Preprint:Under Review 1–19

Undersampling for Fairness: Achieving More Equitable Predictions in Diabetes and Prediabetes

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

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

Abstract

The use of machine learning in healthcare has grown rapidly in recent years, with the potential to improve diagnosis, treatment, and patient outcomes. However, issues of bias and fairness in these models must be addressed to ensure equitable treatment for all patients regardless of their gender. In this study, we evaluated the effectiveness of undersampling to balance the dataset both in terms of positive-negative samples and male-female gender groups and create a more robust and accurate machine learning model in predicting diabetes and prediabetes. In this study, we applied multiple machine learning classifiers to the Behavioral Risk Factor Surveillance System (BRFSS 2015) dataset, which is inherently imbalanced. We tested four algorithms - Logistic Regression, Random Forest, K-Nearest Neighbors, and Multilayer Perceptron - using both the original and the balanced datasets. Our results indicate that balancing the dataset through undersampling improves the performance of all four algorithms in predicting diabetes. Specifically, Logistic Regression showed an increase in Precision from 0.53 to 0.74, while Random Forest showed an increase in F1-score from 0.25 to 0.75. K-Nearest Neighbors showed an increase in Recall from 0.19 to 0.73, and Multilayer Perceptron showed a significant increase in F1-score from 0.38 to 0.78. Moreover, our findings reveal that undersampling can improve fairness by mitigating gender bias from machine learning models, as measured by the Disparate Impact Ratio (DIR). The experiments depict a change in DIR for most algorithms when trained on the

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

balanced dataset, for logistic regression, random forest, and k-nearest neighbour DIR decreases from 1.16 to 1.12, 1.14 to 1.08, and 1.05 to 1.02 respectively, indicating that the models are becoming fairer towards both gender groups. Our study demonstrates that undersampling can be a promising step towards creating more balanced, fair, and accurate machine learning models for gender subgroups.

1. Introduction

Artificial Intelligence (AI) has been increasingly integrated into the healthcare system and has the potential to assist clinicians and medical doctors to take better care of patients Lin (2020); Liaw (2020); Davenport (2018). It has been a long since machine learning (ML) algorithms have been employed to help clinicians diagnose and treat a range of diseases, from deathly cancer to heart disease. With the advancement of data collection technology, the electronic health record (EHR), an electronic version of a patient's medical history, provides a vast amount of data that can be used to train these data-based algorithms RS (2016). However, similar to other fields, the data in the medical field is highly imbalanced resulting in biased learning Afrose et al. (2022). As a matter of fact, the EHR data can be biased too, which could lead to unfair and potentially harmful decisions being made by machine learning algorithms as the models pick up the pattern from the provided data Ninareh Mehrabi (2021); Raji and Buolamwini (2019); Ghassemi (2020); Juhn Y (2020). Therefore, it is essential to evaluate and address potential biases in the EHR data used to train machine learning algorithms, especially before using them in practice where human lives are at stake.

Bias in machine learning models can be introduced in many different ways. For example, certain subgroups in the dataset may be underrepresented, leading to insufficient training data for the algorithms Afrose et al. (2022). This underrepresentation of such subgroups could result in biases in the algorithm's predictions and decision-making, potentially leading to harm to the users. Additionally, algorithms may be biased because of factors such as the way they are modelled and trained, underlying assumptions, or the type of dataset used to train them Zhang et al. (2022). These biases can lead to unfair decisions that may result in poorer outcomes for certain subgroups of patients which can have a serious impact on their lives.

Predicting chronic diseases like diabetes is one area where bias can have an adverse impact De Silva et al. (2020). Millions of people worldwide suffer from diabetes, a chronic condition that needs ongoing care to avoid long-term complications. The majority of the approximately 422 million people with diabetes worldwide reside in low- and middle-income nations and diabetes is directly responsible for 1.5 million deaths annually according to who. Over the past few decades, there has been a consistent rise in both the incidence and prevalence of diabetes Krug (2016). A critical first step in the prevention and treatment of disease is the identification of patients who are at risk of developing diabetes. With encouraging results, machine learning algorithms have been used to predict the onset of diabetes from EHR data Nguyen et al. (2019). However, if these algorithms 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.

Many techniques are utilized to create a balanced dataset and undersampling is one of them. The undersampling technique has long been recognized for balancing an imbalanced

dataset in machine learning. It involves randomly removing some of the majority class samples to create a balanced dataset, thus preventing the model from being biased towards the majority class Goyal (2022); Idris et al. (2019); Thomas and Bell (2003); Cartus et al. (2020). However, our study introduces a novel application of undersampling in addressing gender disparities in diabetes and prediabetes prediction, which has not been previously explored in the literature. In particular, our study shows that undersampling can be used to balance the gender distribution in the diabetes and prediabetes datasets, and significantly improve the performance and fairness of machine learning models. By balancing the dataset, the model becomes less biased towards the majority gender and can better capture the subtle differences in the minority group. This approach can help to ensure that both genders receive equitable healthcare, especially in situations where one gender may be at higher risk for a particular disease. It not only helps to address gender disparities but also promotes fairness and equity in the development and deployment of machine learning models. By leveraging this technique, we can improve the accuracy and reliability of machine learning models while reducing the risk of perpetuating biases and disparities in healthcare.

We used four widely-used machine learning classifiers on the naturally unbalanced Behavioral Risk Factor Surveillance System (BRFSS 2015) dataset cdc to analyze the effect of bias. This dataset is selected because it contains data on a variety of health behaviours and conditions, including diabetes, and is publicly available. The dataset is highly unbalanced because there are significantly fewer samples of the minority class (diabetes) than there are of the majority class (non-diabetes). Surprisingly, the gender groups are imbalanced as well in the prediabetes subset. As a result, all four classifiers are affected by the data imbalance.

Our results reveal that Logistic Regression and Multilayer Perceptron show different accuracies for predicting diabetes among males and females, suggesting the presence of bias in the original dataset. To ensure fairness in decision-making, a strategy such as resampling is effective. In this study, we use the undersampling technique to create a dataset which proved to be effective in mitigating bias, especially for gender subgroups. We evaluate our proposed approach by using a fairness metric called disparate impact ratio (DIR) Feldman et al. (2015). The experiment result shows a positive gain in DIR for all models trained on the resampled dataset which is indicative of fairness.

The major contributions can be summarized as follows:

- Our study identified significant imbalances in the BRFSS 2015 dataset pertaining to type-2 diabetes and prediabetes, particularly in relation to gender disparities.
- We proposed and implemented undersampling technique to address the bias and create more balanced datasets for machine learning models.
- Our work rigorously evaluated the performance and fairness of four popular machine learning algorithms, demonstrating the effectiveness of our approach in improving accuracy and mitigating prediction unfairness.

Our study demonstrates the effectiveness of the undersampling technique in mitigating bias, especially for gender, in imbalanced datasets. This approach can be applied in other healthcare AI applications, where imbalanced datasets are prevalent. Ultimately, our study aims to improve patient outcomes by ensuring that healthcare decisions made by machine

learning algorithms are fair and unbiased. The contribution of this study is to highlight the importance of addressing and mitigating bias in healthcare AI applications.

In the following sections, we provide a detailed explanation of the methods used in our study, present the results, and discuss the implications of our findings.

1.1. Generalizable Insights about Machine Learning in the Context of Healthcare

We address one of the major issues which can arise while using healthcare data to train machine learning models. Because of data imbalance, the model can learn bias from the dataset. We found that a very popular dataset in the healthcare domain, the Behavioral Risk Factor Surveillance System (BRFSS) dataset, can contain bias. For example, if we consider the patients who are diagnosed with diabetes, the number of false cases is significantly larger than the positive cases which supports the general population statistics however, the data imbalance can negatively affect the machine learning models. Moreover, the prediabetes subset contains imbalanced gender-group which eventually induces bias in the model. We claim that balancing the training dataset can effectively unbias the machine learning model. We considered two different scenarios to justify our claim.

First, the number of positive diabetes cases (who are diagnosed with diabetes) is around 35k and false diabetes cases (who are not diagnosed with diabetes) is more than 123k which is significantly larger. We show that the four commonly used machine learning models (Logistic Regression, Random Forest, K-th Nearest Neighbour, and Multilayer Perceptron) which are trained on the original imbalanced dataset perform poor in detecting positive cases compared to the models which are trained on the balanced dataset.

Second, we consider the scenario of prediabetes (Have you ever been told by a doctor or other health professional that you have pre-diabetes or borderline diabetes?). Similarly, the positive and negative prediabetes cases are not equal which affects the model's performance. However, unlike diabetes, the number of male and female cases is not equal which results in gender bias. We use a similar downsampling technique to create a balanced dataset in terms of gender. The resampled dataset reduces the gender bias from the models. We also evaluate our work using a fairness metric called disparate impact ratio (DIR). The models which are trained on the gender-balanced dataset are fairer in terms of gender than the models which are trained on the original imbalanced dataset. our method of downsampling the original dataset is effective for reducing biases from the machine learning models.

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 2015 dataset was collected from all 50 states in the U.S. and the District of Columbia and included responses from 441,456 participants which makes it one of the largest publicly available datasets related to people's health. The survey collected information on various health behaviours and risk factors

Table 1: Selected relevant features from BRFSS dataset for predicting diabetes and prediabetes

#	Feature Description	Variable Name
1	High Blood Pressure	_RFHYPE5
2	High Cholesterol	TOLDHI2
3	Cholesterol check	_CHOLCHK
4	BMI	_BMI5
5	Smoking	SMOKE100
6	Stroke	CVDSTRK3
7	Coronary heart disease	_MICHD
8	Physical Activity	_TOTINDA
9	Diet-Fruit	_FRTLT1
10	Diet-Vegetables	_VEGLT1
11	Alcohol Consumption	_RFDRHV5
12	Insurance-Coverage	HLTHPLN1
13	Failed to visit doctor	MEDCOST
14	Health General	GENHLTH
15	Mental Health	MENTHLTH
16	Physical Health	PHYSHLTH
17	Difficulty walking or climbing stairs	DIFFWALK
18	gender	SEX
19	Age	_AGEG5YR
20	Education	EDUCA
21	Income	INCOME2

such as tobacco use, physical activity, alcohol consumption, existing chronic diseases, and mental health. The survey also gathered demographic information such as age, gender, race/ethnicity, education, and income which sometimes can be helpful to make important correlations and even causation. The BRFSS 2015 dataset includes 330 variables, making it a valuable resource for studying the health status of U.S. adults.

The dataset is grouped by age, gender, and race/ethnicity to ensure a representative sample of the U.S. population. The mean age of participants was 47 years, with a range from 18 to 99 years. The majority of participants were female (55.4%). The BRFSS 2015 dataset is widely used in public health research and policy-making to inform interventions and strategies aimed at improving the health of the United States population. It 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 for example, this dataset contains missing values which pose a challenge for the machine learning algorithms. Moreover, some subgroups are highly imbalanced, for example, positive and negative samples for diabetes subgroups, male and female samples for prediabetes subgroups, and so on.

2.2. Data Preprocessing

At first, we have 441,456 records and 330 variables. Each record contains an individual's BRFSS survey responses. According to the literature on the risk factors for diabetes, Collins et al. (2011); Xie et al. (2019), we selected 21 predictor variables for analysis. The selected variables are listed in Table 1. These 21 variables are used for detecting both diabetes and prediabetes.

Around 100k instances in the raw dataset contain missing values. Being a large dataset, we dropped the instances with the missing values resulting in 343k instances with 21 features. Some of the survey questions were refused by the patients which are equivalent to missing values. So, we filtered out these variables as well resulting in dropping another 90k instances. The preprocessed dataset contains around 253k instances and no missing values. Here we build two different scenarios, one for predicting diabetes and another for predicting prediabetes. We process the datasets differently for these two scenarios.

Now for diabetes cases, the number of negative cases is 213k and the number of positive cases is 35k. This makes the raw diabetes dataset very imbalanced. To make it balanced we downsampled the negative cases to match the number of positive cases. We call it balanced diabetes cases. The balanced diabetes dataset is also fair for gender groups. The number of positive male cases is 16k and the number of positive female cases is 15k which is fairly balanced.

However, for prediabetes, we dropped the diabetes cases. The dataset for prediabetes contains 5782 positive cases and 40k negative cases. Which makes the dataset highly imbalanced. Even for the gender subgroup the dataset is imbalanced. The number of positive female cases (3621) is greater than positive male cases (2161). We call it the raw-prediabetes dataset. To make the dataset even for both gender groups we downsample the positive female cases to match the size of the male cohort which is 2161. We call it the balanced prediabetes dataset.

2.3. Machine Learning Models

In this study we want to show the effect of balancing the dataset by downsampling. So, we selected 5 commonly used machine learning models such as logistic regression (LR) Peng et al. (2002). random forest (RF) Breiman (2001), k-nearest neighbor (KNN) Peterson (2009), and multilayer perceptron (MLP) Gardner and Dorling (1998). For logistic regression, we utilized the powerful liblinear equation to establish the decision boundary between positive and negative classes. In random forest, we carefully selected 100 trees or estimators to achieve optimal performance. Similarly, we set the number of neighbours to 5 in k-nearest neighbours for precise classification. In multilayer perceptron, we strategically employed a single hidden layer with 100 neurons, each activated by the robust relu function. To ensure efficient and effective training, we utilized the adam optimizer for gradient descent. Our parameter selection process was rigorous, and we believe that our findings provide valuable insights into the optimal configuration of these models for various applications. We train these four models on the original dataset and the balanced dataset to evaluate the effectiveness of the balancing dataset by downsampling.

2.4. 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.

$$Accuracy = \frac{TruePositives + TrueNegatives)}{TruePositives + FalsePositives + TrueNegatives + FalseNegatives}$$

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.

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives}$$

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.

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives}$$

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.

$$F1score = 2 * Precision * RecallPrecision + Recall$$

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.

Disparate Impact Ratio Disparate Impact Ratio Feldman et al. (2015) (DIR) is a measure of fairness in machine learning models. It is defined as the ratio of the percentage of protected class members (e.g., gender, race) that are predicted to belong to the positive class (e.g., selected) to the percentage of non-protected class members that are predicted to belong to the positive class. A value of 1 indicates no disparate impact, while a value greater than 1 indicates disparate impact against the protected class.

$$\text{Disparate impact ratio} = \frac{Pr(Y = 1|unprivileged)}{Pr(Y = 1|privileged)}$$

In the context of a healthcare dataset where gender is imbalanced, the disparate impact ratio can be used to assess if the predictive model or algorithm is equally effective for both males and females. This is important because biases in healthcare datasets can lead to disparities in treatment and outcomes for different gender groups. By examining the disparate impact ratio, researchers and healthcare professionals can identify potential biases and work towards addressing them to ensure that healthcare outcomes are fair and equitable for all patients, regardless of gender.

3. Result

We performed experiments with two settings. We state two major research questions. First, the impact of positive and negative class imbalance on diabetes prediction. And second,

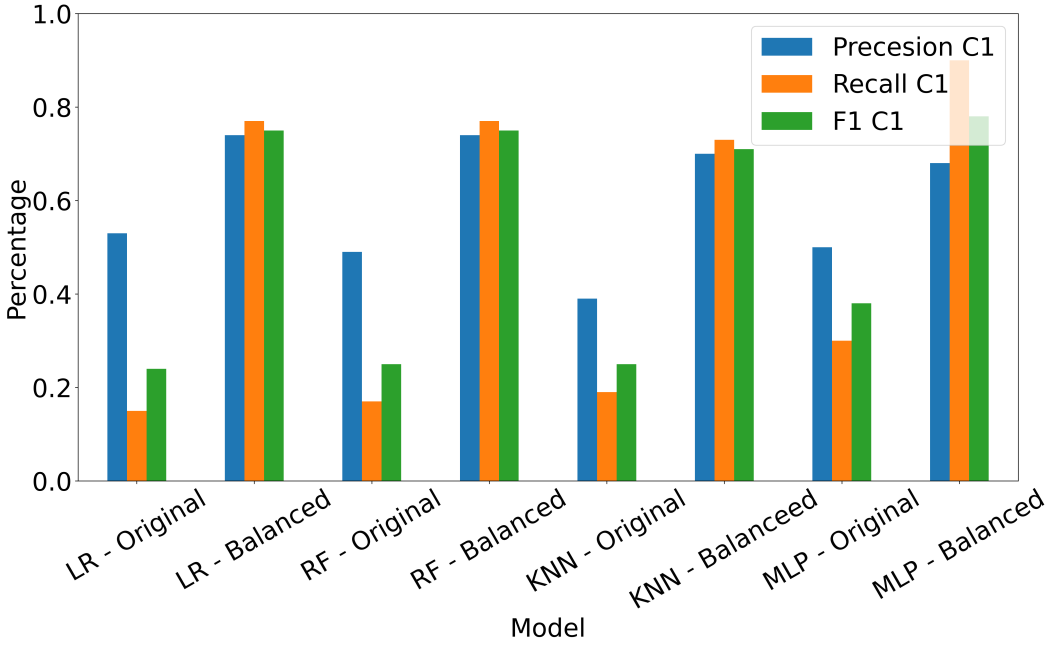

Figure 1: Precision, recall, and F1 score of positive class for four machine learning models trained on the original dataset and the resampled balanced dataset. For each model, the three scores are grouped together. The models trained on the balanced dataset demonstrate improved performance significantly.

the impact of gender imbalance on prediabetes prediction. All the results presented here are averages of at least five trials with a maximum standard deviation of 0.12.

3.1. Diabetes Prediction

To test the effectiveness of this technique, we evaluated four machine learning algorithms - Logistic Regression (LR), Random Forest (RF), K-Nearest Neighbors (KNN), and Multilayer Perceptron (MLP) - using both the original and the balanced datasets. We measured the performance of each algorithm using three common performance metrics - Precision, Recall, and F1-score - for the minority class (C1) which is demonstrated in figure 1.

Our results indicate that balancing the dataset through undersampling improves the performance of all four algorithms. Logistic Regression showed an increase in Precision from 0.53 to 0.74, while Random Forest showed an increase in F1-score from 0.25 to 0.75. K-Nearest Neighbors showed an increase in Recall from 0.19 to 0.73, and Multilayer Perceptron showed a significant increase in F1-score from 0.38 to 0.78.

These results demonstrate the effectiveness of undersampling as a technique to create a more balanced and fair machine learning model. With the increasing use of machine learning in healthcare, it is imperative to address issues of bias and fairness in these models. The use of undersampling to balance the dataset and create a more robust and accurate model is a promising step towards achieving this goal.

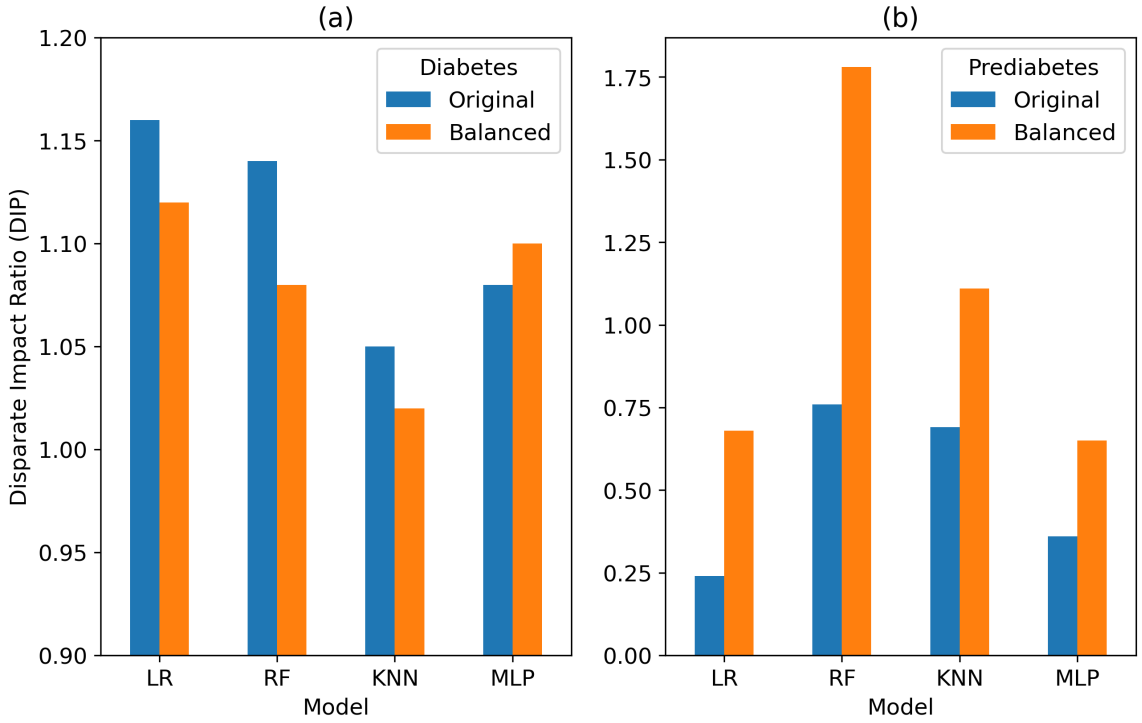

Figure 2: The figure shows, disparate impact ratio (DIR) for (a) Diabetes prediction and (b) Prediabetes prediction. All the models demonstrate improve DIR as they change towards 1 for diabetes. For prediabetes, LR, KNN, and MLP show improved DIR, however, only RF couldn't mitigate the gender bias

Gender disparities in healthcare data are a pervasive issue, with females often regarded as the privileged group, while males are relegated to the unprivileged group. The AI Fairness 360 framework has emphasized the significance of the Disparate Impact Ratio (DIR) as a measure of model fairness in the context of imbalanced gender groups. A higher DIR signifies better results, and a DIR value around 1 denotes equitable privileges for both groups. Conversely, a DIR value below or above 1 signifies a model that favours one group over the other. The data in figure 2 represents the disparate impact ratio (DIR) for four machine learning algorithms in predicting diabetes. The DIR is a measure of fairness in machine learning models, which indicates the extent to which the model favours one group over another. A DIR of 1 indicates that the model is equally fair to both groups. The data shows that the original models for all four algorithms have a DIR greater than 1, indicating that they favour the privileged group. However, when the models are trained on a balanced dataset with an equal representation of both groups, the DIR decreases for all algorithms. Specifically for diabetes prediction (figure 2-a), the DIR for logistic regression and random forest decreases from 1.16 to 1.12 and from 1.14 to 1.08, respectively. For the K-nearest neighbour, the DIR decreases from 1.05 to 1.02, while for the multilayer perceptron, the DIR decreases from 1.08 to 1.1. These results suggest that balancing the

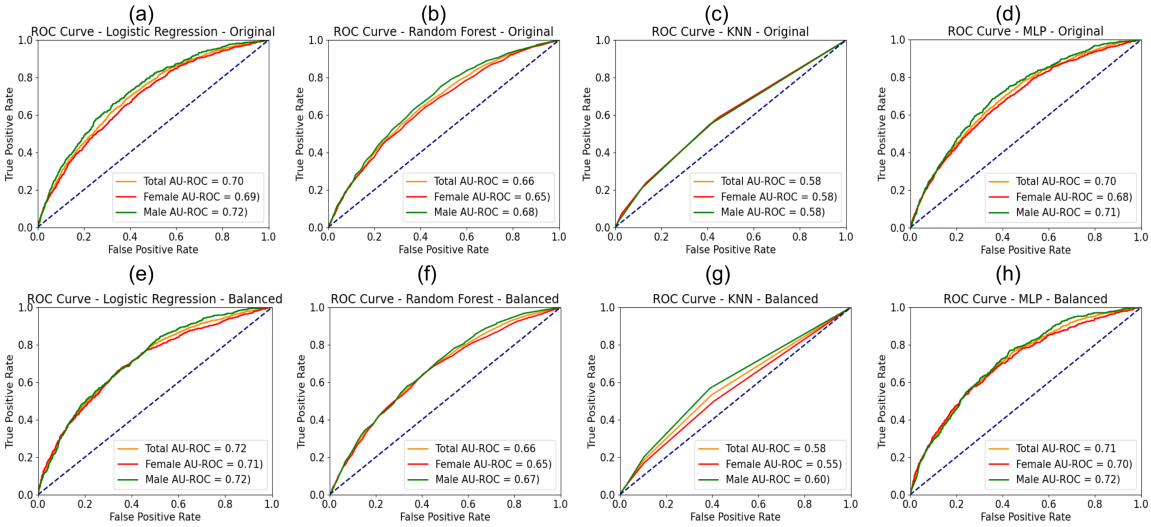

Figure 3: This figure shows the area under (AU) receiver operating curve (ROC) curve of the whole group, male subgroup, and female subgroup for the original and balanced dataset for diabetes prediction. (a), (b), (c), and (d) show AU-ROC of LR, RF, KNN, and MLP respectively trained on the original dataset. (e), (f), (g), and (h) show AU-ROC of LR, RF, KNN, and MLP respectively trained on the resampled balanced dataset. LR and MLP show improvement as the male and female AU-ROC are close and the whole group AU-ROC increased. RF and KNN show the same performance for both settings.

dataset can improve the fairness of machine learning models in predicting diabetes. The decrease in DIR indicates that the models are becoming fairer towards both groups. Thus, the undersampling technique used to balance datasets can make machine learning models fair and robust in predicting diabetes.

Figure 3, shows the area under the ROC (Receiver Operating Characteristic) curve. ROC curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. TPR is also known as sensitivity, recall or hit rate, while FPR is also known as the false alarm rate. The area under the curve (AUC) of the ROC curve is a commonly used metric to compare different binary classifiers, where a higher AUC indicates a better overall performance of the classifier. The ROC curve is widely used in medical diagnosis, signal detection, and machine learning evaluation.

Figure 3 demonstrates that three out of four models' performance improved when trained on the balanced (downsampled) dataset. The AU-ROC of KNN improved the most from 0.71 to 0.76, though this model shows the lowest performance but a greater impact of data balancing. The AU-ROC curve is also plotted for both gender groups (male and female). It shows that in all cases the AU-ROC increased for these groups which make the data balancing technique using undersampling very effective.

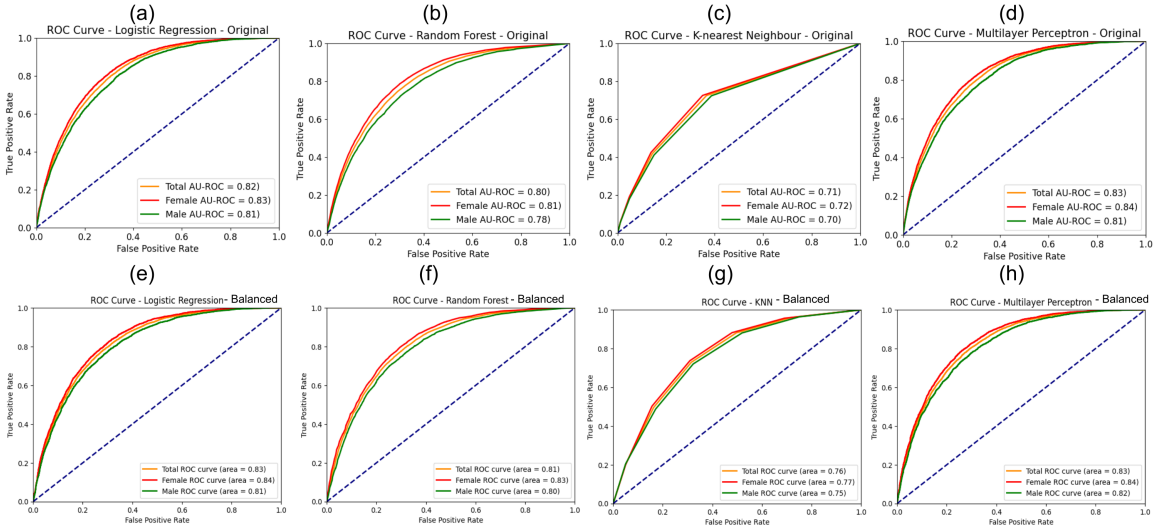

Figure 4: This figure shows the area under (AU) receiver operating curve (ROC) curve of the whole group, male subgroup, and female subgroup for the original and balanced dataset for prediabetes prediction. (a), (b), (c), and (d) show AU-ROC of LR, RF, KNN, and MLP respectively trained on the original dataset. (e), (f), (g), and (h) show AU-ROC of LR, RF, KNN, and MLP respectively trained on the resampled balanced dataset. LR and MLP show improvement as the male and female AU-ROC are close and the whole group AU-ROC increased. RF and KNN show the same performance for both settings.

3.2. Prediabetes Prediction

In our study, we employed four models to evaluate the predictive performance of prediabetes occurs in males and females using the original and balanced BRFSS dataset. Our findings revealed (figure 2-b) that the DIR of logistic regression and MLP increased from 0.24 to 0.68 and 0.36 to 0.65, respectively. KNN had a DIR of 1.11, indicating a fairer model for the minority gender group. However, the DIR for random forest rose from 0.76 to 1.78, highlighting its preference for the privileged group. Our results demonstrate the importance of measuring model fairness in healthcare data to ensure equitable treatment of all groups, regardless of gender.

Our findings, as presented in figure 4, reveal compelling insights into the effectiveness of resampled balanced datasets in mitigating prediction unfairness across different subgroups.

Notably, our logistic regression model exhibited superior prediction performance in detecting diabetes among males in the original BRFSS dataset, with an AUC 3% higher than that of females. However, our resampled balanced dataset produced ROC curves that were nearly identical for males and females, indicating a significant reduction in prediction unfairness. Encouragingly, this pattern was consistent across three out of four models we tested, including random forest and multilayer perceptron models. In contrast, the k-nearest

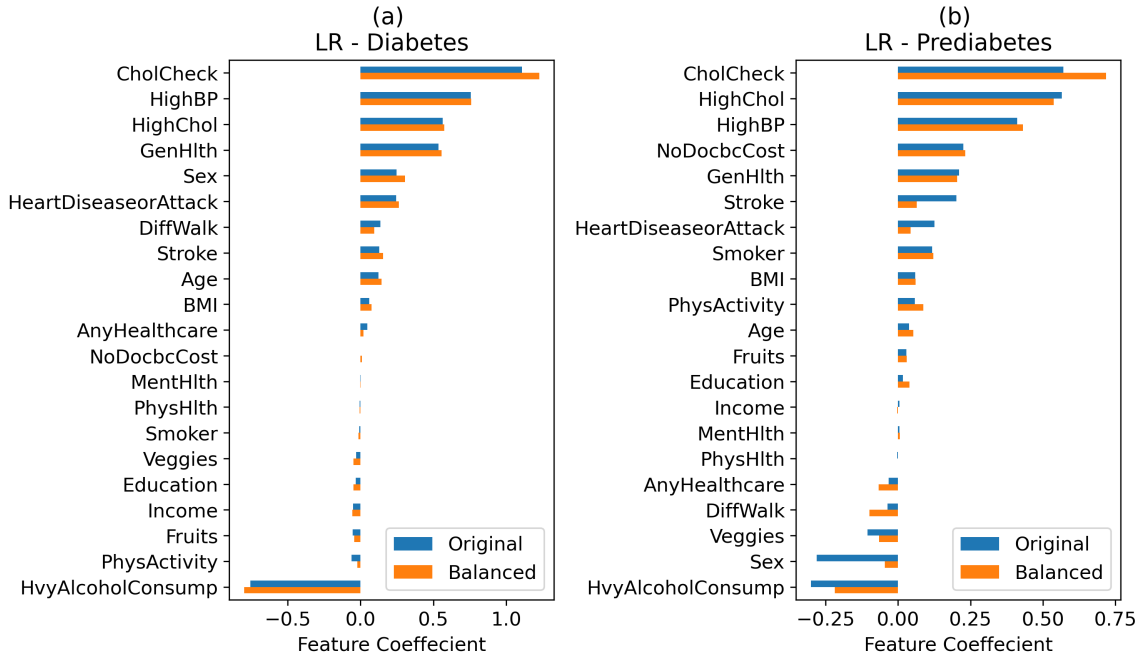

Figure 5: Logistic regression (LR) feature coefficient (equivalent to feature importance) of (a) the diabetes prediction model and (b) the prediabetes prediction model. In both cases, the cholesterol-related variables, high blood pressure, and high alcohol consumption are regarded as the top important factors for diabetes and prediabetes. The prediabetes shows a spike for gender (sex) in the balanced model as the model is resampled on gender group.

neighbours model yielded slightly unfair results in the resampled dataset, possibly due to its reliance on neighbouring sample points for prediction without explicit training.

In this study, we selected four widely used machine learning models among which the logistic regression is fully interpretable. To analyze the change in the impact of different features before and after the resampling, the feature coefficients are plotted in figure 5. It is observed that the feature-importance of almost all of the variables is changing the resampled model. Colestrol-related variables and alcohol consumption variables show a positive change which supports the literature. However, in the prediabetes prediction model, the gender variable demonstrates the highest change. It is because gender is resampled in the balanced model.

Our findings underscore the importance of addressing gender imbalance in healthcare datasets and highlight the potential of resampled balanced datasets in achieving more equitable and accurate predictions. This study paves the way for a more inclusive and effective approach to healthcare analytics, ultimately leading to improved health outcomes for all.

4. Related Work

Machine learning (ML) in healthcare has gained significant attention in recent years due to the vast potential it holds for improving patient outcomes and reducing healthcare costs. However, there are various challenges in deploying machine learning models in healthcare, including data imbalance and fairness concerns. This section provides an overview of the related work in the field of machine learning in healthcare, with a focus on data imbalance and fairness.

4.1. Data Imbalance

One of the major challenges in machine learning 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.

Oversampling involves creating more instances of the minority class to balance the dataset, while undersampling involves reducing the number of instances in the majority class. Several studies have shown the effectiveness of these techniques in improving the performance of machine learning models in healthcare. For example, Goh et al. (2021); Rezapour et al. (2023); Rezapour and Hansen (2022) used oversampling to balance a dataset of electronic health records for predicting heart failure, resulting in improved model performance. On the other hand, Goyal (2022); Idris et al. (2019); Thomas and Bell (2003); Cartus et al. (2020); Liu et al. (2008) used undersampling to balance a dataset for achieving higher accuracy in various domains.

4.2. Fairness

In addition to data imbalance, fairness is another important concern in machine learning on healthcare. There is a growing recognition that machine learning models can perpetuate and amplify existing biases and discrimination, leading to unfair outcomes for certain groups. This is particularly problematic in healthcare, where fairness is crucial for ensuring equal access to care and avoiding disparities.

Several approaches have been proposed to address fairness in machine learning on healthcare. Meng et al. (2022) introduced the interpretability of machine learning 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 machine learning 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.

4.3. BRFSS Dataset

In recent years, BRFSS has been used extensively as a data source for machine learning studies in healthcare due to its large sample size, a wide range of variables, 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 machine learning 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. Apart from these, the BRFSS dataset is being 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).

5. Discussion

Our study has demonstrated the importance of addressing bias and fairness in machine learning models used in healthcare, especially in the context of gender disparities. We have shown that the use of undersampling to balance the dataset for different subgroups can significantly improve the performance and fairness of machine learning models in predicting both diabetes and prediabetes. Our results suggest that undersampling can effectively improve the disparate impact ratio (DIR) for most of the machine learning algorithms, indicating that these models become fairer towards both groups. Additionally, we have shown that the use of balanced datasets can mitigate prediction unfairness across different subgroups, reducing the impact of gender disparities on model performance. Moreover, interpretable machine learning models such as logistic regression, random forests, 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 and prediabetes. Specifically, our analysis has revealed that cholesterol-related attributes and high alcohol consumption have a significant impact on the likelihood of developing these conditions. 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 machine learning models for predicting diabetes and prediabetes. By incorporating cholesterol-related attributes and alcohol consumption into these models, we can significantly improve their accuracy and reliability, ultimately leading to better healthcare outcomes for patients. Furthermore, our study offers important insights into the underlying biology of these conditions, suggesting that there may be a strong connection between lipid metabolism and glucose regulation. In addition to these novel findings, our study also underscores the importance of utilizing large and diverse datasets to fully capture the complex relationships between risk factors and disease outcomes. Our analysis was conducted using the BRFSS 2015 dataset, which includes data from over 400,000 respondents across all 50 US states. By leveraging this rich dataset, we were able to identify previously unrecognized risk factors for diabetes and prediabetes and develop machine learning models that accurately predict these conditions across a range of demographic groups.

Our findings have important implications for the development and deployment of machine learning models in healthcare as fairness is a very important factor. As the use of these models becomes more widespread, it is imperative that they are designed and validated in a way that ensures equitable treatment of all groups, regardless of gender or other demographic factors. By employing techniques such as undersampling to create more balanced and fair datasets, we can increase the accuracy and reliability of machine learning models while reducing the risk of perpetuating biases and disparities in healthcare.

It will be important to continue exploring and refining methods for addressing bias and fairness in machine learning models. This includes developing new techniques for dataset balancing and exploring the use of other performance metrics beyond the traditional precision, recall, and F1-score. By leveraging the power of machine learning to improve healthcare outcomes while promoting fairness and equity, we can help ensure that everyone has access to the high-quality care they deserve.

Limitations It is important to acknowledge that while undersampling can help to mitigate the impact of gender disparities from machine learning models, there may be other sources of bias that need to be addressed, for example, biases associated with race, ethnicity, socioeconomic status, or geographic location. All of these factors play an important role and could have a significant impact on the performance and fairness of these models. This paper solely focuses on mitigating a particular disparity which is gender disparity, but it is essential that future research in this area continues to explore and refine methods for addressing bias and fairness by taking into account the complexities and nuances of different demographic factors and other subgroups, especially minority groups. By doing so, we can help to ensure that these models are effective, reliable, and equitable for all individuals, regardless of their background or circumstances which is one of the main goals we want to achieve in any domain of research. Moreover, in this paper, we use mostly white-box machine learning models for better explainability and interoperability as the healthcare AI should be reliable and trustworthy. However, these models sometimes might not capture the underlying complex relationships or patterns from real-life data, like deep neural networks do. However, deep neural networks are not fully interpretable and sometimes lack trustworthiness. So, research into making deep learning models more interpretable and explainable should be a high priority, especially in the healthcare system where we want to achieve high accuracy and reliability at the same time.

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].

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.

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.

Michael Feldman, Sorelle Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact, 2015.

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.

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

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.

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.

Binh P Nguyen, Hung N Pham, Hop Tran, Nhung Nghiem, Quang H Nguyen, Trang TT Do, Cao Truong Tran, and Colin R Simpson. Predicting the onset of type 2 diabetes using wide and deep learning with electronic health records. *Computer methods and programs in biomedicine*, 182:105055, 2019.

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.

Leif E Peterson. K-nearest neighbor. *Scholarpedia*, 4(2):1883, 2009.

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.

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.

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.