International Research Journal of Multidisciplinary Scope (IRJMS), 2026;7(3):658-675 IRJMS

Original Article | ISSN (O): 2582-631X DOI: 10.47857/irjms.2026.v07i03.010810

Feature Correlation and Class Balancing Effects on Random Forest Accuracy in Diabetes Risk Prediction Using the Teboul Health Indicators Dataset

Lavanya AL[1]*, Rama Mohan Babu Gatram[2]

[1]Department of Computer Science and Engineering, Dr. Y.S. Rajasekhar Reddy University College of Engineering and Technology, Acharya Nagarjuna University, Guntur andhra Pradesh, India, [2]Department of Computer Science and Engineering (Artificial Intelligence and Machine Learning), R.V.R. and J.C. College of Engineering, Chowdavaram andhra Pradesh, Guntur, India. *Corresponding Author's Email: mukthipudi@gmail.com, lavanyapramod408@gmail.com

Abstract

Early detection of diabetes is essential for reducing disease burden and long-term healthcare costs. This study presents a Random Forest-based predictive framework for diabetes risk assessment, using the Teboul Health Indicators Dataset, which is derived from the Centers for Disease Control and Prevention (CDC) Behavioral Risk Factor Surveillance System (BRFSS 2015). The proposed framework incorporates feature correlation analysis and evaluates the effects of four class-balancing strategies, namely under-sampling, oversampling, Synthetic Minority Oversampling Technique (SMOTE) and class weighting, on classification performance. The dataset, comprising 253,680 records and 21 health-related attributes, was preprocessed using normalization and stratified random sampling. Correlation analysis identified General Health (GenHlth), High Blood Pressure (HighBP), Body Mass Index (BMI) and Age as the most influential predictors of diabetes risk. Experimental results demonstrated that the under-sampling approach achieved the best trade-off between precision and sensitivity, yielding an F1-score of 0.426, a recall of 0.777 and an area under the receiver operating characteristic curve (AUC) of 0.808, outperforming the other balancing methods. The findings indicate that appropriate class balancing improves sensitivity and fairness in diabetes classification, while Random Forest provides model interpretability through feature-importance analysis. Future work will focus on integrating explainable artificial intelligence (XAI) techniques and hybrid ensemble balancing methods to further enhance transparency and predictive performance in healthcare analytics.

Keywords: Class Balancing, Diabetes Prediction, Feature Correlation, Healthcare Analytics, Random Forest, Teboul Health Indicators Dataset.

Introduction

Diabetes mellitus is a major global public health challenge, affecting an estimated 537 million adults worldwide in 2021, with projections reaching 643 million by 2030 and 783 million by 2045, according to the according to reports from the International Diabetes Federation and the World Health Organization (1, 2). The disease significantly contributes to morbidity, mortality and healthcare costs, accounting for nearly 12% of global health expenditure and ranking among the top ten causes of death worldwide. Developing countries, especially in Asia and Africa, are experiencing the fastest growth in diabetes rates, driven mainly by urbanization, sedentary lifestyles and poor dietary habits (3).

Conventional diagnostic and surveillance methods-though clinically effective-often detect diabetes at later stages, when complications like neuropathy, nephropathy and cardiovascular diseases have already appeared. As a result, machine learning (ML) predictive modelling has become a promising approach for early risk detection and preventive care. Large health surveillance datasets, such as the Behavioural Risk Factor Surveillance System (BRFSS) (4), provide opportunities for machine learning models to identify hidden relationships among behavioural and physiological factors, thereby supporting targeted interventions and improved disease management. The integration of predictive frameworks into healthcare systems can assist practitioners and policymakers in disease prevention, resource allocation and the reduction of the global diabetes burden. According to the World Health Organization (WHO), the number of adults with diabetes has nearly quadrupled since 1980, reaching more than 460 million in 2021 (5). Approximately 6.7 million deaths were attributed

(Received 24th February 2026; Accepted 23rd June 2026; Published 11th July 2026)

to diabetes in the same year, mainly due to cardiovascular, renal and neuropathic complications, with nearly 80% of the burden occurring in low- and middle-income countries (LMICs), where limited access to prevention and early diagnosis accelerates disease progression (6).

Diabetes also imposes a substantial economic burden, with global healthcare expenditures reaching USD 966 billion in 2021, representing a 316% increase over the previous 15 years (7). Since major risk factors, including obesity, physical inactivity, unhealthy diet and smoking, are modifiable (8), early detection and data-driven prevention strategies are essential.

Although conventional screening methods based on biochemical tests and self-reported surveys are effective, they are resource-intensive and primarily focused on diagnosis rather than prevention (9). Machine learning techniques enable the analysis of complex, high-dimensional health data to identify individuals at risk before symptom onset (10). Applied to large population datasets such as BRFSS, predictive analytics can uncover behavioural and demographic risk patterns that support early intervention. Such approaches are also aligned with the United Nations Sustainable Development Goal 3, which aims to ensure healthy lives and promote well-being for all ages (11).

Various machine learning algorithms, including logistic regression, decision trees, support vector machines and ensemble methods, have been widely applied to diabetes prediction. Studies using structured datasets such as PIDD and NHANES, as well as large-scale datasets like BRFSS, have mainly focused on improving accuracy and comparing algorithms (12). However, the effects of data-level interventions, particularly class balancing and feature correlation analysis, on model sensitivity and interpretability remain underexplored. To address this gap, the present study proposes a Random Forest-based framework incorporating feature correlation analysis and class balancing techniques to enhance early risk detection and classification fairness.

A comprehensive review of machine learning applications in diabetes research highlighted the increasing use of supervised learning algorithms for diagnosis and complication prediction (13). Random Forest achieved an accuracy of 85% on the Pima Indians Diabetes dataset due to its ability to capture nonlinear relationships (14). Machine learning models trained on electronic health records outperformed conventional risk scoring systems in identifying undiagnosed diabetes (15). Hybrid Decision Tree–Naive Bayes models achieved 88% accuracy while improving interpretability (16), whereas ensemble approaches such as XGBoost and AdaBoost demonstrated superior Precision and Recall compared with individual classifiers (17). Deep neural networks applied to NHANES data achieved accuracies exceeding 90%, although interpretability remained limited (18).

Diabetes prediction using BRFSS data has also been widely explored. Logistic Regression and Decision Tree models applied to BRFSS 2015 data identified Age, BMI, hypertension and physical inactivity as major predictors (19). Comparative studies involving multiple machine learning algorithms have demonstrated that ensemble methods such as Random Forest provide robust performance and are particularly suitable for high-dimensional health datasets (20). Ensemble models trained on BRFSS 2020 data improved Sensitivity and reduced false negatives (21). Feature importance analysis further identified BMI, Age and physical activity as the most significant predictors, consistent with clinical evidence (22).

Beyond diabetes, BRFSS has been used to predict obesity, stroke and heart disease, demonstrating its versatility for chronic disease analytics (23–25). Gradient Boosting and Logistic Regression models applied to BRFSS 2021 data achieved an AUC of 0.86, while highlighting challenges related to redundant variables and missing values. These findings emphasize both the value and complexity of BRFSS data and underline the importance of preprocessing and feature correlation analysis (26).

Recent deep learning approaches, including Improved Inception-Capsule Networks and Deep Kronecker Neural Networks optimized using Sand Cat Swarm Optimization, have demonstrated high predictive performance for disease prediction (27, 28). However, these methods require substantial computational resources and offer limited interpretability. In contrast, Random Forest-based frameworks provide a more transparent

and computationally efficient alternative for large-scale diabetes prediction using BRFSS data. Despite the success of Random Forest models, limited attention has been given to the effects of class balancing and feature correlation analysis on model sensitivity and interpretability, particularly for BRFSS datasets. This study addresses this gap through a systematic evaluation of balancing techniques on the Teboul Health Indicators dataset to improve fairness, interpretability and clinical relevance in diabetes prediction.

Research Contributions

The main contributions of this study are fourfold:

(a) Development of a Random Forest framework incorporating feature correlation analysis and class balancing techniques.
(b) Comparative evaluation of under-sampling, oversampling, SMOTE and class weighting methods.
(c) Identification of key predictors associated with diabetes risk.
(d) Assessment of clinically relevant performance metrics, including recall and area under the ROC curve (AUC).

Methodology

Data Description

The research was based on the Behavioural Risk Factor Surveillance System (BRFSS) 2015 data (29), otherwise referred to as the Teboul Health Indicators Dataset, which is a massive health survey by the Centers for Disease Control and Prevention (CDC) in the United States (30). The dataset consists of about 253,680 records of adult respondents throughout the country, each of which is described by 21 predictor variables and one binary target variable, Diabetesbinary. This target represents the cases of people who are diagnosed to have diabetes and 1 represents a case of diabetes and 0 represents a non-diabetic case.

Table 1: Attribute Categorization of the Teboul Health Indicators (BRFSS 2015) Dataset

Attribute Type	Count	Examples
Numerical	4	BMI, Age, PhysHlth, MentHlth
Categorical/Binary	17	HighBP, HighChol, CholCheck, Smoker, Stroke, HeartDiseaseorAttack, PhysActivity, Fruits, Sex, Veggies, Education, HvyAlcoholConsump, AnyHealthcare, NoDocbcCost, GenHlth, DiffWalk, Income
Target	1	Diabetes_binary (0/1)

Note: BMI: Body Mass Index, PhysHlth: Physical Health, MentHlth: Mental Health, HighBP: High Blood Pressure, HighChol: High Cholesterol, CholCheck: Cholesterol Check, HeartDiseaseorAttack: History of Heart Disease or Heart Attack, PhysActivity: Physical Activity, HvyAlcoholConsump: Heavy Alcohol Consumption, AnyHealthcare: Access to Healthcare Coverage, NoDocbcCost: Could Not Visit a Doctor Due to Cost, GenHlth: General Health Status, DiffWalk: Difficulty Walking, Diabetes_binary: Diabetes status (0: non-diabetic, 1: diabetic).

Table 1 presents a summary of the attributes contained in the Teboul Health Indicators dataset used in this study. The dataset consists of four numerical variables and seventeen binary and categorical variables representing demographic, behavioural and clinical health characteristics. The numerical attributes include BMI, Age, Physical Health (PhysHlth) and Mental Health (MentHlth), whereas the categorical variables describe lifestyle factors, healthcare access and comorbid conditions. The target variable, Diabetes_binary, indicates the presence [1] or absence [0] of diabetes. These diverse attributes provide a comprehensive basis for diabetes risk prediction.

Table 2 presents the descriptive statistics of the four numerical features included in the dataset, namely BMI, Age, Physical Health (PhysHlth) and Mental Health (MentHlth). All variables contain 253,680 observations, indicating complete data availability. The average BMI of the respondents was 28.38, while the mean values of Age, PhysHlth and MentHlth were 8.03, 4.24 and 3.18, respectively. The wide range and standard deviation values suggest considerable variability among individuals in terms of physical and mental health conditions. The rest of the features are binary or nominal and coded 0 or 1 and the distributions are presented in the following table in the form of frequencies. This summary gives a clear image of the quantitative and qualitative information about the data.

Table 2: Statistical Summary of Numerical Features

	BMI	Age	PhysHlth	MentHlth
count	253680.00	253680.00	253680.0	253680.0
mean	28.382364	8.032119	4.242081	3.184772
std	6.608694	3.054220	8.717951	7.412847
min	12.000000	1.000000	0.000000	0.000000
25%	24.000000	6.000000	0.000000	0.000000
50%	27.000000	8.000000	0.000000	0.000000
75%	31.000000	10.00000	3.000000	2.000000
max	98.000000	13.00000	30.00000	30.00000

Note: BMI: Body Mass Index, PhysHlth: Physical Health, MentHlth: Mental Health.

Table 3 presents the distribution of the multicategory variables in the dataset. Among these variables, Income has eight categories, Education has six categories and General Health (GenHlth) consists of five categories. These attributes represent socioeconomic and health status characteristics and provide additional information for diabetes risk assessment. Income, Education and General Health (GenHlth) are ordinal categorical variables. Therefore, their distributions were not summarized using counts of 1s and percentages, unlike binary variables.

Table 3*:* Distribution of Multicategory Variables

Variable	Number of Categories
Income	8
Education	6
GenHlth	5

Note: GenHlth: General Health Status.

Table 4*:* Frequency Distribution of Binary and Categorical Variables

Feature	Count of 1s	Percentage (%)
AnyHealthcare	241263	95.11
Veggies	205841	81.14
PhysActivity	191920	75.65
Fruits	160898	63.43
Smoker	112423	44.32
Sex	111706	44.03
HighBP	108829	42.9
HighChol	107591	42.41
DiffWalk	42675	16.82
HeartDiseaseorAttack	23893	9.42
NoDocbcCost	21354	8.42
HvyAlcoholConsump	14256	5.62
Stroke	10292	4.06

Note: AnyHealthcare: Access to Healthcare Coverage, Veggies: Vegetable Consumption, PhysActivity: Physical Activity, Fruits: Fruit Consumption, Smoker: Smoking Status, Sex: Biological Sex, HighBP: High Blood Pressure, HighChol: High Cholesterol, DiffWalk: Difficulty Walking, HeartDiseaseorAttack: History of Heart Disease or Heart Attack, NoDocbcCost: Could Not Visit a Doctor Due to Cost, HvyAlcoholConsump: Heavy Alcohol Consumption, Stroke: History of Stroke.

Table 4 presents the frequency distribution of the binary variables in the dataset. Most respondents reported having access to healthcare (95.11%), consuming vegetables (81.14%) and engaging in physical activity (75.65%). Approximately 43% of the participants had high blood pressure and high cholesterol, while smaller proportions reported stroke (4.06%), heavy alcohol consumption (5.62%) and heart disease or heart attack (9.42%). These variables capture important behavioural and clinical risk factors associated with diabetes.

Figure 1 illustrates the overall experimental workflow adopted in this study for diabetes prediction using the Teboul Health Indicators Dataset derived from BRFSS 2015. The process begins with dataset acquisition, followed by data preprocessing and stratified train-test splitting using an 80:20 ratio. Various class balancing techniques, including random under-sampling, random oversampling, SMOTE and class weight

adjustment, are subsequently applied to address class imbalance.

Feature correlation analysis is then performed to identify significant predictors, after which a Random Forest model is trained for diabetes classification. The trained model is used to generate predictions and its performance is evaluated using Accuracy, Precision, Recall, F1-score and Area Under the ROC Curve (AUC). Finally, a comparative analysis of the different balancing approaches is conducted to determine the most effective strategy and interpret the results in terms of predictive performance and clinical relevance.

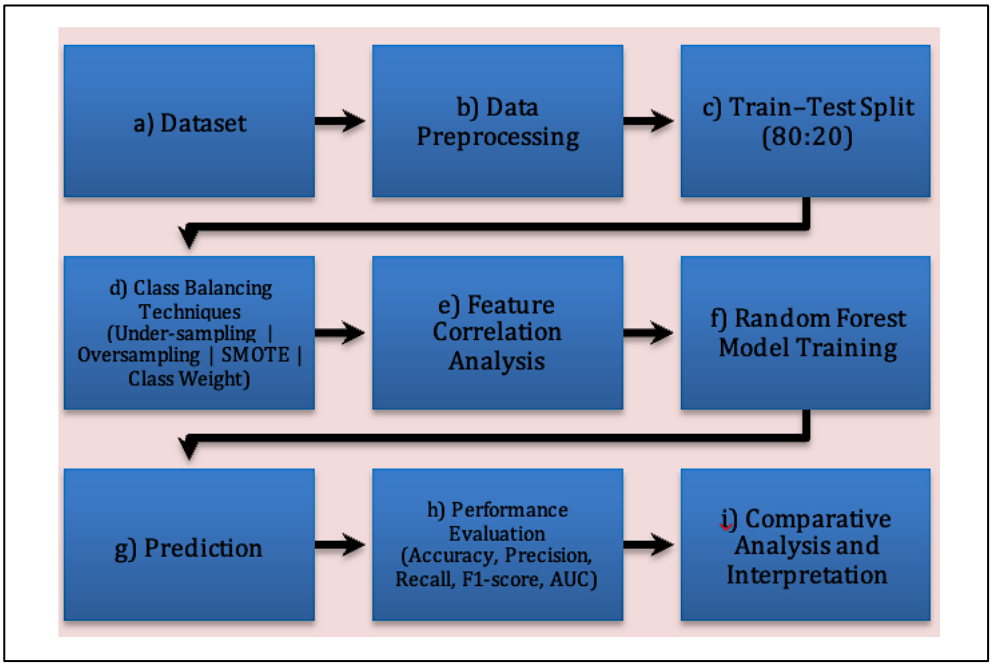

Figure 1: Experimental Design of the Proposed Diabetes Prediction Framework

Preprocessing

The dataset was subjected to a comprehensive preprocessing workflow to ensure consistency, reliability and suitability for machine learning analysis. Before model development, a missing value analysis was performed on all 21 predictor variables and the target variable to ensure data quality and reproducibility. The inspection revealed that no missing values were present in the Teboul Health Indicators Dataset derived from BRFSS 2015. Consequently, no imputation or deletion procedures were required and all 253,680 records were retained for subsequent analysis. The absence of missing values ensured data consistency and eliminated potential bias associated with missing data treatment. Categorical and binary variables were verified to ensure correct 0/1 encoding and label encoding was applied where necessary to convert qualitative variables into numerical representations suitable for machine learning algorithms.

Continuous numerical variables, namely Body Mass Index (BMI), Age, Physical Health (PhysHlth) and Mental Health (MentHlth), were standardized using the StandardScaler algorithm to reduce differences in feature scales and prevent individual variables from disproportionately influencing model learning. Subsequently, the pre-processed dataset was divided into training and testing sets using an 80:20 stratified sampling approach, thereby preserving the original proportions of diabetic and non-diabetic samples and minimizing sampling bias (31). As a result, the dataset remained complete, consistently scaled and representative of the underlying population, providing a robust foundation for feature correlation analysis, class balancing and model training.

Feature Selection

Three complementary analytical procedures were applied to select features, each of which is designed to describe various sides of the relationships between predictor variables and the target outcome (Diabetes_binary). An exploratory study to study the linear dependencies and pairwise associations

between attributes was initially conducted using correlation analysis and heatmap visualization. Meanwhile, the model-driven selection with nonlinear interactions and hierarchical dependencies has been implemented using the built-in feature importance technique of the Random Forest.

First, correlation analysis revealed that General Health (GenHlth), High Blood Pressure (HighBP) and Difficulty Walking (DiffWalk), Body Mass Index (BMI), High Cholesterol (HighChol) and Age were the most significantly positively related variables with the target variable. The visualization of a heatmap provided in Figure 2 also proved that HighBP and HighChol clinical indicators were strongly correlated with each other and the outcome of diabetes. On the contrary, behavioural variables, including Physical Activity and Dietary Habits (Fruits, Veggies), were found to have weaker associations.

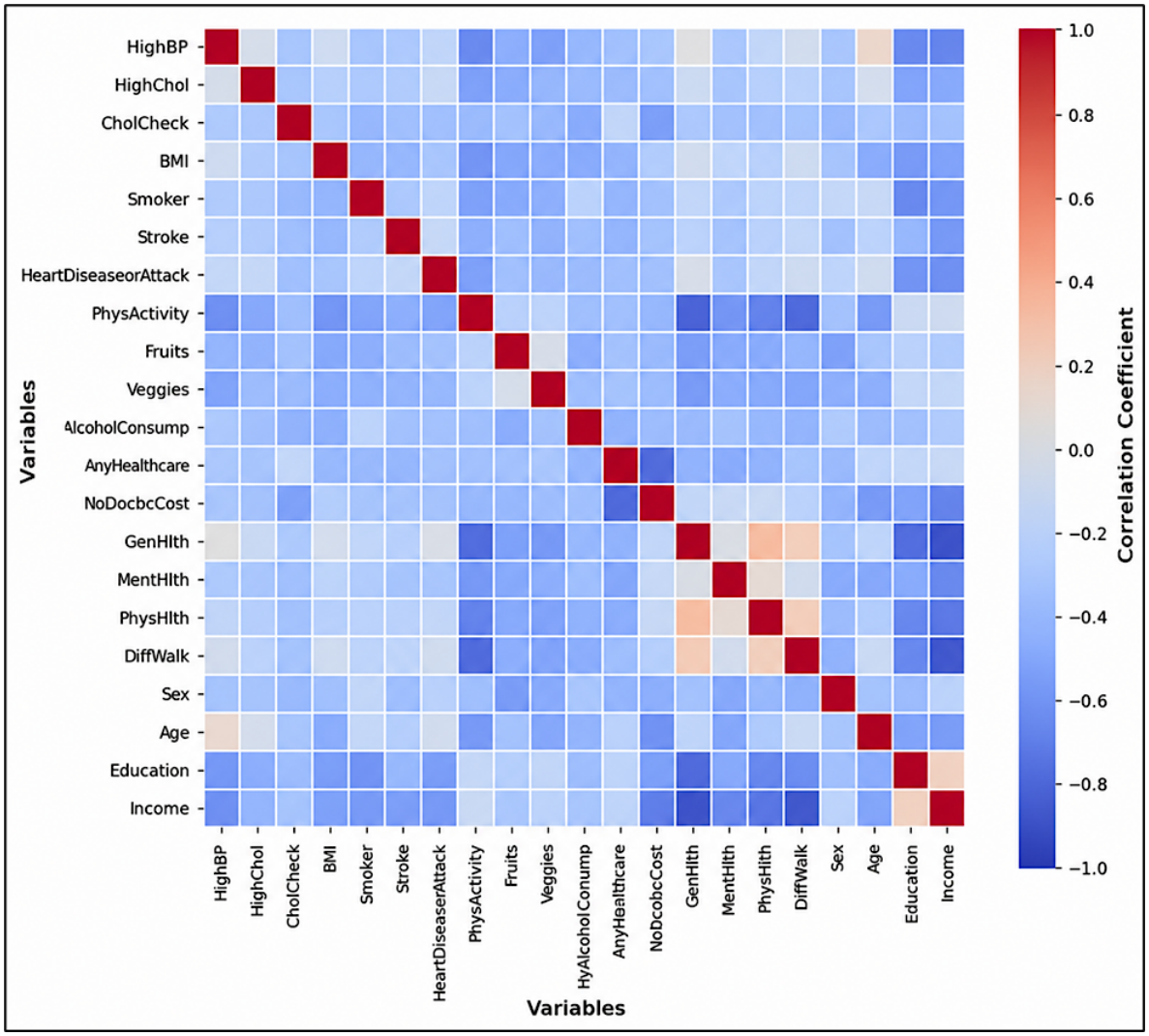

Figure 2: Correlation Heatmap of the Original BRFSS 2015 Health Indicators Dataset

The random forest feature importance metric was employed as the main model-based selection criterion in order to include both the linear and nonlinear interaction of the features. It is a method that measures the importance of each feature across a set of decision trees, enabling a clear ranking of the importance of variables. The analysis revealed that HighBP, GenHlth, HighChol, BMI, Age, Income, DiffWalk, Education, Sex and PhysActivity were the top ten predictors that are the most significant features in predicting diabetes.

All these features make it a balanced combination of physiological, demographic and behavioural signs, which has interpretability and computer efficiency. They were thus stored to be used at later times of class balancing and model training. Table 5 presents the feature importance scores obtained from the Random Forest classifier. Body Mass Index (BMI) was identified

as the most influential predictor with an importance score of 0.1839, followed by Age [0.1220] and Income [0.0987]. Physical Health (PhysHlth), General Health (GenHlth) and Education also contributed substantially to diabetes prediction. Overall, the results indicate that physiological, demographic and socioeconomic factors collectively play an important role in determining diabetes risk.

Table 5: The Values of Feature Importance, which were Calculated Using the Random Forest Classifier, were then Used to Apply Class Balancing Techniques

Feature	Importance
BMI	0.183869
Age	0.12197
Income	0.098667
PhysHlth	0.084372
GenHlth	0.071503
Education	0.070192
MentHlth	0.064209
HighBP	0.043104
Smoker	0.033159
Fruits	0.033059
Sex	0.027639
HighChol	0.02658
Veggies	0.026231
PhysActivity	0.025991

Note: BMI: Body Mass Index, PhysHlth: Physical Health, GenHlth: General Health Status, MentHlth: Mental Health, HighBP: High Blood Pressure, Smoker: Smoking Status, Fruits: Fruit Consumption, HighChol: High Cholesterol, Veggies: Vegetable Consumption, PhysActivity: Physical Activity

In order to have a better understanding of how individual predictors help in detecting diabetes, feature distributions of the top 10 most important attributes were visualized before class balancing, which is presented in Figure 3. The frequency of distribution of a given feature in diabetic and non-diabetic individuals is presented in each subplot. The graphical difference between the two distributions highlights the discriminatory ability of every feature. To further explain the influence of the individual predictors on the classification of diabetes, the distributions of the feature-wise were visualized for the top 10 most important attributes before class balancing, as shown in Figure 3. Each subplot demonstrates the frequency of a given feature of diabetic and non-diabetic people. The distinct separation between the two distributions highlights the discriminative ability of each feature.

Interestingly, attributes such as High Blood Pressure (HighBP), General Health (GenHlth), Body Mass Index (BMI) and Age demonstrate some noticeable rightward shift in the diabetic group, which proves their strong positive correlation with risk of diabetes. On the other hand, the separation of variables like Income and Education is moderate and this means that the variable has an indirect but strong effect on the prevalence of diabetes due to the socioeconomic and lifestyle aspects. These visual patterns support the quantitative results from the Random Forest feature importance ranking and correlation analysis, reinforcing the interpretability and clinical relevance of the chosen predictors.

Class Balancing and Correlation Analysis: Distribution of Resampled Training Class

To determine the influence of various balancing strategies on the dataset and their relationships with the features, various resampling techniques were used on the training subset and the independent test set remained the same to evaluate the dataset without bias. The distribution of the classes in the original training data was highly skewed, with 174,667 non-diabetic (class 0) and 28,277 diabetic (class 1) samples and this made 202,944 samples.

In order to overcome this imbalance, a number of resampling techniques were employed. Random under-sampling technique was used to equalize the majority class to the minority level, making 28,277 samples in each of the classes and 56,554 records in total (32, 33). Random Oversampling, on the other hand, repeated the

instances of the minority class until the minority class and the majority class each had 174,667 samples, which made the training set 349,334 records (34).

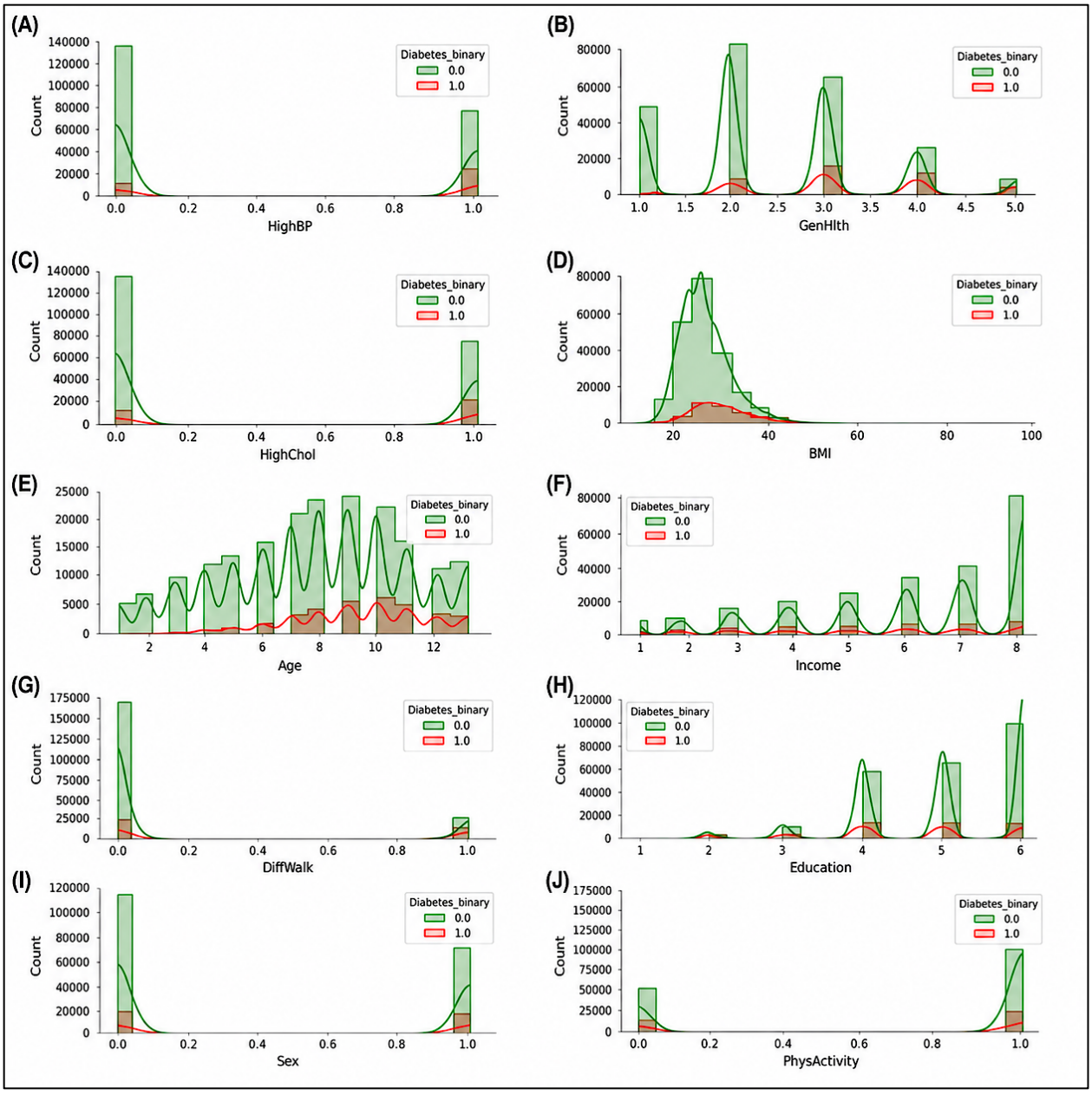

Figure 3: Distribution of Selected Health Indicators Across Diabetes Classes. Histograms with Kernel Density Estimates Compare Non-Diabetic [0] and Diabetic [1] Populations for - (A) Highbp, (B) Genhlth, (C) Highchol, (D) BMI, (E) Age, (F) Income, (G) Diffwalk, (H) Education, (I) Sex, (J) Physactivity, Illustrating Their Discriminatory Potential for Diabetes Prediction

The SMOTE (Synthetic Minority Oversampling Technique) algorithm formed synthetic minority samples, with an equal number of samples per class of 174,667 and the same number of samples in total training. Lastly, the Class Weight approach (35) kept the original data distribution but adjusted for imbalance by assigning inverse-proportional weights to the minority class during model training.

After resampling, correlation analyses were done on each method to test the effect of class balancing on feature-feature and feature-target correlation. The resulting correlation matrices gave an insight into how synthetic and random resampling changed statistical relationships between the health indicators and the outcome variable of diabetes, which formed the basis of evaluating the

stability and robustness of the features selected among balancing strategies.

Top-10 Feature Correlations with the Target Variable

Correlation analysis was performed to examine the association between each predictor and the target variable (Diabetes_binary) under different class-balancing strategies. In the original imbalanced dataset, General Health (GenHlth) (r = 0.2939) and High Blood Pressure (HighBP) (r = 0.2634) exhibited the strongest correlations, followed by Difficulty Walking (DiffWalk) (r = 0.2198), Body Mass Index (BMI) (r = 0.2192) and High Cholesterol (HighChol) (r = 0.2007). HeartDiseaseorAttack (r = 0.1780), Age (r = 0.1772), Physical Health (PhysHlth) (r = 0.1722), Stroke (r = 0.1063) and Mental Health (MentHlth) (r = 0.0698) also showed positive associations with diabetes.

After Random Under-sampling, stronger correlations were observed because of the balanced class distribution. GenHlth and HighBP increased to 0.4082 and 0.3802, respectively, while HighChol, BMI, Age and DiffWalk exhibited correlations ranging from 0.2720 to 0.2880. Similar trends were obtained with Random Oversampling, where GenHlth (r = 0.4089), HighBP (r = 0.3795), BMI (r = 0.2919), HighChol (r = 0.2867) and Age (r = 0.2761) remained the dominant predictors.

SMOTE further strengthened these relationships, producing the highest correlations among all balancing techniques. GenHlth (r = 0.4189), HighBP (r = 0.4101) and HighChol (r = 0.3147) exhibited the strongest associations, followed by Age (r = 0.2801), DiffWalk (r = 0.2750), PhysHlth (r = 0.2140) and HeartDiseaseorAttack (r = 0.2052). In contrast, the Class Weight approach preserved the correlation structure of the original dataset because it modifies algorithm weights without altering the sample distribution.

Overall, the same variables consistently emerged as the most influential predictors across all balancing strategies. GenHlth, HighBP, BMI, HighChol, Age and DiffWalk showed the strongest associations with diabetes, while class balancing, particularly SMOTE, has enhanced the strength of these relationships, indicating improved representation of the minority class.

Feature–feature Correlation Change Analysis

Besides the evaluation of feature-target correlations, this study also analysed the effect of class balancing on the internal relationships of independent variables by comparing the feature-feature correlation, as shown in Table 6. In both resampling strategies, the change in correlation coefficients between all pairs of features in absolute terms was calculated against the original dataset and it measured the impact of data augmentation or data reduction on the inter-feature dependencies.

Table 6: Summary of Feature–feature Correlation Changes Across Class Balancing Methods

Method Name	Max_Abs_Change	Mean_Abs_Change
Original	0.0000	0.000000
Under-sample	0.0463	0.009754
Oversample	0.0488	0.009754
Smote	0.0630	0.014353
Class_Weight	0.0000	0.000000

Note: Max_Abs_Change: Maximum Absolute Change, Mean_Abs_Change: Mean Absolute Change, Smote: Synthetic Minority Over-sampling Technique, Class_Weight: Class Weighting Method

The initial data was used as a base. The maximum absolute change was 0.0000 and the average absolute change was 0.000000, which showed that the correlation structure did not change. Following the application of Random Under-sampling, it was found that there was modest variation, with the largest change of 0.0463 and a mean change of 0.009976, which suggests that there was minor yet significant changes in the relationships between the features as a result of the reduction of the sample. Random Oversampling also demonstrated the same effect, with a maximum change of 0.0488 and a mean change of 0.009754, indicating that even duplicating minority samples has a minimal effect on inter-feature dependencies.

The SMOTE (Synthetic Minority Oversampling Technique) technique led to the greatest change of

features, feature relationships (maximum change of 0.0630, average change of 0.014353), than any other method tested. This is in keeping with the synthetic data generation process of SMOTE, which bases its work on interpolating new samples in the feature space and thus alters linear relationships among the variables. On the other hand, the Class Weight system that does not distort the original data distribution indicated no significant changes (maximum and average = 0.0000), which does not influence the interdependence of features.

In general, the analysis indicates that despite the fact that all resampling techniques lead to minor shifts in feature correlations, the degree of distortion is small (less than 0.015 on average) and would probably not impact the stability and interpretability of the model. Nevertheless, SMOTE generates the greatest deviations and therefore, one should be careful with the interpretation of correlation-based insights in synthetically balanced datasets

Changes in Feature-target Correlations Relative to the Original Dataset

In order to estimate the influence of class balancing on the statistical correlations between predictors and the target, the absolute difference of feature-target correlation values was computed at each resampling technique compared to the original dataset. The purpose of this analysis was to compare the effects of data augmentation or data reduction on the apparent strength of the relationship between individual health indicators and diabetes occurrence.

Out of all methods, SMOTE (Synthetic Minority Oversampling Technique) gave the highest total variation, with its greatest absolute change of 0.1467. High Blood Pressure (HighBP), General Health (GenHlth), High Cholesterol (HighChol), Age and Body Mass Index (BMI) recorded the most significant positive changes with SMOTE (HighBP +0.1467, General Health +0.1250, High Cholesterol +0.1140, Age +0.1029 and Body Mass Index +0.0743). These findings indicate that synthetic oversampling enhanced the likelihood of these important predictors to rise in the balanced data distribution.

Random Under-sampling method exhibited the greatest absolute change of 0.1168, primarily because of the increase in HighBP [0.1168], GenHlth [0.1143], Age [0.0977], HighChol [0.0880] and BMI [0.0690]. The same pattern was observed with the Random Oversampling technique, with a maximum change of 0.1161, indicating similar positive changes in the same features. On the other hand, the Class Weight method that preserves the original class composition did not exhibit any quantifiable difference [0.0] in feature-target correlations as was anticipated.

Although the greatest positive changes were detected in the group of clinical and physiological variables, some socioeconomic and behavioural attributes depicted moderate negative changes in the resampling methods. These were Income [0.06], Education [0.048], Physical Activity (PhysActivity) [0.03 to 0.045], Heavy Alcohol Consumption (HvyAlcoholConsump) [0.03 to 0.05] and dietary variables such as Veggies and Fruits, which had slight reductions in the strength of correlation. Such decreases imply that lifestyle and socioeconomic variables were slightly less strongly related to diabetes status after data balancing, maybe due to the resampled minority group having a higher proportion of clinical markers.

In general, the findings indicate that resampling does not alter the identity of the significant predictive features but has a significant impact on their correlation coefficients, particularly with SMOTE, which brings in the most drastic improvement in correlation. This raises the necessity to do feature selection followed by balancing classes to prevent synthetic data generation, which inflates feature importance.

Feature Importance Comparison Before and After Class Balancing

A correlation-based analysis of the relationships between the features was followed by a model-driven feature importance analysis that furthered the insight into the relevance of variables beyond the linear relationship. Although correlation analysis identifies direct statistical relationships, it lacks the ability to identify feature interactions, as well as nonlinear effects. Thus, feature importance using Random Forests was used to determine the contribution of each attribute to the performance of the model when utilizing different methods of class-balancing. The step provided a broader perspective on balancing strategies on the

significance and stability of predictors in the learning process.

Figure 4 indicates the relative importance of the best features in all balancing methods, with a focus on regular patterns and changes due to the generation of synthetic data. In all the combinations, HighBP, BMI, GenHlth, Age and PhysHlth were always the most significant predictors of diabetes risk. Nevertheless, significant variations in the size of their significance were noticed. The SMOTE technique, specifically, made the samples of CholCheck and HvyAlcoholConsump more significant, presumebly, because it is a data interpolation technique that synthetically interpolates the minority-class data samples. Simultaneously, the Fruits and Veggies, as well as other behavioural variables, were slightly less significant due to Random Under-sampling. In spite of these numerical variations, the general order of medically significant predictors remained quite similar between the methods, which is evidence of the strength of the feature selection of Random Forest. The visualization proves that the size, but not the identity, of the largest feature predictors is primarily influenced by the class balancing.

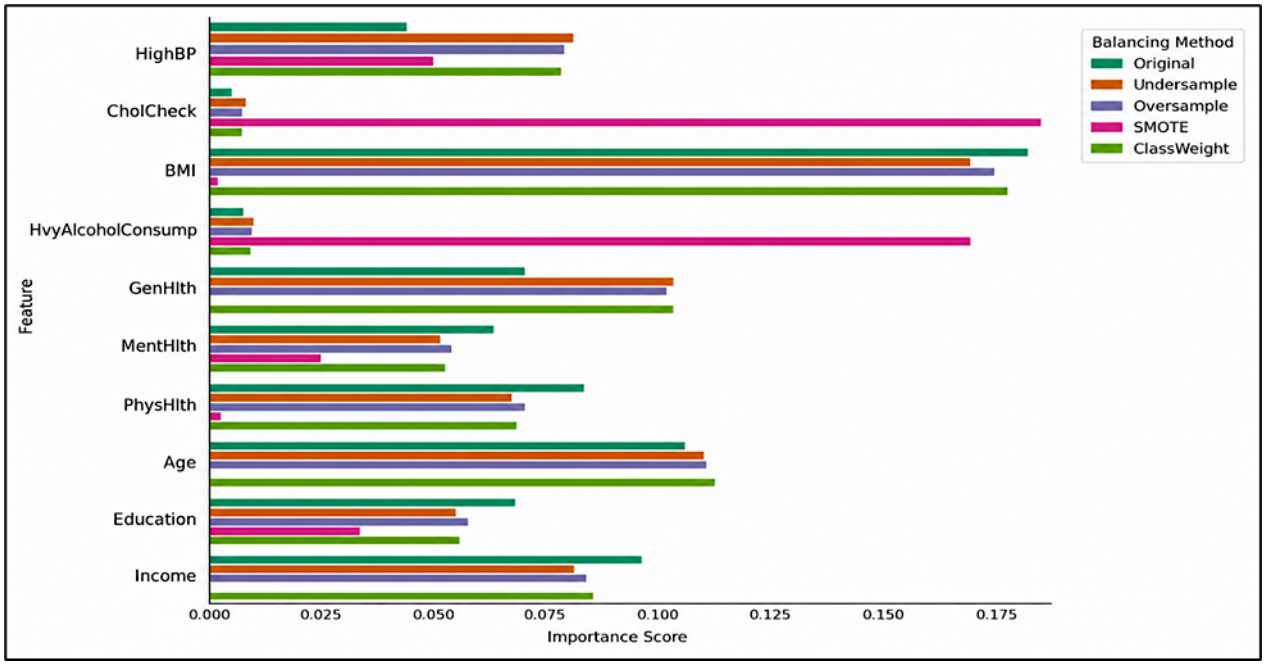

Figure 4: Comparison of Random Forest Feature Importance Scores Before and After Applying Different Class-balancing Methods

Random Forest Model

The main predictive model was the Random Forest algorithm that was applied to the data after the analysis of the class-balancing effect and the correlation among the features in order to estimate the risk of diabetes and determine key health indicators. The primary classifier that was used to categorize diabetic [1] and non-diabetic [0] people was the Random Forest (RF) classifier, which was used to perform binary classification. It was set to 150 estimators (trees), a maximum depth of 20 and at least two samples per split and the sqrt feature selection method per node to regulate the complexity of the tree and enhance generalization. A 5-fold scheme was adopted in the process of hyperparameter tuning through the utilization of the GridSearchCV. The grid of four significant hyperparameters, including number of estimators [100, 150, 200], maximum depth [10, 20, 30], minimum samples per split [2, 3, 5] and feature selection criterion ["sqrt", "log2"] was tested and 54 combinations were tried on five folds and 270 model evaluations were performed per class-balancing setup. Each iteration took place in five parts of equal size, four of which were trained and one of which was used as a validation. To make performance measures reliable and minimize bias, all folds were averaged to obtain the performance measures. The optimization criterion was the F1-score to trade off Precision and Recall, which is a vital strategy in reducing false negatives in medical predictions. The final model was selected as the most successful configuration in terms of the maximum mean F1-

score of the folds. This data optimization approach offered reliable and reproducible results on all balancing strategies and was computationally efficient when dealing with large-scale health data, including the Teboul Health Indicators that are based on the BRFSS.

Results

This subchapter provides and discusses an experimental finding of the application of the Random Forest classifier on the BRFSS-based Teboul Health Indicators dataset with five class-balancing strategies, including Original (Unbalanced), Random Under-sampling, Random Oversampling, SMOTE and Class Weighting. The optimized model parameters that were found in the process of Grid Search Cross-Validation (150 trees, max depth = 20, min samples split = 2, max features = "sqrt") were used to evaluate each method. As shown in Table 7, performance measures Accuracy, Precision, Recall, F1-Score and AUC-ROC show that although accuracy is high in all configurations, Random Under-sampling has the highest balance between Sensitivity and overall performance as indicated by the highest Recall [0.7774], F1-Score [0.4259] and AUC [0.8076].

Table 7: Performance Comparison of Random Forest Classifier Across Different Class Balancing Method

Method	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Original	0.859705	0.490328	0.175697	0.258696	0.796921
Under-Sample	0.707919	0.293245	0.777479	0.425865	0.807612
Over-Sample	0.843385	0.417560	0.314189	0.358573	0.792626
SMOTE	0.855881	0.460870	0.202433	0.281305	0.795340
Class_Weight	0.856827	0.459051	0.154619	0.231323	0.794689

Note: AUC-ROC: Area Under the Receiver Operating Characteristic Curve, F1-score: harmonic mean of Precision and Recall, SMOTE: Synthetic Minority Over-sampling Technique

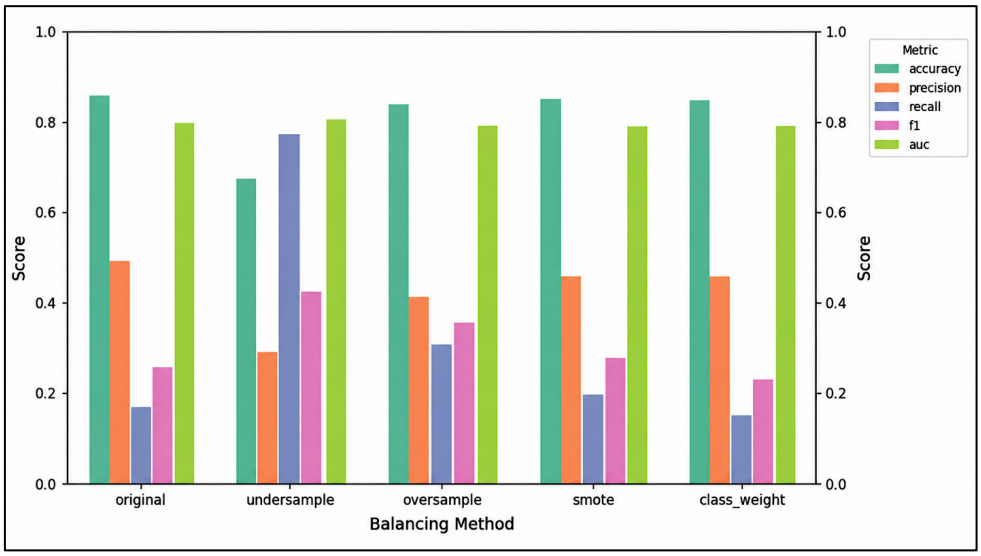

Figure 5: Performance Comparison of Balancing Methods

Figure 5 provides a comparison of the performance of various class-balancing methods based on random forest. Although the original dataset demonstrated the best accuracy [0.86], under-sampling demonstrated the best recall [0.78] and f1-score [0.43], which suggests that it is more sensitive to diabetic cases.

The values of Area Under the Receiver Operating Characteristic (AUC-ROC) of all balancing methods were within a very small range of 0.79-0.81, which implied that the Random Forest classifier was a reliable discriminative model among balancing methods. Such stability indicates that the model architecture is well-adapted to the dataset and is capable of distinguishing between the diabetic and non-diabetic cases even when it is trained on the data that does not follow the same distribution.

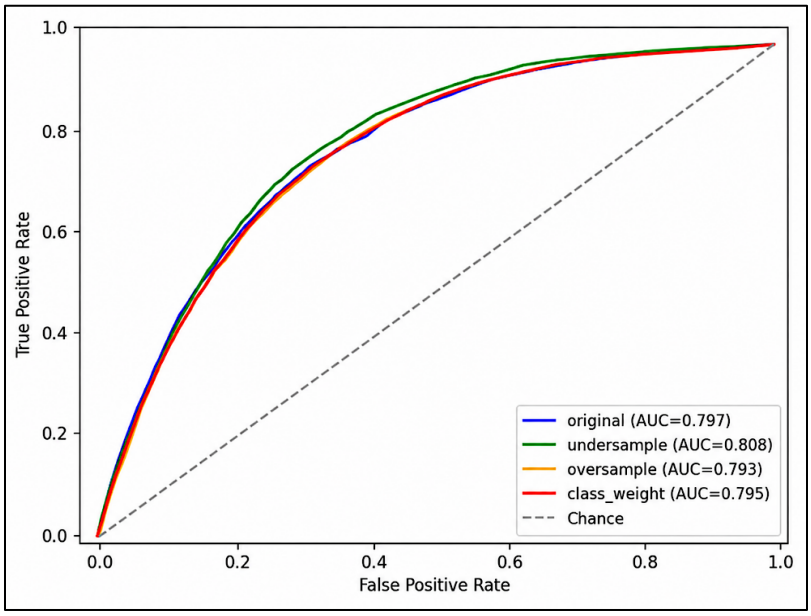

Figure 6: ROC Curves of the Random Forest Classifier Under Different Class-balancing Methods

The Random Under-sampling method has the largest value of AUC of 0.808, a little higher than the original [0.797], oversampling [0.793], SMOTE [0.795] and class-weighted [0.795] models, which are presented in Figure 6. Though the numerical change appears not to be very significant, it is statistically significant, as it proves that under-sampling increased the capacity of the model to rank positive instances better than negative instances at all potential classification thresholds.

The AUC metric conceptually is used to quantify the total separation between the two classes. The larger the AUC, the better the balance between the True Positive Rate (Recall) and False Positive Rate across thresholds. Consequently, the marginally greater AUC of the under-sampled model supports its strength and generalization, which supports the idea that the developed model is not only more efficient in identifying cases of diabetes but also has a stable differentiation of positive and negative results.

Discussion

The present study demonstrated that although the original Random Forest model achieved the highest accuracy [0.8597], accuracy alone was insufficient for evaluating performance because of the highly imbalanced nature of the dataset. Previous studies have emphasized that overall accuracy may provide misleading results in minority-class prediction problems and that evaluation measures such as Recall and F1-score are more suitable for medical applications involving imbalanced data (32–35). In the original dataset, the model achieved a Recall of only 0.1757, indicating that most diabetic cases were missed despite the high accuracy. Similar limitations of accuracy-based evaluation have been reported in diabetes prediction and imbalanced learning studies (17, 20, 21).

The experimental results revealed a clear trade-off between Precision and Recall. The original imbalanced model achieved the highest Precision [0.49], indicating fewer false positives, but this was accompanied by poor Sensitivity. In contrast, Random Under-sampling substantially improved Recall to 0.777 while reducing Precision to 0.29. Such an inverse relationship between Precision and Recall is commonly observed in classification systems and has been reported in previous studies on diabetes prediction and imbalanced datasets (17, 21, 32, 34). From a clinical perspective, false negatives are generally more detrimental than false positives because missed diabetic cases may delay diagnosis and treatment. Therefore, prioritizing Sensitivity over overall accuracy is considered appropriate for disease screening and preventive healthcare applications (20, 21, 35).

Among the balancing techniques evaluated, Random Under-sampling provided the most balanced classification performance, achieving the highest F1-score [0.426]. Although under-sampling reduced the size of the majority class, it enabled the model to learn more representative decision boundaries and improved the detection of diabetic individuals. Previous studies have

similarly demonstrated the effectiveness of balanced learning approaches and ensemble models in improving minority-class prediction and reducing classification bias (17, 20, 21). Furthermore, the importance of Recall and F1-score in imbalanced classification has been highlighted in studies on SMOTE and class imbalance learning. These findings suggest that balanced data distributions contribute to improved fairness and clinical relevance of predictive models.

Overall, the results indicate that Random Under-sampling offers a clinically meaningful trade-off between Precision and Recall and provides superior Sensitivity for early diabetes detection. In healthcare screening systems, minimizing false negatives is often more important than maximizing overall accuracy, since undetected diabetic cases may lead to delayed intervention and increased risk of complications. Therefore, the proposed Random Forest framework with class balancing represents a practical and interpretable approach for population-level diabetes risk prediction. A comparison of the present findings with previous studies and their clinical implications is summarized in Table 8. Random under-sampling produced the highest Recall and F1-score, demonstrating its effectiveness in improving minority-class detection, which is consistent with earlier studies on class imbalance and sampling techniques. These findings emphasize that prioritizing sensitivity over accuracy is more appropriate for diabetes screening, where reducing false negatives is crucial for early diagnosis and timely intervention.

Table 8: Comparison of the Present Findings with Previous Studies and Their Clinical Implications

Performance Aspect	Present Study Observation	Supporting Literature	Interpretation
Accuracy	The original imbalanced model achieved the highest accuracy [0.8597] but low Recall [0.1757].	Class imbalance has been extensively investigated in previous studies (32, 35).	Accuracy alone can be misleading in imbalanced datasets because it favours the majority class.
Precision	The highest precision (0.49) was obtained with the original dataset.	Ensemble learning approaches have demonstrated improved predictive performance and sensitivity in earlier studies (17, 21).	High Precision reduces false positives but may increase false negatives.
Recall (Sensitivity)	Random Under-sampling achieved the highest Recall [0.777].	Previous studies have reported the effectiveness of machine learning techniques for handling imbalanced diabetes datasets (20, 21, 32).	High Recall is clinically important because missed cases of diabetes can delay treatment.
F1-score	Random Under-sampling achieved the highest F1-score [0.426], almost twice that of the original model [0.259].	Synthetic oversampling and class imbalance problems have been widely studied in the literature (34, 35).	F1-score provides a balanced assessment by considering both Precision and Recall.
Effect of Class Imbalance	Original data favoured the majority class and resulted in poor minority-class detection.	Various sampling and imbalance-handling strategies have been proposed in earlier research (32, 33, 35).	Imbalanced datasets require specialized handling to avoid biased predictions.
Effect of Under-sampling	Improved Recall and F1-score while reducing overall accuracy.	Improved classification performance using ensemble methods has been reported previously (17, 21).	A reduction in accuracy is acceptable when sensitivity to positive cases is increased.
Precision–Recall Trade-off	Higher Recall was associated with lower Precision.	Resampling techniques have been widely employed to address class imbalance problems (32, 34).	This trade-off is common in medical classification problems.

Clinical Significance	Prioritizing Sensitivity over Accuracy provides better early detection of diabetes.	Previous investigations have demonstrated the effectiveness of machine learning approaches for imbalanced healthcare datasets (20, 21, 35).	False negatives are more harmful than false positives in healthcare screening.
Overall Best Method	Random Under-sampling provided the most balanced performance.	Sampling-based approaches for mitigating class imbalance have been extensively explored in previous studies (32–34).	Suitable for diabetes screening applications where balanced classification is required.

Clinical Applicability and Sensitivity-precision Trade-off

The proposed Random Forest model combined with Random Under-sampling is particularly suitable for population-level diabetes screening applications, where the primary objective is the early identification of individuals at risk rather than definitive diagnosis. The model achieved a Recall of 0.777, indicating that nearly 78% of diabetic cases were correctly detected. Although this improvement in Sensitivity was accompanied by a reduction in precision of 0.29, such a trade-off is considered acceptable in preventive healthcare settings. A lower Precision implies that some non-diabetic individuals may be incorrectly flagged as high risk; however, these false positives can subsequently be verified through clinical examinations and laboratory tests. In contrast, false negatives are more concerning because undetected diabetic individuals may experience delayed diagnosis and increased risk of complications. Therefore, prioritizing Sensitivity over Precision is clinically justified for screening systems, where the cost of missing true cases outweighs the inconvenience associated with additional follow-up tests. Similar observations have been reported in previous studies on diabetes prediction and imbalanced learning, which emphasized that sensitivity-oriented models provide greater clinical utility than models optimized solely for overall accuracy (20, 21, 32–35). Consequently, the proposed framework is realistically applicable as an initial risk stratification tool to assist healthcare professionals in identifying high-risk individuals and supporting early intervention, rather than as a replacement for conventional diagnostic procedures.

Conclusion

In this paper, a machine learning system based on the Random Forest was built and tested to predict diabetes based on the Teboul Health Indicators Dataset, which was obtained using the CDC BRFSS 2015 survey. The study highlighted the effects of the methods of balancing the classes - such as under-sampling, oversampling, SMOTE and weighting classes on the predictive performance and interpretability of features. These findings showed that the under-sampling method offered the most balanced Sensitivity and generalization with the highest Recall [0.777] and F1-score [0.426] and a competitive AUC [0.808]. In spite of the fact that the initial lopsided dataset gave the most precise result [0.86], it had a low Recall, which means that it did not identify an adequate percentage of diabetic cases. Here, the trade-off between Precision and Sensitivity in imbalanced health datasets is illustrated. The importance of features and correlation analyses proved that the following variables were always the most significant predictors of diabetes: General Health (GenHlth), High Blood Pressure (HighBP), Body Mass Index (BMI) and Age. The use of the balancing technique slightly modified the correlation coefficients of the secondary attributes, like Education, Income and Physical Activity, indicating that resampling may affect the nonlinear feature-target associations. In general, the under-sampling plus Random Forest yielded a powerful, understandable and computationally efficient method of early risk of diabetes detection.

Although the presented framework using Random Forest demonstrated high predictive accuracy and interpretability, there are a number of limitations that should be mentioned. First, the study was done based on one dataset (Teboul Health Indicators/BRFSS 2015), which can restrict the generalizability of the findings to other population groups or demographics. Second, the research

employed synthetic and statistical methods of balancing (under-sampling, oversampling, SMOTE and class weighting) that manipulate the training distribution; consequently, some inherent variability and characteristics of the minority group could have been reduced. Third, the model did not take into account any temporal or clinical laboratory characteristics of the data, which would enhance the accuracy of the diagnosis. Lastly, although Random Forest provides feature importance scores, advanced explainability techniques such as SHAP (SHapley Additive Explanations) and LIME were not incorporated in the present study, limiting the interpretability of individual predictions. These constraints give more chances to develop multi-source data integration, high-level balancing ensembles and explainable AI systems in the future to develop more clinically sound and transparent diabetes prediction systems.

Future research will investigate novel ensemble and hybrid balancing approaches that are dynamically changed to class ratios during the training of the model. These measures are portentous in increasing the accuracy and reliability of diabetes prediction models even further. Moreover, the explainable AI (XAI) approaches like SHAP or LIME may enhance the model transparency and assist in clinical decision-making. Classifiers based on deep learning and longitudinal validation using multi-year BRFSS data will also be an option to improve the level of temporal generalization and predictive accuracy. Such an all-inclusive view of the future of work brings hope for the ever-growing advancement in the sphere of diabetes prediction.

Abbreviations

BMI: Body Mass Index, GenHlth: General Health, HighBP: High Blood Pressure, HighChol: High Cholesterol, MentHlth: Mental Health, PhysActivity: Physical Activity, PhysHlth: Physical Health, XAI: Explainable AI

Acknowledgement

None.

Author Contributions

All the authors have contributed substantially to the research and preparation of this manuscript.

Conflict of Interest

The corresponding author declares that there is no conflict of interest on behalf of all authors.

Data Availability

The datasets used in this study are obtained from publicly available data portals. The compiled data can be made available from the corresponding author upon a reasonable request.

Declaration of Artificial Intelligence (AI) Assistance

During the preparation of this manuscript, the authors used Grammarly for language editing. ChatGPT was used to assist in language refinement and formatting. These tools were employed solely to improve clarity and language presentation. All outputs were critically reviewed and verified by the authors, who assume full responsibility for the final manuscript.

Ethics Approval

Not Applicable.

Funding

This research did not get any funding.

References

1. International Diabetes Federation (IDF). IDF Diabetes Atlas. 10th ed. Brussels, Belgium: International Diabetes Federation; 2021. https://diabetesatlas.org/
2. World Health Organization (WHO). Global Report on Diabetes. Geneva: World Health Organization; 2021. https://www.who.int/publications/i/item/9789241565257
3. Chen M, Zhang S, Zhang L. Urbanization and the epidemiology of diabetes: a systematic review. Glob Health Res Policy. 2020;5(1):45-53. doi:10.1186/s41256-020-00162-8
4. Centres for Disease Control and Prevention (CDC). Behavioural Risk Factor Surveillance System (BRFSS). Atlanta (GA): CDC; 2015. https://www.cdc.gov/brfss
5. World Health Organization (WHO). Diabetes fact sheet. Geneva: World Health Organization; 2021. https://www.who.int/news-room/fact-sheets/detail/diabetes
6. Kengne AP, Ramachandran A. Feasibility of prevention of type 2 diabetes in low- and middle-income countries. Diabetologia. 2024;67:763-772. doi:10.1007/s00125-023-06085-1
7. International Diabetes Federation. Global Health Expenditure on Diabetes 2021. Brussels, Belgium: International Diabetes Federation; 2021.

https://diabetesatlas.org/resources/idf-diabetes-atlas-reports/
8. Siegel KR, Bullard KM, Imperatore G, Ali MK, Albright A, Mercado CI, Li R, Gregg EW. Prevalence of major behavioural risk factors for type 2 diabetes. Diabetes Care. 2018;41(5):1032-1039. doi:10.2337/dc17-1775
9. Jonas DE, Crotty K, Yun JDY, Middleton JC, Feltner C, Taylor-Phillips S, Barclay C, Dotson A, Baker C, Balio CP, Voisin CE, Harris RP. Screening for prediabetes and type 2 diabetes: updated evidence report and systematic review for the US Preventive Services Task Force. JAMA. 2021;326(8):744-760. doi:10.1001/jama.2021.10403
10. Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine learning and data mining methods in diabetes research. Comput Struct Biotechnol J. 2017;15:104-116. doi:10.1016/j.csbj.2016.12.005
11. United Nations. Transforming our world: the 2030 agenda for sustainable development. New York: United Nations; 2015. https://sdgs.un.org/2030agenda
12. Petridis PD, Kristo AS, Sikalidis AK, Kitsas IK. A Review on Trending Machine Learning Techniques for Type 2 Diabetes Mellitus Management. Informatics. 2024;11(4):70. doi:10.3390/informatics11040070
13. Contreras I, Vehi J. Artificial Intelligence for Diabetes Management and Decision Support: Literature Review. J Med Internet Res. 2018;20(5):e10775. doi:10.2196/10775
14. Sisodia D, Sisodia S. Prediction of diabetes using classification algorithms. Procedia Comput Sci. 2018; 132:1578-1585. doi:10.1016/j.procs.2018.05.122
15. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine learning improve cardiovascular risk prediction using routine clinical data? PLoS One. 2017;12(4):e0174944. doi:10.1371/journal.pone.0174944
16. Edeh MO, Khalaf OI, Tavera CA, Tayeb S, Ghouali S, Abdulsahib GM, Richard-Nnabu NE, Louni A. A classification algorithm-based hybrid diabetes prediction model. Front Public Health. 2022; 10:829519. doi:10.3389/fpubh.2022.829519
17. Wang L, Wang X, Chen A, Che H. Prediction of type 2 diabetes risk and its effect evaluation based on the XGBoost model. Healthcare (Basel). 2020;8(3):247. doi:10.3390/healthcare8030247
18. Wu H, Yang S, Huang Z, He J, Wang X. Type 2 Diabetes Mellitus Prediction Model Based on Data Mining. Inform Med Unlocked. 2018;10:100-107. doi:10.1016/j.imu.2017.12.006
19. Zou Q, Qu K, Luo Y, Yin D, Ju Y, Tang H. Predicting Diabetes Mellitus with Machine Learning Techniques. Front Genet. 2018;9:515. doi:10.3389/fgene.2018.00515
20. Maniruzzaman M, Kumar N, Abedin MM, Islam MS, Suri HS, El-Baz AS, Suri JS. Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm. Comput Methods Programs Biomed. 2017;152:23-34. doi:10.1016/j.cmpb.2017.09.004
21. Perveen S, Shahbaz M, Guergachi A, Keshavjee K. Performance Analysis of Data Mining Classification Techniques to Predict Diabetes. Procedia Comput Sci. 2016;82:115-121. doi:10.1016/j.procs.2016.04.016
22. Xie Z, Nikolayeva O, Luo J, Li D. Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques. Prev Chronic Dis. 2019;16:E130. doi:10.5888/pcd16.190109
23. Tompra KV, Papageorgiou G, Tjortjis C. Strategic machine learning optimization for cardiovascular disease prediction and high-risk patient identification. Algorithms. 2024;17(5):178. doi:10.3390/a17050178
24. Heo J, Yoon JG, Park H, Kim YD, Nam HS, Heo JH. Machine Learning-Based Model for Prediction of Outcomes in Acute Stroke. Stroke. 2019;50(5):1263-1265. doi:10.1161/STROKEAHA.118.024293
25. Mohan S, Thirumalai C, Srivastava G. Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. IEEE Access. 2019;7:81542-81554. doi:10.1109/ACCESS.2019.2923707
26. Meng XH, Huang YX, Rao DP, Zhang Q, Liu Q. Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. Kaohsiung J Med Sci. 2013;29(2):93-99. doi:10.1016/j.kjms.2012.08.016
27. Nayak MSS. Improved inception-capsule deep learning model with enhanced feature selection for early prediction of heart disease. Sci Rep. 2025;15: 18551. doi:10.1038/s41598-025-18551-4
28. Nayak MSS, Syed H. Cardiac disorder classification: an efficient novel deep Kronecker neural network with sand cat swarm optimisation algorithm for feature selection. Int J Biomed Eng Technol. 2025; 47(2):134-164. doi:10.1504/IJBET.2025.144812
29. Teboul A. Diabetes Health Indicators Dataset. Kaggle, 2019. https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset
30. Centers for Disease Control and Prevention (CDC). Behavioral Risk Factor Surveillance System annual data 2015. Atlanta (GA): U.S. Department of Health and Human Services; 2015. https://www.cdc.gov/brfss/annual_data/annual_2015.html
31. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011; 12:2825-2830. http://jmlr.org/papers/v12/pedregosa11a.html
32. He H, Garcia EA. Learning from imbalanced data. IEEE Trans Knowl Data Eng. 2009;21(9):1263-1284. doi:10.1109/TKDE.2008.239
33. Lemaitre G, Nogueira F, Aridas CK. imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. J Mach Learn Res. 2017;18(17):1-5. http://jmlr.org/papers/v18/16-365.html

34. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321-357. doi:10.1613/jair.953

35. Buda M, Maki A, Mazurowski MA. A systematic study of the class imbalance problem in convolutional neural networks. Neural Netw. 2018;106:249-259. doi:10.1016/j.neunet.2018.07.011

How to Cite: Lavanya AL, Gatram RMB. Feature Correlation and Class Balancing Effects on Random Forest Accuracy in Diabetes Risk Prediction Using the Teboul Health Indicators Dataset. Int Res J Multidiscip Scope. 2026;7(3):658-675. DOI: 10.47857/irjms.2026.v07i03.011810