Healthcare Analytics 5 (2024) 100297

Contents lists available at ScienceDirect

Healthcare Analytics

journal homepage: www.elsevier.com/locate/health

An investigation of machine learning algorithms and data augmentation techniques for diabetes diagnosis using class imbalanced BRFSS dataset

Mohammad Mihrab Chowdhury [a,*], Ragib Shahariar Ayon [b], Md Sakhawat Hossain [a]

[a] Department of Mathematics and Statistics, Texas Tech University, Lubbock, TX, USA
[b] Department of Electronics and Telecommunication Engineering, Rajshahi University of Engineering and Technology, Rajshahi, Bangladesh

ARTICLE INFO

Dataset link: https://www.cdc.gov/brfss/index.html

Keywords:
Diabetes
Imbalanced data
Machine learning
Classification
Nominal data
Sampling
Behavioral Risk Factor Surveillance System (BRFSS)

ABSTRACT

Diabetes is a prevalent chronic condition that poses significant challenges to early diagnosis and identifying at-risk individuals. Machine learning plays a crucial role in diabetes detection by leveraging its ability to process large volumes of data and identify complex patterns. However, imbalanced data, where the number of diabetic cases is substantially smaller than non-diabetic cases, complicates the identification of individuals with diabetes using machine learning algorithms. This study focuses on predicting whether a person is at risk of diabetes, considering the individual's health and socio-economic conditions while mitigating the challenges posed by imbalanced data. We employ several data augmentation techniques, such as oversampling (Synthetic Minority Over Sampling for Nominal Data, i.e.SMOTE-N), undersampling (Edited Nearest Neighbor, i.e. ENN), and hybrid sampling techniques (SMOTE-Tomek and SMOTE-ENN) on training data before applying machine learning algorithms to minimize the impact of imbalanced data. Our study sheds light on the significance of carefully utilizing data augmentation techniques without any data leakage to enhance the effectiveness of machine learning algorithms. Moreover, it offers a complete machine learning structure for healthcare practitioners, from data obtaining to machine learning prediction, enabling them to make informed decisions.

1. Introduction

Communicable diseases like coronavirus, dengue fever, hepatitis, HIV/AIDS, and chickenpox have garnered global attention due to their potential for rapid cross-border transmission and profound impact on public health. Historically responsible for pandemics like COVID-19, Spanish flu, MARS, and cholera, these diseases have prompted extensive research and response efforts [1–6]. In contrast, non-communicable diseases (NCDs), including chronic illnesses, have received less recognition despite their significant global health implications [7–9]. Rooted in genetic, physiological, environmental, and behavioral factors, NCDs progress gradually and are not contagious, often affecting specific populations within regions [10,11]. Nonetheless, the prevalence of NCDs is surging, becoming the foremost cause of death and disability worldwide [12–15]. Diabetes, cardiovascular disease, chronic lung disease, and cancer are the predominant NCDs, collectively accounting for most NCD-related mortality [13,16].

Diabetes, in particular, holds a prominent place within the NCD landscape. It ranks as the seventh leading cause of death in the United States [17], with over 37.3 million Americans affected in 2019, i.e., approximately one in every ten people [18,19]. Alarmingly, many individuals with diabetes or pre-diabetes remain unaware of their condition, with 1 in 5 people suffering from diabetes and 8 in 10 people with pre-diabetes [18,19]. Such underdiagnosis is concerning, given that individuals with diabetes are more vulnerable to seasonal and emerging diseases like COVID-19. Around 39.7% of hospitalized COVID-19 patients have diabetes as an underlying condition, rising to 46.5% for patients aged 50 to 64 [20–22].

Moreover, those with diabetes face a 60% higher risk of early mortality than those without diabetes and increased susceptibility to complications such as blindness, kidney failure, heart attacks, strokes, and limb amputation [13,17]. Over the last two decades, diabetes prevalence has doubled in the USA, raising significant concerns [19]. Financially, people diagnosed with diabetes incur substantial medical expenditures, averaging $16,752 per year, around $9601 of which are attributed to diabetes, making it an economic burden [23–25].

In this context, the timely identification of individuals at risk of diabetes is crucial for effective preventive measures. However, mass testing for diabetes would be costly, time-consuming, and overwhelming for healthcare facilities. Machine learning emerges as a pivotal tool in diabetes detection, capitalizing on its capacity to process intricate datasets and discern complex patterns. Consequently, our study's core objective is to comprehensively explore the intricate nexus between

* Corresponding author.
E-mail address: mu.chowdhury@ttu.edu (M.M. Chowdhury).

https://doi.org/10.1016/j.health.2023.100297
Received 23 August 2023; Received in revised form 15 December 2023; Accepted 23 December 2023
Available online 30 December 2023

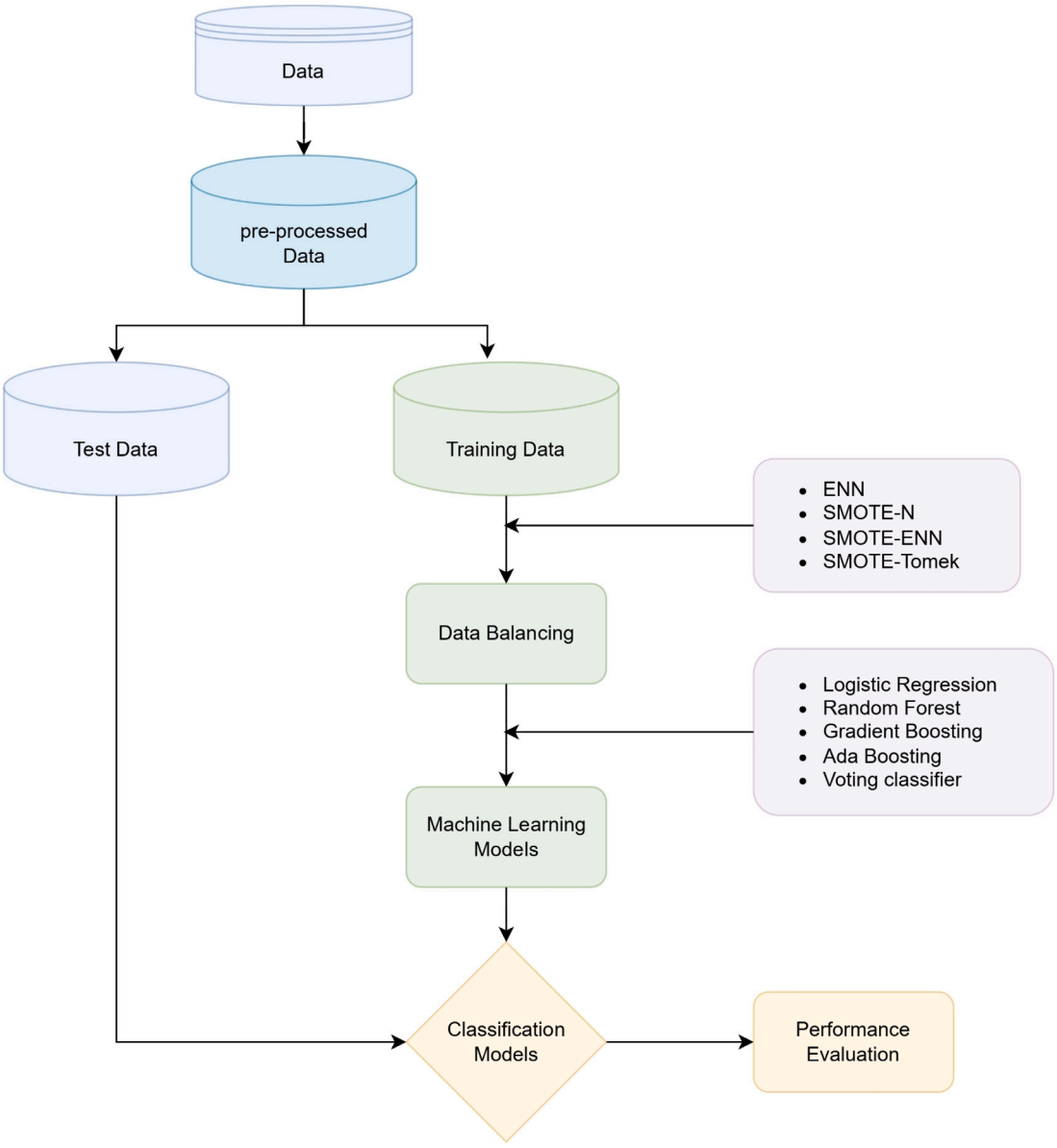

Fig. 1. Illustration of the study process from data collection to result interpretation.

health, socioeconomic factors, and diabetes through accessible data and develop a machine-learning approach.

During the analysis phase, a notable challenge surfaced: a substantial class imbalance within the dataset, a significant hurdle for achieving accurate results via machine learning algorithms. Existing literature underscores the adverse impact of dataset imbalance on algorithm performance, with established methods demonstrating subpar performance in identifying minority classes when data leakage is absent [26–28]. We employ various sampling techniques and ensemble machine learning algorithms to address this imbalance [29,30], ensuring no data leakage occurs at any study stage. Valuing these techniques and identifying the most effective strategy for handling imbalanced data within the context of machine learning algorithms form critical dimensions of our investigation. Impressively, all sampling techniques have yielded superior recall values compared to the original dataset. Our methodology notably highlights the superior performance of the Editest Nearest Neighbors (ENN) sampling technique across all metrics. Moreover, our findings align with existing research, indicating that higher BMI, elevated blood pressure levels, and advanced age correlate with elevated diabetes risk [31–37].

2. Study design and data wrangling

The step-by-step workflow, as illustrated in Diagram Fig. 1, outlines the various stages and processes involved in conducting our study. The diagram provides a visual representation of how the study progresses from data collection to analysis and interpretation of the results.

2.1. Data overview

This study employs the 2021 Behavioral Risk Factor Surveillance System (BRFSS) dataset, sourced from telephone surveys, encompassing USA residents' health behaviors, conditions, and socioeconomic aspects [38]. The BRFSS-2021 dataset holds 438,693 records featuring 303 attributes. For diabetes classification, our binary approach excludes gestational diabetes, focusing on Type 1 and Type 2 diabetes [39–41]. The latter is omitted due to its transient nature, which is linked explicitly to pregnancy [42]. Gestational diabetes is reversible; it typically resolves after childbirth, unlike Type 1 and Type 2 diabetes, which require lifelong management. Type 1 diabetes results from immune-driven insulin deficiency [43,44], while Type 2 diabetes emerges from

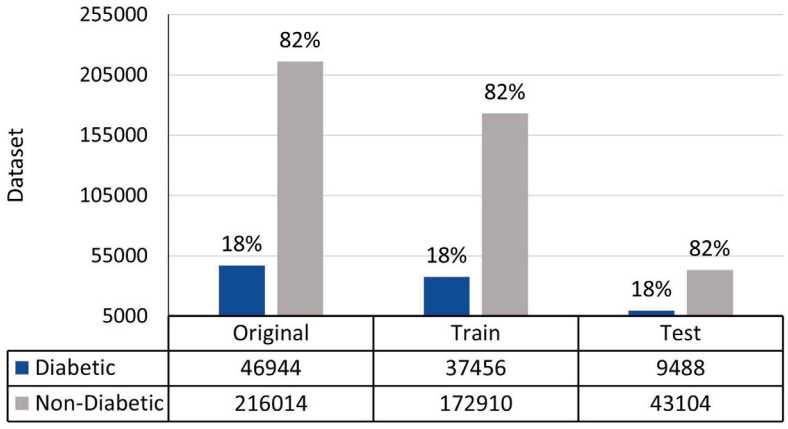

Fig. 2. Visual representation of before and after splitting the dataset.

Table 1
Diabetes classification question and response.

Target question	Response	ID
"Ever told you had diabetes?"	Answered "yes"	1
	Answered "yes" but, "Female told only during pregnancy"	2
	No (Diabetes)	3
	No (Pre-diabetes or borderline diabetes)	4
	Don't know/Not Sure	7
	Refused	9

insulin resistance [45,46]. Approximately 90%–95% of cases are Type 2, while 5%–10% are Type 1 [19,39,41]. This study adopts a simplified binary categorization approach by grouping Type 1 and Type 2 diabetes under the variable "diabetes".

2.1.1. Data preprocessing

We meticulously curate the most pertinent features for our study through an extensive review of existing literature [7,10,47–52], resulting in a selection of 20 variables. Our independent variables encompass BMI, AGE, Income, Smoking, Blood Pressure, Cholesterol, Heart Disease, Asthma, Kidney Disease, Marital Status, Education, General Health Condition, Exercise, Arthritis, Depression, Food and Vegetable Consumption, Sex, and Diabetes as the dependent variable. The target variable inquired whether respondents had ever been told they had diabetes, with response options including Yes (1), Yes (But female told only during pregnancy) (2), No (3), No (Pre-diabetes or borderline diabetes) (4), Don't know/Not Sure (7), Refused (9), and Blank (Table 1).

To align the dataset for machine learning analysis, we conduct preprocessing by eliminating missing values and disregarding instances where respondents indicate "Don't Know/Not Sure", "Refused", or left fields "Blank". Furthermore, we exclude gestational diabetes due to its temporary and reversible nature. Pregnant women are also removed from our study set to mitigate biases. Consequently, our study considers only yes (1) without gestational diabetes and no (3) without pregnant women. The dataset's diabetic and non-diabetic percentages are presented in Table 2, with age-specific distributions in Table 3. Notably, the dataset demonstrates a minimal impact of diabetes on young individuals, aligning with broader literature indicating increased diabetes risk for individuals over 45 years of age [17]. Consequently, we focus on individuals aged 40 and above to harmonize our dataset with existing literature and address dataset imbalances. This dataset preparation culminates in a dataset size of (262,958, 21), providing a solid foundation for subsequent data preprocessing and analysis.

2.1.2. Splitting the dataset

The organized dataset undergoes partitioning into training and testing subsets, utilizing an 80% training and 20% testing ratio. For this purpose, we employ Python's model selection library's default test-train split command [53,54]. This command not only shuffles the dataset but also employs a stratified approach, thereby preserving the proportional representation of each diabetes class observed in the original dataset (See in Fig. 2 and Table 4).

Table 2
Structure of the dataset.

Type	Size	Ratio
Diabetic	48,581	15%
Non-Diabetic	265,332	85%
Total	313,913	100%

2.2. Balancing techniques

In Table 4, the imbalanced nature of our dataset's target variable is evident, with diabetic patients representing only 18.0%. To address this class imbalance, we implement four distinct sampling techniques on the training data tailored to nominal data. These techniques encompass oversampling, undersampling, and hybrid approaches. We apply these techniques to balance the dataset and enhance our models' performance.

2.3. Encoding nominal values

Given that all dataset features are of nominal type, using the values directly will mislead machine learning models. Due to this, we apply one-hot encoding to the dataset. This approach creates distinct categories based on unique values in each column, subsequently expanding the number of columns to 92.

3. Theoretical overview and definitions

3.1. Sampling techniques

For the oversampling technique, we adapt the Synthetic Minority Over-sampling Technique for nominal data (SMOTE-N) [29], given the

Table 3
Age structure of the dataset.

Age	Criteria	Percentage	Age	Criteria	Percentage
18–24	Diabetic	0.02	55–59	Diabetic	0.17
	Non-Diabetic	0.98		Non-Diabetic	0.83
25–29	Diabetic	0.02	60–64	Diabetic	0.19
	Non-Diabetic	0.98		Non-Diabetic	0.81
30–34	Diabetic	0.03	65–69	Diabetic	0.20
	Non-Diabetic	0.97		Non-Diabetic	0.80
35–39	Diabetic	0.05	70–74	Diabetic	0.23
	Non-Diabetic	0.95		Non-Diabetic	0.77
40–44	Diabetic	0.07	75–79	Diabetic	0.24
	Non-Diabetic	0.93		Non-Diabetic	0.76
45–49	Diabetic	0.11	80 or older	Diabetic	0.20
	Non-Diabetic	0.89		Non-Diabetic	0.80
50–54	Diabetic	0.13	Don't know/	Diabetic	0.13
	Non-Diabetic	0.87	Refused/Missing	Non-Diabetic	0.87

Table 4
Before and after splitting.

Before split			After split			
Type	Size	Percentage	Category	Type	Size	Percentage
Diabetic	46,944	18%	Train	Diabetic	37,456	18%
Non-Diabetic	216,014	82%		Non-Diabetic	172,910	82%
Total	262,958	100%	Test	Diabetic	9488	18%
				Non-Diabetic	43,104	82%

nominal nature of our data. As for undersampling, we utilize the Edited Nearest Neighbors algorithm (ENN) [55]. We employ SMOTE-Tomek, which combines SMOTE-N with Tomek Links, and SMOTE-ENN, which integrates SMOTE-N with ENN [29] as hybrid techniques. The specific details of these sampling techniques are briefly outlined in the following section.

3.1.1. Synthetic Minority Over-sampling Technique (SMOTE)

SMOTE is a widely acknowledged technique for addressing imbalanced data in classification problems [29,56,57]. It operates by generating synthetic instances for the minority class through interpolation between existing examples rather than through replacement-based oversampling [29]. For each minority instance x_i, SMOTE constructs N synthetic examples by interpolating with its K nearest neighbors x_j, incorporating a parameter λ within the range of 0 to 1. This interpolation process is represented as $x_{new} = x_i + \lambda(x_j - x_i)$, where x_i and x_j denote feature value vectors.

3.1.2. Edited Nearest Neighbor (ENN)

ENN, or Edited Nearest Neighbors, is an enhanced classification algorithm derived from k-Nearest Neighbors (k-NN), designed to eliminate noisy and mislabeled instances from the training dataset. This refinement improves classification accuracy, rendering ENN particularly valuable for addressing imbalanced datasets [55,58].

3.1.3. SMOTE-ENN and SMOTE-TOMEK

SMOTE-ENN emerges as the fusion of SMOTE and ENN techniques, effectively tackling imbalanced data classification challenges. By merging SMOTE's minority class oversampling with ENN's majority class undersampling, the method harmonizes the dataset's distribution. This technique was formulated by [59] as a robust approach for handling class imbalance.

Similarly, SMOTE-TOMEK is another amalgamation technique employed in imbalanced data classification scenarios. It is a potent strategy renowned for its efficacy in addressing imbalanced datasets, particularly when faced with noisy or overlapping instances. By integrating SMOTE and Tomek Links, this technique effectively navigates imbalanced scenarios to improve model performance.

3.2. Machine learning algorithms

Employing a diverse set of machine learning algorithms, including Logistic Regression, Gradient Boosting, AdaBoost, and Random Forest classifiers, we strive to leverage their distinct strengths for enhanced predictions. The top three performing algorithms are aggregated using a voting classifier to optimize overall predictive accuracy for each sampling technique. This collaborative approach aims to bolster the accuracy and robustness of our predictions across different sampling strategies. The following section provides a concise overview of these algorithmic components.

3.2.1. Logistic Regression

Logistic Regression (LR) models binary responses using a set of independent predictors. Let P_i denote the probability of a patient responding "Yes" and $1 - P_i$ the probability of responding "No". The LR model equation is formulated as follows:

$$ln\left[\frac{P_i}{1-P_i}\right] = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} \quad (1)$$

This equation computes the natural logarithm of the odds ratio of responses, with $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ representing the model coefficients. Subsequently, the equation is transformed as:

$$\frac{P_i}{1-P_i} = exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}) \quad (2)$$

This further leads to:

$$P_i = \frac{exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi})}{1 + exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi})} \quad (3)$$

3.2.2. Random forest

Initially introduced by [60], Random Forest operates as a tree-based ensemble prediction model. This prediction model builds multiple decision trees using randomly selected predictor variables and training datasets. The independent variables are represented as $X = (X_1, X_2, \ldots, X_k)$, while Y signifies the response variable. The main goal is to predict Y by establishing a prediction function $f(X)$ [61].

The model minimizes the loss function $L(Y, f(X))$ to determine the prediction function. In classification tasks, the commonly used loss function is the zero–one loss.

$$L(Y, f(X)) = \begin{cases} 0 & if \quad Y = f(X) \\ 1 & otherwise. \end{cases} \tag{4}$$

3.2.3. Gradient boosting

Gradient Boosting, a widely adopted ensemble method for classification and regression tasks, is initially introduced by [62]. This technique constructs an additive model by sequentially incorporating weak learners, enhancing the model's overall performance [63]. By iteratively focusing on the mistakes made by prior learners, each subsequent learner aims to correct and improve upon the prior ones, effectively creating a powerful ensemble model that effectively leverages each learner's strengths.

3.2.4. AdaBoost (Adaptive Boosting)

AdaBoost, introduced by [64], is an iterative algorithm designed to enhance the performance of weak classifiers, also known as base classifiers. AdaBoost improves data classification capabilities by adapting these base classifiers' errors iteratively. This process contributes to a reduction in both bias and variance. The algorithm's strength lies in its ability to continuously train and refine the classifiers, resulting in an ensemble model that leverages the individual strengths of these classifiers, ultimately yielding improved overall classification accuracy [65].

3.2.5. Voting classifier

Voting classifiers belong to the ensemble machine learning category, where the outcomes of multiple distinct classifiers are combined to yield a final prediction [66]. This approach harnesses collective knowledge to bolster prediction accuracy and reliability. By amalgamating predictions from each classifier, the voting classifier employs a majority rule mechanism to make a final prediction, opting for the class label that accumulates the highest number of votes. This collaborative approach enhances the model's overall predictive power, benefiting from its constituent classifiers' diverse insights.

3.3. Evaluation metrics

During our evaluation, we focus on four essential metrics: precision, recall, accuracy, and AUC–ROC [67]. In the context of imbalanced data, recall gains significance by accurately pinpointing positive instances within the minority class, which is crucial in real-world applications. Unlike accuracy, which can be deceptive in such datasets, robust recall guarantees adept detection and prediction of minority class instances. This section briefly elaborates on the evaluation metrics' descriptions.

3.3.1. Precision

Precision addresses how accurate identifications are, expressed as the percentage of correct optimistic predictions out of all predicted positives. Also known as positive predictive value (PPV), it is calculated by dividing true positives by the sum of true positives and false positives.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{5}$$

3.3.2. Recall (Sensitivity)

Recall quantifies the proportion of true positives correctly detected, reflecting the ratio of true positives to all instances that should have been identified as positive. In binary classification, recall corresponds to sensitivity.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{6}$$

3.3.3. Accuracy

Accuracy gauges the percentage of correct classifications has achieved by a trained model, calculated by dividing the sum of true negatives and true positives by the total instances in the dataset. This metric is particularly effective for balanced classification tasks with relatively even class representation.

$$Accuracy = \frac{True\ Negative + True\ Positive}{True\ Negative + True\ Positive + False\ Negative + False\ Positive} \tag{7}$$

3.3.4. AUC-ROC

The Area Under the Receiver Operating Characteristic Curve (AUC–ROC) evaluates binary classification model performance. It measures the model's ability to differentiate between positive and negative classes by plotting True Positive Rate (TPR) against False Positive Rate (FPR). A perfect classifier yields an AUC of 1.0, while a random classifier has an AUC of 0.5. The AUC–ROC value is computed by integrating the TPR-FPR curve over the entire FPR range from 0 to 1.

$$AUC - ROC = \int_0^1 TPR(FPR), dFPR \tag{8}$$

3.4. Hyper parameter tuning

To optimize model performance, we leverage GridSearchCV for hyperparameter tuning on the training data. This method systematically explores various hyperparameter combinations within a predefined grid, enabling the models to be trained and assessed for the most favorable configuration that maximizes performance metrics. Through this approach, we fine-tune our models, identifying optimal hyperparameters that enhance accuracy and better generalization when handling unseen data. For this, we acknowledge the High-Performance Computing Center (HPCC) at Texas Tech University for providing computational resources that have contributed to the research results reported within this paper. URL: http://www.hpcc.ttu.edu.

4. Result

Diabetes, a chronic condition affecting millions globally, arises from the body's inability to regulate blood sugar levels, resulting in elevated glucose levels in the bloodstream. This chronic condition causes severe complications, including nerve and kidney damage, vision issues, and cardiovascular disease. Timely identification of individuals at risk of developing diabetes is essential for effective intervention. Traditional testing methods for diabetes, while valuable, can be costly, time-consuming, and may not capture the full complexity of risk factors. In contrast, machine learning offers distinct advantages by leveraging existing data to provide cost-effective and efficient solutions. It is particularly crucial for early detection, where subtle risk factors may play a significant role. However, the challenge of imbalanced data, where the number of diabetic cases is limited compared to non-diabetic cases, poses significant difficulties in training machine learning algorithms. Overcoming this hurdle is the key to accurate diabetes prediction using machine learning algorithms.

For the study, we have considered the BRFSS dataset which is heavily imbalanced (Table 2). To address this data imbalance, we have employed three sampling techniques: oversampling, undersampling, and hybrid sampling techniques on the training data (Figs. 3, 4). We employ the SMOTE-N Technique to oversample the minority class (individuals with diabetes), ENN to reduce the number of samples in the majority class (people without diabetes) and SMOTE-Tomek and SMOTE-ENN as hybrid techniques, to achieve a balanced dataset. We apply several machine learning algorithms to each sampling technique, including logistic regression, random forest, adaBoost, and gradient boost to observe their performance in terms of different sampling techniques (Table 6). Following an initial evaluation of AUC scores and

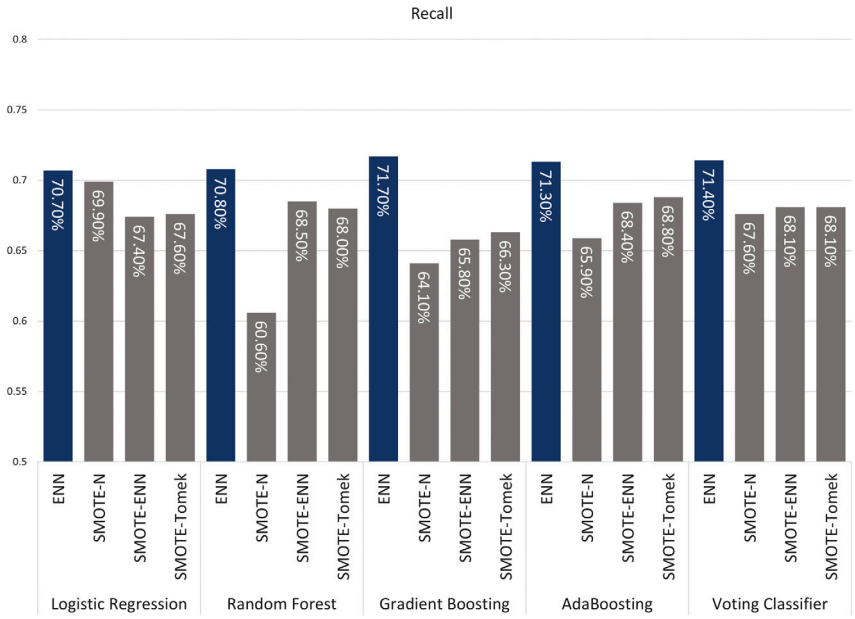

Fig. 3. Recall for different sampling techniques: Edited Nearest Neighbors (ENN), Synthetic Minority Over-sampling Technique (SMOTE-N), SMOTE-Tomek, and SMOTE-ENN for logistic regression, random forest, gradient boosting, and AdaBoost.

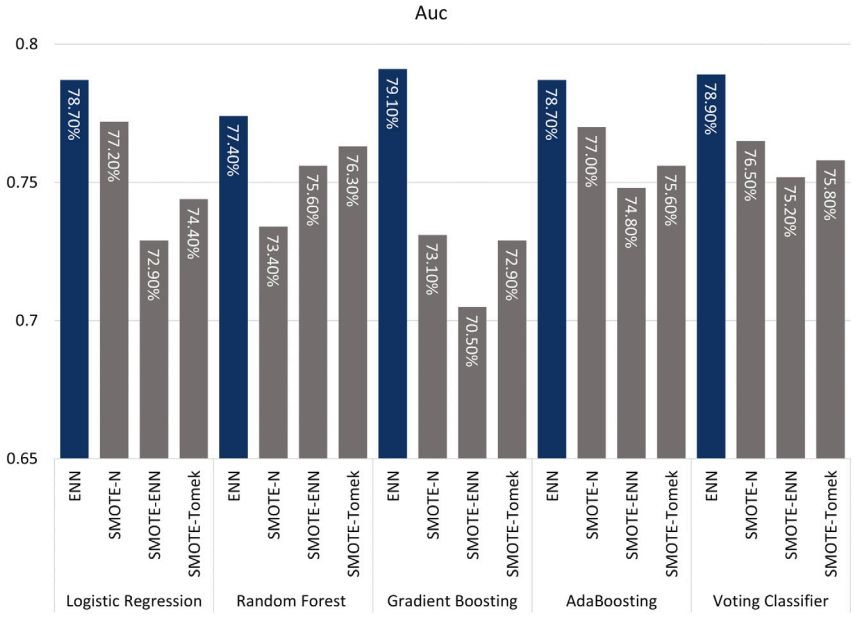

Fig. 4. AUC for ENN, SMOTE-N, SMOTE-Tomek, and SMOTE-ENN for logistic regression, random forest, gradient boosting, and AdaBoost.

recall metrics, we fine-tune hyperparameters using GridSearchCV for these algorithms, utilizing the TTU High-Performance Computing Center (HPCC). Model performance evaluation includes precision, recall, accuracy, and AUC scores (Table 6, Fig. 4). Specifically, our focus lies on recall, as it is paramount in disease detection due to the inherent imbalance of positive cases (Fig. 3) [67].

The challenge of imbalanced data becomes evident when comparing the model performance before and after applying the sampling techniques (Fig. 5). Applying logistic regression, random forest, gradient boosting, and adaptive boosting to the raw data reveals relatively high accuracy scores, ranging from 81.7% to 83.0% (Table 5). However, this high accuracy is accompanied by relatively low recall scores, ranging from 57.4% to 58% (Table 5). This discrepancy underscores the misclassification of positive cases as negative, adversely affecting timely diagnosis and patient care. Misclassification can lead to missed interventions and an increased risk of complications.

Table 5
Before implementing the data balancing techniques.

Models	Precision	Recall	Accuracy	AUC
Logistic regression	0.718	0.577	0.83	0.787
Random forest	0.663	0.574	0.817	0.754
Gradient boosting	0.722	0.575	0.83	0.789
AdaBoosting	0.717	0.58	0.83	0.787

Our employment of various data-balancing techniques to address this limitation, resulting in substantial improvements, particularly in recall (Table 6). For the undersampling technique, the recall is improved by 13%, 13.4%, 14.2%, 13.3% compared to the original result

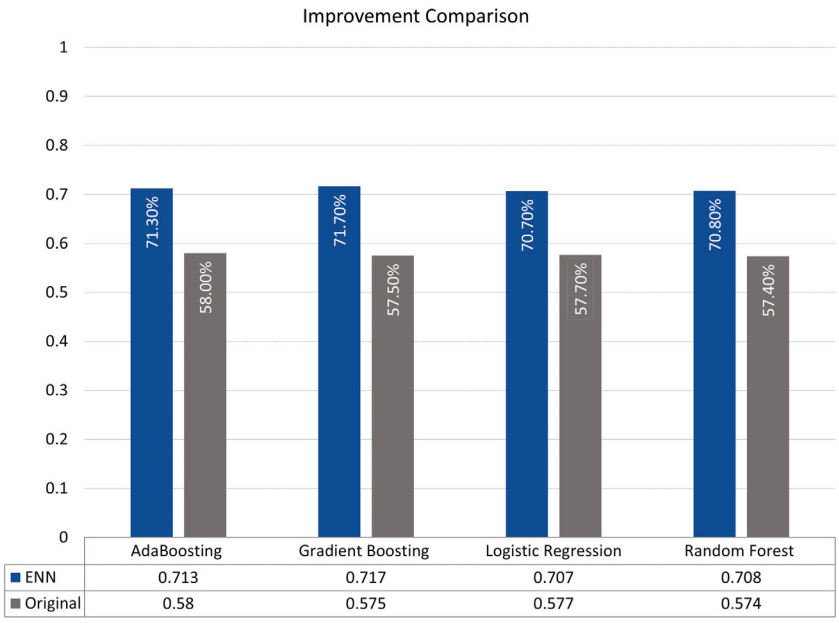

Fig. 5. Comparison of improvement of recall value between original and sampling technique (ENN) for different machine learning algorithms.

Table 6
After implementing different sampling strategies.

Sampling strategy	Models	Precision	Recall	Accuracy	AUC
ENN	Logistic regression	0.623	0.707	0.625	0.787
	Random forest	0.629	0.708	0.691	0.774
	Gradient boosting	0.635	0.717	0.703	0.791
	AdaBoosting	0.632	0.713	0.694	0.787
SMOTE-N	Logistic regression	0.627	0.699	0.706	0.772
	Random forest	0.618	0.606	0.783	0.734
	Gradient boosting	0.63	0.641	0.775	0.731
	AdaBoosting	0.652	0.659	0.791	0.770
SMOTE-ENN	Logistic regression	0.609	0.674	0.679	0.729
	Random forest	0.624	.685	0.718	0.756
	Gradient boosting	0.602	0.658	0.684	0.705
	AdaBoosting	0.61	0.684	0.642	0.748
SMOTE-Tomek	Logistic regression	0.617	0.676	0.710	0.744
	Random forest	0.63	0.68	0.742	0.763
	Gradient boosting	0.613	0.663	0.719	0.729
	AdaBoosting	0.623	0.688	0.711	0.756

for linear regression, random forest, gradient boosting, adaBoosting respectively. In the same order for machine learning algorithms, the recall is improved by 12.2%, 3.2%, 6.6%, 7.9% for oversampling technique and improved by 9.7%, 11.1%, 8.3%, 10.4% for the SMOTE-ENN, 0.099%, 10.6%, 8.8%, 10.8% for the SMOTE-Tomek respectively. Overall among all the strategies, the ENN undersampling technique stands out, enhancing recall by approximately 14.2% when paired with gradient boosting. When adaboosting and logistic regression, pairs with ENN, also yield significant recall improvements of 13.3% and 13.1%, respectively (Fig. 5). The effectiveness of these techniques emphasizes the importance of data balancing, particularly for minority class identification.

Moreover, we extended the study by employing ensemble methods, such as the soft voting classifier, to achieve a robust and stable model by combining top-performing algorithms based on recall. For instance, when employing the ENN sampling method, the voting classifier attains a recall of 71.4% and an AUC of 78.9% (Table 7), highlighting its capability to identify positive cases and differentiate between classes more precisely. Other techniques, including SMOTE-ENN and SMOTE-N, yield promising results, further underscoring the value of ensemble methods for enhancing model accuracy in imbalanced datasets (see Fig. 6).

Table 7
Voting Classifier.

Sampling strategy	Precision	Recall	Accuracy	AUC
ENN	0.635	0.714	0.706	0.789
SMOTE-N	0.635	0.676	0.757	0.765
SMOTE-ENN	0.617	0.681	0.699	0.752
SMOTE-Tomek	0.625	0.681	0.728	0.758

Furthermore, we also try to identify essential risk factors for diabetes classification through feature selection and analysis from our study. Key factors include age, BMI, and blood pressure, consistent with prior research. We emphasize that machine learning techniques offer a significant advantage over traditional testing methods because they leverage existing data for cost-effective and efficient early detection. Our findings indicate that while under-sampling techniques exhibit superior outcomes, hybrid techniques provide comparable results. This indicates their potential for effective diabetes classification in terms of imbalanced health data. In the case of limited computational resources, even running logistic regression can provide valuable initial insights. Ultimately, our study underscores the pivotal role of machine learning in enhancing early diabetes detection using existing data, given the challenges posed by imbalanced datasets.

5. Conclusion

The implications of our study carry profound significance within the context of the prevailing global burden of diabetes. With millions of individuals affected worldwide, diabetes has substantial social, economic, and healthcare ramifications. According to data from the International Diabetes Federation, an estimated 463 million adults between the ages of 20 and 79 were grappling with diabetes in 2019, and this number is projected to escalate to an astounding 700 million by 2045 [68]. Beyond the personal toll of diabetes on physical health and well-being, the disease significantly strains healthcare systems and economies globally. It is predicted that annual healthcare expenditures linked to diabetes will surpass $800 billion by 2045 [69].

In light of this pressing issue, the imperative for proactive diabetes prevention and effective management strategies becomes clear. This necessitates the need for our study to identify and mitigate pivotal risk factors for the disease and the refinement of diagnostic and treatment

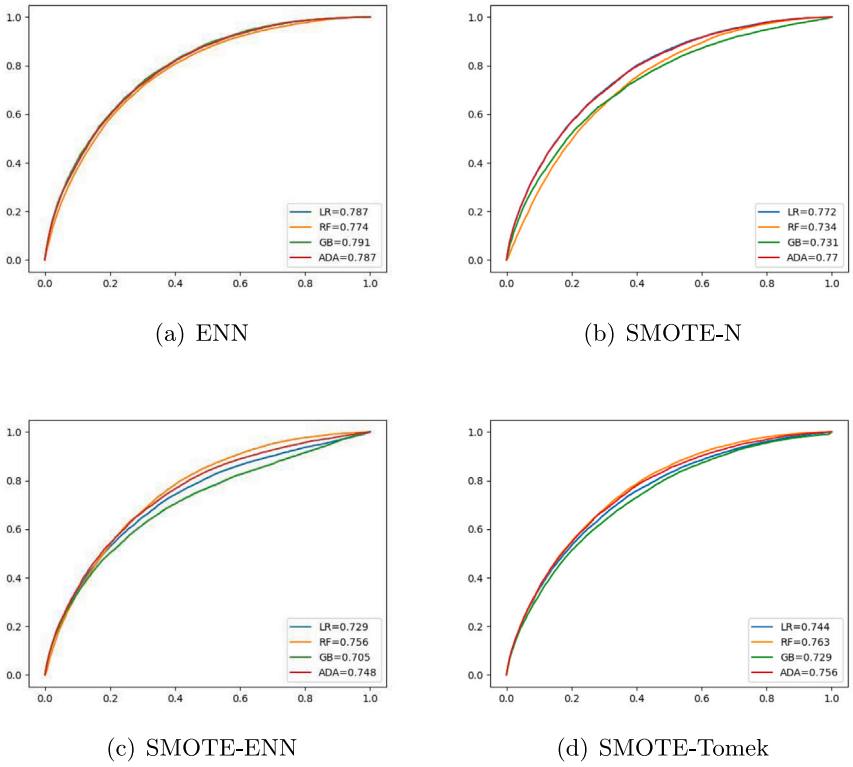

Fig. 6. Comparative evaluation of Area Under the Receiver Operating Characteristic (AUC-ROC) curves for different sampling techniques and the corresponding machine learning algorithms.

approaches. Our framework has shown how machine learning algorithms and data augmentation techniques can play a significant role in the early identification of individuals at risk of diabetes by shedding light on the intricate web of underlying causes. Our study delves into the application of data augmentation techniques such as ENN, SMOTE-N, SMOTE-ENN, and SMOTE-Tomek. To analyze the impact of different sampling techniques, we then explore four different machine learning algorithms, namely logistic regression, adaboost, gradient boost and random forest for each sampling technique. The study suggests that ENN with gradient boosting performs best among all others. Additionally, the study findings highlight the potential effectiveness of our strategy in the case of imbalanced data, as it yields more precise and potent machine-learning models when employed wisely (Figs. 5, 3). However, it is imperative to be aware of the potential pitfalls. For instance, indiscriminate application to the entire dataset could inadvertently introduce biases or lead to over-fitting, undermining the model's capacity to generalize to new and unseen data. After conducting a thorough literature review, we found that using a sampling technique on the entire raw data may result in data leakage and ultimately lead to inflated metric values [70–74]. However, when appropriately used, our strategy enhances the outcome of severely imbalanced data compared to the initial and existing results [75]. Also, it is imperative to note that comparing results to existing literature may only sometimes yield fruitful results. It is due to the fact that the parameters of the machine learning algorithm depend on the dataset's unique characteristics, including the set of dependent and independent variables. As a result, changes in the dataset can affect the outcome [76,77]. Therefore, it is essential to consider how much improvement in prediction is happening compared to the raw data when making comparisons. Moreover, some of these methods may have a computational and temporal cost, particularly when dealing with extensive datasets. As a result, a prudent evaluation of these factors is crucial when determining the most suitable technique for a given project.

In summation, the implications of our study resonate deeply with the global burden of diabetes. Our approach not only advances the understanding of sampling techniques in the context of imbalanced datasets but also provides a practical guide for application. Building on our findings, future research could explore the integration of our insights with other machine learning models where imbalanced data poses a significant challenge. Therefore, our holistic approach promises to contribute significantly to the global effort in combating not just diabetes but also various chronic diseases, improving healthcare outcomes worldwide.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data is publicly available in CDC website. https://www.cdc.gov/brfss/index.html.

Declaration of Generative AI and AI-assisted technologies in the writing process

While preparing this work, the authors used https://chat.openai.com to improve language and readability. After using this tool, the authors reviewed and edited the content as needed and takes full responsibility for the publication's content.

Funding

The authors received no funding for this study.

Code availability

The data is available at [38] and the code will be provided upon reasonable request.

References

[1] J.M. Van Seventer, N.S. Hochberg, Principles of infectious diseases: transmission, diagnosis, prevention, and control, Int. Encyclopedia Public Health (2017) 22.

[2] N. Kenworthy, M. Thomann, R. Parker, From a global crisis to the 'end of AIDS': New epidemics of signification, Glob. Public Health 13 (8) (2018) 960–971.

[3] A. Zumla, A.N. Alagaili, M. Cotten, E.I. Azhar, Infectious diseases epidemic threats and mass gatherings: refocusing global attention on the continuing spread of the middle east respiratory syndrome coronavirus (MERS-CoV), BMC Med. 14 (1) (2016) 1–4.

[4] M.H. Green, Taking "pandemic" seriously: Making the black death global, Medieval Globe 1 (1) (2015) 27–61.

[5] M.R. Islam, T. Oraby, A. McCombs, M.M. Chowdhury, M. Al-Mamun, M.G. Tyshenko, C. Kadelka, Evaluation of the United States COVID-19 vaccine allocation strategy, PLoS One 16 (11) (2021) e0259700.

[6] M.M. Chowdhury, M.R. Islam, M.S. Hossain, N. Tabassum, A. Peace, Incorporating the mutational landscape of SARS-COV-2 variants and case-dependent vaccination rates into epidemic models, Infect. Dis. Modell. 7 (2) (2022) 75–82.

[7] A. Budreviciute, S. Damiati, D.K. Sabir, K. Onder, P. Schuller-Goetzburg, G. Plakys, A. Katileviciute, S. Khoja, R. Kodzius, Management and prevention strategies for non-communicable diseases (NCDs) and their risk factors, Front. Public Health (2020) 788.

[8] H. Frumkin, A. Haines, Global environmental change and noncommunicable disease risks, Annu. Rev. Public Health 40 (2019) 261–282.

[9] WHO, Noncommunicable Diseases, World Health Organization, 2023, Accessed March 22. Accessed at https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases.

[10] S. Supakul, H.Y. Park, B.N. Nguyen, K.B. Giang, Prevalence differences in major non-communicable diseases in a low-middle income country: a comparative study between an urban and a rural district in Vietnam, J. Global Health Sci. 1 (2) (2019).

[11] J.J. Bigna, J.J. Noubiap, The rising burden of non-communicable diseases in sub-Saharan Africa, Lancet Glob. Health 7 (10) (2019) e1295–e1296.

[12] S.H. Habib, S. Saha, Burden of non-communicable disease: global overview, Diabetes Metab. Syndr.: Clin. Res. Rev. 4 (1) (2010) 41–47.

[13] CDC, Global noncommunicable diseases fact sheet, 2023, Accessed March 24. Accessed at https://www.cdc.gov/globalhealth/healthprotection/resources/fact-sheets/global-ncd-fact-sheet.html#:~:text=Noncommunicable.

[14] WHO, Global Health Estimates: Life Expectancy and Leading Causes of Death and Disability, World Health Organization, 2023, Accessed March 22. Accessed at https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates.

[15] J. Divers, E.J. Mayer-Davis, J.M. Lawrence, S. Isom, D. Dabelea, L. Dolan, G. Imperatore, S. Marcovina, D.J. Pettitt, C. Pihoker, et al., Trends in incidence of type 1 and type 2 diabetes among youths—selected counties and Indian reservations, United States, 2002–2015, Morb. Mortal. Wkly. Rep. 69 (6) (2020) 161.

[16] WHO, Noncommunicable Diseases, World Health Organization, 2023, Accessed March 22. Accessed at https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases.

[17] CDC, What is Diabetes? Center for Disease Control, 2023, Accessed March 22. Accessed https://www.cdc.gov/diabetes/basics/diabetes.html#:~:text=Diabetes.

[18] CDC, Diabetes Basics, Center for Disease Control and Prevention, 2023, Accessed March 22. Accessed at https://www.cdc.gov/diabetes/basics/index.html.

[19] CDC, Diabetes Fast Facts, Center for Disease Control and Prevention, 2023, Accessed March 22. Accessed at https://www.cdc.gov/diabetes/basics/quick-facts.html.

[20] CDC, Diabetes and COVID-19, Center for Disease Control and Prevention, 2023, Accessed March 22. Accessed at https://www.cdc.gov/diabetes/library/reports/reportcard/diabetes-and-covid19.html.

[21] S. Kastora, M. Patel, B. Carter, M. Delibegovic, P.K. Myint, Impact of diabetes on COVID-19 mortality and hospital outcomes from a global perspective: An umbrella systematic review and meta-analysis, Endocrinol. Diabetes Metab. 5 (3) (2022) e00338.

[22] A. Rajpal, L. Rahimi, F. Ismail-Beigi, Factors leading to high morbidity and mortality of COVID-19 in patients with type 2 diabetes, J. Diabetes 12 (12) (2020) 895–908.

[23] A.D. Association, Economic costs of diabetes in the US in 2017, Diabetes Care 41 (5) (2018) 917–928.

[24] S. Chen, M. Kuhn, K. Prettner, D.E. Bloom, The macroeconomic burden of noncommunicable diseases in the United States: Estimates and projections, PLoS One 13 (11) (2018) e0206702.

[25] A.D. Association, The Cost of Diabetes, American Diabetes Association, 2023, Accessed June 22, 2023. Accessed https://diabetes.org/about-us/statistics/cost-diabetes.

[26] A. Fernández, S. García, M. Galar, R.C. Prati, B. Krawczyk, F. Herrera, Learning from Imbalanced Data Sets, Vol. 10, Springer, 2018.

[27] H. Kaur, H.S. Pannu, A.K. Malhi, A systematic review on imbalanced data challenges in machine learning: Applications and solutions, ACM Comput. Surv. 52 (4) (2019) 1–36.

[28] I. Ul Hassan, R.H. Ali, Z. Ul Abideen, T.A. Khan, R. Kouatly, Significance of machine learning for detection of malicious websites on an unbalanced dataset, Digital 2 (4) (2022) 501–519.

[29] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357.

[30] A. Anand, G. Pugalenthi, G.B. Fogel, P. Suganthan, An approach for classification of highly imbalanced data using weighting and undersampling, Amino Acids 39 (2010) 1385–1391.

[31] K.S. Leong, J.P. Wilding, Obesity and diabetes, Best Pract. Res. Clin. Endocrinol. Metab. 13 (2) (1999) 221–237.

[32] N. Gray, G. Picone, F. Sloan, A. Yashkin, The relationship between BMI and onset of diabetes mellitus and its complications, South. Med. J. 108 (1) (2015) 29.

[33] A.S. Group, Effects of intensive blood-pressure control in type 2 diabetes mellitus, N. Engl. J. Med. 362 (17) (2010) 1575–1585.

[34] L.S. Geiss, D.B. Rolka, M.M. Engelgau, Elevated blood pressure among US adults with diabetes, 1988–1994, Amer. J. Prev. Med. 22 (1) (2002) 42–48.

[35] C.J. Caspersen, G.D. Thomas, L.A. Boseman, G.L. Beckles, A.L. Albright, Aging, diabetes, and the public health system in the United States, Amer. J. Public Health 102 (8) (2012) 1482–1497.

[36] R.S. Ahima, Connecting obesity, aging and diabetes, Nat. Med. 15 (9) (2009) 996–997.

[37] J.E. Morley, Diabetes and aging: epidemiologic overview, Clin. Geriatr. Med. 24 (3) (2008) 395–405.

[38] CDC, Behavioral Risk Factor Surveillance System, Center for Disease Control and Prevention, 2023, Accessed March 22. Accessed at https://www.cdc.gov/brfss/index.html.

[39] CDC, Type 2 diabetes, 2023, Accessed March 22. Accessed at https://www.cdc.gov/diabetes/basics/type2.html.

[40] CDC, About Prediabetes & Type 2 Diabetes, Center for Disease Control and Prevention, 2023, Accessed March 22. Accessed at https://www.cdc.gov/diabetes/prevention/about-prediabetes.html.

[41] A.D. Association, Diagnosis and classification of diabetes mellitus, Diabetes Care 33 (Supplement_1) (2010) S62–S69.

[42] T.A. Buchanan, A.H. Xiang, et al., Gestational diabetes mellitus, J. Clin. Invest. 115 (3) (2005) 485–491.

[43] A. Katsarou, S. Gudbjörnsdottir, A. Rawshani, D. Dabelea, E. Bonifacio, B.J. Anderson, L.M. Jacobsen, D.A. Schatz, Å. Lernmark, Type 1 diabetes mellitus, Nat. Rev. Dis. Primers 3 (1) (2017) 1–17.

[44] G.S. Eisenbarth, Type I diabetes mellitus, New England J. Med. 314 (21) (1986) 1360–1368.

[45] A. Astrup, N. Finer, Redefining type 2 diabetes: 'diabesity' or 'obesity dependent diabetes mellitus'? Obes. Rev. 1 (2) (2000) 57–59.

[46] S. Chatterjee, K. Khunti, M.J. Davies, Type 2 diabetes, Lancet 389 (10085) (2017) 2239–2251.

[47] G. Robertson, E.D. Lehmann, W. Sandham, D. Hamilton, Blood glucose prediction using artificial neural networks trained with the AIDA diabetes simulator: a proof-of-concept pilot study, J. Electr. Comput. Eng. 2011 (2011) 2.

[48] A. Dinh, S. Miertschin, A. Young, S.D. Mohanty, A data-driven approach to predicting diabetes and cardiovascular disease with machine learning, BMC Med. Inform. Decis. Mak. 19 (1) (2019) 1–15.

[49] F. Hill-Briggs, N.E. Adler, S.A. Berkowitz, M.H. Chin, T.L. Gary-Webb, A. Navas-Acien, P.L. Thornton, D. Haire-Joshu, Social determinants of health and diabetes: a scientific review, Diabetes Care 44 (1) (2021) 258–279.

[50] V. Shriraam, S. Mahadevan, P. Arumugam, Prevalence and risk factors of diabetes, hypertension and other non-communicable diseases in a tribal population in south India, Indian J. Endocrinol. Metab. 25 (4) (2021) 313.

[51] D. Asiimwe, G.O. Mauti, R. Kiconco, Prevalence and risk factors associated with type 2 diabetes in elderly patients aged 45–80 years at kanungu district, J. Diabetes Res. 2020 (2020) 1–5.

[52] Z. Ullah, F. Saleem, M. Jamjoom, B. Fakieh, F. Kateb, A.M. Ali, B. Shah, et al., Detecting high-risk factors and early diagnosis of diabetes using machine learning methods, Comput. Intell. Neurosci. 2022 (2022).

[53] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[54] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, G. Varoquaux, API design for machine learning software: experiences from the scikit-learn project, in: ECML PKDD Workshop: Languages for Data Mining and Machine Learning, 2013, pp. 108–122.

[55] R. Alejo, J.M. Sotoca, R.M. Valdovinos, P. Toribio, Edited nearest neighbor rule for improving neural networks classifications, in: Advances in Neural Networks-ISNN 2010: 7th International Symposium on Neural Networks, ISNN 2010, Shanghai, China, June 6-9, 2010, Proceedings, Part I 7, Springer, 2010, pp. 303–310.

[56] J. Burez, D. Van den Poel, Handling class imbalance in customer churn prediction, Expert Syst. Appl. 36 (3) (2009) 4626–4636.

[57] H. He, E.A. Garcia, Learning from imbalanced data, IEEE Trans. Knowl. Data Eng. 21 (9) (2009) 1263–1284.

[58] D.L. Wilson, Asymptotic properties of nearest neighbor rules using edited data, IEEE Trans. Syst. Man Cybern. (3) (1972) 408–421.
[59] G. Batista, R. Prati, M.-C. Monard, A study of the behavior of several methods for balancing machine learning training data, SIGKDD Explor. 6 (2004) 20–29, http://dx.doi.org/10.1145/1007730.1007735.
[60] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[61] A. Cutler, D.R. Cutler, J.R. Stevens, Random forests, Ensemble Machine Learning: Methods and Applications, Springer, 2012, pp. 157–175.
[62] J.H. Friedman, Greedy function approximation: a gradient boosting machine, Ann. Statist. (2001) 1189–1232.
[63] T. Hastie, R. Tibshirani, J.H. Friedman, J.H. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Vol. 2, Springer, 2009.
[64] Y. Freund, R.E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, J. Comput. System Sci. 55 (1) (1997) 119–139, http://dx.doi.org/10.1006/jcss.1997.1504, URL https://www.sciencedirect.com/science/article/pii/S002200009791504X.
[65] Y. Wu, Y. Ke, Z. Chen, S. Liang, H. Zhao, H. Hong, Application of alternating decision tree with AdaBoost and bagging ensembles for landslide susceptibility mapping, Catena 187 (2020) 104396, http://dx.doi.org/10.1016/j.catena.2019.104396.
[66] M. Beyeler, Machine Learning for OpenCV, Packt Publishing Ltd, 2017.
[67] N. Japkowicz, Why question machine learning evaluation methods, in: AAAI Workshop on Evaluation Methods for Machine Learning, Citeseer, 2006, pp. 6–11.
[68] P. Saeedi, I. Petersohn, P. Salpea, B. Malanda, S. Karuranga, N. Unwin, S. Colagiuri, L. Guariguata, A.A. Motala, K. Ogurtsova, 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. 157 (2019) 107843.
[69] R. Williams, S. Karuranga, B. Malanda, P. Saeedi, A. Basit, S. Besançon, C. Bommer, A. Esteghamati, K. Ogurtsova, P. Zhang, et al., Global and regional estimates and projections of diabetes-related health expenditure: Results from the international diabetes federation diabetes atlas, Diabetes Res. Clin. Pract. 162 (2020) 108072.
[70] I.E. Tampu, A. Eklund, N. Haj-Hosseini, Inflation of test accuracy due to data leakage in deep learning-based classification of OCT images, Sci. Data 9 (1) (2022) 580.
[71] G.F. Silva, T.P. Fagundes, B.C. Teixeira, A.D. Chiavegatto Filho, Machine learning for hypertension prediction: a systematic review, Curr. Hypertens. Rep. 24 (11) (2022) 523–533.
[72] N. Jagan Mohan, R. Murugan, T. Goel, Deep learning for diabetic retinopathy detection: Challenges and opportunities, in: Next Generation Healthcare Informatics, Springer, 2022, pp. 213–232.
[73] P. Jamuna Devi, B. Kavitha, Data leakage and data wrangling in machine learning for medical treatment, in: Data Wrangling: Concepts, Applications and Tools, Wiley Online Library, 2023, pp. 91–107.
[74] C.L.A. Navarro, J.A. Damen, T. Takada, S.W. Nijman, P. Dhiman, J. Ma, G.S. Collins, R. Bajpai, R.D. Riley, K.G. Moons, et al., Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review, BMJ 375 (2021).
[75] Z. Xie, O. Nikolayeva, J. Luo, D. Li, Peer reviewed: building risk prediction models for type 2 diabetes using machine learning techniques, Prev. Chronic Dis. 16 (2019).
[76] C.A. James, K.M. Wheelock, J.O. Woolliscroft, Machine learning: the next paradigm shift in medical education, Acad. Med. 96 (7) (2021) 954–957.
[77] M. Rowe, An introduction to machine learning for clinicians, Acad. Med. 94 (10) (2019) 1433–1436.