Decision Tree in Improving Prediction Accuracy for

Diabetes Mellitus. *Communications in Computer and Information Science, 188*, 537-545. doi:10.1007/978-3-642-22389-1_47

Alehegn, M., Joshi, R., & Alehegn, M. (2017). Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble approach. *International Research Journal of Engineering and Technology, 4*(10), 426-435. Retrieved from https://api.semanticscholar.org/CorpusID:212507018

Anand, A., & Shakti, D. (2015). Prediction of diabetes based on personal lifestyle indicators. *2015 1st International Conference on Next Generation Computing Technologies (NGCT-2015)*, 673-676. doi:10.1109/NGCT.2015.7375206

Esmaily, H., Tayefi, M., Doosti, H., Ghayour-Mobarhan, M., Nezami, H., & Amirabadizadeh, A. (2018). A Comparison between Decision Tree and Random Forest in Determining the Risk Factors Associated with Type 2 Diabetes. *Journal of Research in Health Sciences, 18*(2), e00412-e00412. Retrieved from http://jrhs.umsha.ac.ir/Article/3777

Fiarni, C., Sipayung, E. M., & Maemunah, S. (2019). Analysis and Prediction of Diabetes Complication Disease using Data Mining Algorithm. *Procedia Computer Science, 161*, 449-457. doi:10.1016/j.procs.2019.11.144

Francesco Rubino 1, R. L.-L. (2023). Lancet Diabetes & Endocrinology Commission on the Definition and Diagnosis of Clinical Obesity. *The Lancet Diabetes and Endocrinology, 11*(4), 226 - 228. doi:10.1016/S2213-8587(23)00058-X

(2016). *Global Report on Diabetes.* Geneva: World Health Organisation. Retrieved from https://www.who.int/publications/i/item/9789241565257

Guo, Y., Li, Y., Wang, G., & Liu, X. (2014). Application of artificial neural network to predict individual risk of type 2 diabetes mellitus. *Journal of Zhengzhou University:Medical Sciences, 49*(3), 180-183.

Husain, A., & Khan, M. H. (2018). Early Diabetes Prediction Using Voting Based Ensemble Learning. *Advances in Computing and Data Sciences, 905*, 95-103. doi:10.1007/978-981-13-1810-8_10

Kandhasamy, J. P., & Balamurali, S. (2015). Performance Analysis of Classifier Models to Predict Diabetes Mellitus. *Procedia Computer Science, 47*, 45-51. doi:10.1016/j.procs.2015.03.182

Kandhasamy, J. P., & Balamurali, S. (2015). Performance Analysis of Classifier Models to Predict Diabetes Mellitus. *Procedia Computer Science, 47*, 45-51. doi:10.1016/j.procs.2015.03.182

Kaur, H., & Kumari, V. (2020). Predictive modelling and analytics for diabetes using a machine learning approach. *Applied Computing and Informatics, 18*(1/2), 90-100. doi:10.1016/j.aci.2018.12.004

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. *Computational and Structural Biotechnology Journal, 15*, 104-116. doi:10.1016/j.csbj.2016.12.005

Kishor, A., & Chakraborty, C. (2021). Early and accurate prediction of diabetics based on FCBF feature selection and SMOTE. *International Journal of System Assurance Engineering and Management, 15*, 4649–4657. doi:10.1007/s13198-021-01174-z

Kumari, S., Kumar, D., & Mittal, M. (2021). An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. *International Journal of Cognitive Computing in Engineering, 2*, 40-46. doi:10.1016/j.ijcce.2021.01.001

Lukmanto, R. B., Suharjito, Nugroho, A., & Akbar, H. (2019). Early Detection of Diabetes Mellitus using Feature Selection and Fuzzy Support Vector Machine. *Procedia Computer Science, 157*, 46-54. doi:10.1016/j.procs.2019.08.140

Mujumdar, A., & V, V. (2019). Diabetes prediction using Machine learning algorithms. *Procedia Computer Science, 165*, 292-299. doi:10.1016/j.procs.2020.01.047

Nai-arun, N., & Moungmai, R. (2015). Comparison of Classifiers for the Risk of Diabetes Prediction. *Procedia Computer Science, 69*, 132-142. doi:10.1016/j.procs.2015.10.014

Nilashi, M., Ibrahim, O., Dalvi, M., & Ahmadi, H. (2017). Accuracy Improvement for Diabetes Disease Classification: A Case on a Public Medical Dataset. *Fuzzy Information and Engineering, 9*(3), 345-357. doi:10.1016/j.fiae.2017.09.006

Ogurtsova, K., Guariguata.Leonor, Barengo, N. C., Ruiz, P. L.-D., Sacree, J. W., & Sacree, S. (2022). IDF diabetes Atlas: Global estimates of undiagnosed diabetes in adults for 2021. *Diabetes Research and Clinical Practice, 183*, 109118. doi:10.1016/j.diabres.2021.109118

Perveen, S., Shahbaz, M., Guergachi, A., & Karim, K. (2016). Performance Analysis of Data Mining Classification Techniques to Predict Diabetes. *Procedia Computer Science, 82*, 115-121. doi:10.1016/j.procs.2016.04.016

Sisodia, D., & Sisodia, D. S. (2018). Prediction of Diabetes using Classification Algorithms. *Procedia Computer Science, 132*, 1578-1585. doi:10.1016/j.procs.2018.05.122

Sudharsan, B., Peeples, M., & Shomali, M. (2014). Hypoglycemia Prediction Using Machine Learning Models for Patients With Type 2 Diabetes. *Journal of Diabetes Science and Technology, 9*(1), 86-90. doi:10.1177/1932296814554260

Vijayan, V. V., & Ravikumar, A. (2014). Study of data mining algorithms for prediction and diagnosis of diabetes mellitus. *International Journal of Computer Applications, 95*(17), 12-16. doi:10.5120/16685-6801

Wu, H., Yang, S., Huang, Z., He, J., & Wang, X. (2018). Type 2 diabetes mellitus prediction model based on data mining. *Informatics in Medicine Unlocked, 10*, 100–107. doi:10.1016/j.imu.2017.12.006

Xu, W., Zhang, J., Zhang, Q., & Wei, X. (2017). Risk prediction of type II diabetes based on random forest model. *Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)*, 382-386. doi:10.1109/AEEICB.2017.7972337

Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting Diabetes Mellitus With Machine Learning Techniques. *Frontiers in Genetics, 9*, 515. doi:10.3389/fgene.2018.00515