Enhancing Fairness and Accuracy in Type 2 Diabetes Prediction through Data Resampling

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

Machine learning (ML) methodologies have gained significant traction in the realm of healthcare due to their capacity to enhance diagnosis, treatment, and patient outcomes. Nevertheless, mitigating bias within these models is imperative to ensure equitable healthcare regardless of demographic factors such as age, gender, and ethnicity. This study explores the effectiveness of various sampling strategies for balancing imbalanced datasets in the context of improving the accuracy of type 2 diabetes prediction. The investigation leverages multiple ML classifiers and applies them to the inherently imbalanced Behavioral Risk Factor Surveillance System (BRFSS) datasets. Three distinct ML algorithms, namely Logistic Regression, Random Forest, and Multilayer Perceptron, are assessed on both the original and resampled datasets. The study reveals that dataset balancing through undersampling and oversampling techniques significantly enhances the models' sensitivity and balanced accuracy by at least 52% and 15%. However, it is observed that certain methods such as SMOTE, ADASYN, Tomek Links, Edited Nearest Distance, and Near Miss do not notably improve model sensitivity. Furthermore, this pattern of performance enhancement holds consistent when tested across multiple years of datasets (2021, 2019, 2017, and 2015). The analysis underscores that models trained on raw, imbalanced datasets exhibit subpar sensitivity across various subgroups, particularly among the White population (Sensitivity 0.17). The adoption of subgroup-based resampling techniques effectively ameliorates sensitivity and balanced accuracy by at least 45% and 10% respectively. Notably, the study identifies blood pressure, kidney disease, cholesterol levels, and BMI are the most important indicators of type 2 diabetes. This research underscores the potential of the resampling technique as a promising approach to developing more equitable, balanced, and accurate ML models, especially when addressing different disparities in healthcare outcomes.

Keywords: Fairness, Diabetes, Machine Learning, Healthcare, Sampling Algorithm, BRFSS dataset

1. Introduction

Machine learning (ML) has been increasingly integrated into the healthcare system because of its potential to assist clinicians and medical doctors to take better care of patients Lin (2020); Liaw (2020); Davenport (2018). With advances of data collection technology, the electronic health record (EHR), an electronic version of a patient's medical history, can be used to train these data-based algorithms RS (2016).

© 2023 T.S. Pias, Y. Su, X. Tang, H. Wang, S. Faghani & D.(. Yao.

However, the EHR data suffers from imbalance, which could lead to unfair and potentially harmful decisions being made by ML algorithms as the models pick up the pattern from the provided data Afrose et al. (2022); Ninareh Mehrabi (2021); Raji and Buolamwini (2019); Ghassemi (2020); Juhn Y (2020). Bias in ML models can stem from various sources, like underrepresented subgroups in the dataset, resulting in inadequate training data for algorithms Afrose et al. (2022) and biased selection of algorithms Zhang et al. (2022). Therefore, it is essential to evaluate and address potential biases in the EHR data used to train ML algorithms, especially before using them in practice where human lives are at stake.

A critical first step in the prevention and treatment of diseases is the identification of patients who are at risk of developing diabetes. However, if these models are biased, they might fail to identify patients with a high risk of developing diabetes, missing out on chances to prevent and treat the condition. In the realm of ML model evaluation, the conventional metric of accuracy has long been held in high regard, often regarded as the definitive gauge of a model's overall performance. Nevertheless, particularly in healthcare applications, where the imbalanced distribution of medical conditions is commonplace, the sole reliance on accuracy can be misleading and potentially detrimental to patient outcomes Afrose et al. (2022). Also the literature of ML in healthcare suggests to empathises on positive class recall more than overall accuracy Gupta and Sedamkar (2020); Gupta et al. (2021); Shinde and Singh (2023). If a model achieves high accuracy by predominantly classifying cases as negative, it might inadvertently overlook genuine instances of the disease, leading to delayed diagnoses and compromised patient well-being. This paper underscores the paramount importance of emphasizing recall or the true positive rate, as a central performance indicator for healthcare ML models.

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. The prevalence of diabetes, specifically type 2 diabetes, has been steadily increasing over the past few decades Krug (2016); Xie et al. (2019). In our paper, we focus on addressing bias in predicting type 2 diabetes. We used multiple widely-used ML classifiers on the naturally imbalanced Behavioral Risk Factor Surveillance System (BRFSS) datasets from years 2021, 2019, 2017, and 2015 cdc to assess the potential unfairness. The datasets are highly imbalanced because there are significantly fewer samples of the minority class (diabetes) than there are of the majority class (non-diabetes). As a result, all ML classifiers are affected by the data imbalance. To solve this issue, our study introduces a novel combination of random undersampling and oversampling to address minority disparities in diabetes prediction, enhancing the fairness and performance of ML models. This approach balances common demographic groups in diabetes datasets, reduces bias, and promotes equitable healthcare, improving model accuracy and reliability while mitigating biases and disparities in healthcare.

Our results reveal that LR and MLP show inaccuracies in predicting diabetes among different subgroups, suggesting the presence of bias in the original dataset. Moreover, the models trained on raw (imbalanced) datasets demonstrate high accuracy of at least 85%, however, a maximum sensitivity of 22% and balanced accuracy of 59%, which is not acceptable in the domain of healthcare. The LR model shows low sensitivity of detecting diabetes among different subgroups, for example, sensitivity of only 1%, 5%, 7%, 9%, and 14% among 30-54 age groups. To ensure fairness in decision-making, we use the resampling technique to create datasets which proved to be effective in mitigating bias. LR trained on resampled dataset achieved 71% to 78% sensitivity among 30-54 age groups. Moreover, the model significantly enhanced the models' overall sensitivity and balanced accuracy by at least 52% and 15% and improve subgroup sensitivity and balanced accuracy by at least 45% and 10% respectively. This study demonstrates the improvement of diabetes detection rate by carefully resampling the original dataset.

Our major contributions are: (1) our study identified significant imbalances in the BRFSS 2015, 2017, 2019, and 2021 datasets pertaining to type-2 diabetes, particularly in relation to demographics subgroups; (2) we compared different sampling strategies and proposed a combination of undersampling and oversampling techniques to address the bias and create more balanced subgroup datasets; (3) our work rigorously evaluated the performance and fairness of three popular ML algorithms, demonstrating the effectiveness of our approach in improving detection accuracy and mitigating prediction unfairness; and (4) we also conducted an investigation into the relationships between various variables that influence the prediction of type 2 diabetes.

Our study demonstrates the effectiveness of the carefully crafted resampling technique in mitigating bias in imbalanced datasets. This approach can be applied in other healthcare-ML applications, where imbalanced datasets are prevalent. Ultimately, our study aims to improve patient outcomes by ensuring that healthcare decisions made by ML algorithms are fair and unbiased. The contribution of this study is to highlight the importance of addressing and preventing unfairness in healthcare-ML applications.

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 400k participants which make it one of the largest publicly available datasets related to people's health. Each record 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.

The BRFSS dataset is a valuable resource for identifying health disparities and evaluating the effectiveness of public health programs and policies. However, the datasets are not free from challenges. The missing values pose a negative impact on ML algorithms. Some subgroups are highly imbalanced, such as demographic subgroups. To ensure a representative sample of the U.S. population, we selected four BRFSS datasets from 2015, 2017, 2019, and 2021 (the most recent), respectively across seven years to assess the potential unfairness in healthcare data. Moreover, we investigate the datasets by the most common demographic characteristics including age, gender, and race/ethnicity (White, Black, American Indian, Asian, Native Hawaiian, Other race, Multiracial, and Hispanic).

Table 1: Number of cases in preprocessing stages of BRFSS datasets. N/A: Not available

Type	2021 BRFSS	2019 BRFSS	2017 BRFSS	2015 BRFSS
Raw dataset	438,693	418,268	450,016	441,456
Containing N/A	113,015	83,057	90,520	117,293
Refused to answer	125,542	124,337	121,366	103,773
Processed dataset	200,136	210,874	238,130	220,390
Diabetes positive	169,296	178,103	202,117	187,701
Diabetes negative	30,840	32,771	36,013	32,689

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). Table 1 shows the details of the four preprocessed datasets. 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. 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 A5 and A6. We chose datasets depending on the availability of these selected variables. The BRFSS datasets from 2020, 2018, and 2016 lack some crucial variables, while the datasets from 2021, 2019, 2017, and 2015 encompass all the selected variables. Therefore, we opted for the BRFSS datasets from 2021, 2019, 2017, and 2015 for our study. The distribution of all selected variables are plotted in the Appendix Figure A3.

2.3. Sampling Algorithms

A frequently employed technique for addressing the challenge of data imbalance is the utilization of sampling methods. To ensure the appropriate sampling strategy, 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); (7) NearMiss, selecting instances from the majority class based on their distance to instances in the minority class Mani and Zhang (2003); (8) SMOTE undersampling, combining SMOTE and random undersampling to balance the dataset; and (9) Cost-sensitive Learning (CSL) Thai-Nghe et al. (2010), modifying the algorithm's cost function to penalize misclassifying the minority class more. Nonetheless, current sampling techniques are not tailored to tackle subgroup biases, as they uniformly sample the entire minority class. These methods do not discern between various demographic subgroups, such as Black patients or patients within the age range of 30 to 40. Therefore, we present an innovative resampling approach that combines both undersampling and oversampling strategies, guided by specific data statistics.

2.4. Machine Learning Models

We selected three commonly used ML models including logistic regression (LR) Peng et al. (2002), random forest (RF) Breiman (2001), and multilayer perceptron (MLP) Gardner and Dorling (1998) to evaluate the effectiveness of the sampling strategy. 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. 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 for at least five times and ensured consistent performance (standard deviation less than 2 of all experiments). To examine the potential unfairness in the datasets, we calculated class-based accuracy, recall, precision, F1-score, AUROC, AU-PRC, and MCC (Appendix A). 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. Each year's dataset undergoes a random division into training (80%) and testing sets (20%) 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 various experiments, ensuring the robustness and comparability of our results. All the ML models were developed in Python and executed in the Google Colab platform.

3. Results

In the conducted experiments, the impact of class imbalance on diabetes prediction was investigated. Surprisingly, when the models were trained on the unaltered (imbalanced) dataset, poor detection accuracy was observed. Specifically, a sensitivity rate of 21% was achieved by the LR model, 23% by the MLP, and 18% by the RF model. These results emphasize the significant challenges posed by class imbalance in the context of diabetes prediction.

3.1. Performance of Sampling Algorithms

The LR model was trained using various datasets, including the original BRFSS 2021 dataset and synthetic datasets generated through eight distinct sampling algorithms and a cost-sensitive learning approach. However, for the sake of maintaining consistency and facilitating meaningful comparisons, the same test dataset was utilized to evaluate the model's performance across all experiments. Figure 1 illustrates the confusion matrices for diabetes and non-diabetes predictions using LR . The diagonal entries

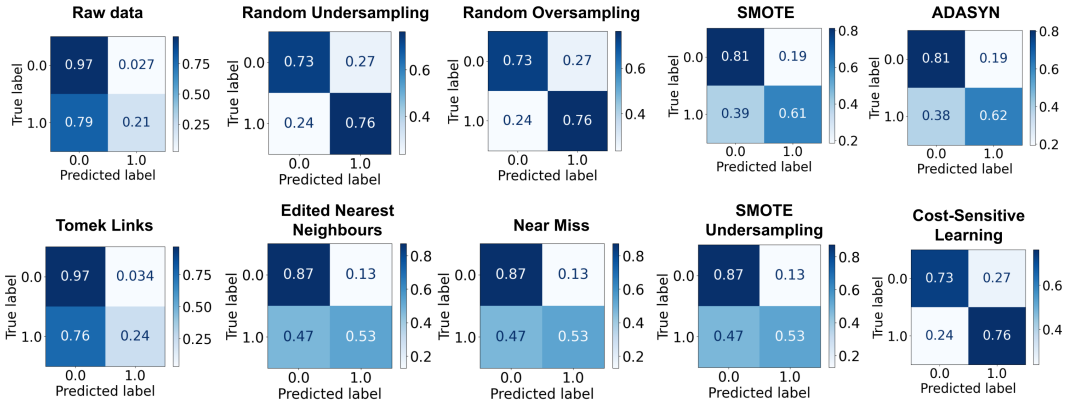

Figure 1: Confusion metrics of the LR model trained on various resampled training sets (specified in each plot's title), in addition to the raw dataset and a cost-sensitive learning approach. In this representation, "0" corresponds to negative diabetes cases, and "1" signifies positive diabetes cases. The X-axis represents the true labels, while the Y-axis denotes the predicted labels generated by the LR model. Specific metrics depicted include "0-0" for specificity, "1-1" for sensitivity, "0-1" for the missed detection rate (false negatives), and "1-0" for the false alarm rate (false positives).

in each confusion matrix represent the recall of class 0 (non-diabetes) or specificity and the recall of class 1 (diabetes) or sensitivity. Notably, the model trained on the raw dataset exhibited a notably low diabetes detection accuracy, standing at 21%. The application of the Tomek link approach didn't significantly enhance this accuracy, reaching only 23%. In contrast, the edited nearest neighbour, near miss, and SMOTE-undersampling algorithms demonstrated an improved detection accuracy of 53%. The SMOTE and ADASYN algorithms achieved even higher sensitivity rates, reaching 61% and 62%, respectively. Among all the methods evaluated, random undersampling, random oversampling, and cost-sensitive learning emerged as the most successful, achieving the highest diabetes detection accuracy, which stood at 76%. These three algorithms also displayed a balanced performance when considering both sensitivity (76%) and specificity (73%).

Then we leveraged the random undersampling strategy on the raw datasets to obtain balanced datasets for diabetes for across year. The balanced datasets are also fair for demographic groups.

3.2. Performance Across Datasets

To ensure the consistency of our resampling techniques, we opted for undersampling, which consistently exhibited the highest detection accuracy and was tested across multiple years of the dataset. We carefully selected 12 distinct performance metrics to provide a comprehensive view of our models' capabilities, and the detailed definitions for each metric are available in Appendix A.

Figure 2 presents a performance comparison for LR , RF, and MLP models when applied to both the original and undersampled datasets from the years 2021, 2019, 2017, and 2015. When these models were trained on the imbalanced dataset, LR achieved sensitivity rates of 21-22%, RF 18-19%, and MLP 23-24%. In contrast, training the models on a balanced dataset achieved notable improvements: LR achieved a sensitivity of 76-78%, RF 78-80%, and MLP 76%. Additionally, when assessing a composite performance metric like balanced accuracy, we observed improvements of at least 15% for LR , 16% for RF, and 12% for MLP. This indicates that the models not only exhibit strong accuracy in detecting diabetes but also maintain a balanced performance across both classes. Furthermore, crucial performance metrics such as the F1-score for both classes, precision for class 0, and MCC all demonstrated enhancements. Notably, the AU-ROC and AU-PRC remained consistent for both raw and balanced datasets. It's worth mentioning that the overall accuracy declined, which was expected since the test set predominantly comprised non-diabetes samples. The LR model was also

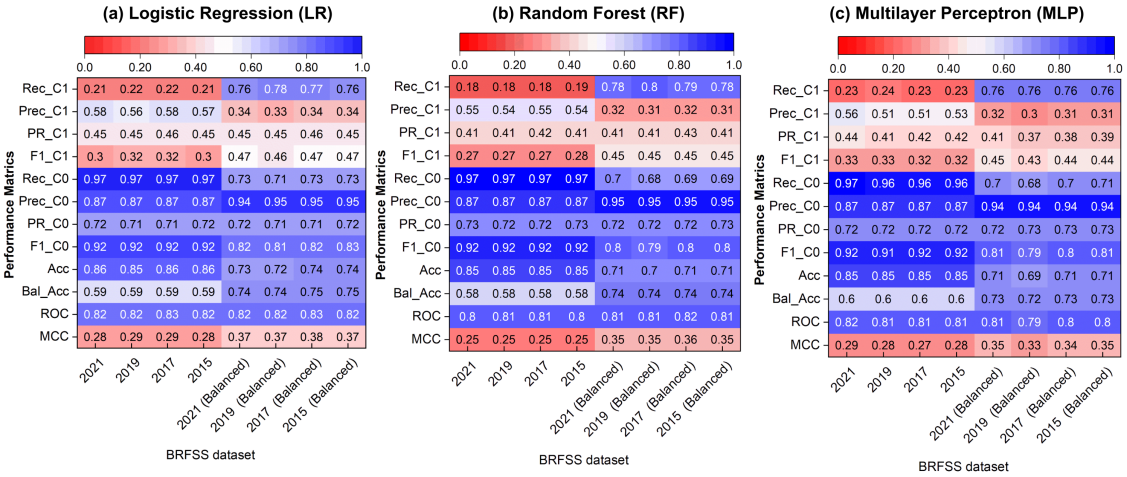

Figure 2: Performance of ML models trained on raw and balanced training set derived from the 2021 BRFSS dataset and tested on 2021, 2019, 2017, and 2015 BRFSS test set. The title of each plot represents the model. The x axis represents the test set and model nature (trained on the raw or balanced training set). The y-axis represents performance metrics where Rec_C1, Prec_C1, PR_C1, F1_C1, Rec_C0, Prec_C0, PR_C0, F1_C0, Acc, Bal_Acc, ROC, MCC stand for Recall Class 1, Precision Class 1, Area Under the Precision-Recall Curve Class 1, F1 score Class 1, Recall Class 0, Precision Class 0, Area Under the Precision-Recall Curve Class 0, F1 score Class 0, Accuracy, Balanced Accuracy, Area under the ROC Curve, Matthews Correlation Coefficient, respectively.

trained over a merged training set from all four years BRFSS training sets, however, the performance still remains the same as training using a single year's imbalanced train set, indicating no advantage of increasing data using multiple years dataset (Appendix Figure 1). This demonstrates the effectiveness of the undersampling algorithm, the performance of ML models was greatly improved on the balanced datasets compared with the imbalanced datasets.

3.3. Performance Over Subgroups

Previous experiments have demonstrated that resampling the entire dataset can enhance the accuracy of diabetes detection. However, it's crucial to recognize that improved overall accuracy doesn't necessarily equate to fairness across all subgroups, such as Black or Asian patients. Therefore, we conducted further evaluations of model performance, considering demographic features including gender (male and female), age groups (30-34, 35-49, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79, and 80-99), and ethnicity (White, Black, American Indian, Asian, Hawaiian, Other race, Multicultural, and Hispanic).

The subgroup-specific performance is visualized in Figure 3. In Figure 3a, we observe the performance of the LR model trained on the raw (imbalanced) training set outlined in the Appendix Table A2 which derived from raw BRFSS 2021 dataset outlined in Appendix Table A1. Notably, this model exhibits a biased performance, with the highest detection rate (35%) among the American Indian population and a significantly lower detection accuracy (16%) for the Asian population, indicating a clear discrepancy in the model's behavior. Surprisingly, the detection accuracy is even lower for the age groups ranging from 30 to 55, ranging from only 1% to 14%. Figure 3b showcases the performance of the LR model trained on a balanced (undersampled) training set. In this case, the performance of all ethnic and gender subgroups notably improves. However, age groups 30 to 54 still exhibit comparatively poor detection rates (Recall of C1). This is attributed to the fact that the ratio of positive and negative cases in those age-related subgroups is not close to unity, as detailed in Appendix Table A3. Balancing these subgroups through undersampling would further reduce their total numbers which might dilute the impact of these

subgroups. Consequently, we employ oversampling techniques to augment the number of positive cases, aiming to achieve a balanced ratio within these subgroups. The resulting resampled training set, as illustrated in Appendix Table A4, is then used to train the LR model, which exhibits improved detection accuracy compared to previous models 3c. The same experiments were conducted using the MLP model and the outcome was similar (Appendix Figure A2) to the LR model.

3.4. Risk Factors Association

Among the three ML models employed, LR stands out due to interpretability, allowing us to establish meaningful associations between diabetes and contributing factors. In LR, model coefficients serve as key instruments for comprehending the strength and direction of the relationships between individual features and the log-odds of the binary outcome. These coefficients play a pivotal role in elucidating how each feature influences the likelihood of an event occurring.

Figure 4 presents an overview of the coefficients associated with different features. Notably, variables related to blood pressure, kidney disease, and cholesterol exhibit positive correlations with diabetes, suggesting that elevated levels of these factors are more likely to cause diabetes. Conversely, factors such as routine checkups and regular exercise demonstrate a negative correlation, indicating that these behaviours can potentially reduce the risk of diabetes.

On the other hand, features like education level, income level, and homeownership appear to have relatively limited influence on the likelihood of having diabetes, as their corresponding coefficients do not exhibit significant impacts. This insight derived from LR coefficients provides valuable guidance for understanding the relative importance of various factors in the context of diabetes prediction.

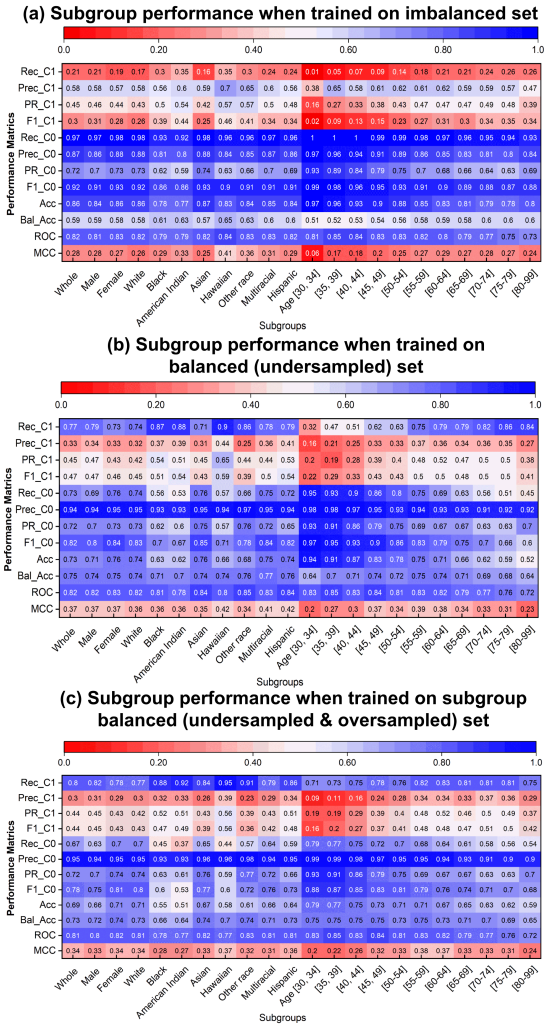

Figure 3: The performance metrics of LR model trained on (a) raw (imbalanced) training set, (b) balanced (undersampled) training set, and (c) balanced-subgroup (undersampled and oversampled) training set from 2021 BRFSS dataset. All three models are tested using the 2021 BRFSS test set. The x-axis represents different gender, ethnic, and age-based subgroups. The y axis represents different performance metrics.

4. Related Work

This section talks about the related work in the field of ML in healthcare, with a focus on data imbalance and fairness.

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 dif-

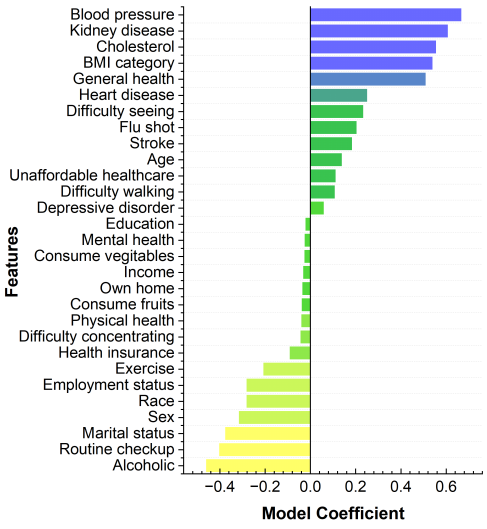

Figure 4: LR (trained on balanced 2021 BRFSS training set) coefficients of all features used in training. The one-hot encoded features (Race, Marital status, and Employment status) coefficients are averaged.

ferences 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 using the BRFSS dataset. The BRFSS dataset has been used for many other purposes 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 for achieving 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 on 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. 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.

5. Discussion

Our study has demonstrated the importance of addressing fairness issues in leveraging ML models in healthcare, especially in the context of minority disparities. We have shown that the use of resampling techniques to balance the dataset for different subgroups can significantly improve the performance and fairness of ML models in predicting diabetes. Our results suggest that a combination of undersampling and oversampling can effectively improve the detection accuracy (sensitivity) of diabetes for all three ML models, indicating that these models become fairer towards all subgroups. Moreover, interpretable ML models such as LR , RFs, etc can be used to explain the results and practitioners can learn from big data.

Our study offers a novel insight into the risk factors that contribute to the development of diabetes. Specifically, our analysis has revealed that high blood pressure, kidney disease, obesity and cholesterol have a significant impact on the likelihood of developing

type 2 diabetes. This finding is particularly notable given that previous research has primarily focused on more traditional risk factors such as obesity, family history, and age. Our study highlights the importance of considering a broader range of risk factors in the development of ML models for predicting diabetes. By incorporating factors such as blood pressure and cholesterol-related attributes into these models, we can significantly improve their accuracy and reliability, ultimately leading to better healthcare outcomes for patients. Our analysis was conducted using the BRFSS datasets collected from 2021, 2019, 2017, and 2015. By leveraging the rich datasets, we were able to identify previously unrecognized risk factors for diabetes and develop ML models that accurately predict these conditions across a range of demographic groups.

It will be important to continue exploring and refining methods for addressing bias and fairness in ML models. This includes developing new techniques for dataset balancing and exploring the use of other performance metrics beyond the traditional accuracy and AUC-ROC values. By leveraging the power of ML to improve healthcare outcomes while promoting fairness and equity, we can help ensure that everyone has access to the high-quality care they deserve.

References

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

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

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

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.

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.

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.

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.

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.

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/.

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.

Apoorva Gupta, Ashutosh Anand, and Yasha Hasija. Recall-based machine learning approach for early detection of cervical cancer. In *2021 6th International Conference for Convergence in Technology (I2CT)*, pages 1–5. IEEE, 2021.

Shiwani Gupta and RR Sedamkar. Machine learning for healthcare: Introduction. *Machine learning with health care perspective: Machine learning and healthcare*, pages 1–25, 2020.

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.

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.

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

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.

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.

Inioluwa Deborah Raji and Joy Buolamwini. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 429–435, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450363242. doi: 10.1145/3306618.3314244. URL https://doi.org/10.1145/3306618.3314244.

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.

Evans RS. Electronic health records: then, now, and in the future. *Yearbook of medical informatics*, 20(Supp1):S48–61, 2016. URL 10.15265/IYS-2016-s006.

Abhitay Shinde and Artika Singh. Enhancing diabetes detection using machine learning: A focus on optimizing recall performance. In *2023 Second International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT)*, pages 1–8. IEEE, 2023.

Nguyen Thai-Nghe, Zeno Gantner, and Lars Schmidt-Thieme. Cost-sensitive learning methods for imbalanced data. In *The 2010 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2010. doi: 10.1109/IJCNN.2010.5596486.

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.

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.

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

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.

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. Performance Metrics

Accuracy Accuracy is a measure of how well a model correctly identifies the classes in the dataset. It is defined as the ratio of the number of correctly predicted instances to the total number of instances in the dataset.

Recall Recall (also known as sensitivity) is the fraction of the total number of positive instances that were correctly predicted by the model. It is a measure of the model's ability to identify all positive instances in the dataset.

Precision Precision is the fraction of the total number of positive predictions made by the model that is actually correct. It is a measure of the model's ability to make accurate positive predictions.

F1-score F1-score is a measure of the performance of a classification model that takes into account both precision and recall. It is the harmonic mean of precision and recall and is a measure of the balance between precision and recall.

AU-ROC AU-ROC (area under the receiver operating characteristic curve) is a metric used to evaluate the performance of binary classification models. It is a plot of the true positive rate (sensitivity) against the false positive rate (1-specificity) at different probability thresholds. The area under the curve (AUC) of the ROC plot is a measure of the model's ability to distinguish between the positive and negative classes.

AU-PRC AU-PRC (area under precision-recall curve) is another metric used to evaluate the performance of binary classification models. It is a plot of the precision versus the recall. Different from AU-ROC, AU-PRC is class balanced-dependent.

MCC MCC (Matthews correlation coefficient) is a classification metric used to measure the difference between the predicted labels and true labels.

C1 & C0 Class 1 (C1) represents the minority class, i.e., diabetes. Similarly, class 0 (C0) represents the majority class, i.e., no diabetes.

Appendix B. Combined training results

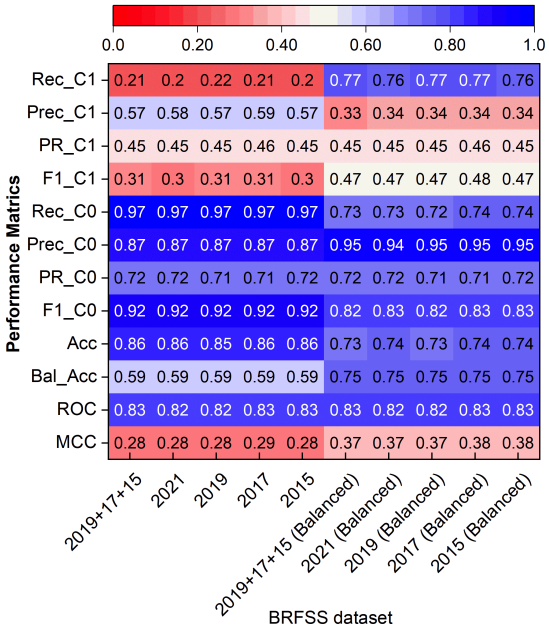

Figure 1: Performance of LR model trained on combined training sets derived from the three-year BRFSS dataset and tested on combined BRFSS, 2021, 2019, 2017, and 2015 BRFSS test sets.

Appendix C. Dataset Statistics

Table A1: Subgroup statistics of raw BRFSS 2021 dataset

Subgroup	Total	Negative	Positive	Ratio
Age groups				
30-34	11188	10864	324	0.0298
35-39	13878	13290	588	0.0442
40-44	15724	14661	1063	0.0725
45-49	16077	14466	1611	0.1114
50-54	19357	16836	2521	0.1497
55-59	21757	18155	3602	0.1984
60-64	24672	20140	4532	0.2250
65-69	25526	20422	5104	0.2499
70-74	23121	17891	5230	0.2923
75-79	14740	11257	3483	0.3094
80-99	14096	11314	2782	0.2459
Ethnic groups				
White	159313	136565	22748	0.1666
Black	13808	10551	3257	0.3087
American Indian	2985	2219	766	0.3452
Asian	4199	3633	566	0.1558
Native Hawaiian	772	598	174	0.2910
Other race	1453	1198	255	0.2129
Multiracial	3751	3142	609	0.1938
Hispanic	13855	11390	2465	0.2164
Gender groups				
Male	95988	79844	16144	0.2022
Female	104148	89452	14696	0.1643

Table A2: Subgroup statistics of raw BRFSS 2021 training set

Subgroup	Total	Negative	Positive	Ratio
Age groups				
30-34	8848	8589	259	0.0302
35-39	11171	10687	484	0.0453
40-44	12506	11640	866	0.0744
45-49	12940	11648	1292	0.1109
50-54	15426	13426	2000	0.1490
55-59	17396	14523	2873	0.1978
60-64	19663	16055	3608	0.2247
65-69	20531	16387	4144	0.2529
70-74	18577	14390	4187	0.2910
75-79	11718	8949	2769	0.3094
80-99	11332	9090	2242	0.2466
Ethnic group				
White	127564	109295	18269	0.1672
Black	11034	8408	2626	0.3123
American Indian	2418	1797	621	0.3456
Asian	3329	2874	455	0.1583
Native Hawaiian	619	487	132	0.2710
Other race	1162	942	220	0.2335
Multiracial	2978	2489	489	0.1965
Hispanic	11004	9092	1912	0.2103
Gender groups				
Male	76844	63913	12931	0.2023
Female	83264	71471	11793	0.1650

Table A3: Subgroup statistics of balanced (undersampled majority class) BRFSS 2021 training set

Subgroup	Total	Negative	Positive	Ratio
Age groups				
30-34	1807	1548	259	0.1673
35-39	2490	2006	484	0.2413
40-44	3007	2141	866	0.4045
45-49	3372	2080	1292	0.6212
50-54	4374	2374	2000	0.8425
55-59	5531	2658	2873	1.0809
60-64	6588	2980	3608	1.2107
65-69	7124	2980	4144	1.3906
70-74	6872	2685	4187	1.5594
75-79	4426	1657	2769	1.6711
80-99	3857	1615	2242	1.3882
Ethnic group				
White	38206	19937	18269	0.9163
Black	4160	1534	2626	1.7119
American Indian	962	341	621	1.8211
Asian	987	532	455	0.8553
Native Hawaiian	222	90	132	1.4667
Other race	394	174	220	1.2644
Multiracial	962	473	489	1.0338
Hispanic	3555	1643	1912	1.1637
Gender groups				
Male	24582	11651	12931	1.1099
Female	24866	13073	11793	0.9021

Table A4: Subgroup statistics of resampled (undersampled and oversampled) BRFSS 2021 training set

Subgroup	Total	Negative	Positive	Ratio
Age groups				
30-34	3096	1548	1548	1.0000
35-39	4012	2006	2006	1.0000
40-44	4282	2141	2141	1.0000
45-49	4160	2080	2080	1.0000
50-54	4748	2374	2374	1.0000
55-59	5531	2658	2873	1.0809
60-64	6588	2980	3608	1.2107
65-69	7124	2980	4144	1.3906
70-74	6872	2685	4187	1.5594
75-79	4426	1657	2769	1.6711
80-99	3857	1615	2242	1.3882
Ethnic group				
White	38206	19937	18269	0.9163
Black	4160	1534	2626	1.7119
American Indian	962	341	621	1.8211
Asian	987	532	455	0.8553
Native Hawaiian	222	90	132	1.4667
Other race	394	174	220	1.2644
Multiracial	962	473	489	1.0338
Hispanic	3555	1643	1912	1.1637
Gender groups				
Male	24582	11651	12931	1.1099
Female	24866	13073	11793	0.9021

Appendix D. Subgroup performance - MLP

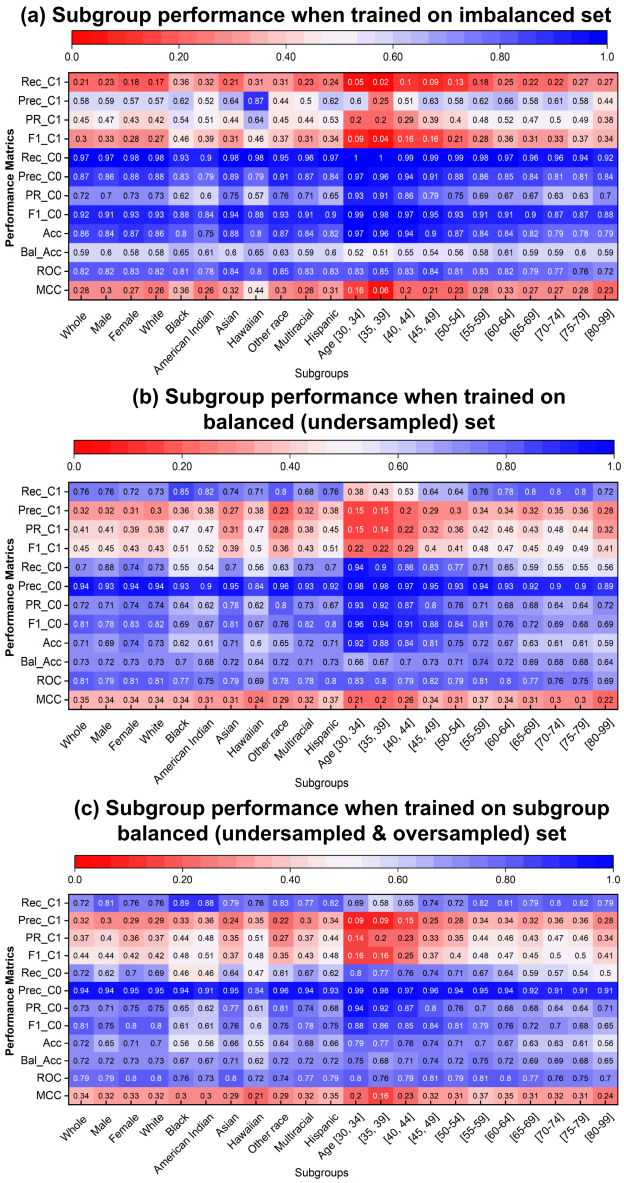

Figure A2: The performance metrics of MLP model trained on (a) raw (imbalanced) training set, (b) balanced (undersampled) training set, and (c) balanced-subgroup (undersampled and oversampled) training set from 2021 BRFSS dataset. All three models are tested using the 2021 BRFSS test set. The x-axis represents different gender, ethnic, and age-based subgroups. The y axis represents different performance metrics.

Appendix E. Feature details

Table A5: Selected relevant features and corresponding variable names from BRFSS datasets for predicting diabetes

#	Features	BRFSS 2021	BRFSS 2019	BRFSS 2017	BRFSS 2015
1	Blood pressure	BPHIGH6	BPHIGH4	BPHIGH4	BPHIGH4
2	Cholesterol	TOLDHI3	TOLDHI2	TOLDHI2	TOLDHI2
3	BMI category	_BMI5CAT	_BMI5CAT	_BMI5CAT	_BMI5CAT
4	Stroke	CVDSTRK3	CVDSTRK3	CVDSTRK3	CVDSTRK3
5	Heart disease	_MICHD	_MICHD	_MICHD	_MICHD
6	Consume fruits	_FRTLT1A	_FRTLT1A	_FRTLT1A	_FRTLT1
7	Consume vegitables	_VEGLT1A	_VEGLT1A	_VEGLT1A	_VEGLT1
8	Alcoholic	DRNKANY5	DRNKANY5	DRNKANY5	DRNKANY5
9	Health insurance	_HLTHPLN	HLTHPLN1	HLTHPLN1	HLTHPLN1
10	Unaffordable healthcare	MEDCOST1	MEDCOST	MEDCOST	MEDCOST
11	General health	GENHLTH	GENHLTH	GENHLTH	GENHLTH
12	Mental health	MENTHLTH	MENTHLTH	MENTHLTH	MENTHLTH
13	Physical health	PHYSHLTH	PHYSHLTH	PHYSHLTH	PHYSHLTH
14	Difficulty walking	DIFFWALK	DIFFWALK	DIFFWALK	DIFFWALK
15	Sex	SEXVAR	SEXVAR	SEX	SEX
16	Age	_AGEG5YR	_AGEG5YR	_AGEG5YR	_AGEG5YR
17	Education	EDUCA	_EDUCAG	_EDUCAG	EDUCA
18	Income	INCOME3	INCOME2	INCOME2	INCOME2
19	Race	_RACE	_RACE	_RACE	_RACE
20	Routine checkup	CHECKUP1	CHECKUP1	CHECKUP1	CHECKUP1
21	Flu shot	FLUSHOT7	FLUSHOT7	FLUSHOT6	FLUSHOT6
22	Employment status	EMPLOY1	EMPLOY1	EMPLOY1	EMPLOY1
23	Marital status	MARITAL	MARITAL	MARITAL	MARITAL
24	Kidney disease	CHCKDNY2	CHCKDNY2	CHCKIDNY	CHCKIDNY
25	Depressive disorder	ADDEPEV3	ADDEPEV3	ADDEPEV2	ADDEPEV2
26	Own home	RENTHOM1	RENTHOM1	RENTHOM1	RENTHOM1
27	Exercise	EXERANY2	EXERANY2	EXERANY2	EXERANY2
28	Difficulty seeing	BLIND	BLIND	BLIND	BLIND
29	Difficulty concerntrating	DECIDE	DECIDE	DECIDE	DECIDE
30	Diabetes	DIABETE4	DIABETE4	DIABETE3	DIABETE3

Table A6: 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: $<10K, 2$:10–15K, 3: 15–20 K, 4: 20–25 K, 5: 25–35 K, 6: 35–50 K, 7: 50–75 K, 8: ¿75K
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

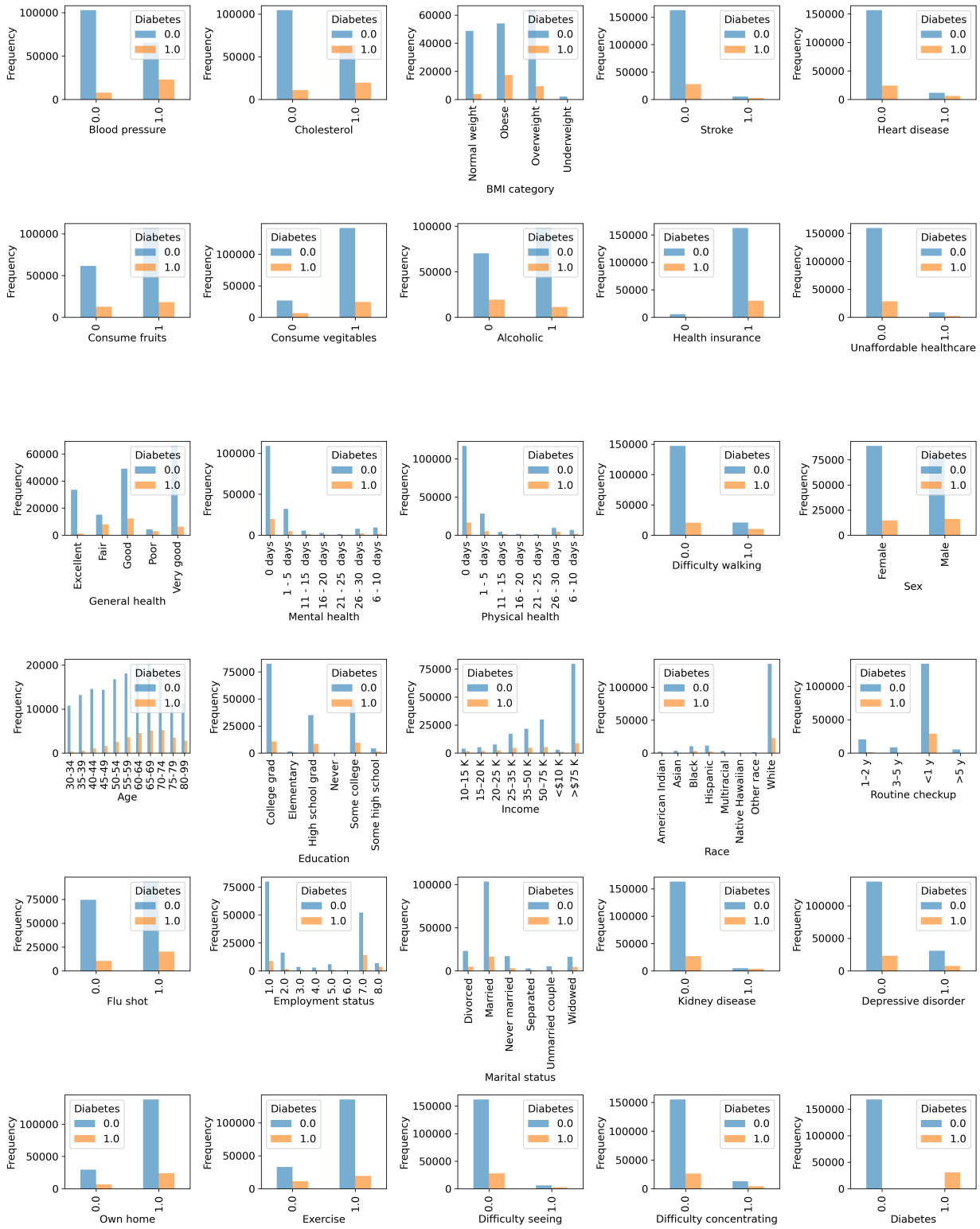

Figure A3: BRFSS 2021 data distribution