Bias Propagation in Health AI: Measuring Pre-Training Bias and Its Effect on Machine Learning Model Outcomes

Diego Dimer Rodrigues[1], **Mariana Recamonde-Mendoza**[1,2]

[1]Institute of Informatics, Universidade Federal do Rio Grande do Sul (UFRGS),
Porto Alegre - RS, Brazil

[2]Bioinformatics Core, Hospital de Clínicas de Porto Alegre (HCPA),
Porto Alegre - RS, Brazil

{ddrodrigues,mrmendoza}@inf.ufrgs.br

Abstract. *Machine learning (ML) has become an essential tool in healthcare, supporting diagnosis, prognosis, and treatment decisions. However, biases present in pre-training data can compromise both model performance and fairness, disproportionately affecting underrepresented groups. This study systematically examines the impact of four pre-training bias metrics on the accuracy of three ML models across four health-related datasets. Our findings show that more data does not necessarily translate to better performance, particularly when data imbalance and bias are present. Moreover, pre-training bias metrics are associated with accuracy disparities, underscoring the importance of proactive bias assessment to develop more equitable ML models in healthcare.*

1. Introduction

Machine learning (ML) is increasingly used in various domains, including Health, where it plays a crucial role in assisting diagnosis, prognosis, and treatment decisions [Chen et al. 2021]. The ability of ML models to analyze vast amounts of medical data has led to significant advancements in precision medicine, disease prediction, and patient care. However, despite these benefits, ML algorithms are susceptible to biases that may arise from the data they are trained on, leading to unfair and potentially harmful outcomes [Obermeyer et al. 2019, Juhn et al. 2022]. Bias in healthcare models can result in disparities in predictive performance across different demographic groups, reinforcing existing health inequalities and affecting the reliability of clinical decision-making.

To address this issue, it is essential to systematically evaluate the presence of biases prior to training predictive models. Pre-training bias metrics provide a quantitative way to assess potential risks in datasets before they affect model outcomes. These metrics can reveal structural imbalances related to protected attributes, such as race, gender, or age, that might otherwise go unnoticed until after deployment.

This work aims to measure the risk of bias in health datasets using four pre-training bias metrics. By analyzing the relationship between these metrics and the performance of supervised ML algorithms, we investigate whether pre-existing biases in the data translate into variations in model outcomes. Specifically, our experiments assess predefined protected attributes within four health-related datasets, examining variations in accuracy and F1 scores across different demographic groups. Understanding these relationships can help identify datasets that may require bias mitigation strategies before model training, ultimately contributing to the development of more equitable and trustworthy AI systems in healthcare.

2. Theoretical background

ML has been widely applied in various fields, including healthcare, where supervised models are used for classification tasks such as disease diagnosis and risk prediction. However, the quality and representativeness of the training data play a crucial role in model performance. Biases present in the data can lead to imbalanced predictions, disproportionately affecting certain demographic groups and potentially reinforcing existing health disparities. These biases often arise from differences in data distribution across protected attributes, such as race, gender, age, or socioeconomic status, which are characteristics legally or ethically recognized as requiring fairness considerations. When ML models are trained on biased data, they may produce systematically different outcomes for different protected groups, leading to unfair and potentially harmful consequences in healthcare decision-making [Caton and Haas 2024].

Pre-training bias metrics have been developed to quantify inequalities in datasets before model training. These metrics assess the distribution of labels in relation to protected attributes, helping identify disparities that could impact model performance. In this study, we utilize four pre-training bias metrics:

- **Class Imbalance (CI)**: measures class representation in a dataset, ranging from -1 to 1. Values near zero indicate balance, positive values show an overrepresentation of the advantaged group, and negative values indicate an overrepresentation of the disadvantaged group.
- **Kullback-Leibler (KL) Divergence**: measures the divergence between the label distributions of two facets. It ranges from 0 to $+\infty$, with values near zero indicating similar distributions and higher values indicating more significant divergence.
- **Kolmogorov-Smirnov (KS)**: measures the maximum divergence between labels in the distribution for different facets. It ranges from 0 to 1, with values near zero indicating evenly distributed labels and values near one indicating imbalanced labels.
- **Conditional Demographic Disparity in Labels (CDDL)**: measures the proportion of negative outcomes in a specific facet of a dataset. It ranges from -1 to 1, with positive values indicating demographic disparity favoring the advantaged group and negative values indicating disparity favoring the disadvantaged group.

3. Related Work

Many libraries and open-source tools have been developed to address fairness in ML. For example, the AIF 360 library [Bellamy et al. 2018] provides various techniques for identifying and mitigating bias. Fairness Measures [Zehlike et al. 2017] offers multiple methods for classifying fairness, bias, and discrimination in any *csv* dataset. Additionally, FairLens [Synthesized.io 2023] is a Python library that automatically detects and quantifies bias in data, generating reports on data quality.

Studies have shown that word embedding algorithms can encode biases related to marginalized populations, perpetuating social inequalities. For instance, [Zhang et al. 2020] demonstrated that these models associated African Americans and

Black individuals with prisons, while linking Whites and Caucasians with hospitals in a fill-in-the-blank task. Bias in AI models trained on health data can arise from protected attributes such as ethnicity and race. Several studies have explored methods to evaluate and mitigate these biases. [Júnior et al. 2022] analyzed the COMPAS dataset and found that the classifier disproportionately predicted recidivism for Black individuals, favoring other groups. [Noseworthy et al. 2020] emphasized the importance of reporting model performance separately for all protected attributes to ensure transparency. [Park et al. 2021] found that bias mitigation algorithms were more effective than simply removing the protected attribute suspected of causing bias. Similarly, [Mandhala et al. 2022] used pre-training bias metrics to evaluate models for three datasets, concluding that mitigation strategies improved performance for disadvantaged groups.

While significant research has focused on defining and mitigating bias, a gap remains in studies assessing the direct impact of pre-training bias on the performance of ML models, particularly in the healthcare domain. This work aims to address this gap by analyzing the correlation between pre-training bias metrics and the performance of supervised ML algorithms.

4. Methodology

To investigate the impact of pre-training bias on ML performance, we designed a systematic approach involving data collection, preprocessing, bias manipulation, model training, and correlation analysis. Our methodology involves the following steps:

1. **Data Collection:** We gathered four datasets from different sources, each containing protected attributes relevant to bias analysis. The datasets, along with their sources and protected attributes, are detailed in Table 1.
2. **Data Preprocessing:** We cleaned the datasets by removing missing values and encoding categorical variables.
3. **Computing Cramer's V Coefficient:** We calculated Cramer's V coefficient to identify correlations between categorical features and the target variable, which was used in the computation of the CDDL metric.
4. **Bias Manipulation:** We artificially modified the training sets by adjusting data distributions to either increase or decrease bias levels.
5. **Model Training:** We trained three supervised ML models—Logistic Regression, Decision Tree, and Random Forest—on each dataset variation. For robustness, we performed 10 train-test splits per dataset variation.
6. **Pre-training Bias Metrics Calculation:** We calculated the CI, KL Divergence, KS, and CDDL metrics for each dataset, averaging over the ten-fold splits.
7. **Performance Analysis:** We examined the relation between the computed pre-training bias metrics and the accuracy of the trained models to assess how dataset bias influences predictive performance.

The ML models were implemented using the Scikit-learn library [Pedregosa et al. 2011], an open-source ML framework for Python. All hyperparameters remained constant across the four datasets to ensure comparability.

Table 1. Datasets used in this work

Source	Dataset Name	Protected Attributes	Number of instances
[Newman et al. 1998]	Heart	Sex	303
[Maslej et al. 2022]	IntersectionalBias	Sex, Race	11000
[Teboul 2023]	Diabetes	Sex, Age, Education, Income	253680
[Tasci et al. 2022]	Glioma	Gender, Race	840

5. Experiments and Results

5.1. Heart Dataset

The first dataset was the Heart dataset, a real-life scenario collected from patients, consisting of around 300 instances, used to predict heart disease. For the **highly imbalanced** version of this dataset, we removed 85% of women with attribute 'thal' equals to 2 and negative output, 80% of women with 'thal' equals to 3 and negative output, 80% of women with attribute 'cp' equals to 2 and negative output, 80% of women with 'cp' equals to 0 and negative output, and 20% of instances with positive output. These changes resulted in an increase in the values for all metrics. However, CDDL was still above 0.4 for the two correlated variables. In the **equally balanced** version, we managed to get all metrics close to zero. The complete report of the metrics is in Table 2.

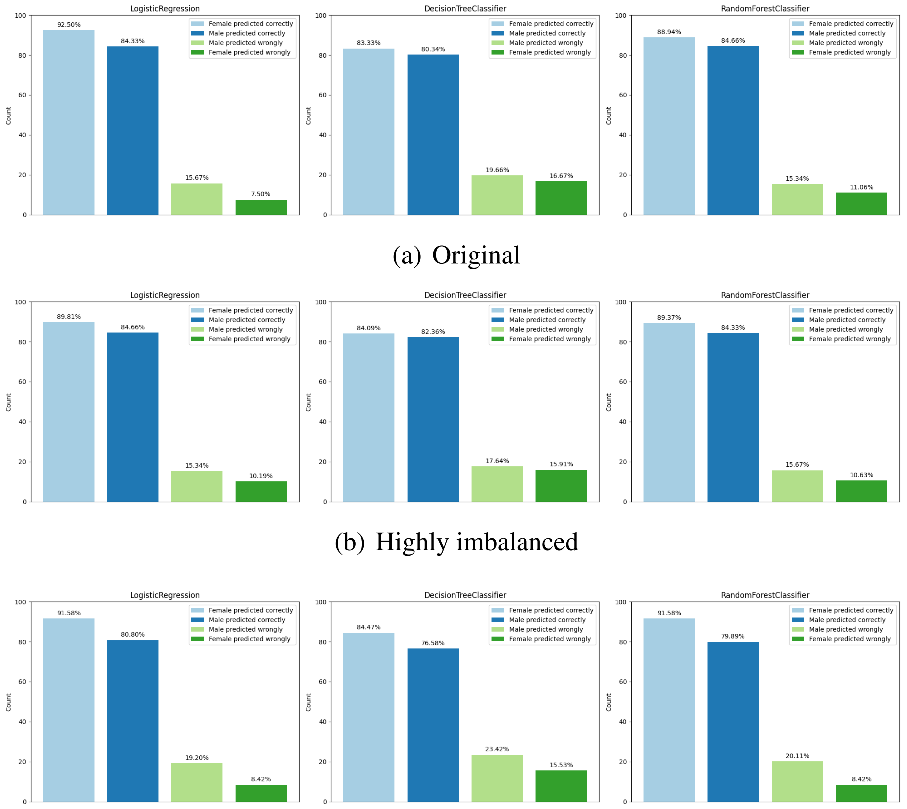
Table 2. Pre-training metrics for the Heart dataset

	Original	High Imbalance	Equal Balance
Class Imbalance (sex)	0.357	0.548	0.000
KL Divergence (sex)	0.202	1.346	0.000
KS (sex)	0.299	0.524	0.000
CDDL (sex, cp)	0.291	0.375	0.086
CDDL (sex, thal)	0.108	0.275	-0.123

The performance for each dataset variation and the mean dataset size used for the training are in Table 3. The results do not indicate the impact of the pre-training metrics, as seen in Figure 1, since false positive and false negative rates were similar across dataset variations. Feature importance analysis, shown in Figure 2, suggests that the varying attributes may not have significantly influenced the decision-making process.

Table 3. Performance results on the protected attributes for the Heart dataset

	Attribute	Class	Accuracy	F1-Score	Mean Train Set Size
Original Dataset	Sex	Female	86.486	0.909	241.0
		Male	79.608	0.774	
High Imbalance	Sex	Female	85.946	0.912	211.2
		Male	80.627	0.786	
Equal Balance	Sex	Female	87.748	0.919	155.0
		Male	73.490	0.746	



(a) Original

(b) Highly imbalanced

(c) Equally balanced

Figure 1. Prediction results for the feature ‘sex’ on the Heart dataset.

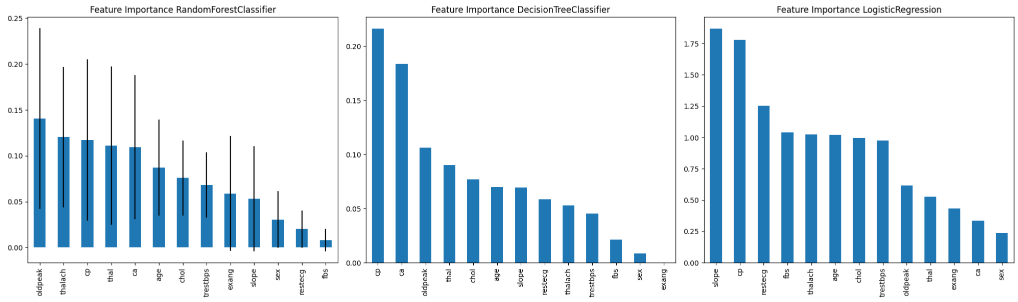


Figure 2. Feature Importance for the original version of the Heart dataset

5.2. Intersectional-bias Dataset

The second dataset was artificially created to predict a diagnosis (schizophrenia or depression) [Maslej et al. 2022]. The dataset contained two protected attributes, ‘sex’ (with “female” as the unprivileged class) and ‘race’ (with “non-white” as the unprivileged class). To achieve the metrics’ values reported in Table 4, in the **high imbalance** version, we removed 95% of instances of non-white with negative output, 50% of non-white with positive output, 80% of women with negative output, and 95% of women with positive output. We note that the metrics for the ‘sex’ attribute did not get above 0.5; increasing them would lead to even less data in the training set, deviating from the metrics’ analysis.

Prediction results for the protected attribute are reported in Table 5.2. We iden-

Table 4. Pre-training metrics for the Intersectional bias dataset

	Original	High Imbalance	Equal Balance
Class Imbalance (Sex)	-0.103	0.755	0.000
KL Divergence (Sex)	0.077	0.474	0.000
KS (Sex)	0.195	0.459	0.000
CDDL (Sex, Rumination)	-0.184	-0.207	0.005
Class Imbalance (Race)	-0.268	0.235	0.000
KL Divergence (Race)	0.018	0.938	0.000
KS (Race)	0.096	0.503	0.000
CDDL (Race, Rumination)	0.079	0.500	-0.007

tified that the performance, especially in the 'sex' attribute, was affected in the highly imbalanced version, with "female" having close to 9% less accuracy. However, this could also be caused by the number of instances in the train set size since the equally balanced that had better performance was more than two times larger, and the original was even larger.

Table 5. Performance results on the protected attributes for the Intersectional Bias dataset

	Attribute	Class	Accuracy	F1-Score	Mean Train Set Size
Original Dataset	Sex	Female	80.994	0.778	5280.0
		Male	95.207	0.962	
	Race	Non-White	85.849	0.872	
		White	90.090	0.892	
High Imbalance	Sex	Female	71.423	0.710	1529.7
		Male	93.594	0.950	
	Race	Non-White	77.317	0.820	
		White	88.512	0.870	
Equal Balance	Sex	Female	79.932	0.780	3114.4
		Male	94.317	0.954	
	Race	Non-White	83.927	0.858	
		White	90.723	0.900	

5.3. Glioma Dataset

The Glioma dataset contains information about 20 genes (mutated or not) for the classification of brain tumors that can be classified into LGG (Lower-Grade Glioma) or GBM (Glioblastoma Multiforme). This dataset also contains two protected attributes: Gender (with "female" being the unprivileged) and Race (with "non-white" being the unprivileged). To make the **highly imbalanced** version on this dataset, we removed 70% of women with Grade 1, 75% of men with Grade 0, 75% of non-white people with Grade 0, and 55% of white people with Grade 1. The pre-training bias metrics for this dataset are provided in Table 6.

In this dataset, we observed a different pattern: the highly imbalanced version had the worst performance compared to the other two, with around three times more data

Table 6. Pre-training metrics for the Glioma dataset

	Original	High Imbalance	Equal Balance
Class Imbalance (Gender)	-0.158	-0.415	-0.197
KL Divergence (Gender)	0.008	0.601	0.028
KS (Gender)	0.062	0.454	0.118
CDDL (Gender, Age_at_diagnosis)	-0.061	-0.535	0.064
Class Imbalance (Race)	0.822	0.857	-0.032
KL Divergence (Race)	0.080	0.994	0.000
KS (Race)	0.199	0.638	0.000
CDDL (Race, Age_at_diagnosis)	0.021	0.054	0.001

Table 7. Performance results on the protected attributes for the Glioma dataset

	Attribute	Class	Accuracy	F1-Score	Mean Train Set Size
Original Dataset	Gender	Female	86.022	0.857	670.0
		Male	85.659	0.834	
	Race	Non-White	74.537	0.796	
		White	86.936	0.855	
High Imbalance	Gender	Female	64.785	0.442	332.0
		Male	82.316	0.801	
	Race	Non-White	66.204	0.739	
		White	72.504	0.598	
Equal Balance	Gender	Female	82.258	0.817	93.4
		Male	83.236	0.801	
	Race	Non-White	67.824	0.732	
		White	84.049	0.821	

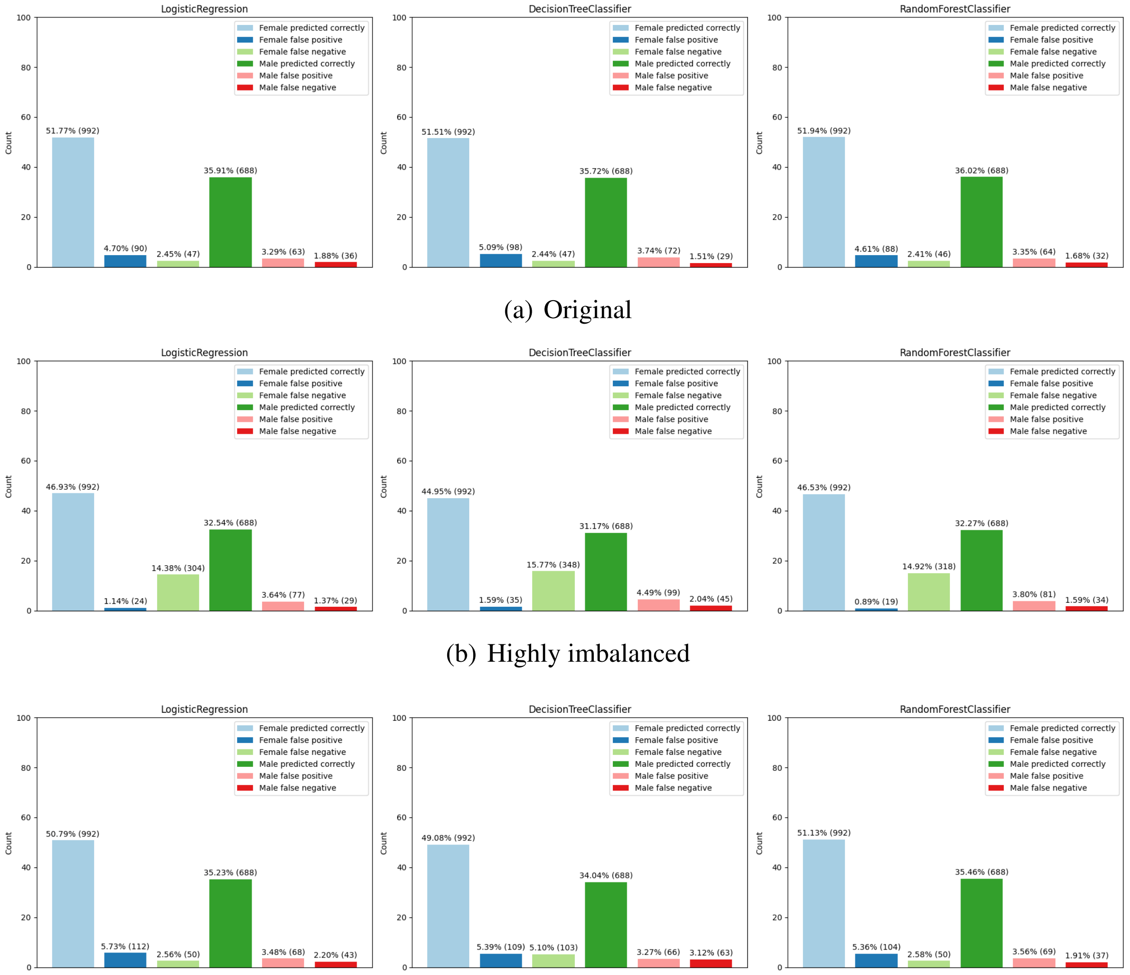
than the equally balanced version, and the performance affected primarily females and non-white people, as reported in Table 7. As we see in Figure 3, especially in 3(b), we can infer that all models were prone to wrongly predicting the negative class for women, which would result in misdiagnosis for specific groups. The importance of each variable for the particular models, as reported in Figure 4, does not show the protected attributes as important features and shows *Age_at_diagnosis* (which is correlated to the protected attributes) as important to the decision-making process in the decision trees.

5.4. Diabetes Dataset

The diabetes dataset contains 70,692 survey responses. The target variable Diabetes_binary has 2 classes: 0 is for no diabetes, and 1 is for prediabetes or diabetes. This dataset has 21 features, and for the experiments, we considered Sex and Age for the analysis of pre-training bias.

In the **highly imbalanced** version of this dataset, we removed 85% of women with negative output, 85% of men with positive output, 55% of people aged between 18 and 24 with positive output, 45% of people with age above 24 with negative output, 20% of women with low blood pressure with negative output and 20% of men with high blood pressure with positive output, respectively. The report on the metrics' values is in Table 8.

The performance results for the two protected attributes are in Table 9. In this



(a) Original

(b) Highly imbalanced

(c) Equally balanced

Figure 3. Prediction results for the feature ‘Gender’ on the Glioma dataset.

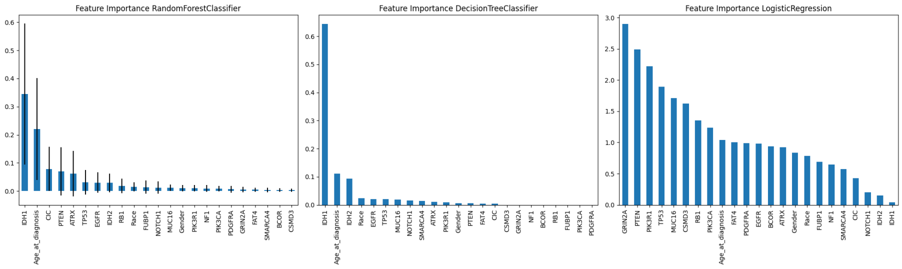


Figure 4. Feature Importance for the original version of the Glioma dataset

dataset, we see that the high values from the metrics do translate to the performance on the unprivileged classes, especially for females, with an accuracy 30% lower than the other variations, even with the highly imbalanced dataset containing 53 times more data than the equally balanced. In the age attribute, the opposite occurred; accuracy was lower in the privileged class (age above 24). F1-scores were also considered very low for this dataset, which is indicative of the small number of responses with a positive outcome, given that the dataset is inherently unbalanced in the target attribute, with 194377 and 35097 responses without diabetes and with diabetes, respectively. Analyzing the importance of each feature for this dataset, as shown in Figure 5, we can see that protected attributes were not the most important features in the decision-making process for the original and equally balanced versions of this dataset. However, for the unbalanced version, these attributes had higher importance rates (except for the Logistic Regression model), which directly reflected in the performance for the unprivileged groups.

Table 8. Pre-training metrics for the Diabetes dataset

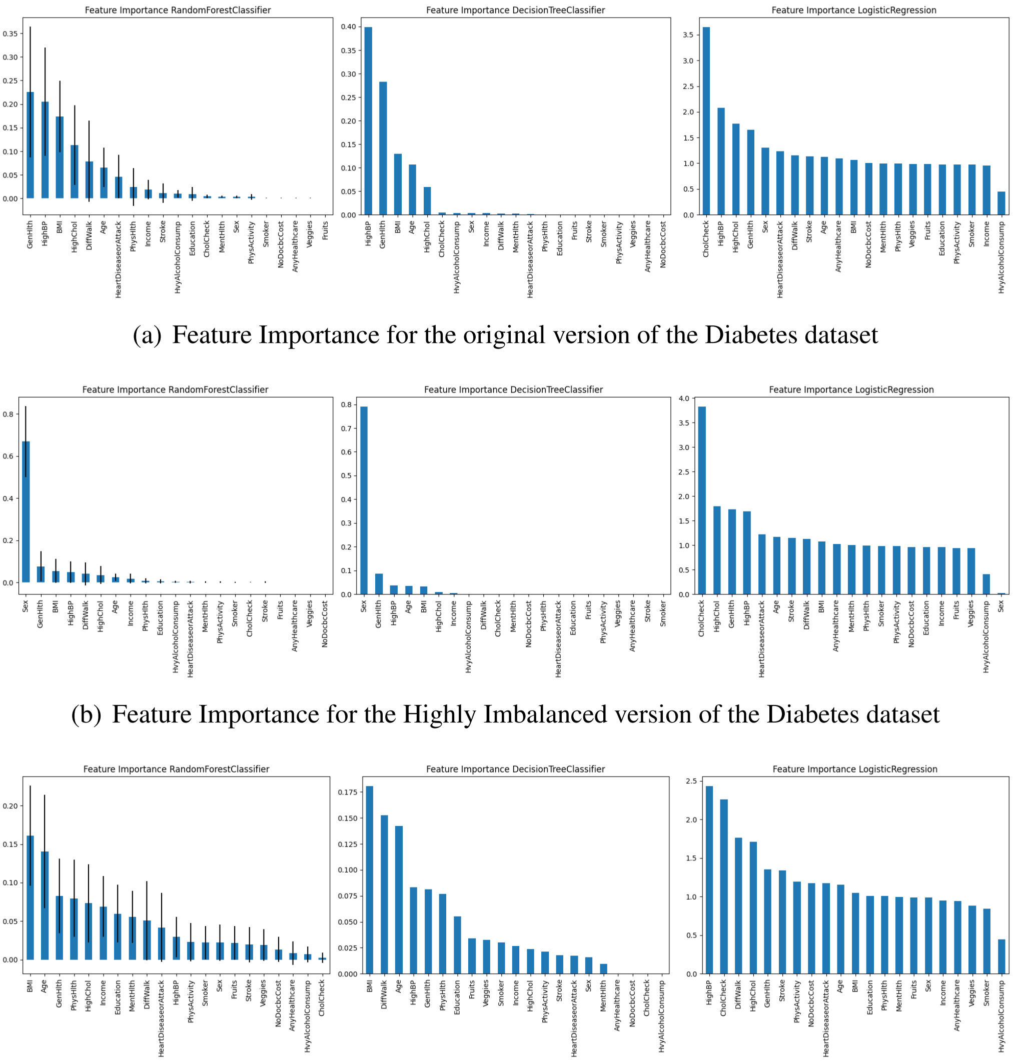
	Original	High Imbalance	Equal Balance
Class Imbalance (Sex)	-0.122	0.247	0.000
KL Divergence (Sex)	0.002	1.064	0.000
KS (Sex)	0.024	0.681	0.003
CDDL (Sex, HighBP)	-0.044	0.744	-0.005
Class Imbalance (Age)	0.952	0.905	0.000
KL Divergence (Age)	0.241	0.900	0.159
KS (Age)	0.142	0.314	0.136
CDDL (Age, HighBP)	-0.022	-0.057	-0.397

Table 9. Performance results on the protected attributes for the Diabetes dataset

	Attribute	Class	Accuracy	F1-Score	Mean Train Set Size
Original Dataset	Sex	Female	86.205	0.181	183579.0
		Male	83.719	0.159	
	Age	Unprivileged	83.245	0.213	
		Privileged	85.120	0.167	
High Imbalance	Sex	Female	53.063	0.361	53216.9
		Male	83.352	0.008	
	Age	Unprivileged	70.975	0.339	
		Privileged	66.299	0.304	
Equal Balance	Sex	Female	84.570	0.233	1002.4
		Male	82.276	0.207	
	Age	Unprivileged	80.790	0.229	
		Privileged	83.575	0.217	

6. General Discussion

Our analysis of the four datasets provides insights into the impact of pre-training bias metrics on model performance across the three trained models. The findings in Sections 5.4 and 5.3 indicate that increasing the dataset size without considering fairness implications can lead to the accumulation of biased data, which, in turn, may reduce model accuracy—especially for historically underrepresented groups. Notably, the observed bias effects were consistent across all three models, suggesting that the bias influenced performance regardless of the algorithm used. We also note that unbalanced data regarding the target attribute in the prediction doesn't affect the bias, as seen in Section 5.4, where the F1-Scores were low, whereas we could still see the impact on the protected classes.



(a) Feature Importance for the original version of the Diabetes dataset

(b) Feature Importance for the Highly Imbalanced version of the Diabetes dataset

(c) Feature Importance for the Equally Balanced version of the Diabetes dataset

Figure 5. Feature Importance for the Diabetes dataset

Examining feature importance across dataset variations revealed that, in all cases except for the Heart Dataset (discussed in Section 5.1), the highly imbalanced versions assigned significant weight to protected attributes, ranking them among the top five most important features. This effect was particularly pronounced in tree-based models (Decision Tree and Random Forest). These findings emphasize that bias in the training data can strongly influence model decision-making.

To ensure transparency and reproducibility, we provide the complete set of charts, tables, code, and instructions for replicating our experiments on GitHub.[1]. This study emphasizes that we use empirical modifications on datasets to artificially introduce or reduce data bias. While this approach does not alter the fundamental nature of the data, whether it is artificial or real, it does change how the data reflects the society from which it was gathered. In real-world scenarios, where the validation set (or where the model is

[1] https://github.com/diegodimer/SBCAS2025

applied) might not reflect the data from where it was trained, we provide the pre-training metrics as a way of assessing the data quality, but it is up to the application developers to define whether the bias is harmful or not.

7. Conclusion

This work assessed the impact of pre-training bias on the performance of ML algorithms in health datasets. We analyzed four pre-training bias metrics: Class Imbalance (CI), Kullback-Leibler (KL) Divergence, Kolmogorov-Smirnov (KS), and Conditional Demographic Disparity in Labels (CDDL). Our experiments on four datasets showed that high values of these metrics often correlate with lower performance for underprivileged groups. In particular, for the Glioma and Diabetes datasets, we found that dataset size did not necessarily correlate with better performance—highly imbalanced datasets, despite being larger, performed worse than their balanced counterparts. Additionally, the impact of pre-training bias was more pronounced in datasets with larger sample sizes, and real-world datasets exhibited stronger bias effects compared to artificially generated ones, as observed in the IntersectionalBias dataset.

These findings highlight the importance of addressing bias before training models. Pre-training bias metrics can serve as early indicators of dataset quality, helping mitigate bias before computational resources are invested in model training. Future research directions include: (i) extending bias investigations beyond protected attributes by analyzing the full feature set; (ii) conducting sensitivity analyses to quantify the numerical impact of bias metrics; (iii) incorporating model explainability techniques, such as PCA and SHAP values, to better interpret biased predictions; (iv) developing new performance metrics and mitigation strategies to further reduce bias and its effects on pre-training metrics; and (v) analyze if the models can benefit from artificially changing the data prior to training, aiming to better reflect the real-world scenarios.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Funding Code 001, and the Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS), through projects 21/2551-0002052-0 (Project MARCS) and 22/2551-0000390-7 (Project CIARS). M. Recamonde-Mendoza is partially funded by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) [308075/2021-8].

References

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., and Zhang, Y. (2018). Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv*.

Caton, S. and Haas, C. (2024). Fairness in machine learning: A survey. *ACM Computing Surveys*, 56(7):1–38.

Chen, I. Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K., and Ghassemi, M. (2021). Ethical machine learning in healthcare. *Annual Review of Biomedical Data Science Annu. Rev. Biomed. Data Sci*, 2021:123–144.

Juhn, Y. J., Ryu, E., Wi, C.-I., King, K. S., Malik, M., Romero-Brufau, S., Weng, C., Sohn, S., Sharp, R. R., and Halamka, J. D. (2022). Assessing socioeconomic bias in machine learning algorithms in health care: a case study of the houses index. *Journal of the American Medical Informatics Association*, 29(7):1142–1151.

Júnior, R. L. I., Silveira, L., de Faria, V. C. N., and Lorena, A. C. (2022). Justiça nas previsões de modelos de aprendizado de máquina: um estudo de caso com dados de reincidência criminal. In *Anais do XIX Encontro Nacional de Inteligência Artificial e Computacional*, pages 636–647. SBC.

Mandhala, V. N., Bhattacharyya, D., Midhunchakkaravarthy, D., and Kim, H. J. (2022). Detecting and mitigating bias in data using machine learning with pre-training metrics. *Ingenierie des Systemes d'Information*, 27:119–125.

Maslej, M., Sikstrom, L., Reslan, D., and Wang, Y. (2022). Intersectional-bias-assessment.

Newman, D., Hettich, S., Blake, C., and Merz, C. (1998). UCI repository of machine learning databases.

Noseworthy, P. A., Attia, Z. I., Brewer, L. P. C., Hayes, S. N., Yao, X., Kapa, S., Friedman, P. A., and Lopez-Jimenez, F. (2020). Assessing and mitigating bias in medical artificial intelligence: The effects of race and ethnicity on a deep learning model for ecg analysis. *Circulation: Arrhythmia and Electrophysiology*.

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453.

Park, Y., Hu, J., Singh, M., Sylla, I., Dankwa-Mullan, I., Koski, E., and Das, A. K. (2021). Comparison of methods to reduce bias from clinical prediction models of postpartum depression. *JAMA Network Open*, 4.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Synthesized.io (2023). Fairlens: A toolkit for fair and interpretable machine learning. `https://github.com/synthesized-io/fairlens`. Accessed: 2024-10-12.

Tasci, E., Zhuge, Y., Kaur, H., Camphausen, K., and Krauze, A. V. (2022). Hierarchical voting-based feature selection and ensemble learning model scheme for glioma grading with clinical and molecular characteristics. *International Journal of Molecular Sciences*, 23(22).

Teboul, A. (2023). Diabetes health indicators dataset. Accessed: 2024-12-30.

Zehlike, M., Castillo, C., Bonchi, F., Hajian, S., and Megahed, M. (2017). Fairness measures: Datasets and software for detecting algorithmic discrimination. `https://fairnessmeasures.github.io`.

Zhang, H., Lu, A. X., Abdalla, M., McDermott, M., and Ghassemi, M. (2020). Hurtful words. In *ACM CHIL 2020 - Proceedings of the 2020 ACM Conference on Health, Inference, and Learning*, pages 110–120. Association for Computing Machinery, Inc.