Proceedings of Machine Learning Research LEAVE UNSET:1–17, 2024 Submitted LEAVE UNSET; Published LEAVE UNSET

Enhancing Fairness and Accuracy in Diagnosing Type 2 Diabetes in Young Population

Tanmoy Sarkar Pias TANMOYSARKAR@VT.EDU
Computer Science, Virginia Tech, Blacksburg, Virginia, USA

Yiqi Su YIQISU@VT.EDU
Computer Science, Virginia Tech, Blacksburg, Virginia, USA

Xuxin Tang XUXINTANG@VT.EDU
Computer Science, Virginia Tech, Blacksburg, Virginia, USA

Haohui Wang HAOHUIW@VT.EDU
Computer Science, Virginia Tech, Blacksburg, Virginia, USA

Shahriar Faghani FAGHANI.SHAHRIAR@MAYO.EDU
Radiology, Mayo Clinic, Rochester, USA

Danfeng (Daphne) Yao DANFENG@VT.EDU
Computer Science, Virginia Tech, Blacksburg, Virginia, USA

Abstract

While type 2 diabetes is predominantly found in the elderly population, recent publications indicates an increasing prevalence in the young adult population. Failing to predict it in the minority younger age group could have significant adverse effects on their health. The previous work acknowledges the bias of machine learning models towards different gender and race groups and proposes various approaches to mitigate it. However, prior work has not proposed any effective methodologies to predict diabetes in the young population which is the minority group in the diabetic population. In this paper, we identify this deficiency in traditional machine learning models and implement double prioritization (DP) bias correction techniques to mitigate the bias towards the young population when predicting diabetes. Deviating from the traditional concept of one-model-fits-all, we train customized machine-learning models for each age group. The DP model consistently improves recall of diabetes class by 26% to 40% in the young age group (30-44). Moreover, the DP technique outperforms 7 commonly used whole-group sampling techniques such as random oversampling, SMOTE, and AdaSyns techniques by at least 36% in terms of diabetes recall in the young age group. We also analyze the feature importance to investigate the source of bias in the original model.

Data and Code Availability We use a publicly available dataset called Behavioral Risk Factor Surveillance System (BRFSS) from 2021 CDC. To reproduce the result, the anonymised code has been attached as supplementary files. The code will be uploaded to a public repository upon publication.

Institutional Review Board (IRB) Our research does not require IRB approval.

1. Introduction

Predicting chronic diseases like diabetes is crucial, given the substantial global burden of diabetes-related complications and deaths, particularly in low- and middle-income countries De Silva et al. (2020); WHO. At the present rate of expansion, the International Diabetes Federation predicts that by 2045, a staggering 693 million individuals globally will be affected by diabetes Cho et al. (2018). The prevalence of diabetes, specifically type 2 diabetes, has been steadily increasing over the past few decades Krug (2016); Xie et al. (2019). While diabetes has traditionally been associated with the elderly population, recent studies indicate a rising prevalence of diabetes among the younger population as well Lascar et al. (2018). According to CDC, by 2060 type 2 diabetes cases might increase by about 70-700% in the young population CDC (2023). Young adults

© 2024 T.S. Pias, Y. Su, X. Tang, H. Wang, S. Faghani & D.(. Yao.

with diabetes face a significantly elevated risk of early health complications and even premature death compared to their counterparts without diabetes CDC (2022). Younger individuals diagnosed with diabetes face an increased risk of developing early and severe complications, encompassing both microvascular (retinopathy, neuropathy, ulceration, nephropathy) and macrovascular (cardiovascular, cerebrovascular, peripheral vascular) diseases McCoy et al. (2019); UK; NIH (2021).

Machine learning (ML) has been increasingly integrated into the healthcare system Lin (2020); Liaw (2020); Davenport (2018) because of its potential to assist clinicians and medical doctors in taking better care of patients. Accordingly, machine learning models have proved to be successful in detecting diabetes Xie et al. (2019). However, the health data can be imbalanced, which could potentially lead machine learning models to learn patterns with existing bias from the provided data Ninareh Mehrabi (2021); Ghassemi (2020); Juhn Y (2020); Zhang et al. (2022); Gervasi et al. (2022); Chen et al. (2018). Human biases, if not addressed, can exacerbate and perpetuate inequalities in the performance of algorithms in different subgroups Xu et al. (2022); Chen (2022); Seyyed-Kalantari et al. (2020), particularly in historically underserved populations like female patients Chen et al. (2019), black patients, or those with low socioeconomic status Seyyed-Kalantari et al. (2021). Moreover, AI models existing in the healthcare domain can show faithfulness issues Xie et al. (2023). It is of high importance that we evaluate the machine learning model before deployment ensuring social fairness Chen et al. (2021a); McDermott et al. (2023).

This paper identifies that the traditional machine learning models, trained on imbalance BRFSS (with only 15% representing the diabetes population) dataset CDC, tend to misdiagnose diabetes more frequently in the younger population (30-44 years) compared to other subgroups. The recall of the positive class (Rec_C1) in the 30-34 age group is a mere 30%, whereas the positive class (Rec_C1) is 68-72% in the gender group and 66-84% in the ethnic group. Moreover, we bring attention to the fact that solely relying on AU_ROC can be misleading – while the overall group's AU_ROC is 82%, the age group 30-34 exhibits a seemingly high 84% AU_ROC that masks the poor diabetes detection rate (recall C1) within that specific age group.

We introduce a novel application of double prioritized bias correction technique from a prior work Afrose et al. (2022) and successfully mitigated bias from young groups. To the best of our knowledge, this technique to mitigate bias from the young diabetic population has never been used previously.

Our major contributions are:

1. Our study identified significant imbalances in the young diabetic group in the BRFSS datasets which is the major cause of the machine learning models' poor detection performance in the young group
2. We applied 7 different sampling strategies such as SMOTE, AdaSyns, Random oversampling, Random undersampling, Near Miss, Edit distance, and Tomek-Links and showed that these whole group-based sampling techniques failed to mitigate the bias from the young group.
3. We demonstrated that using whole group-based performance metrics and subgroup-based AUROC can be misleading. We analyze the importance of using recall of class 1 when determining the performance.
4. We utilized the double prioritized bias correction technique from the prior work Afrose et al. (2022) and successfully mitigated bias from young groups.

In this study, we train diverse machine learning models, incorporating a transparent model like logistic regression, to uncover the roots of bias and missed detections. This is achieved through a thorough analysis of association coefficients responsible for shaping predictions.

2. Methods

2.1. Dataset

The Behavioral Risk Factor Surveillance System (BRFSS) is an annual health survey conducted in the United States to monitor various health behaviours such as cardiovascular diseases, chronic diseases, diabetes, obesity and other risk factors that contribute to the leading causes of death and disability CDC. The BRFSS datasets have been collected from all 50 states in the U.S. and the District of Columbia and included responses from over 400,000 participants which makes it one of the largest publicly available datasets related to public health. Each record

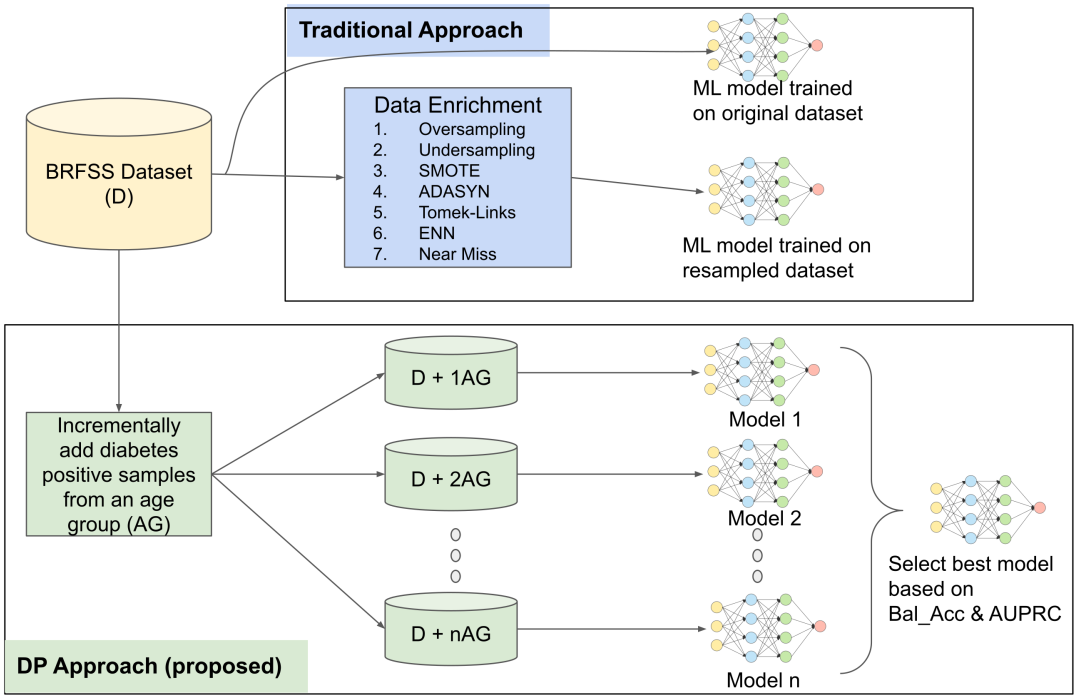


Figure 1: The double prioritized (DP) bias correction approach in contrast to traditional approaches, enriches the minority age groups. DP creates new datasets by replicating diabetic samples from a minority age group. n machine learning models are trained on each of the n datasets. The best model is selected based on performance metric balance accuracy (Bal_Acc) and area under precision and recall curve (AUPRC)

contains an individual's BRFSS survey responses on various health behaviours and risk factors such as tobacco use, physical activity, alcohol consumption, existing chronic diseases, and mental health. The survey also gathered demographic information such as age, gender, and race/ethnicity, which are helpful to explore important correlations and even causation.

This survey is conducted every year, and the CDC make it publicly available for research. In this study, we selected the BRFSS dataset from 2021, which contains more than 438,000 subject information and 330+ attributes from each subject. The BRFSS dataset is a valuable resource for identifying health disparities and evaluating the effectiveness of public health programs and policies. However, the dataset is not free from challenges because a large portion of attribute values are missing (25%) and this dataset can be a highly imbalanced data imbalance (diabetes class 15%).

2.2. Data Preprocessing

According to the literature on the risk factors for diabetes, Collins et al. (2011); Xie et al. (2019), we selected 30 variables out of 300+ variables including diabetes. The respondents who were under the age of 30 were excluded from the study, as they were more likely to have type 1 diabetes Xie et al. (2019); Sullivan et al. (2005); Noble et al. (2011). We create age groups spanning 5 years starting from 30 up to 80+ years. The respondents who showed diabetes during pregnancy and who had prediabetes were excluded from the study as well. If the respondent refused to answer (or missing answers) any questions related to 30 selected variables, were also excluded.

We discarded 113,000 subject information which contains at least one missing value and 125,000 subjects who refused to answer at least one questionnaire. This leaves with 200,000 subject information where approximately 169,000 subjects are diabetes negative and 30,000 subjects are positive.

As we mentioned before, this dataset is not balanced for the age, gender, and racial subgroups. In the age group 30-34, the number of negative cases is 10,864 but the number of positive cases is only 324. The age groups 35-39 and 40-44 are also highly imbalanced as the positive-negative ratio is only 4.4% and 7.2%.

The selected variables fall into three distinct types: nominal, ordinal, and binary. Nominal variables lack any inherent order, ordinal variables possess a meaningful order, and binary variables exclusively hold two distinct values. For example, the presence of high blood pressure, cholesterol, or heart disease can be represented by a binary variable. BMI category, education level, and income level are considered ordinal variables. On the other hand, marital status and race are nominal variables. Binary variables are represented using a binary encoding, where 1 signifies a positive outcome, and 0 represents a negative outcome. Ordinal variables are encoded using integers, preserving their meaningful order. Nominal variables are transformed into one-hot encoding for appropriate representation. The selected variables and their categories are listed in Appendix Table 1. The distribution of all selected variables are plotted in the Appendix Figure 10.

2.3. Double prioritized bias correction

The double prioritized (DP) bias correction method prioritizes a specific subgroup, the young age group in this case, that suffers from data imbalance. DP model incrementally replicates data points of the minority class (diabetes positive indicated by class 1) and chooses an optimal unit of replication based on the model performance.

This technique replicates all samples of diabetes class from the young population up to n times. Each time the duplicated sample units are merged with the original dataset, including the original unit. In our experiment, we set n to 30 and selected the best n empirically based on model performance.

Each training set containing the original set and the duplicated n (1 to 30) units is used to train a single model. In this way, we train 31 models and select the best-performing model based on balanced accuracy and the area under the curve (AUC) of the minority class (C1) precision-recall curve, denoted by PRC_C1. We identify the top three machine learning models with the highest balanced accuracy values and select the model that gives the highest PR_C1 for that particular age group. This selected model is used only for predicting diabetes for that particular group e.g., the DP model trained with duplicated age-group 30-34 samples is used to predict new patients from age-group 30-34 years.

2.4. Sampling Algorithms

A frequently employed technique for addressing the challenges of data imbalance is the utilization of sampling methods. To evaluate the effectiveness of the sampling algorithms detecting diabetes in young groups, we first compared multiple sampling algorithms, including (1) random oversampling, randomly duplicating instances from the minority class Van Hulse et al. (2007); (2) random undersampling, randomly removing instances from the majority class Van Hulse et al. (2007); (3) SMOTE (Synthetic Minority Over-sampling Technique), creating synthetic samples for the minority class by interpolating between existing instances Chawla et al. (2002); (4) ADASYN (Adaptive Synthetic Sampling), similar to SMOTE, but generating more synthetic samples for difficult-to-learn instances He et al. (2008); (5) Tomek Links, removing instances from the majority class that are close to instances in the minority class tom (1976); (6) ENN (Edited Nearest Neighbors), removing instances from the majority class that are misclassified based on the nearest neighbors from both classes Wilson (1972); and (7) NearMiss, selecting instances from the majority class based on their distance to instances in the minority class Mani and Zhang (2003). These sampling techniques have been proven to be effective in the literature, however, these techniques are not tailored to tackle subgroup biases. As a result, the diabetes is mis-detected in the young population. Therefore, we utilize new concept of using one model for a single group which deviates from the traditional one model-fits-all ideology. Moreover, we combine this technique with a double prioritized bias correction technique Afrose et al. (2022) where duplicate minority samples are controlled empirically.

2.5. Machine Learning Models

We selected three commonly used ML models including logistic regression (LR) Peng et al. (2002), random forest (RF) Breiman (2001), Adaboost, and multilayer perceptron (MLP) Gardner and Dorling (1998) to evaluate the effectiveness of the sampling strategy. We purposely selected simple models such

as logistic regression because of their interpretability. It is very important for the science community, especially for healthcare to create interpretable models to find out the root cause of a prediction, however, McDermott et al. (2021) evaluated 511 scientific papers across different ML domains and identified a notable deficiency in reproducibility metrics, including dataset and code accessibility in clinical ML domain papers. Moreover, complex and high-end models which have recently been applied to the healthcare domain might pose difficulties in reproducibility Chen et al. (2021b). Using simple models like logistic regression will pave the way both for explaining the results and make it easy for other researchers to reproduce the model and the same results.

In the RF, we selected 100 trees or estimators to achieve optimal performance. In our MLP configuration, we used a single hidden layer with 100 ReLU-activated neurons and optimized training with the Adam gradient descent optimizer for efficiency and effectiveness. We repeated the training process at least five times and ensured consistent performance (standard deviation less than 2 of all experiments).

For LR, we utilized the liblinear equation to establish the decision boundary between positive and negative classes. A LR -trained model can be effectively employed to elucidate the underlying influence of each attribute within the dataset. This analysis can provide valuable insights into feature importance, offering a comprehensive understanding of how each variable contributes to the model's predictions. We investigate the features responsible for biased prediction from the original model. The positive coefficient represents the positive correlation of the corresponding feature with the prediction and the negative coefficient means negative correlation. The value of the coefficient represents the strength of the correlation. The logistic regression estimates the probability from the coefficients and the corresponding feature value which makes it a white box and fully interpretable. The coefficients of one-hot encoded features are averaged because only one option of the one-hot encoded feature contributes to the prediction while other options are zeroed out.

To examine the potential unfairness in the datasets, we calculated class-based accuracy, recall, precision, F1-score, AU_ROC, AU-PRC, and Matthews correlation coefficient (MCC). Nonetheless, our primary focus lies in prioritizing the recall of the positive class, also known as sensitivity, as the detection of diabetes holds paramount importance in our specific case. The dataset undergoes a random division into training (70%), validation sets (15%) and testing sets (15%) through the utilization of the Python sklearn library. The training dataset serves as the foundation for model training, while the testing dataset remains consistent across all the experiments, ensuring the robustness and comparability of our results. The performance of the models in each experiment is recorded as the average value of three independent trials. All the ML models were developed in Python and executed in the Google Colab platform.

2.6. Model Calibration and Threshold Tuning

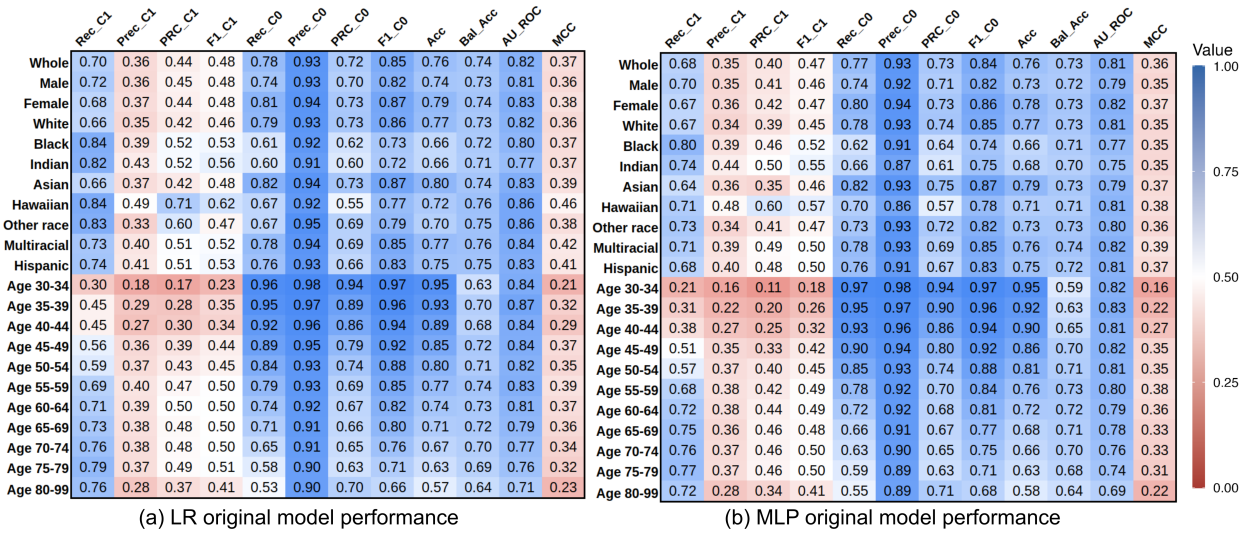
Before mapping probabilities into labels, we calibrate the predicted probabilities for each model on the validation set. Calibration involves adjusting the distribution of probabilities to enhance accuracy. We calibrate the output probabilities using the Isotonic Regression technique. We then perform threshold tuning to find the optimal threshold based on balanced accuracy and the F1_C1 score. Threshold tuning is the process of selecting an optimal threshold value in binary classification models. In these models, predictions are typically expressed as probabilities, and a threshold is applied to determine the class label. Adjusting this threshold can impact the trade-off between false positives and false negatives.

At first, we identify the top 3 thresholds that produce the highest F1_C1 scores and then we finally select the optimal threshold that generates the highest balanced accuracy for all samples in the validation dataset. For some subgroups, only a few samples (¡100) are present in the validation set. Selecting the threshold based on subgroup samples may overfit the validation set. So, we select a threshold based on the whole validation performance. We calculated the optimal threshold to be 0.195 from 10 independent trials by averaging the best thresholds from each trial.

3. Result

3.1. Performance of Original Model

The machine learning models trained on the original training set show different performances for the minority diabetes-positive group (C1) and the majority-negative group (C0). Figure 2 shows the complete performance metrics of the original logistic regression model. Whole group recall C1 is 0.70 whereas recall



	Rec_C1	Prec_C1	PRC_C1	F1_C1	Rec_C0	Prec_C0	PRC_C0	F1_C0	Acc	Bal_Acc	AU_ROC	MCC
Whole	0.70	0.36	0.44	0.48	0.78	0.93	0.72	0.85	0.76	0.74	0.82	0.37
Male	0.72	0.36	0.45	0.48	0.74	0.93	0.70	0.82	0.74	0.73	0.81	0.36
Female	0.68	0.37	0.44	0.48	0.81	0.94	0.73	0.87	0.79	0.74	0.83	0.38
White	0.66	0.35	0.42	0.46	0.79	0.93	0.73	0.86	0.77	0.73	0.82	0.36
Black	0.84	0.39	0.52	0.53	0.61	0.92	0.62	0.73	0.66	0.72	0.80	0.37
Indian	0.82	0.43	0.52	0.56	0.60	0.91	0.60	0.72	0.66	0.71	0.77	0.37
Asian	0.66	0.37	0.42	0.48	0.82	0.94	0.73	0.87	0.80	0.74	0.83	0.39
Hawaiian	0.84	0.49	0.71	0.62	0.67	0.92	0.55	0.77	0.72	0.76	0.86	0.46
Other race	0.83	0.33	0.60	0.47	0.67	0.95	0.69	0.79	0.70	0.75	0.86	0.38
Multiracial	0.73	0.40	0.51	0.52	0.78	0.94	0.69	0.85	0.77	0.76	0.84	0.42
Hispanic	0.74	0.41	0.51	0.53	0.76	0.93	0.66	0.83	0.75	0.75	0.83	0.41
Age 30-34	0.30	0.18	0.17	0.23	0.96	0.98	0.94	0.97	0.95	0.63	0.84	0.21
Age 35-39	0.45	0.29	0.28	0.35	0.95	0.97	0.89	0.96	0.93	0.70	0.87	0.32
Age 40-44	0.45	0.27	0.30	0.34	0.92	0.96	0.86	0.94	0.89	0.68	0.84	0.29
Age 45-49	0.56	0.36	0.39	0.44	0.89	0.95	0.79	0.92	0.85	0.72	0.84	0.37
Age 50-54	0.59	0.37	0.43	0.45	0.84	0.93	0.74	0.88	0.80	0.71	0.82	0.35
Age 55-59	0.69	0.40	0.47	0.50	0.79	0.93	0.69	0.85	0.77	0.74	0.83	0.39
Age 60-64	0.71	0.39	0.50	0.50	0.74	0.92	0.67	0.82	0.74	0.73	0.81	0.37
Age 65-69	0.73	0.38	0.48	0.50	0.71	0.91	0.66	0.80	0.71	0.72	0.79	0.36
Age 70-74	0.76	0.38	0.48	0.50	0.65	0.91	0.65	0.76	0.67	0.70	0.77	0.34
Age 75-79	0.79	0.37	0.49	0.51	0.58	0.90	0.63	0.71	0.63	0.69	0.76	0.32
Age 80-99	0.76	0.28	0.37	0.41	0.53	0.90	0.70	0.66	0.57	0.64	0.71	0.23

(a) LR original model performance

	Rec_C1	Prec_C1	PRC_C1	F1_C1	Rec_C0	Prec_C0	PRC_C0	F1_C0	Acc	Bal_Acc	AU_ROC	MCC
Whole	0.68	0.35	0.40	0.47	0.77	0.93	0.73	0.84	0.76	0.73	0.81	0.36
Male	0.70	0.35	0.41	0.46	0.74	0.92	0.71	0.82	0.73	0.72	0.79	0.35
Female	0.67	0.36	0.42	0.47	0.80	0.94	0.73	0.86	0.78	0.73	0.82	0.37
White	0.67	0.34	0.39	0.45	0.78	0.93	0.74	0.85	0.77	0.73	0.81	0.35
Black	0.80	0.39	0.46	0.52	0.62	0.91	0.64	0.74	0.66	0.71	0.77	0.35
Indian	0.74	0.44	0.50	0.55	0.66	0.87	0.61	0.75	0.68	0.70	0.75	0.35
Asian	0.64	0.36	0.35	0.46	0.82	0.93	0.75	0.87	0.79	0.73	0.79	0.37
Hawaiian	0.71	0.48	0.60	0.57	0.70	0.86	0.57	0.78	0.71	0.71	0.81	0.38
Other race	0.73	0.34	0.41	0.47	0.73	0.93	0.72	0.82	0.73	0.73	0.80	0.36
Multiracial	0.71	0.39	0.49	0.50	0.78	0.93	0.69	0.85	0.76	0.74	0.82	0.39
Hispanic	0.68	0.40	0.48	0.50	0.76	0.91	0.67	0.83	0.75	0.72	0.81	0.37
Age 30-34	0.21	0.16	0.11	0.18	0.97	0.98	0.94	0.97	0.95	0.59	0.82	0.16
Age 35-39	0.31	0.22	0.20	0.26	0.95	0.97	0.90	0.96	0.92	0.63	0.83	0.22
Age 40-44	0.38	0.27	0.25	0.32	0.93	0.96	0.86	0.94	0.90	0.65	0.81	0.27
Age 45-49	0.51	0.35	0.33	0.42	0.90	0.94	0.80	0.92	0.86	0.70	0.82	0.35
Age 50-54	0.57	0.37	0.40	0.45	0.85	0.93	0.74	0.88	0.81	0.71	0.81	0.35
Age 55-59	0.68	0.38	0.42	0.49	0.78	0.92	0.70	0.84	0.76	0.73	0.80	0.38
Age 60-64	0.72	0.38	0.44	0.49	0.72	0.92	0.68	0.81	0.72	0.72	0.79	0.36
Age 65-69	0.75	0.36	0.46	0.48	0.66	0.91	0.67	0.77	0.68	0.71	0.78	0.33
Age 70-74	0.76	0.37	0.46	0.50	0.63	0.90	0.65	0.75	0.66	0.70	0.76	0.33
Age 75-79	0.77	0.37	0.46	0.50	0.59	0.89	0.63	0.71	0.63	0.68	0.74	0.31
Age 80-99	0.72	0.28	0.34	0.41	0.55	0.89	0.71	0.68	0.58	0.64	0.69	0.22

(b) MLP original model performance

Figure 2: Performance of (a) Logistic Regression and (b) MLP model trained and tested on the original 2021 BRFSS dataset. The x-axis represents performance metrics where Rec, Prec, PRC, F1, Acc, Bal_Acc, ROC, MCC represents Recall, Precision, Area Under the Precision-Recall Curve, F1 score, Accuracy, Balanced Accuracy, Area under the ROC Curve, Matthews Correlation Coefficient, respectively. C1 and C0 stand for class 1 (diabetes positive) and class 0 (diabetes negative) respectively. The Y-axis represents the whole population, gender groups, ethnic groups, and age groups.

C0 is 0.78. For both MLP and RF models, recall C1 is 0.68 and recall C0 is 0.77.

For the gender group, recall C1 is lower than recall C0. For males, recall_C1 is 0.72 and recall-C0 is 0.74. Female recall_C1 is 0.68 and recall-C0 is 0.81. The ethnic groups such as White, Asian, and multicultural show recall_C1 of 0.66, 0.66, and 0.73 respectively which is comparatively lower than recall-C0. On the other hand, Black, Indian, and Hawaiian show recall_C1 of 0.84, 0.82, and 0.84 respectively which is higher than the corresponding group recall-C0.

However, the young age group suffers the most from missed detection of diabetes. For age groups [30-34], [35-39], and [40-44] recall C1 is only 0.30, 0.45, and 0.45 respectively. This means out of diabetic 100 patients 70 patients are mis-detected. On the other hand, these three age groups show very high recall C0. The Adaboost and random forest models also show similar performance, represented in figure 9.

In imbalanced datasets, commonly employed metrics like AUC_ROC and accuracy can be misleading and do not accurately represent the performance of the minority class. Despite potentially poor performance in the minority class, these metrics might indicate falsely elevated values. The AU_ROC of the whole population, age groups [30-34], [35-39], and [40-44] are 0.82, 0.84, 0.87, and 0.84. However, these age groups show very poor recall_C1, F1_C1, and MCC which is not reflected with AU_ROC.

In contrast to the AU_ROC and accuracy metrics which can be overly optimistic, we use Recall_C1 to measure the true detection rate reflecting the performance of the model.

3.2. DP model improves detection accuracy

The DP model for three age groups improves the diabetes detection accuracy significantly. Figure 4 shows the performance difference between the DP model and the original model in terms of positive recall and balanced accuracy. Diff Rec_C1 means subtracting the recall of class 1 of the original model from the DP model. DP model for age groups [30-34], [35-39], and [40-44] improves the positive recall by 41% (SD 6.4%), 32% (SD 6.4%), and 24% (SD 2.8%) respectively. This means the DP models are better at detecting diabetes in the young population.

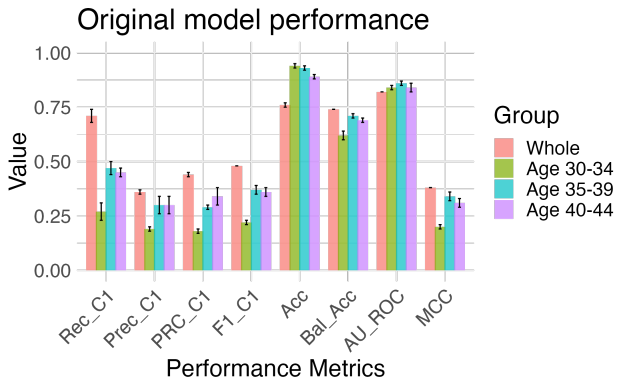


Figure 3: Performance of original Logistic Regression model applied to the whole population and minority age group.

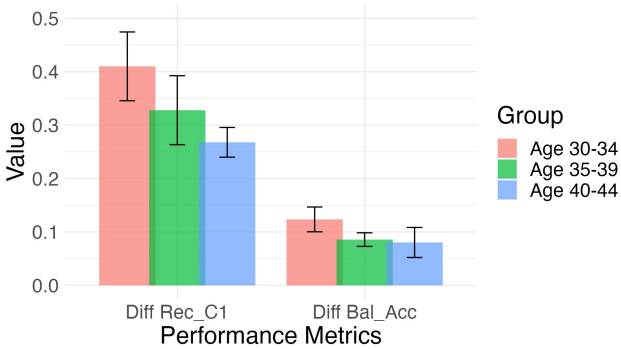


Figure 4: Performance difference DP and original model Logistic Regression model. "Diff Rec_C1" means subtracting the recall of class 1 of the original model from the DP model and "Diff Bal_Acc" means subtracting the balanced accuracy of the original model from the DP model. Positive values indicate performance improvement from the original model. The error bars represent the standard deviation of the experiment results.

On the other hand, Diff Bal_Acc means subtracting the balanced accuracy of the original model from the DP model. The balanced accuracy (average recall of both C0 and C1 classes) is also improved by 13% (SD 2.3%), 10.5% (SD 1%), and 7.7% (SD 2.8%) for the DP model.

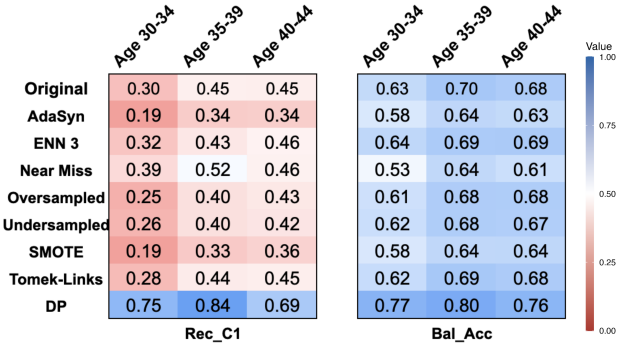


Figure 5: Comparing whole group sampling method performance with DP using class 1 recall (left sub-figure) and Balanced accuracy (right sub-figure). In each sub-figure, each row corresponds to a sampling method. Each column represents a subgroup that a model is evaluated on. The last rows show the group's performance in their corresponding DP models.

3.3. Whole-population sampling

The whole group-based sampling approach doesn't improve the detection rate in the young group Figure 5. Moreover, AdaSyn (Rec_C1 0.19), SMOTE (Rec_C1 0.19), Tomek-Link (Rec_C1 0.28), Random oversampling (Rec_C1 0.25) and undersampling (Rec_C1 0.26) methods decrease the original detection accuracy (Rec_C1 0.30) for age group [30-34]. A similar performance decrease is observed in the other two age groups [35-39] and [40-44]. Near Miss and ENN merely improve the recall C1 by 9% and 2%. On the other hand, the respective DP models (i.e. trained on age group [30-34] and applied to age group [30-34]) improve the positive recall by at least 24%.

The whole group sampling methods also show poor performance in terms of balanced accuracy. For example, for the age group [35-39] all of the methods decrease the balanced accuracy from the original value. On the other hand, the DP model increases balanced accuracy by at least 10%. The performance of all the mentioned sampling algorithms is represented in the supplementary figure 9.

3.4. Cross-group performance

DP model trained for a specific age group is applied to all other age groups Figure 6. For example, a DP

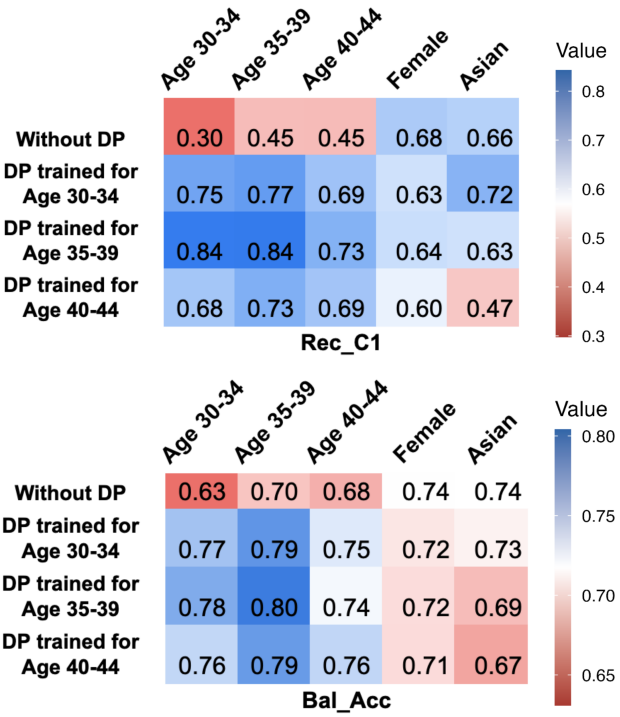


Figure 6: Cross-group performance analysis using class 1 recall and Balanced accuracy. In each subfigure, each row corresponds to a DP model trained for a specific subgroup. Each column represents a subgroup that a model is evaluated on. The first rows show the group's performance in the original model.

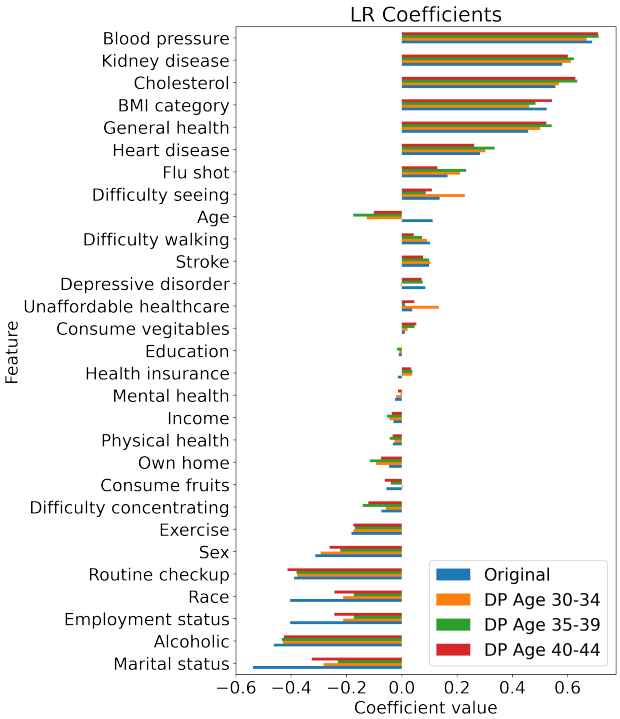


Figure 7: Coefficient values associated with each feature from the original model and DP models. The one hot encoded feature (Marital status, Employment status, and Race) coefficients are averaged.

model trained for age group [30-34], is tested on age groups [30-34], [35-39], and [40-44]. We also compare the whole group's performance with the DP models. The original model shows very poor recall C1 for all three age groups. Interestingly, the DP model trained for the age group [30-34] can also be applied to age groups [35-39] and [40-44]. DP model for age group [35-39] shows the highest recall C1 of 73% - 84% for all age groups. The balanced accuracy is also high when one dp model is applied to another age group.

On the other hand, DP model trained for the young age group, shows poor performance when applied for gender or ethnic group for which the dp model is not trained for. The recall of class 1 goes down by 8% and 19% when the DP model for age group [40-44] is applied to the female and Asian group respectively. The balanced accuracy also declines if age group DP models are applied to female or Asian group.

3.5. Feature analysis

One of the major motivations for selecting a logistic regression model is its interpretability. Beam (2021) showed that deep learning models are mostly black-box and cannot always be correctly interpretable.

Figure 7 shows the logistic regression model coefficients of the original and the DP models. It shows that the original model had a strong correlation with the subject's race (-0.40), employment status (-0.40), and marital status (-0.54). The DP model reduces the strength of these attributes by at least 19%. Moreover, the age was positively correlated in the original model while it gets negatively correlated in the DP models. Other attribute coefficients show minor changes.

4. Related Work

4.1. BRFSS Dataset

In recent years, BRFSS has been extensively used in ML studies in healthcare due to its high quality, comprehensive information, large sample size, and national representation. Althoff et al. (2022) demonstrated disparate associations between food environment and diet using the BRFSS dataset. Jain (2010) developed regression models to predict height, weight and obesity. Edwards et al. (2017) examined differences between male and female caregivers by demographics and health-related quality of life. Oladayo et al. (2023) used ML algorithms to investigate factors associated with complete edentulism among older adults in the United States. Ullah et al. (2022) detected high-risk factors and early diagnosis of diabetes. The BRFSS dataset has been used for many prediction and association tasks Quiñones et al. (2021); Mamun et al. (2022); Fang et al. (2021); Nguyen et al. (2016); Ekpo et al. (2023).

4.2. Data Imbalance

One of the major challenges in ML in healthcare is data imbalance, where the number of instances in each class is significantly different He and Garcia (2009); Li et al. (2017); Razzaghi et al. (2016); Chiu et al. (2021); Matsuo et al. (2020); Yang et al. (2022). This can lead to biased models that perform poorly on underrepresented classes. Various techniques have been proposed to address data imbalance in healthcare, such as oversampling, undersampling, and cost-sensitive learning.

Several studies have shown the effectiveness of these techniques in improving the performance of ML models in healthcare. For example, Goh et al. (2021); Rezapour et al. (2023); Rezapour and Hansen (2022) used oversampling to balance an EHR dataset for predicting heart failure, resulting in improved model performance. Goyal (2022); Idris et al. (2019); Thomas and Bell (2003); Cartus et al. (2020); Liu et al. (2008) used undersampling to balance a dataset to achieve higher accuracy in various domains.

4.3. Fairness

In healthcare, addressing fairness concerns in machine learning is vital due to the potential for models to perpetuate biases and discrimination, impacting equitable access to diagnosis and treatment. Several approaches have been proposed to address fairness in ML in healthcare. Meng et al. (2022) introduced the interpretability of ML in predicting hospital mortality from the MIMIC-IV dataset Johnson et al. (2023). Afrose et al. (2022) introduced double prioritized bias correction to improve the fairness of the ML model in predicting in-hospital mortality risk using the MIMIC-III dataset Johnson et al. (2016). Grote and Keeling (2022) provide the building blocks for an account of algorithmic bias and its normative relevance in medicine. AI are able to predict race from medical images Gichoya et al. (2022) or notes Adam et al. (2022). A recent study Ricci Lara et al. (2022) analyzes the current state of the field, identifying strengths and highlighting areas of vacancy, challenges and opportunities regarding the fairness issue for medical imaging. However, none of the previous machine learning work acknowledged the poor diabetes detection accuracy in the young population. In this paper, we propose an effective application of the DP algorithm to solve the problem by improving the model for the young group.

5. Discussion

The machine learning trained on the original BRFSS dataset shows poor diabetes prediction performance in the young group. Because the diabetes data is limited in the original dataset the traditional machine learning model tends to pick the general statistics to build a prediction model. As a result, the models show poor performance in the minor young group. Whole group-based data enrichment such as sampling methods cannot overcome the problem of poor performance as it doesn't enhance the young group. Some of them such as SMOTE, a popular sampling method which is well known for removing bias from imbalanced datasets, decrease the performance. This is because these sampling methods are not equipped to reduce disparate ratios in the minority group. Moreover, we tested multiple and diverse machine learning models such as logistic regression, multilayer perceptron, adaboost, and random forest. However, all the models show consistently poor performance in this scenario.

Finally, we utilized a recently proposed method called the double prioritized bias correction technique (DP) to make the model effective and useful for the young age group. By replicating the minor group population dynamically, the technique improves the model's performance. However, one DP model can be

targeted for one particular group. The results show that the DP model trained for the young age group is not applicable for both gender or ethnic groups. However, this limitation can easily be overcome by using multiple DP models for each minor group.

We also investigated the root cause of bias in the original model by visualizing the coefficients of the logistic regression model. Being a white box model, we can easily understand the feature importance and the correlation of each feature with the prediction probability. The original model shows a strong correlation with non-medical attributes such as marital status, employment status, and race. The DP model decreases the strength of correlation with these factors significantly. Moreover, the original model is positively correlated with age. It means the diabetes-positive probability increases with age. However, this positive correlation was one of the key factors why the original shows poor performance in the young age group.

In conclusion, we identified a major deficiency in the traditional machine learning and commonly used sampling approach when it comes to diagnosing diabetes in the young age group which is the minority population in this case. We successfully utilized the DP method to overcome this issue and improved detection performance significantly. We strongly believe this approach will be effective in correcting bias issues in diabetes diagnostics and will pave the way for effective precision medicine.

References

Two modifications of cnn. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-6(11):769–772, 1976. doi: 10.1109/TSMC.1976.4309452.

Hammaad Adam, Ming Ying Yang, Kenrick Cato, Ioana Baldini, Charles Senteio, Leo Anthony Celi, Jiaming Zeng, Moninder Singh, and Marzyeh Ghassemi. Write it like you see it: Detectable differences in clinical notes by race lead to differential model recommendations. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 7–21, 2022.

Sharmin Afrose, Wenjia Song, Charles B Nemeroff, Chang Lu, and Danfeng Yao. Subpopulation-specific machine learning prognosis for underrepresented patients with double prioritized bias correction. *Communications medicine*, 2(1):111, 2022.

Tim Althoff, Hamed Nilforoshan, Jenna Hua, and Jure Leskovec. Large-scale diet tracking data reveal disparate associations between food environment and diet. *Nature communications*, 13(1):267, 2022.

Ghassemi M Oakden-Rayner L Beam. Al the false hope of current approaches to explainable artificial intelligence in health care lancet digit. *Health*, 3 (11):e745, 2021.

Leo Breiman. Random forests. *Machine learning*, 45: 5–32, 2001.

Abigail R Cartus, Lisa M Bodnar, and Ashley I Naimi. The impact of undersampling on the predictive performance of logistic regression and machine learning algorithms: a simulation study. *Epidemiology (Cambridge, Mass.)*, 31(5):e42, 2020.

CDC. CDC - BRFSS — cdc.gov. https://www.cdc.gov/brfss/. [Accessed 23-Apr-2023].

CDC. Making the leap from type 1 teen to adult, Jan 2022. URL https://www.cdc.gov/diabetes/basics/type-1-juvenile-to-adult.html#:~:text=Teens%20and%20young%20adults%20with.

CDC. Diabetes in young people is on the rise, Feb 2023. URL https://www.cdc.gov/diabetes/resources-publications/research-summaries/diabetes-young-people.html.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

Irene Chen, Fredrik D Johansson, and David Sontag. Why is my classifier discriminatory? *Advances in neural information processing systems*, 31, 2018.

Irene Y Chen. *Machine Learning Approaches for Equitable Healthcare*. PhD thesis, Massachusetts Institute of Technology, 2022.

Irene Y Chen, Peter Szolovits, and Marzyeh Ghassemi. Can ai help reduce disparities in general medical and mental health care? *AMA journal of ethics*, 21(2):167–179, 2019.

Irene Y Chen, Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, and Marzyeh Ghassemi. Ethical machine learning in healthcare. *Annual review of biomedical data science*, 4:123–144, 2021a.

Irene Y Chen, Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, and Marzyeh Ghassemi. Ethical machine learning in healthcare. *Annual review of biomedical data science*, 4:123–144, 2021b.

Ting-Wei Chiu, Yu-Lin Tsai, and Shun-Feng Su. Automatic detect lung node with deep learning in segmentation and imbalance data labeling. *Scientific Reports*, 11(1):11174, 2021.

Nam H Cho, JE Shaw, Suvi Karuranga, Yafang Huang, JD da Rocha Fernandes, AW Ohlrogge, and BIDF Malanda. Idf diabetes atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. *Diabetes research and clinical practice*, 138:271–281, 2018.

Gary S Collins, Susan Mallett, Omar Omar, and Ly-Mee Yu. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. *BMC medicine*, 9(1):1–14, 2011.

Hongsermeier-T. Mc Cord K. A. Davenport, T. H. Using ai to improve electronic health record. *Harvard Business Review*, 12, 2018.

Kushan De Silva, Wai Kit Lee, Andrew Forbes, Ryan T Demmer, Christopher Barton, and Joanne Enticott. Use and performance of machine learning models for type 2 diabetes prediction in community settings: A systematic review and meta-analysis. *International journal of medical informatics*, 143: 104268, 2020.

Valerie J Edwards, Lynda A Anderson, William W Thompson, and Angela J Deokar. Mental health differences between men and women caregivers, brfss 2009. *Journal of women & aging*, 29(5):385–391, 2017.

Raphael Henshaw Ekpo, Victor Chukwudi Osamor, Ambrose A Azeta, Excellent Ikeakanam, and Beatrice Opeyemi Amos. Machine learning classification approach for asthma prediction models in children. *Health and Technology*, pages 1–10, 2023.

Min Fang, Yingru Chen, Rui Xue, Huihui Wang, Nilesh Chakraborty, Ting Su, and Yuyan Dai. A hybrid machine learning approach for hypertension risk prediction. *Neural Computing and Applications*, pages 1–11, 2021.

Matt W Gardner and SR Dorling. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric environment*, 32(14-15):2627–2636, 1998.

Stephanie S Gervasi, Irene Y Chen, Aaron Smith-McLallen, David Sontag, Ziad Obermeyer, Michael Vennera, and Ravi Chawla. The potential for bias in machine learning and opportunities for health insurers to address it: Article examines the potential for bias in machine learning and opportunities for health insurers to address it. *Health Affairs*, 41 (2):212–218, 2022.

Naumann T.-Schulam P. Beam A. L. Chen I. Y. Ranganath R. Ghassemi, M. A review of challenges and opportunities in machine learning for health. *AMIA Summits on Translational Science Proceedings*, 200(01):191–200, 2020. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7233077/.

Judy Wawira Gichoya, Imon Banerjee, Ananth Reddy Bhimireddy, John L Burns, Leo Anthony Celi, Li-Ching Chen, Ramon Correa, Natalie Dullerud, Marzyeh Ghassemi, Shih-Cheng Huang, et al. Ai recognition of patient race in medical imaging: a modelling study. *The Lancet Digital Health*, 4(6):e406–e414, 2022.

Kim Huat Goh, Le Wang, Adrian Yong Kwang Yeow, Hermione Poh, Ke Li, Joannas Jie Lin Yeow, and Gamaliel Yu Heng Tan. Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. *Nature communications*, 12(1):711, 2021.

Somya Goyal. Handling class-imbalance with knn (neighbourhood) under-sampling for software defect prediction. *Artificial Intelligence Review*, 55 (3):2023–2064, 2022.

Thomas Grote and Geoff Keeling. On algorithmic fairness in medical practice. *Cambridge Quarterly of Healthcare Ethics*, 31(1):83–94, 2022.

Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pages 1322–1328. Ieee, 2008.

Adnan Idris, Aksam Iftikhar, and Zia Ur Rehman. Intelligent churn prediction for telecom using gp-adaboost learning and pso undersampling. *Cluster Computing*, 22(Suppl 3):7241–7255, 2019.

RB Jain. Regression models to predict corrected weight, height and obesity prevalence from self-reported data: data from brfss 1999–2007. *International Journal of Obesity*, 34(11):1655–1664, 2010.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3 (1):1–9, 2016.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Benjamin Moody, Brian Gow, Li-wei H Lehman, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10 (1):1, 2023.

Liu H Juhn Y. Artificial intelligence approaches using natural language processing to advance ehr-based clinical research. *J Allergy Clin Immunol*, 145(2): 463–469, 2020. URL https://doi:10.1016/j.jaci.2019.12.897.

Etienne G Krug. Trends in diabetes: sounding the alarm. *The Lancet*, 387(10027):1485–1486, 2016.

Nadia Lascar, James Brown, Helen Pattison, Anthony H Barnett, Clifford J Bailey, and Srikanth Bellary. Type 2 diabetes in adolescents and young adults. *The lancet Diabetes & endocrinology*, 6(1): 69–80, 2018.

Jinyan Li, Lian-sheng Liu, Simon Fong, Raymond K Wong, Sabah Mohammed, Jinan Fiaidhi, Yunsick Sung, and Kelvin KL Wong. Adaptive swarm balancing algorithms for rare-event prediction in imbalanced healthcare data. *PloS one*, 12(7): e0180830, 2017.

Liyanage H.-Kuziemsky C. et al. Liaw, S. Ethical use of electronic health record data and artificial intelligence: Recommendations of the primary care informatics working group of the international medical informatics association. *Yearbook of medical informatics*, 29(01):051–057, 2020. URL 10.1055/s-0040-1701980.

Chen-J. S. Chiang M. F. Hribar M. R. Lin, W. C. Applications of artificial intelligence to electronic health record data in ophthalmology. *Translational vision science technology*, 9(2):13–13, 2020. URL https://doi.org/10.1167/tvst.9.2.13.

Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2):539–550, 2008.

Muntasir Mamun, Md Milon Uddin, Vivek Kumar Tiwari, Asm Mohaimenul Islam, and Ahmed Ullah Ferdous. Mlheartdis: Can machine learning techniques enable to predict heart diseases? In *2022 IEEE 13th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, pages 0561–0565. IEEE, 2022.

Inderjeet Mani and I Zhang. knn approach to unbalanced data distributions: a case study involving information extraction. In *Proceedings of workshop on learning from imbalanced datasets*, volume 126, pages 1–7. ICML, 2003.

Hidetoshi Matsuo, Mizuho Nishio, Tomonori Kanda, Yasuyuki Kojita, Atsushi K Kono, Masatoshi Hori, Masanori Teshima, Naoki Otsuki, Ken-ichi Nibu, and Takamichi Murakami. Diagnostic accuracy of deep-learning with anomaly detection for a small amount of imbalanced data: discriminating malignant parotid tumors in mri. *Scientific Reports*, 10 (1):19388, 2020.

Rozalina G McCoy, Renée SM Kidney, Danette Holznagel, Tina Peters, and Vimbai Madzura. Challenges for younger adults with diabetes. *Minnesota medicine*, 102(2):34, 2019.

Matthew BA McDermott, Shirly Wang, Nikki Marinsek, Rajesh Ranganath, Luca Foschini, and Marzyeh Ghassemi. Reproducibility in machine learning for health research: Still a ways to go. *Science Translational Medicine*, 13(586):eabb1655, 2021.

Matthew BA McDermott, Bret Nestor, and Peter Szolovits. Clinical artificial intelligence: Design principles and fallacies. *Clinics in Laboratory Medicine*, 43(1):29–46, 2023.

Chuizheng Meng, Loc Trinh, Nan Xu, James Enouen, and Yan Liu. Interpretability and fairness evalua-

tion of deep learning models on mimic-iv dataset. *Scientific Reports*, 12(1):7166, 2022.

Quynh C Nguyen, Dapeng Li, Hsien-Wen Meng, Suraj Kath, Elaine Nsoesie, Feifei Li, and Ming Wen. Building a national neighborhood dataset from geotagged twitter data for indicators of happiness, diet, and physical activity. *JMIR public health and surveillance*, 2(2):e5869, 2016.

GOV NIH. Serious complications from youth-onset type 2 diabetes arise by young adulthood. *National Institutes of Health*, Jul 2021. URL https://www.nih.gov/news-events/news-releases/serious-complications-youth-onset-type-2-diabetes-arise-young-adulthood.

Nripsuta Saxena Kristina Lerman Aram Galstyan Ninareh Mehrabi, Fred Morstatter. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021. URL https://dl.acm.org/doi/abs/10.1145/3457607.

Douglas Noble, Rohini Mathur, Tom Dent, Catherine Meads, and Trisha Greenhalgh. Risk models and scores for type 2 diabetes: systematic review. *Bmj*, 343, 2011.

Abimbola M Oladayo, Hikkaduwa Withanage Miyuraj Harishchandra, Erliang Zeng, Daniel J Caplan, Azeez Butali, and Leonardo Marchini. Using machine learning algorithms to investigate factors associated with complete edentulism among older adults in the united states. *Special care in dentistry*, 2023.

Chao-Ying Joanne Peng, Kuk Lida Lee, and Gary M Ingersoll. An introduction to logistic regression analysis and reporting. *The journal of educational research*, 96(1):3–14, 2002.

Sarah Quiñones, Aditya Goyal, and Zia U Ahmed. Geographically weighted machine learning model for untangling spatial heterogeneity of type 2 diabetes mellitus (t2d) prevalence in the usa. *Scientific reports*, 11(1):6955, 2021.

Talayeh Razzaghi, Oleg Roderick, Ilya Safro, and Nicholas Marko. Multilevel weighted support vector machine for classification on healthcare data with missing values. *PloS one*, 11(5):e0155119, 2016.

Mostafa Rezapour and Lucas Hansen. A machine learning analysis of covid-19 mental health data. *Scientific Reports*, 12(1):14965, 2022.

Mostafa Rezapour, Muhammad Khalid Khan Niazi, and Metin Nafi Gurcan. Machine learning-based analytics of the impact of the covid-19 pandemic on alcohol consumption habit changes among united states healthcare workers. *Scientific Reports*, 13 (1):6003, 2023.

María Agustina Ricci Lara, Rodrigo Echeveste, and Enzo Ferrante. Addressing fairness in artificial intelligence for medical imaging. *nature communications*, 13(1):4581, 2022.

Laleh Seyyed-Kalantari, Guanxiong Liu, Matthew McDermott, Irene Y Chen, and Marzyeh Ghassemi. Chexclusion: Fairness gaps in deep chest x-ray classifiers. In *BIOCOMPUTING 2021: proceedings of the Pacific symposium*, pages 232–243. World Scientific, 2020.

Laleh Seyyed-Kalantari, Haoran Zhang, Matthew BA McDermott, Irene Y Chen, and Marzyeh Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in underserved patient populations. *Nature medicine*, 27 (12):2176–2182, 2021.

Patrick W Sullivan, Elaine H Morrato, Vahram Ghushchyan, Holly R Wyatt, and James O Hill. Obesity, inactivity, and the prevalence of diabetes and diabetes-related cardiovascular comorbidities in the us, 2000–2002. *Diabetes care*, 28(7):1599–1603, 2005.

E Louise Thomas and Jimmy D Bell. Influence of undersampling on magnetic resonance imaging measurements of intra-abdominal adipose tissue. *International journal of obesity*, 27(2):211–218, 2003.

Diabetes UK. Type 2 diabetes and young adults. URL https://www.diabetes.org.uk/diabetes-the-basics/types-of-diabetes/type-2/young-adults.

Zahid Ullah, Farrukh Saleem, Mona Jamjoom, Bahjat Fakieh, Faris Kateb, Abdullah Marish Ali, Babar Shah, et al. Detecting high-risk factors and early diagnosis of diabetes using machine learning methods. *Computational Intelligence and Neuroscience*, 2022, 2022.

Jason Van Hulse, Taghi M Khoshgoftaar, and Amri Napolitano. Experimental perspectives on learning from imbalanced data. In *Proceedings of the 24th international conference on Machine learning*, pages 935–942, 2007.

WHO. Diabetes — who.int. https://www.who.int/health-topics/diabetes#tab=tab_1. [Accessed 23-Apr-2023].

Dennis L Wilson. Asymptotic properties of nearest neighbor rules using edited data. *IEEE Transactions on Systems, Man, and Cybernetics*, (3):408–421, 1972.

Qianqian Xie, Edward J Schenck, He S Yang, Yong Chen, Yifan Peng, and Fei Wang. Faithful ai in healthcare and medicine. *medRxiv*, pages 2023–04, 2023.

Zidian Xie, Olga Nikolayeva, Jiebo Luo, and Dongmei Li. Peer reviewed: Building risk prediction models for type 2 diabetes using machine learning techniques. *Preventing chronic disease*, 16, 2019.

Jie Xu, Yunyu Xiao, Wendy Hui Wang, Yue Ning, Elizabeth A Shenkman, Jiang Bian, and Fei Wang. Algorithmic fairness in computational medicine. *EBioMedicine*, 84, 2022.

Wensheng Yang, Chengsheng Pan, and Yanyan Zhang. An oversampling method for imbalanced data based on spatial distribution of minority samples sd-kmsmote. *Scientific Reports*, 12(1):16820, 2022.

Kuan Zhang, Bardia Khosravi, Sanaz Vahdati, Shahriar Faghani, Fred Nugen, Seyed Moein Rassoulinejad-Mousavi, Mana Moassefi, Jaidip Manikrao M Jagtap, Yashbir Singh, Pouria Rouzrokh, et al. Mitigating bias in radiology machine learning: 2. model development. *Radiology: Artificial Intelligence*, 4(5):e220010, 2022.

Appendix A. ML Model Performance

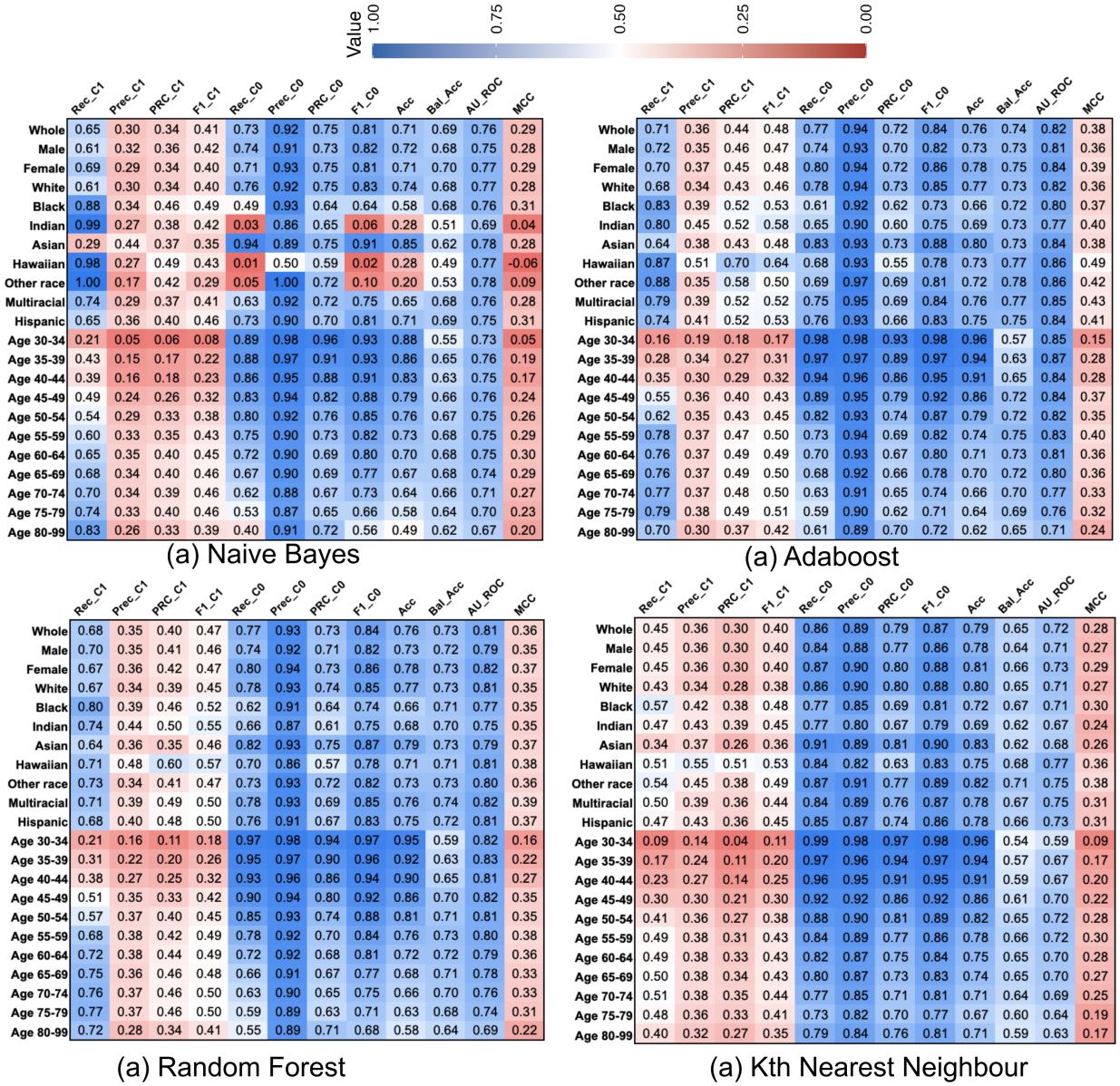


Figure 8: Performance of machine learning models trained and tested on the original 2021 BRFSS dataset. The x-axis represents performance metrics where Rec, Prec, PRC, F1, Acc, Bal_Acc, ROC, MCC represents Recall, Precision, Area Under the Precision-Recall Curve, F1 score, Accuracy, Balanced Accuracy, Area under the ROC Curve, Matthews Correlation Coefficient, respectively. C1 and C0 stand for class 1 (diabetes positive) and class 0 (diabetes negative) respectively. The Y-axis represents the whole population, gender groups, ethnic groups, and age groups.

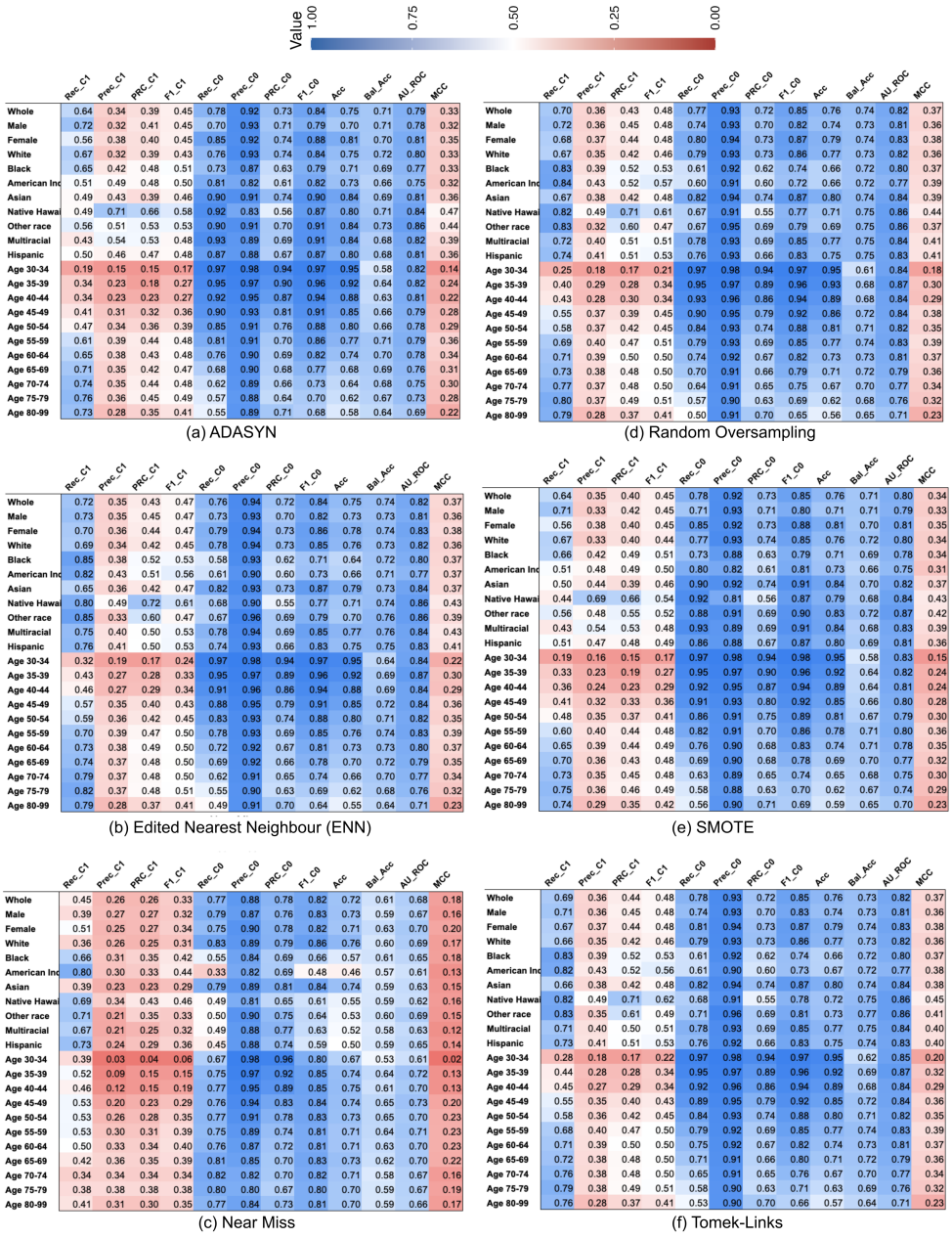


Figure 9: Performance of logistic regression model trained on different resampled datasets. The x-axis represents performance metrics where Rec, Prec, PRC, F1, Acc, Bal_Acc, ROC, MCC represents Recall, Precision, Area Under the Precision-Recall Curve, F1 score, Accuracy, Balanced Accuracy, Area under the ROC Curve, Matthews Correlation Coefficient, respectively. C1 and C0 stand for class 1 (diabetes positive) and class 0 (diabetes negative) respectively. The Y-axis represents the whole population, gender groups, ethnic groups, and age groups.

Appendix B. Feature details

Table 1: Selected features from BRFSS datasets for predicting diabetes

Column name	Question	Values
Blood pressure	Have you ever been told by a doctor, nurse or other health professional that you have high blood pressure?	1: Yes, 0: No
Cholesterol	Have you ever been told by a doctor, nurse or other health professional that your cholesterol is high?	1: Yes, 0: No
BMI category	Body Mass Index (BMI)	1: Underweight, 2: Normal weight, 3: Overweight, 4: Obese
Stroke	(Ever told) (you had) a stroke	1: Yes, 0: No
Heart disease	Respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI)	1: Yes, 0: No
Consume fruits	Consume Fruit 1 or more times per day	1: Yes, 0: No
Consume vegitables	Consume Vegetables 1 or more times per day	1: Yes, 0: No
Alcoholic	Adults who reported having had at least one drink of alcohol in the past 30 days.	1: Yes, 0: No
Health insurance	Adults who had some form of health insurance	1: Yes, 0: No
Unaffordable healthcare	Was there a time in the past 12 months when you needed to see a doctor but could not because you could not afford it?	1: Yes, 0: No
General health	Would you say that in general your health is:	1: Excellent, 2: Very good, 3: Good, 4: Fair, 5: Poor
Mental health	Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good?	0 days: 0 1 - 5 days: 1, 6 - 10 days: 2 11 - 15 days: 3, 16 - 20 days: 4 21 - 25 days: 5, 26 - 30 days: 6
Physical health	Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good?	0 days: 0 1 - 5 days: 1, 6 - 10 days: 2 11 - 15 days: 3, 16 - 20 days: 4 21 - 25 days: 5, 26 - 30 days: 6
Difficulty walking	Do you have serious difficulty walking or climbing stairs?	1: Yes, 0: No
Sex	Are you male or female?	1: Male, 2: Female
Age	Fourteen-level age category	1: 18-24, 2: 25-29, 3: 30-34, 4: 35-39, 5: 40-44, 6: 45-49, 7: 50-54, 8: 55-59, 9: 60-64, 10: 65-69, 11: 70-74, 12: 75-79. 13: 80-99
Education	What is the highest grade or year of school you completed?	1: Never, 2: Elementary, 3: Some high school, 4: High school grad, 5: Some college, 6: College grad
Income	Is your annual household income from all sources	1: <$10 K, 2: $10–$15 K, 3: $15–$20 K, 4: $20–$25 K, 5: $25–$35 K, 6: $35–$50 K, 7: $50–$75 K, 8: $75 K
Race	Race/ethnicity categories	1: White, 2: Black, 3: American Indian, 4: Asian, 5: Native Hawaiian, 6: Other race, 7: Multiracial, 8: Hispanic
Routine checkup	About how long has it been since you last visited a doctor for a routine checkup?	: <1 y, 2: 1–2 y, 3: 3–5 y, 4: >5 y
Flu shot	During the past 12 months, have you had flu vaccine?	1: Yes, 0: No,
Employment status	Are you currently. . . ?	1: Yes, 0: No
Marital status	Are you: (marital status)	1: Married, 2: Divorced, 3: Widowed, 4: Separated, 5: Never married, 6: Unmarried couple
Kidney disease	Were you ever told you had kidney disease?	1: Yes, 0: No
Depressive disorder	(Ever told) (you had) a depressive disorder (including depression, major depression, dysthymia, or minor depression)?	1: Yes, 0: No
Own home	Do you own or rent your home?	1: Own, 0: Rent
Exercise	During the past month, other than your regular job, did you participate in any physical activities or exercises such as running, calisthenics, golf, gardening, or walking for exercise?	1: Yes, 0: No
Diffculty seeing	Are you blind or do you have serious difficulty seeing, even when wearing glasses?	1: Yes, 0: No
Diffculty concentrating	Because of a physical, mental, or emotional condition, do you have serious difficulty concentrating, remembering, or making decisions?	1: Yes, 0: No
Diabetes	(Ever told) (you had) diabetes?	1: Yes, 0: No

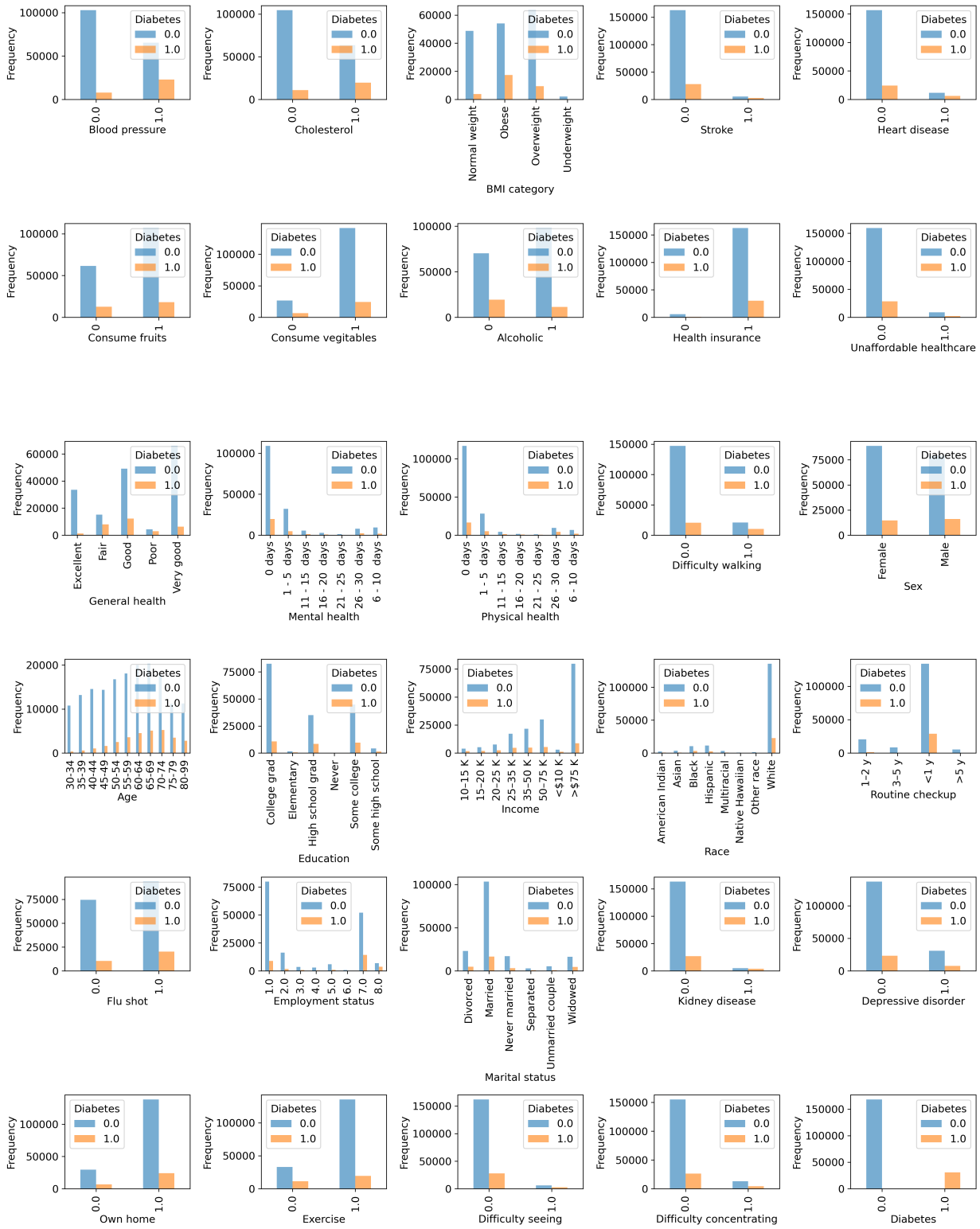


Figure 10: BRFSS 2021 data distribution