

RESEARCH

Open Access



Predicting diabetes in adults: identifying important features in unbalanced data over a 5-year cohort study using machine learning algorithm

Maryam Talebi Moghaddam^{1,2}, Yones Jahani³, Zahra Arefzadeh⁴, Azizallah Dehghan^{1,5}, Mohsen Khaleghi^{6*}, Mehdi Sharafi^{7*} and Ghasem Nikfar⁸

Abstract

Background Imbalanced datasets pose significant challenges in predictive modeling, leading to biased outcomes and reduced model reliability. This study addresses data imbalance in diabetes prediction using machine learning techniques. Utilizing data from the Fasa Adult Cohort Study (FACS) with a 5-year follow-up of 10,000 participants, we developed predictive models for Type 2 diabetes.

Methods We employed various data-level and algorithm-level interventions, including SMOTE, ADASYN, SMOTEENN, Random Over Sampling and KMeansSMOTE, paired with Random Forest, Gradient Boosting, Decision Tree and Multi-Layer Perceptron (MLP) classifier. We evaluated model performance using F1 score, AUC, and G-means—metrics chosen to provide a comprehensive assessment of model accuracy, discrimination ability, and overall balance in performance, particularly in the context of imbalanced datasets.

Results our study uncovered key factors influencing diabetes risk and evaluated the performance of various machine learning models. Feature importance analysis revealed that the most influential predictors of diabetes differ between males and females. For females, the most important factors are triglyceride (TG), basal metabolic rate (BMR), and total cholesterol (CHOL), whereas for males, the key predictors are body Mass Index (BMI), serum glutamate Oxaloacetate Transaminase (SGOT), and Gamma-Glutamyl (GGT). Across the entire dataset, BMI remains the most important variable, followed by SGOT, BMR, and energy intake. These insights suggest that gender-specific risk profiles should be considered in diabetes prevention and management strategies. In terms of model performance, our results show that ADASYN with MLP classifier achieved an F1 score of 82.17 ± 3.38 , AUC of 89.61 ± 2.09 , and G-means of 89.15 ± 2.31 . SMOTE with MLP followed closely with an F1 score of 79.85 ± 3.91 , AUC of 89.7 ± 2.54 , and G-means of 89.31 ± 2.78 . The SMOTEENN with Random Forest combination achieved an F1 score of 78.27 ± 1.54 , AUC of 87.18 ± 1.12 , and G-means of 86.47 ± 1.28 .

*Correspondence:
Mohsen Khaleghi
khaleghiir@gmail.com
Mehdi Sharafi
mehdisharafi_2002@yahoo.com

Full list of author information is available at the end of the article



© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

Conclusion These combinations effectively address class imbalance, improving the accuracy and reliability of diabetes predictions. The findings highlight the importance of using appropriate data-balancing techniques in medical data analysis.

Keywords Imbalanced datasets, Diabetes prediction, Machine learning, Artificial intelligence, Data-level method, Algorithm-level method

Introduction

Type 2 diabetes, is a chronic metabolic disorder characterized by insulin resistance or insufficient insulin production, that significantly contributes to the global burden of disease [1]. It is associated with severe complications, including heart disease, stroke, kidney failure, blindness, and lower-limb amputation. It has been linked to increased risks of dementia, hearing loss, and certain cancers, thereby heightening the risk of premature mortality [2, 3].

The incidence of diabetes is alarmingly on the rise. According to the International Diabetes Federation, the global diabetes population was 382 million in 2013, anticipated to surge to 592 million by 2035 [4]. Similarly, a study highlighted that the prevalence among adults was 6.4% in 2010, affecting 285 million adults, with projections indicating an increase to 7.7%, affecting 439 million adults by 2030. This escalating trend underscores the critical need for robust predictive tools to effectively manage and mitigate the disease's impact [5].

In response, machine learning (ML) techniques are increasingly leveraged to forecast the onset of diabetes and its complications [6]. These methods have shown considerable efficacy in enhancing risk prediction, prognosis, treatment, and management strategies [7, 8]. Popular ML models used in diabetes prediction include Random Forest, K-NN, neural networks, support vector machines, decision trees, and extra trees [7, 9].

A notable challenge in this domain is the prevalence of data imbalance in clinical datasets, which typically include variables like blood sugar and blood pressure [10]. Such imbalance can drastically affect the performance of predictive models, often resulting in biased outcomes and less reliable predictions. This is particularly critical in forecasting the incidence of type 2 diabetes, where the accuracy of predictions can significantly influence preventive and therapeutic measures [11].

To tackle these challenges, our study adopts both data-level and algorithm-level interventions [11–14]. We explore oversampling, under sampling, and hybrid sampling techniques to correct data imbalances [15]. Additionally, we utilize a range of ML algorithms, evaluating their effectiveness through metrics such as the F1 score, AUC, and G-means indices to identify the most proficient approaches in predicting the incidence of diabetes [16].

Related work

Recent studies have addressed the challenges of applying machine learning algorithms to imbalanced datasets, particularly in the prediction of diabetes. These efforts are marked by the development of various resampling methods and algorithmic adjustments to improve predictive performance.

A significant study in 2024 by O. Olawale Awe et al. investigated the use of various resampling algorithms—including random oversampling, SMOTE, ADASYN, random subsampling, Tomek linkages, NearMiss, and others—across four imbalanced datasets related to diabetes, anemia, lung cancer, and obesity. They found that the Repeated Nearest Neighbor Sampling method (RENN) combined with logistic regression achieved the most substantial improvement in predictions [17].

In 2023, Wahyu Nugraha et al. focused on the Pima Indians dataset, employing the SMOTE+Tomek link method along with a decision tree classification algorithm. Their experimental results showed that this combination performed better than using SMOTE without Tomek links [18].

Also in 2023, Hirani Hairani and Dadang Priyanto studied the same Pima Indian dataset using SVM and Random Forest with SMOTE-ENN. They concluded that the Random Forest method with SMOTE-ENN outperformed the SVM method [19].

Karmand et al. (2023) utilized machine learning models for predicting diabetes. In their study, the Gradient Boosting Model (GBM) was identified as the best-performing model compared to others [20].

In 2023, Nematollahi and colleagues explored the relationship between body composition and diabetes using artificial intelligence algorithms. Among the models tested, XGBoost was identified as the best-performing model [21].

M. Sandeep Kumar et al. (2022) used six algorithms including k-nearest Neighbor, Naive Bayes, Support Vector Machines, Random Forest, Logistic Regression, and Decision Trees on the PIMA Indian dataset. They tested both oversampling and undersampling methods. Their results suggested that SVM outperformed other models in dealing with imbalanced data [22].

In 2022, Somieh et al. utilized Deep Neural Network (DNN), Extreme Gradient Boosting (XGBoost), and Random Forest (RF) algorithms on the Tehran Lipid and Glucose Study (TLGS) cohort data, which was notably

imbalanced. Their findings highlighted that undersampling methods yielded superior results compared to other techniques in managing data skewness [23].

MATLOOB KHUSHI et al. (2021) conducted a study using lung cancer datasets, PLCO and NLST, both characterized by imbalance. They employed 23 resampling models alongside hybrid systems, using logistic regression, Random Forest, and LinearSVC to determine the most effective forecasting model. Their results indicated that under-sampling techniques generally exhibited higher standard deviations, while over-sampling resulted in lower variances. Random Forest was identified as having the best predictive ability for the lung cancer datasets used [24].

Masoud Mohammad Hassan et al. (2019) explored six different algorithms—logistic regression, decision tree, K-Nearest Neighbors, Naive Bayes, Support Vector Machine, and artificial neural networks—on the Pima dataset. They implemented the Synthetic Minority Over-sampling Technique (SMOTE) to address data imbalance. The study concluded that SVM benefited most significantly from this resampling method, demonstrating enhanced model performance [25].

These studies collectively illustrate a diverse range of strategies for addressing data imbalance in diabetes prediction, highlighting the effectiveness of different resampling techniques and machine learning algorithms.

Methods and materials

The workflow of our research process, including data collection, preprocessing, handling of data imbalance, classifier selection, and evaluation metrics, is summarized in Fig. 1. This figure provides a comprehensive overview of the methods and procedures employed in this study, guiding the reader through each step from data collection to performance analysis.

Data collection

Study population

This study utilizes data from the Fasa Adult Cohort Study (FACS), initiated in 2014 with baseline data collection completed in 2016. The follow-up phase was initiated in 2016 after the baseline data was collected. During this phase, interviewers monitor participants annually to assess various health outcomes. These outcomes include cardiovascular diseases, diabetes, cancers, fatty liver disease, and others. Notably, this study focuses on the 5-year cumulative incidence of type 2 diabetes. To date, a 5-year follow-up has been conducted for all participants, with an initial cohort size of 10,000 individuals. Data were gathered through detailed interviews conducted by trained personnel, using comprehensive questionnaires. These questionnaires covered a wide range of topics such as demographics, nutritional status, physical activity,

personal habits, and history of chronic or underlying diseases. Additionally, participants underwent blood, urine, and stool tests, along with anthropometric assessments. After applying data preparation and preprocessing, the dataset was reduced to 7408 records. More comprehensive details about the FACS methodology are available in the study's protocol and profile publications [26, 27].

Definition of variables

The primary outcome, or dependent variable, of this study is the 5-year cumulative incidence of Type 2 diabetes. Diagnosis is primarily determined using the A1C test as recommended by the American Diabetes Association, where a result of 6.5% or higher suggests diabetes. Alternatively, a 2-hour plasma glucose (2-h PG) value of 200 mg/dL or higher during an Oral Glucose Tolerance Test (OGTT) also indicates diabetes [28].

The independent variables include demographic factors such as gender, age, occupation, and education level. Health-related variables include the presence of cardiovascular diseases, smoking status, opium use, Basal Metabolic Rate (BMR), Energy and chronic conditions like kidney disorders, fatty liver, and lung disease. Anthropometric measurements taken into account are Body Mass Index (BMI), Waist Hip Ratio (WHR) and weight. Additionally, the study considers Medical equivalent task (MET) levels, socioeconomic status (Assert_index), and lipid profiles, including LDL, HDL, total cholesterol (CHOL), triglycerides (TG), Diastolic blood pressure (DBP), Systolic blood pressure (SBP), Blood Urea Nitrogen (BUN), Creatinine (CERT), Serum Glutamate Oxaloacetate Transaminase (SGOT), Serum Glutamate Pyruvate Transaminase (SGPT), Alkaline Phosphatase (ALP), Gamma-Glutamyl (GGT) and Pulse rate (PR) as potential predictors.

This robust methodological framework is designed to accurately capture the complex interplay of various factors contributing to the incidence of Type 2 diabetes, facilitating a comprehensive analysis of risks associated with the disease.

Data preparation and preprocessing

Data preprocessing is an essential phase in machine learning projects, setting the foundation for the effectiveness of the analysis. In this study, the preprocessing stage was meticulously designed to ensure the data's suitability for the applied machine learning models, focusing on handling missing values and data normalization.

This stage started with a total of 10,000 records. After applying our data cleaning procedures, which removed 1,249 participants who had diabetes at the beginning of the study, 630 participants with blood glucose levels above 120 and 713 records with missing values, the dataset was reduced to 7,408 records.

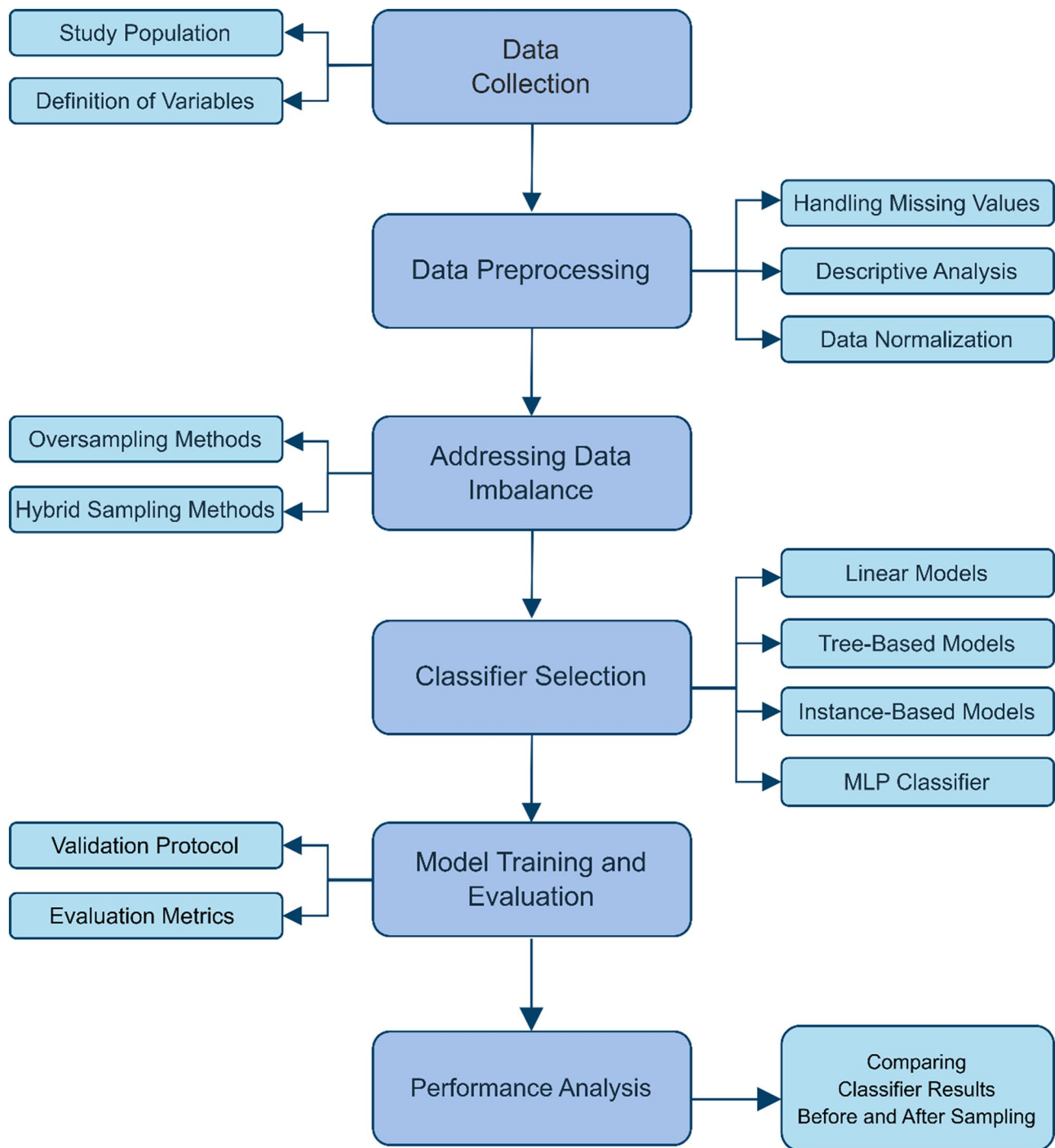


Fig. 1 The workflow algorithm of our research

Handling missing values

In dealing with missing values, our approach was to maintain the integrity and quality of the dataset by removing any rows that contained missing data. This decision was based on the premise that the presence of missing values could introduce bias or inaccuracies into our models' predictions. As mentioned above, 713 records were deleted in this process. Although this method resulted in

a reduction of the dataset size, it ensured that the remaining data was complete, thereby improving the reliability of our analysis. The direct removal of rows with missing data was deemed the most straightforward and effective strategy, considering the dataset's sufficient size and the distribution of missing values across the dataset.

Descriptive analysis

In total, 7408 people were included with a mean age of 46.55 ± 8.89 . 3,806 (%51.38) were male. The 5 years cumulative incidence of type 2 diabetes was (31.8, %95 CI: 27.9–36.1) in 1000 population. The characteristics of the study population are shown in Table 1.

Data normalization

Normalization was the primary step in our data preprocessing, specifically employing min-max scaling. This technique adjusts the values of numeric columns in the dataset to a common scale, between 0 and 1, without distorting differences in the ranges of values or losing information. Min-max scaling is mathematically represented as:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

where X is the original value, X_{min} and X_{max} are the minimum and maximum values for the feature, respectively, and X_{norm} is the normalized value.

This step is crucial, particularly in our context, where the dataset encompasses a wide range of physiological and clinical measures. Normalizing these values ensures that no single feature disproportionately influences the model due to its scale, facilitating a more balanced and effective analysis. Moreover, min-max scaling aids in accelerating the convergence of gradient descent algorithms by ensuring that the feature space is uniformly scaled [29].

Balancing data through sampling methods

Sampling methods are essential for addressing data imbalance, ensuring that classification models maintain high accuracy and sensitivity in diabetes detection. The primary goal of these techniques is to rebalance the dataset to allow equitable learning from both classes [30, 31]. In the sections that follow, we will delve into various sampling methods, such as oversampling, undersampling, and hybrid approaches. These strategies are critical for mitigating the challenges associated with imbalanced datasets in diabetes classification.

Oversampling methods

1. **Random Oversampling:** This method involves randomly duplicating examples from the minority class to balance the class distribution. Although straightforward, it may result in overfitting by potentially replicating noise in the data [32].
2. **SMOTE (Synthetic Minority Oversampling Technique):** Developed by Chawla et al. in SMOTE

creates synthetic samples through interpolation between multiple minority class samples, helping to avoid overfitting by expanding the decision region for the minority class [33].

3. **ADASYN (Adaptive Synthetic Sampling Approach):** This technique, introduced by He et al. ADASYN focuses on generating synthetic data for samples that are difficult to classify, enhancing model generalization [34].
4. **Borderline SMOTE:** This variant targets minority class samples near the decision boundary, aiming to improve classification on challenging cases by generating synthetic samples along the borderline [35].
5. **KMeans SMOTE:** By integrating K-Means clustering with SMOTE, this method clusters the minority class before applying SMOTE within each cluster to produce more contextually relevant synthetic samples [36].
6. **Smotified GAN (SMOTE and GAN):** This innovative approach combines Generative Adversarial Networks (GANs) with SMOTE to produce realistic synthetic samples, enhancing the diversity and authenticity of the data [37].

Hybrid sampling methods

Hybrid sampling strategies merge the benefits of both oversampling and undersampling to balance the dataset while minimizing information loss and reducing the risk of overfitting.

1. **SMOTEENN (SMOTE + Edited Nearest Neighbors):** This method pairs SMOTE with the undersampling technique Edited Nearest Neighbors (ENN) to refine synthetic samples by removing misclassified instances near their nearest neighbors [38].
2. **SMOTETomek:** Similar to SMOTEENN, SMOTETomek combines SMOTE with Tomek links to identify and eliminate close sample pairs from opposing classes, clarifying the decision boundary between classes [39].

Classifier selection post-sampling for diabetes prediction

With the dataset now balanced, we proceed to select suitable machine learning classifiers. As Table 2 illustrates, each algorithm has strengths and challenges that, when carefully paired with the newly equilibrated data, can yield a more accurate prediction model for diabetes. The Table 3 delineates the various classifiers considered for this study, post-sampling. It aids in understanding the potential impact of each classifier's advantages and

Table 1 Bassline characteristics of the study population based on gender

Quantitative variable	Subgroup	Mean SD	P-value
Age	Male	47.29 ± 9.12	< 0.001
	Female	45.87 ± 8.60	
DBP	Male	72.91 ± 10.97	< 0.001
	Female	72.10 ± 10.88	
SBP	Male	107.65 ± 15.18	< 0.001
	Female	106.28 ± 15.19	
PR	Male	70.73 ± 10.02	< 0.001
	Female	76.06 ± 10.38	
MET	Male	45.86 ± 14.42	< 0.001
	Female	38.68 ± 6.84	
Energy	Male	3054.77 ± 1173.33	< 0.001
	Female	2862.42 ± 1111.22	
TG	Male	132.82 ± 86.12	< 0.001
	Female	118.91 ± 66.80	
CHOL	Male	178.39 ± 36.45	< 0.001
	Female	187.96 ± 37.59	
HDL	Male	47.23 ± 14.23	< 0.001
	Female	54.32 ± 16.53	
LDL	Male	104.56 ± 30.67	< 0.001
	Female	109.83 ± 32.11	
WHR	Male	1.39 ± 0.64	< 0.001
	Female	2.84 ± 0.45	
BMR	Male	6628.84 ± 984.17	< 0.001
	Female	5418.65 ± 628.36	
BUN	Male	13.54 ± 3.72	< 0.001
	Female	11.89 ± 3.63	
CERAT	Male	0.98 ± 0.17	0.138
	Female	0.97 ± 0.16	
SGOT	Male	23.86 ± 8.33	< 0.001
	Female	21.30 ± 3.0	
SGPT	Male	25.82 ± 16.52	< 0.001
	Female	20.43 ± 11.75	
ALP	Male	209.60 ± 59.102	< 0.001
	Female	198.42 ± 75.96	
GGT	Male	25.04 ± 19.62	< 0.001
	Female	18.86 ± 20.86	
Education years	Male	6.052 ± 6.052	< 0.001
	Female	4.36 ± 3.47	
Socioeconomic score (Assert index)	Male	0.54 ± 2.41	< 0.001
	Female	0.42 ± 1.70	
Categorical Variable	Male N(%)	Female N(%)	P-value
Have a Diabetes	3,710(97.48)	3,462(96.11)	< 0.001
	No 96(2.52)	140(3.89)	
BMI	797(20.94)	292(8.10)	< 0.001
	1 1,588(41.72)	1,159(32.15)	
2	1,421(37.33)	2,151(59.68)	< 0.001
	3		
Smoking	2,117(55.62)	127(3.53)	< 0.001
	No 1,689(44.38)	3,475(96.47)	
Yes			< 0.001
Drug users	2,017(53.00)	3,583(99.47)	< 0.001
	No 1,789(47.00)	19(0.53)	
Yes			< 0.001

Table 1 (continued)

Quantitative variable	Subgroup	Mean SD	P-value
Marital status	88(2.31)	246(6.83)	< 0.001
1	3,701(97.24)	2,984 (82.84)	
2	5(0.13)	309(8.58)	
3	12(0.32)	63(1.75)	
4			
Use Alcohol	3,376(88.70)	3,601(99.97)	< 0.001
No	430(11.30)	1(0.03)	
Yes			

Table 2 Illustrates, each algorithm has strengths and challenges

Category	Algorithm	Advantages	Disadvantages
Linear Models	Logistic Regression [40]	Simple to implement and interpret. Efficient to train. Good for binary classification.	Assumes linear relationship between variables. Not suitable for complex relationships.
	Support Vector Machine (SVM) [42]	Effective in high dimensional spaces. Memory efficient. Versatile with kernel functions.	Requires careful parameter tuning. Not suitable for large datasets.
	SGD Classifier [47]	Efficient for large-scale problems. Easy to implement and provides a lot of opportunities for code tuning.	Sensitive to feature scaling. Requires a number of hyperparameters
Tree-Based Models	Decision Tree Classifier [43]	Easy to interpret and visualize. Can handle both numerical and categorical data.	Prone to overfitting. Can become unstable with small variations in data.
	Random Forest Classifier [44]	Handles overfitting well. Works well on large datasets. Provides feature importances.	Can be slow to predict. Complex and difficult to interpret.
	AdaBoost Classifier [45]	Improves classification accuracy. Flexible to combine with any learning algorithm.	Sensitive to noisy data and outliers. Can overfit on very complex datasets.
	Gradient Boosting [46]	Highly effective and flexible. Can optimize on different loss functions.	Prone to overfitting without proper tuning. Time-consuming to train.
Instance-Based Models	K-Nearest Neighbors (KNN) [41]	No assumption about data. Simple and effective. Adaptable to any type of data.	Computationally expensive. Performance depends on the number of dimensions.
Probabilistic Models	GaussianNB [48]	Works well with high-dimensional data. Simple and fast.	Assumes that features are independent. Performance can be affected if the independence assumption is not met.
Neural Network Model	MLP Classifier [49, 50]	Capable of modeling complex non-linear relationships and works well with large datasets.	Requires significant computational resources and can be prone to overfitting without proper regularization.

disadvantages within the balanced data context, guiding our selection process towards the most effective algorithm for predicting the incidence of diabetes.

Model training and evaluation metrics

Validation protocol

In this model training, we use an 80–20 train-test split within each fold, where 80% of the data is used for training and 20% is reserved for testing. Additionally, a 5-fold stratified cross-validation approach was utilized, which was repeated 5 times with different shuffles of the data. This method ensures that the class distribution is preserved in each fold, addressing the imbalanced nature of the dataset. Ultimately the average and standard deviation across these 25 runs (5 folds × 5 repetitions) is reported.

Specially, for the MLP classifier, the hyper parameters are configured as follows:

- **Activation function:** Sigmoid.

- **Optimizer:** Adam.
- **Learning rate:** 0.0005.
- **Loss function:** Binary Crossentropy.
- **Epochs:** 150.
- **Batch size:** 64.
- **Validation split:** 20% of the training data.

Evaluation metrics

In predictive modeling for conditions like diabetes where dataset imbalance is prevalent, reliance on standard accuracy metrics can be misleading. Therefore, we employ three alternative metrics: the F1 score, AUC, and G-means. Each metric provides a distinct perspective on model performance, addressing the issues inherent in imbalanced datasets [51]. The **F1 score**,

$$F1\ score = 2 \frac{precision * recall}{precision + recall}$$

Table 3 The optimal classifier sampling combination results

Machine learning models	Resampling techniques	Step	F1 (Mean ± SD)	AUC (Mean ± SD)	G-means (Mean ± SD)
Random Forest Classifier	SMOTEENN	before	70.77 ± 3.57	77.44 ± 2.15	74.03 ± 2.89
		after	78.27 ± 1.54	87.18 ± 1.12	86.47 ± 1.28
MLP Classifier	SMOTEENN	before	42.22 ± 5.39	64.63 ± 2.24	54.88 ± 3.94
		after	71.33 ± 1.99	88.76 ± 1.17	88.52 ± 1.28
Gradient Boosting	SMOTE	before	65.77 ± 1.77	74.65 ± 1.01	70.23 ± 1.46
		after	71.63 ± 2.42	85.47 ± 1.66	84.71 ± 1.88
Random Forest Classifier	SMOTE	before	70.77 ± 3.57	77.44 ± 2.15	74.03 ± 2.89
		after	82.18 ± 2.76	85.97 ± 1.84	84.85 ± 2.14
Decision Tree Classifier	SMOTE	before	62.01 ± 2.95	80.69 ± 1.71	79.33 ± 2.01
		after	63.88 ± 2.84	83.59 ± 2.11	82.84 ± 2.38
MLP Classifier	SMOTE	before	40.60 ± 5.43	63.8 ± 2.3	53.28 ± 4.13
		after	79.85 ± 3.91	89.7 ± 2.54	89.31 ± 2.78
Random Forest Classifier	Random Over Sampling	before	70.77 ± 3.57	77.44 ± 2.15	74.03 ± 2.89
		after	78.56 ± 3.60	82.72 ± 2.51	80.85 ± 3.11
MLP Classifier	Random Over Sampling	before	42.25 ± 2.41	64.64 ± 1.17	54.99 ± 2.14
		after	82.97 ± 2.46	89.25 ± 1.57	88.73 ± 1.75
Gradient Boosting	ADASYN	before	65.77 ± 1.77	74.65 ± 1.01	70.23 ± 1.46
		after	68.23 ± 0.98	85.17 ± 1.29	84.52 ± 1.49
Random Forest Classifier	ADASYN	before	70.77 ± 3.6	77.44 ± 2.15	74.03 ± 2.89
		after	81.24 ± 3.47	86.02 ± 2.42	84.93 ± 2.84
MLP Classifier	ADASYN	before	40.24 ± 4.98	63.52 ± 2.03	52.7 ± 3.6
		after	82.17 ± 3.38	89.61 ± 2.09	89.15 ± 2.31
Gradient Boosting	KMeansSMOTE	before	65.77 ± 1.77	74.65 ± 1.01	70.23 ± 1.46
		after	69.08 ± 4.15	77.04 ± 2.66	73.53 ± 3.52
Random Forest Classifier	KMeansSMOTE	before	70.77 ± 3.57	77.44 ± 2.15	74.03 ± 2.89
		after	74.66 ± 4.36	79.92 ± 2.82	77.28 ± 3.59
MLP Classifier	KMeansSMOTE	before	38.66 ± 4.6	62.84 ± 1.97	51.32 ± 4.03
		after	78.33 ± 6.98	88.25 ± 2.25	87.73 ± 2.42

reflects the model's balance between precision (the proportion of true positives out of all positive predictions) and recall (the proportion of true positives out of actual positive cases). It's an important metric in medical predictions due to the high cost of false negatives and positives. **AUC** represents a model's ability to differentiate classes, ranging from 0.5 (no better than random) to 1 (perfect classification). It is summarized from the ROC curve, which plots sensitivity (true positive rate) against 1-specificity (false positive rate). AUC is favored in imbalanced datasets as it is not influenced by the skew in class distribution. The **G-means** metric

$$G - means = \sqrt{sensitivity * specificity}$$

effectively capturing a model's performance on both minority and majority classes. It ensures that the model is not overly biased toward the predominant class, with higher values indicating a balanced classification performance.

Performance analysis

Interpretation of correlation matrices

The correlation matrices provided offer an intricate look at the relationships between various features for diabetes prediction across three different groups: female, male, and the entire dataset. Below is an analysis of how these relationships manifest and what they reveal about diabetes prediction.

Figure 2 and Figure 3 show the correlation matrix between covariates and the incidence of diabetes in female and male are the same. A high correlation was found between diastolic blood pressure (DBP) and systolic blood pressure (SBP), BMI and BMR, SGPT and SGOT, LDL and CHOL.

Figure 4 demonstrates the correlation matrix between covariates for the prediction of diabetes. The result shows that a high correlation was between Gender and WHR, Smoking and Gender, CHOL and LDL, SPGT and SGOT, diastolic blood pressure (DBP) and systolic blood pressure (SBP).

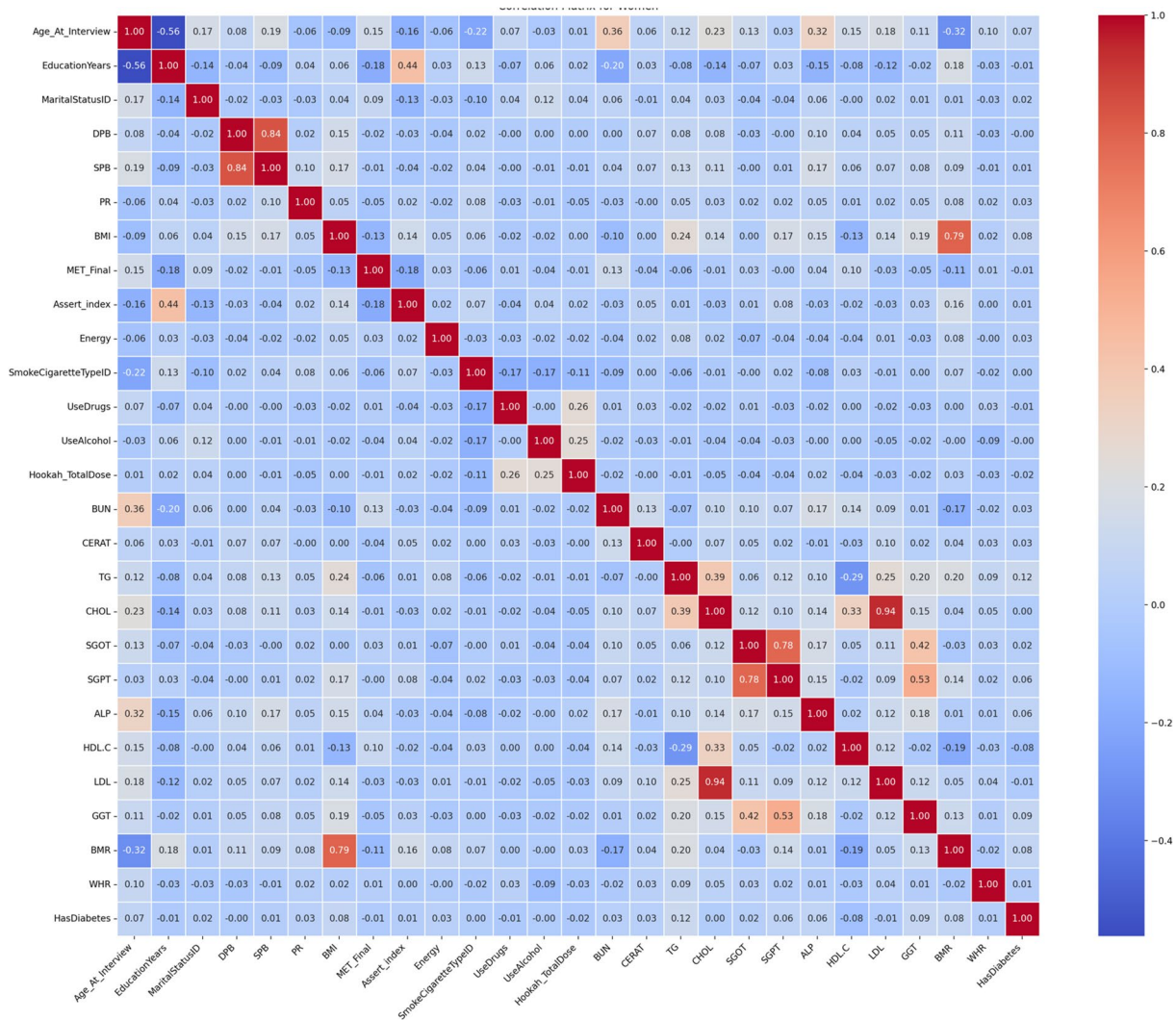


Fig. 2 Correlation matrix for female

Interpretation of feature importance analysis

The feature importance values presented in this section were derived using a Random Forest classifier, a method known for its robustness and ability to capture complex feature interactions [52].

The process began with training the Random Forest model on the dataset. The feature importance, which commonly known as Gini importance, provides an estimate of each feature's contribution to the model's predictive accuracy. To clearly convey the significance of each feature, the importance scores were visualized in Figs. 5, 6 and 7.

To illustrate, these figures provide critical insights into the most influential factors affecting the appearance of diabetes among the different groups analyzed. Below is an interpretation of the feature importance results for male, female, and the entire data.

Figure 5 shows all important features for prediction diabetes in female. This highlights that female with high

level of the Triglyceride (TG), BMR, cholesterol (CHOL), Energy intake, HDL, BMI, GGT, LDL, SGPT, SGOT, PR, MET, ALP, assert index, BUN, Age and Education are the primary drivers of diabetes risk.

Figure 6 indicates the important features that predict diabetes in male. According to the results, the most importance variable is BMI as follow: SGOT, GGT, ALP, socioeconomic status, TG, CHOL, BMR, LDL, HDL-C, SGPT, MET, Energy, Age, BUN and others that show in this figure.

In entire data, the results show that BMI is the most important variable for the prediction of diabetes. Other variables respectively include SGOT, BMR, Energy, GGT, TG, LDL, CHOL, APL, HDL, socioeconomic status, SPGT and Age. Other variables are shown in Fig. 7.

These results highlight that important variables differ in male and female and total population. These insights can help guide targeted prevention and management

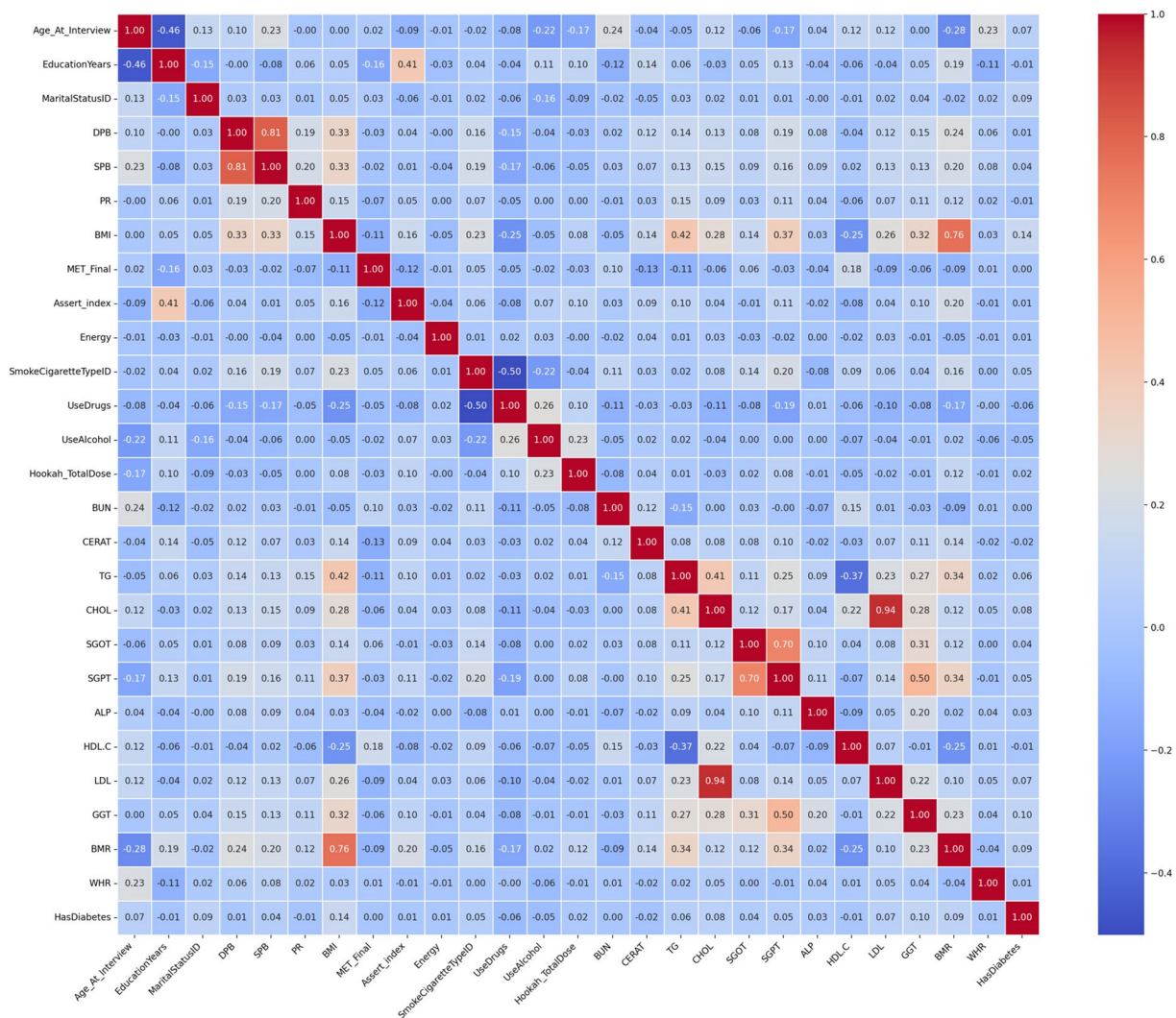


Fig. 3 Correlation matrix for males

strategies for diabetes based on gender-specific risk profiles.

Selection of optimal classifier-sampling combinations

All possible combinations of the introduced resampling methods and classifiers, in Table 2, were implemented on the study data. The acceptable outcomes show in Table 3.

The analysis reveals that resampling techniques, particularly RandomOverSampling, SMOTE, ADASYN, and KMeansSMOTE, significantly enhance the performance of machine learning models in handling class imbalance. The Multi-Layer Perceptron (MLP) consistently demonstrates superior performance across various resampling methods, indicating its robustness and adaptability in complex data scenarios.

Impact of resampling techniques

Random over sampling This technique, particularly when paired with MLP, achieves the highest F1 score of

82.97 ± 2.46 , along with a strong AUC of 89.25 ± 1.57 and G-Mean of 88.73 ± 1.75 . The exceptional performance of Random Over Sampling with MLP suggests that simply increasing the representation of minority class examples can significantly improve model training, especially in neural network-based models. This highlights the potential of Random Over Sampling in scenarios where the model architecture can effectively leverage the additional data without overfitting.

SMOTE and ADASYN Both SMOTE and ADASYN show strong performance improvements, particularly with MLP and Random Forest. For instance, SMOTE with MLP results in an F1 score of 79.85 ± 3.91 , AUC of 89.7 ± 2.54 , and G-Mean of 89.31 ± 2.78 . ADASYN with MLP achieves the highest F1 score among the ADASYN combinations, with an F1 score of 82.17 ± 3.38 , AUC of 89.61 ± 2.09 , and G-Mean of 89.15 ± 2.31 . These techniques not only address the imbalance by generating syn-

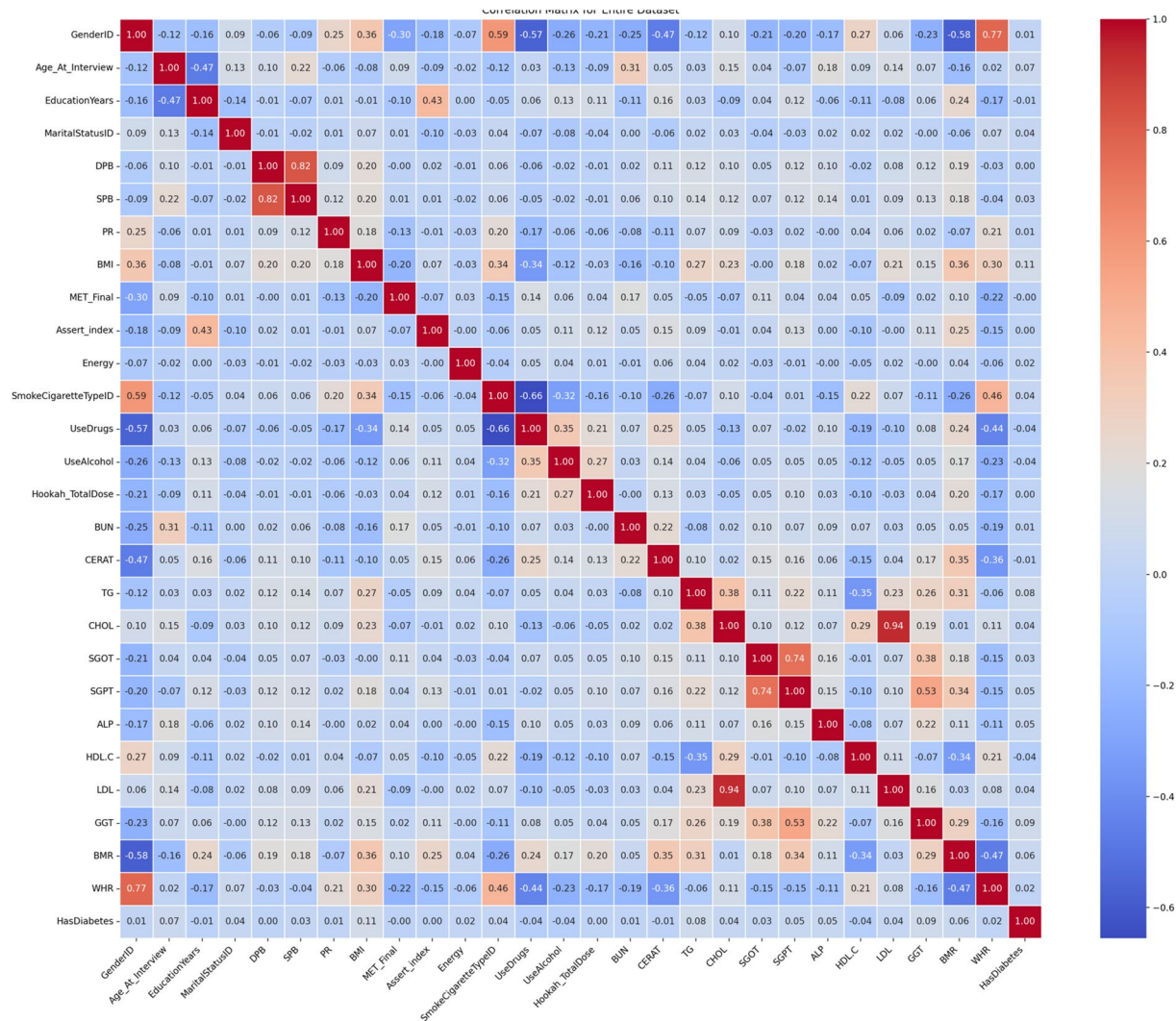


Fig. 4 Correlation matrix between covariates for the prediction of diabetes in entire data

thetic examples but also enhance the model's ability to generalize, as evidenced by the high AUC and G-means scores. The consistent performance gains suggest that the synthetic data generated by these methods provides meaningful and varied examples that help the model better understand the decision boundaries.

KMeansSMOTE The robust results achieved with KMeansSMOTE, especially with MLP, emphasize the importance of intelligently generating synthetic samples. KMeansSMOTE with MLP achieves an F1 score of 78.33 ± 6.98 , AUC of 88.25 ± 2.25 , and G-Mean of 87.73 ± 2.42 . By clustering the data before generating synthetic samples, KMeansSMOTE ensures that the new data points are more representative of the underlying distribution, thereby enhancing model performance in terms of both F1 score and AUC.

Model-specific insights

Multi-Layer Perceptron (MLP) MLP stands out as the most effective model across various resampling techniques, consistently achieving high F1 scores, AUC, and G-means. These results are illustrated in Fig. 8. For example, MLP with Random Over Sampling achieves an F1 score of 82.97 ± 2.46 , AUC of 89.25 ± 1.57 , and G-Mean of 88.73 ± 1.75 . This indicates that neural networks, with their capacity to model complex relationships, benefit significantly from balanced datasets. The adaptability of MLP to various resampling methods underscores its potential as a versatile tool in predictive modeling for imbalanced data.

Random forest model Figure 9 indicates that the Random Forest model also demonstrates substantial improvements with resampling techniques, particularly with Random Over Sampling and ADASYN. For

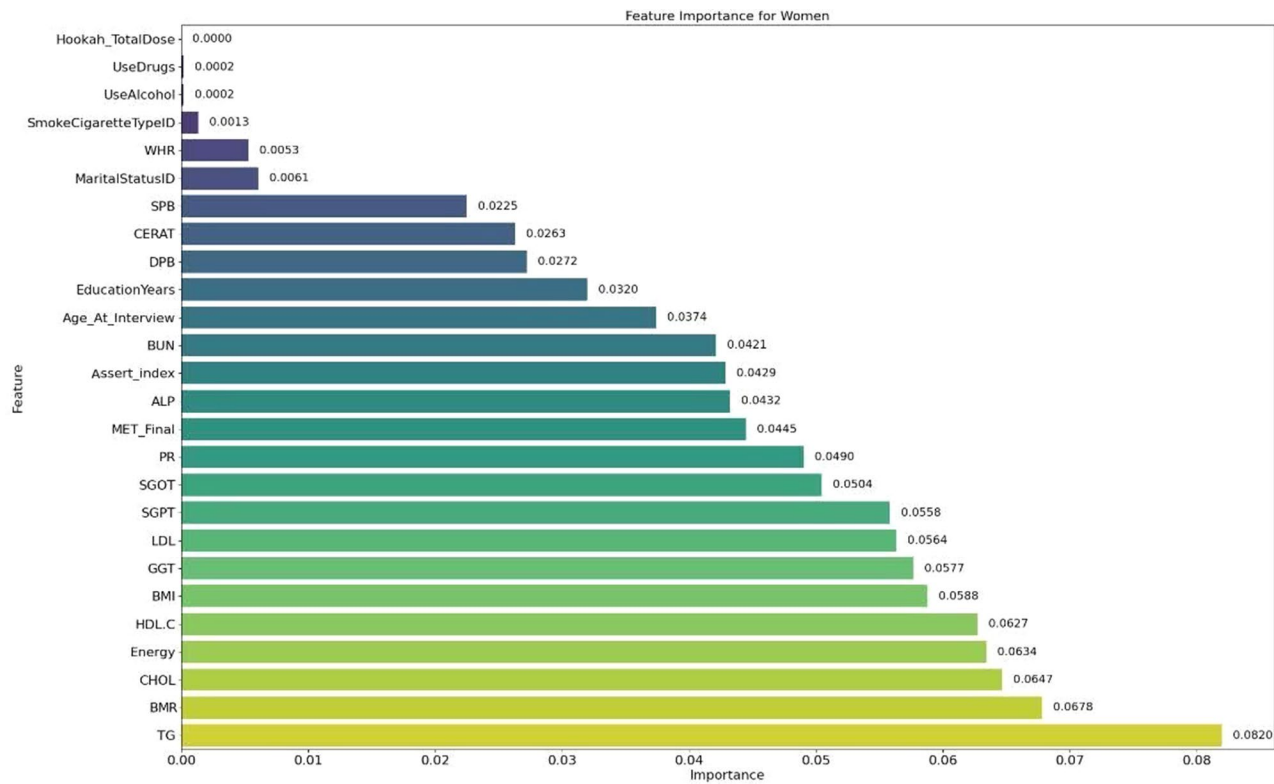


Fig. 5 Shows all important features of female

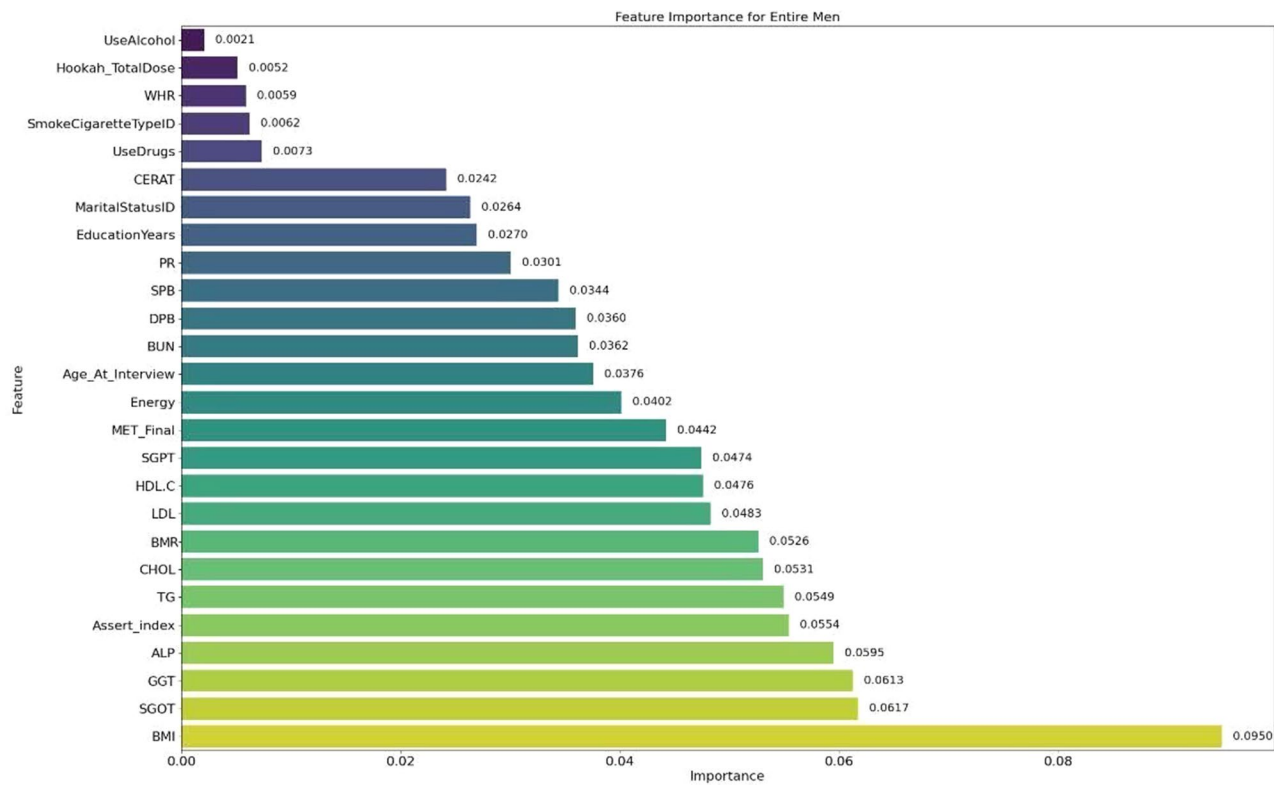


Fig. 6 Shows all important features of male

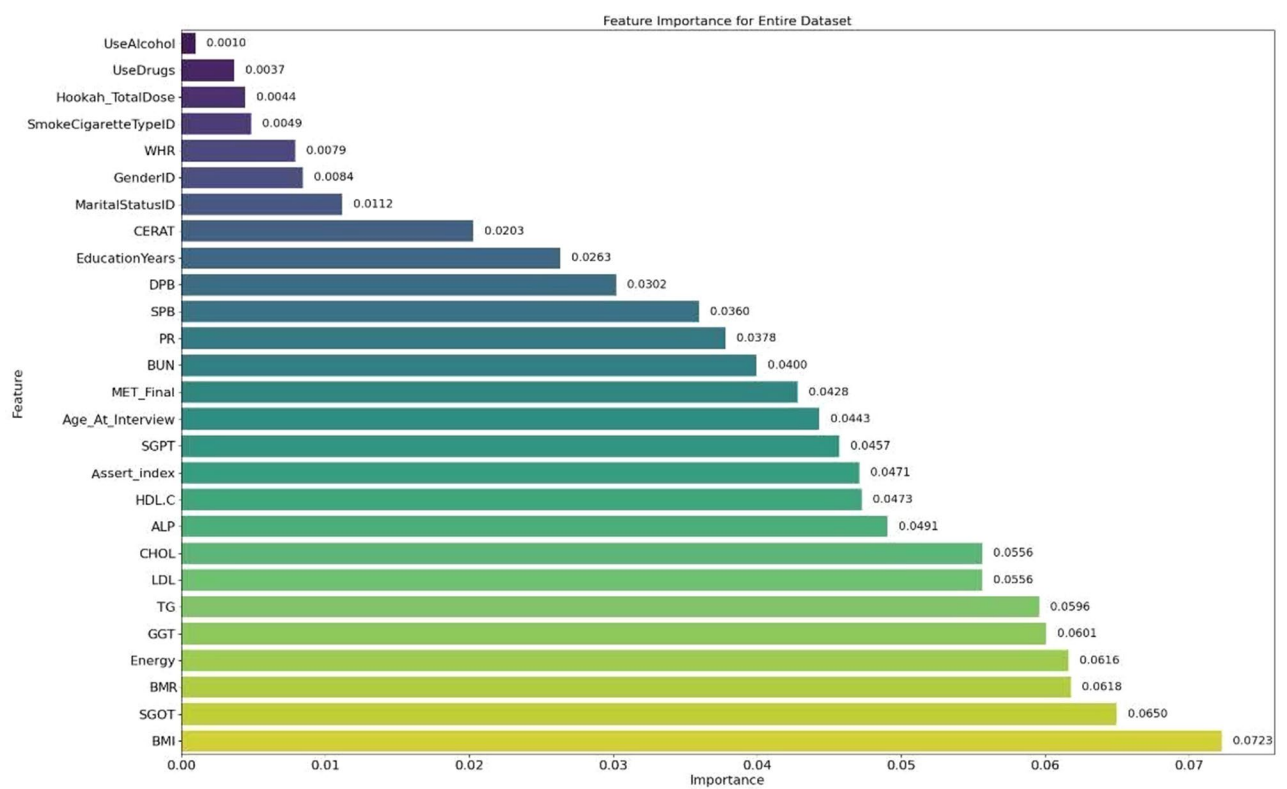


Fig. 7 Shows all important features of entire data

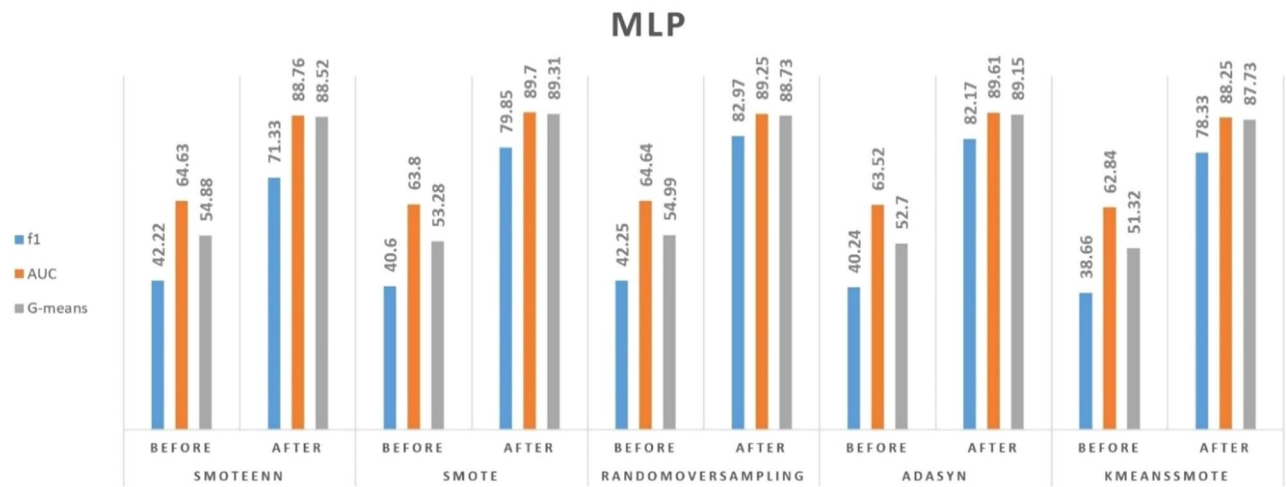


Fig. 8 MLP model results before and after data resampling technique

instance, Random Forest with ADASYN achieves an F1 score of 81.24 ± 3.47 , AUC of 86.02 ± 2.42 , and G-Mean of 84.93 ± 2.84 . The inherent ability of Random Forest to handle variability and reduce overfitting makes it well-suited to benefit from the additional or synthetic samples provided by resampling methods.

Broader implications
Improving predictive reliability The substantial improvements in F1 score, AUC, and G-means across most resampling techniques underscore the critical role of data balancing in predictive modeling. By addressing class imbalance, these methods not only improve the accuracy of predictions but also enhance the reliability and robustness of the models. This is particularly impor-



Fig. 9 RandomForest model results before and after data resampling technique

tant in medical applications where predictive accuracy can directly impact patient outcomes.

Algorithm and technique selection The findings suggest a strategic approach to selecting resampling techniques and machine learning models based on the specific characteristics of the dataset and the desired outcomes. For example, in scenarios where neural networks are preferred, Random Over Sampling or SMOTE with MLP might be the optimal choice. Conversely, for tree-based models, ADASYN might offer significant performance benefits.

To evaluate the performance of the combination of various resampling techniques Receiver Operating Characteristic (ROC) curves are illustrated in Fig. 10.

In Fig. 11, loss trends for different sampling methods before and after balancing the dataset used the MLP classifier trained with the hyper parameters discussed in Sect. 3.5.1. To clarify, the training and validation loss curves steadily decline and converge towards low loss values, indicating that each approach helped reduce data imbalance and enhanced the classifier's ability to learn from and generalize to the validation set. Minor spikes represent the natural fluctuations of complex learning processes. Overall five sampling methods improved the model's predictive accuracy and generalization, as reflected by the consistent loss trends.

Conclusion

This study explored the predictive power of machine learning models combined with advanced data balancing techniques to forecast diabetes incidence in an adult cohort over a 5-year period. Resampling methods like SMOTE, ADASYN, Random Over Sampling, and KMeansSMOTE effectively improved model performance, addressing the challenge of data imbalance.

Post-sampling, most models showed enhanced predictive accuracy, particularly in F1 scores and AUC measures. Random Over Sampling with MLP and ADASYN with MLP were identified as the most effective pairings, achieving significant gains in AUC, F1, and G-means scores. Additionally, the Random Over Sampling with Random Forest combination effectively addressed class imbalance, demonstrating notable improvements in predictive performance.

These findings underscore the importance of balancing techniques in medical data analysis, providing a clear pathway to develop more reliable predictive models. Future research will focus on feature selection methods, particularly leveraging autoencoders for dimensionality reduction and feature extraction. Finally, refining algorithm-level approaches for handling imbalanced data will include integrating ensemble learning with specialized cost-sensitive classifiers that prioritize the minority class. Techniques such as hybrid ensemble methods that combine boosting and bagging, or innovative architectures like one-class neural networks, could be explored for better detection of diabetic cases. Furthermore, incorporating reinforcement learning for adaptive resampling strategies may provide a dynamic approach to data balancing. Also, these findings can help guide targeted prevention and management strategies for prevention and control diabetes based on gender-specific risk profiles.

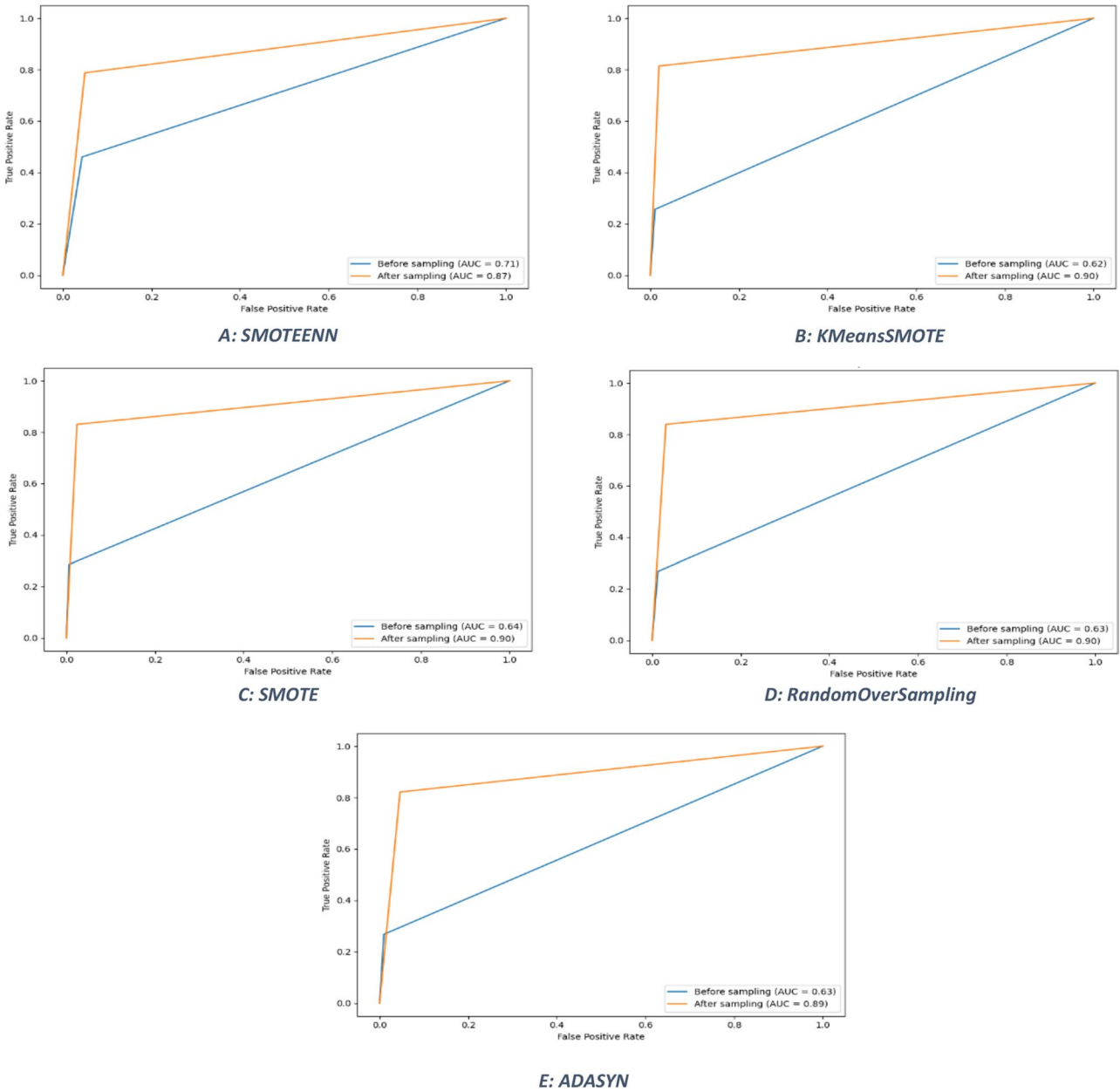


Fig. 10 ROC curves for different sampling methods used with the MLP classifier

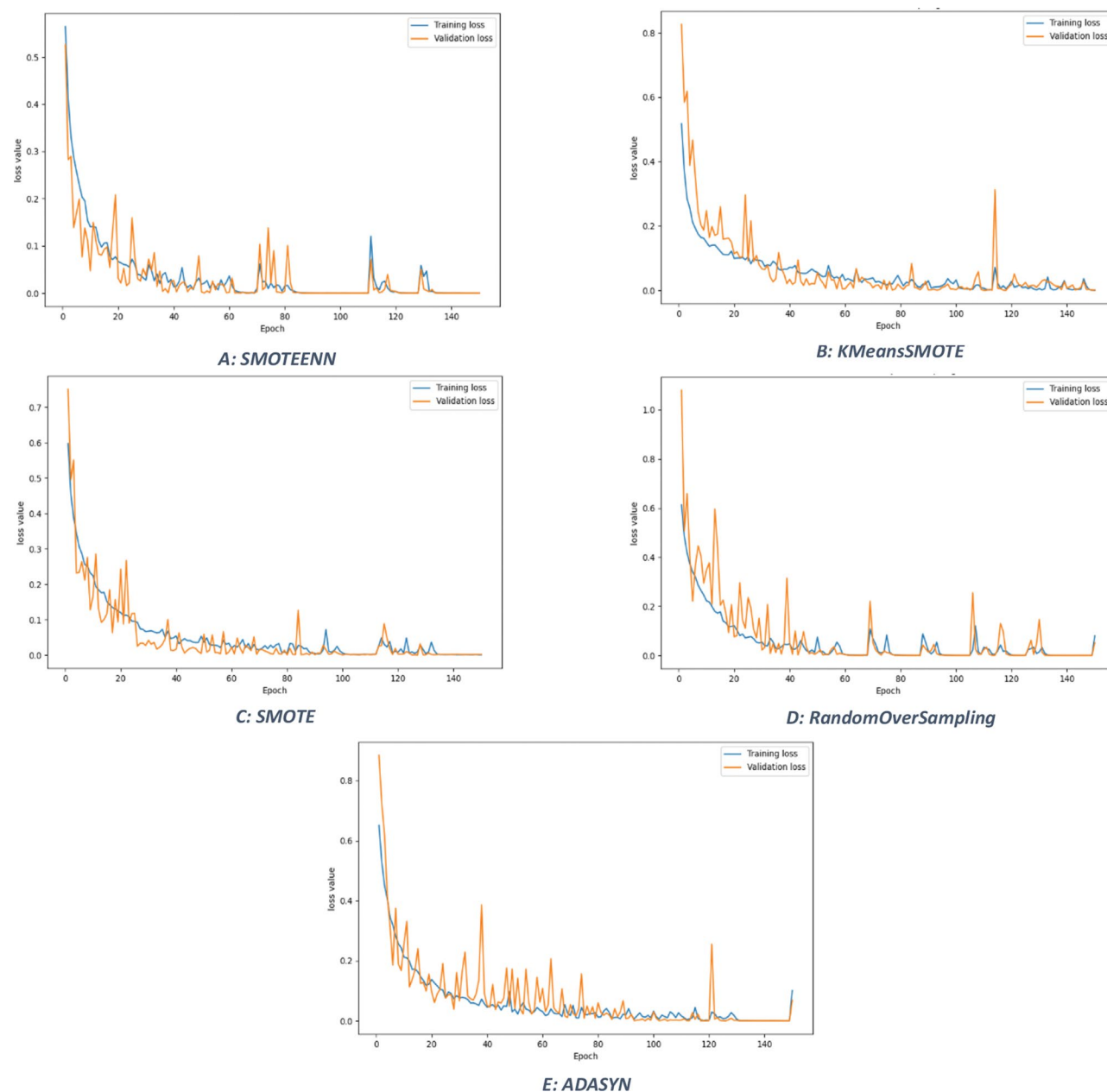


Fig. 11 Loss function for different sampling methods used with the MLP classifier

Acknowledgements

We would also like to thank Fasa University of Medical Sciences for supporting this research.

Author contributions

MT, ZA and YJH: providing the main idea of study and methodology, final analysis, developing the idea and revising the final manuscript, MKH and MSH : developing the idea and revising the final manuscript, contributed to data analysis and revising the final manuscript. ADH and GHN revised the final manuscript. All authors approved the final version of the manuscript that is submitted.

Funding

Fasa University of Medical Sciences.

Data availability

No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate

Ethical issues including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, redundancy, etc. were completely observed by the authors. This study was performed according to the ethical guidelines expressed in the Declaration of Helsinki and the Strengthening of the Reporting of Observational Studies in Epidemiology (STORB) guideline. The study was also approved by the Research Ethics Committee of Fasa University of Medical Sciences (IR.FUMS.REC.1402.172). Informed consent was also waived by the Research Ethics Committee of Fasa University of Medical Sciences (IR.FUMS.REC.1402.172).

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

¹Noncommunicable Diseases Research Center, Fasa University of Medical Sciences, Fasa, Iran

²Student of Biostatistics, Department of Biostatistics and Epidemiology, School of Public Health, Kerman University of Medical Sciences, Kerman, Iran

³Modeling in Health Research Center Institute for Futures Studies in Health, Kerman University of Medical Sciences, Kerman, Iran

⁴Faculty of Data Science and Intelligent Systems, Persian Gulf University, Bushehr, Iran

⁵Department of Epidemiology and Biostatistics, School of Health, Fasa University of Medical Sciences, Fasa, Iran

⁶Department of Mathematics and Computer Science, Fasa Branch, Islamic Azad University, Fasa, Iran

⁷Endocrinology and Metabolism Research Center, Hormozgan University of Medical Sciences, Bandar, Abbas, Iran

⁸Research Development Unit Valiasr Hospital, Fasa University of Medical Sciences, Fasa, Iran

Received: 20 July 2024 / Accepted: 16 September 2024

Published online: 27 September 2024

References

1. Hameed I, Masoodi SR, Mir SA, Nabi M, Ghazanfar K, Ganai BA. Type 2 diabetes mellitus: from a metabolic disorder to an inflammatory condition. *World J Diabetes*. 2015;6(4):598.
2. Kaze AD, Jaar BG, Fonarow GC, Echouffo-Tcheugui JB. Diabetic kidney disease and risk of incident stroke among adults with type 2 diabetes. *BMC Med*. 2022;20(1):127.
3. Sattar N, Presslie C, Rutter MK, McGuire DK. Cardiovascular and kidney risks in individuals with type 2 diabetes: contemporary understanding with Greater emphasis on excess adiposity. *Diabetes Care*. 2024;doi:10.230041.
4. Saeedi P, Petersohn I, Salpea P, Malanda B, Karuranga S, Unwin N, et al. Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: results from the International Diabetes Federation Diabetes Atlas. *Diabetes Res Clin Pract*. 2019;157:107843.
5. Safiri S, Karamzad N, Kaufman JS, Bell AW, Nejadghaderi SA, Sullman MJ, et al. Prevalence, deaths and disability-adjusted-life-years (DALYs) due to type 2 diabetes and its attributable risk factors in 204 countries and territories, 1990–2019: results from the global burden of disease study 2019. *Front Endocrinol*. 2022;13:838027.
6. Dagliati A, Marini S, Sacchi L, Cogni G, Teliti M, Tibollo V, et al. Machine learning methods to predict diabetes complications. *J Diabetes Sci Technol*. 2018;12(2):295–302.
7. Alghamdi T. Prediction of diabetes complications using computational intelligence techniques. *Appl Sci*. 2023;13(5):3030.
8. Dutta A, Hasan MK, Ahmad M, Awal MA, Islam MA, Masud M, et al. Early prediction of diabetes using an ensemble of machine learning models. *Int J Environ Res Public Health*. 2022;19(19):12378.
9. Shin J, Kim J, Lee C, Yoon JY, Kim S, Song S, et al. Development of various diabetes prediction models using machine learning techniques. *Diabetes Metabolism J*. 2022;46(4):650.
10. Lyra S, Leonhardt S, Antink CH, editors. Early prediction of sepsis using random forest classification for imbalanced clinical data. *IEEE; 2019*. 2019 Computing in Cardiology (CinC).
11. He H, Garcia EA. Learning from imbalanced data. *IEEE Trans Knowl Data Eng*. 2009;21(9):1263–84.
12. López V, Fernández A, Moreno-Torres JG, Herrera F. Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. Open problems on intrinsic data characteristics. *Expert Syst Appl*. 2012;39(7):6585–608.
13. Kumar M, Sheshadri H. On the classification of imbalanced datasets. *Int J Comput Appl*. 2012;44(8):1–7.
14. Sun Y, Wong AK, Kamel MS. Classification of imbalanced data: a review. *Int J Pattern Recognit Artif Intell*. 2009;23(04):687–719.
15. Chawla NV, Japkowicz N, Kotcz A. Special issue on learning from imbalanced data sets. *ACM SIGKDD Explorations Newsl*. 2004;6(1):1–6.
16. Liu Q, Zhang M, He Y, Zhang L, Zou J, Yan Y, et al. Predicting the risk of incident type 2 diabetes mellitus in Chinese elderly using machine learning techniques. *J Personalized Med*. 2022;12(6):905.
17. Awe OO, Ojumu JB, Ayanwoye GA, Ojumoola JS, Dias R. Machine Learning Approaches for Handling Imbalances in Health Data Classification. *Sustainable Statistical and Data Science Methods and Practices: Reports from LISA 2020 Global Network*, Ghana, 2022: Springer; 2024. pp. 375–91.
18. Nugraha W, Maulana R, Latifah L, Rahayuningsih PA, Nuralasari N, editors. Over-sampling strategies with data cleaning for handling imbalanced problems for diabetes prediction. *AIP Conference Proceedings*; 2023: AIP Publishing.
19. Hairani Hairani H, Dadang Priyanto D. A New Approach of Hybrid Sampling SMOTE and ENN to the Accuracy of Machine Learning Methods on Unbalanced. *Diabetes Disease Data*. 2023;14(8):585–890. A new approach of hybrid sampling SMOTE and ENN to the accuracy of machine learning methods on unbalanced diabetes disease data.
20. Karmand H, Andishgar A, Tabrizi R, Sadeghi A, Pezesghi B, Ravankhah M, et al. Machine-learning algorithms in screening for type 2 diabetes mellitus: data from Fasa adults Cohort Study. *Endocrinol Diabetes Metabolism*. 2024;7(2):e00472.
21. Nematollahi MA, Askarinejad A, Asadollahi A, Bazrafshan M, Sarejloo S, Moghadami M, et al. A cohort study on the predictive capability of body composition for Diabetes Mellitus using machine learning. *J Diabetes Metabolic Disorders*. 2024;23(1):773–81.
22. Kumar MS, Khan MZ, Rajendran S, Noor A, Dass AS, Prabhu J. Imbalanced classification in diabetics using ensembled machine learning. *Computers Mater Continua*. 2022;72(3):4397–409.
23. Sadeghi S, Khalili D, Ramezankhani A, Mansournia MA, Parsaeian M. Diabetes mellitus risk prediction in the presence of class imbalance using flexible machine learning methods. *BMC Med Inf Decis Mak*. 2022;22(1):36.
24. Khushi M, Shaukat K, Alam TM, Hameed IA, Uddin S, Luo S, et al. A comparative performance analysis of data resampling methods on imbalance medical data. *IEEE Access*. 2021;9:109960–75.
25. Hassan MM, Amiri N. Classification of imbalanced data of diabetes disease using machine learning algorithms. *Age (Years)*. 2019;21(81):3324.
26. Homayounfar R, Farjam M, Bahramali E, Sharafi M, Poustchi H, Malekzadeh R, et al. Cohort Profile: the Fasa adults Cohort Study (FACS): a prospective study of non-communicable diseases risks. *Int J Epidemiol*. 2023;52(3):e172–8.
27. Farjam M, Bahrami H, Bahramali E, Jamshidi J, Askari A, Zakeri H, et al. A cohort study protocol to analyze the predisposing factors to common chronic non-communicable diseases in rural areas: Fasa Cohort Study. *BMC Public Health*. 2016;16:1–8.
28. Ahuja V, Aronen P, Pramodkumar TA, Looker H, Chetrit A, Bloigu AH, et al. Accuracy of 1-Hour plasma glucose during the oral glucose tolerance test in diagnosis of type 2 diabetes in adults: a Meta-analysis. *Diabetes Care*. 2021;44(4):1062–9.
29. Shantal M, Othman Z, Bakar AA. A Novel Approach for Data feature weighting using correlation coefficients and Min–Max Normalization. *Symmetry*. 2023;15(12):2185.
30. Chowdhury MM, Ayon RS, Hossain MS. An investigation of machine learning algorithms and data augmentation techniques for diabetes diagnosis using class imbalanced BRFSS dataset. *Healthc Analytics*. 2024;5:100297.
31. Kovács G. An empirical comparison and evaluation of minority oversampling techniques on a large number of imbalanced datasets. *Appl Soft Comput*. 2019;83:105662.
32. Yang C, Fridgeirsson EA, Kors JA, Reps JM, Rijnbeek PR. Impact of random oversampling and random undersampling on the performance of prediction models developed using observational health data. *J Big Data*. 2024;11(1):7.
33. Ramezankhani A, Pournik O, Shahrabi J, Azizi F, Hadaegh F, Khalili D. The impact of oversampling with SMOTE on the performance of 3 classifiers in prediction of type 2 diabetes. *Med Decis Making*. 2016;36(1):137–44.
34. He H, Bai Y, Garcia EA, Li S, editors. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence); 2008: IEEE.
35. Mohanty MN. *Advances in intelligent computing and communication*. Springer; 2021.
36. Douzas G, Bacao F, Last F. Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. *Inf Sci*. 2018;465:1–20.

37. Sharma A, Singh PK, Chandra R. SMOTified-GAN for class imbalanced pattern classification problems. *IEEE Access*. 2022;10:30655–65.
38. Muntasir Nishat M, Faisal F, Jahan Ratul I, Al-Monsur A, Ar-Rafi AM, Nasrullah SM, et al. A comprehensive investigation of the performances of different machine learning classifiers with SMOTE-ENN oversampling technique and hyperparameter optimization for imbalanced heart failure dataset. *Sci Program*. 2022;2022:1–17.
39. Wang Z, Wu C, Zheng K, Niu X, Wang X. SMOTETomek-based resampling for personality recognition. *IEEE Access*. 2019;7:129678–89.
40. Tu JV. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *J Clin Epidemiol*. 1996;49(11):1225–31.
41. Imandoust SB, Bolandraftar M. Application of k-nearest neighbor (knn) approach for predicting economic events: theoretical background. *Int J Eng Res Appl*. 2013;3(5):605–10.
42. Burbidge R, Buxton B. An introduction to support vector machines for data mining. Keynote papers, young OR12. 2001:3–15.
43. Kalcheva N, Todorova M, Marinova G, editors. Naive Bayes Classifier, Decision Tree and AdaBoost Ensemble Algorithm—Advantages and Disadvantages. Proceedings of the 6th ERAZ Conference Proceedings (part of ERAZ conference collection), Online; 2020.
44. Aria M, Cuccurullo C, Gnasso A. A comparison among interpretative proposals for Random forests. *Mach Learn Appl*. 2021;6:100094.
45. Hao L, Huang G. An improved AdaBoost algorithm for identification of lung cancer based on electronic nose. *Heliyon*. 2023;9(3).
46. Ahn JM, Kim J, Kim K. Ensemble machine learning of gradient boosting (XGBoost, LightGBM, CatBoost) and attention-based CNN-LSTM for harmful algal blooms forecasting. *Toxins*. 2023;15(10):608.
47. Elmogy AM, Tariq U, Ammar M, Ibrahim A. Fake reviews detection using supervised machine learning. *Int J Adv Comput Sci Appl*. 2021;12(1).
48. Singh SK, Taylor RW, Pradhan B, Shirzadi A, Pham BT. Predicting sustainable arsenic mitigation using machine learning techniques. *Ecotoxicol Environ Saf*. 2022;232:113271.
49. Bishop CM. Neural networks for pattern recognition. Oxford University Press; 1995.
50. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. *Nature*. 1986;323(6088):533–6.
51. Susan S, Kumar A. The balancing trick: optimized sampling of imbalanced datasets—A brief survey of the recent state of the art. *Eng Rep*. 2021;3(4):e12298.
52. Manzali Y, Akhiat Y, Abdoulaye Barry K, Akachar E, El Far M. Prediction of Student Performance using Random Forest Combined with Naïve Bayes. *Comput J*. 2024:bxae036.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.