Can Data Complexity Measures Detect Pre-Training Bias in Machine Learning? A Case-Study with Health Data

Gabriel Differoni Leal[1], **Diego Dimer Rodrigues**[1], **Júlia Mombach da Silva**[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

{gabriel.dleal,ddrodrigues,jmsilva,mrmendoza}@inf.ufrgs.br

Abstract. *Bias in healthcare data can negatively impact vulnerable populations and reduce the reliability of predictive models. This work investigates the use of data complexity measures to identify features that may introduce bias during model training with machine learning (ML) algorithms. We analyze a synthetic dataset on schizophrenia and depression and a real dataset on liver disease, evaluating complexity across subgroups defined by protected attributes such as sex and race. The approach is validated against traditional pre-training bias metrics. Preliminary results suggest that data complexity measures can serve as an early indicator of bias, supporting the development of fairer and more transparent predictive models. This framework could inform bias mitigation strategies to improve model fairness in health-related ML applications.*

1. Introduction

Understanding the complexity of a dataset is fundamental to identifying biases in machine learning (ML) models since the complexity metrics are directly related to the quality and diversity of the data, as well as to the model's ability to generalize without learning biased patterns. This topic holds particular significance in healthcare, as these models aid in diagnosing health conditions and recommending treatments. By identifying a non-representative dataset, ineffective treatments for specific population groups can be avoided, thus reducing the risk of medical complications. However, data analysis is often complicated, mainly due to a lack of information, noisy data, missing or poorly updated records, or information with input errors.

This work aims to explore the use of data complexity measures to identify features at-risk of bias in health datasets, comparing the measures with existing pre-training bias metrics to validate our approach. It is beneficial to use complexity measures in this context, as they have shown value to recommend classifiers and pre-processing techniques for a certain dataset, thus offering a new application for their use in a ML pipeline [Lorena et al. 2019]. This proposal evaluates two datasets and their variations based on protected attributes. It assesses complexity metrics and analyzes structure to determine whether certain groups are underrepresented or disadvantaged. Our preliminary results indicate that these complexity measures provide valuable insights and can be used alongside pre-training bias metrics to identify biases for or against specific groups, providing an additional method for evaluating biases in machine learning.

2. Related Work

The study of data complexity metrics has gained significant attention in recent years due to their importance in learning the characteristics of datasets and their influence on classification problems. For example, [Sotoca et al. 2005] focused on describing these measures and relating them to trained classifiers for a given problem, their performance directly related to the characteristics of the data used. The study highlighted the main measures in the literature, separating them into groups that best describe their characteristic and discussing their possible applications in pattern classification. Similarly, [Lorena et al. 2019] analyzed measures derived from training datasets to characterize the complexity of the corresponding classification problems. The implementation of the metrics is available in an R package called ECoL (*Extended Complexity Library*), which allows the calculation of metrics according to an arbitrary input dataset. [Arruda et al. 2020] analyzes these metrics from another perspective, reducing some measures from the literature to the instance level, demonstrating that they can provide a complementary concept on the difficulty of instances.

One critical aspect discussed in the literature on data complexity is the presence of pre-training bias in machine learning, which arises from imbalances or disparities in data representation. Previous work of our group [Rodrigues 2023] investigated the presence of bias in health data using traditional pre-training bias metrics and assesses how it can affect the performance of ML algorithms through modified datasets to introduce or reduce bias metrics and correlate its values with the generated models' performance. However, an analysis of data complexity measures in this domain and a comparison with existing metrics considering the potential to detect biases is a research topic that has not been explored before.

3. Theoretical Background

The data complexity measures employed in this work, as presented in [Lorena et al. 2019], describe the regularities and irregularities contained in a dataset and are used to assess the difficulty of separating instances into their expected classes. We divide the measures into three groups:

1. **Featured-based measures:** Describes the level of information provided by the attributes available to distinguish the classes. Includes measures F1, F1v, F2, F3 and F4.
2. **Linearity measures:** Evaluates the feasibility of separating classes using a linear approach. Includes measures L1, L2 and L3.
3. **Neighborhood measures:** Characterizes the presence and density of equal or different classes in local neighborhoods. Includes measures N1, N2, N3, N4, N5 and N6.

For detailed information on each complexity measure, refer to GitHub[1]. It is important to note that these measures vary between 0 and 1, except for the N6 measure that varies between 0 and $1 - \frac{1}{n}$, with n being the number of instances in the dataset. The closer the value is to the upper limit, the greater the complexity of the data analyzed.

The pre-training bias metrics used in this work follow the choices in our previous study [Rodrigues 2023] and consider for the calculation protected attributes with two

[1] `https://github.com/gdleal11/sbcas2025`

facets: one representing the group favored by the bias, i.e., the historically favored group, while the other represents the disadvantaged group, i.e., the historically disfavored group. For example, if the attribute is sex, the favored group is male, and the disfavored group is female. These metrics are as follows:

1. **Class Imbalance (CI):** Measures the imbalance in the distribution of instances between the groups of the demographic attribute considered. The values vary between -1 and 1, with positive values indicating that instances of the favored group have greater representation than the disadvantaged.
2. **Kullback-Leibler (KL) Divergence:** Measures the difference between label distributions (for the predictions' target attribute) on the protected attribute considered. The range of values for this metric is between 0 and ∞, where values close to 0 mean that the different values for the target attribute are evenly distributed. Positive value means divergence; the larger the value, the more significant the divergence.
3. **Kolmogorov-Smirnov (KS):** Measures the maximum divergence between labels in the distribution for the groups of the demographic attribute considered. The values vary between 0 and 1, with values close to zero indicating that the target attribute is evenly distributed between the groups.
4. **Conditional Demographic Disparity in Labels (CDDL):** Evaluates demographic disparity by checking whether the target attribute is independent of the demographic attribute considered. The values vary between -1 and 1, with positive values indicating demographic disparity and negative values suggesting the opposite.

4. Materials and Methods

4.1. Datasets

The first dataset used for the experiments was the **Intersectional-Bias Dataset** [Maslej et al. 2022]. It was created artificially, containing demographic and clinical attributes of people with a possible diagnosis of schizophrenia or depression. This dataset has two protected attributes, **race** and **sex**, and the **diagnosis** is the target attribute for the prediction task. The dataset has a total of 11,000 instances and 19 attributes.

The second dataset used was the **Indian Liver Patient Dataset** [Ramana and Venkateswarlu 2022]. It comprises data from Andhra Pradesh, India, and contains social and clinical information on patients with a positive or negative diagnosis of liver disease. The protected attribute is **sex**, and the target attribute is **diagnosis**. The dataset has a total of 584 instances with 11 attributes.

To better analyze the complexity measures concerning the protected attributes, we derive artificial datasets from the original: For the Intersectional-Bias Dataset, we create four versions: containing only women (6063 instances), only men (4937 instances), only white people (4026 instances), and only non-white people (6974 instances). For the Indian Liver Patient Dataset, we create two versions: containing only women (142 instances) and only men (441 instances).

4.2. Complexity Measures and Pre-Training Bias Metrics

We calculate the complexity measures and pre-training bias with the generated datasets, as outlined in Section 3. We employ the R package ECoL [Lorena et al. 2019] to obtain

the complexity measures. We utilize the code developed in [Rodrigues 2023] to assess pre-training bias metrics. The target attribute is the input to calculate these metrics in all instances.

The statistical technique PCA (Principal Component Analysis) was applied to investigate better the structure of the datasets used. According to [Karamizadeh et al. 2013], this method provides a tool that allows multidimensional data to be reduced to lower dimensions, keeping most of the information. PCA is valuable for understanding how instances of different classes are distributed in the dataset, enabling the visualization of the overlap between classes.

5. Results

To better analyze our results, we compute the score for each complexity group, as outlined in Section 3. This score uses the value of each complexity measure, following the formula described in Equation 1. We decided to use these values because we believe that together they provide a balanced and complementary view based on the grouped measures. The *Mean Absolute* generally summarizes the average degree of complexity within the group. *Max Absolute* is used to highlight the most critical measure within the group, and is ideal for identifying relevant outliers. *Std Dev* reveals whether there is consistency or heterogeneity in the assessment of complexity within the group. A higher value can indicate that some measures are in disagreement, contributing to greater complexity.

$$\text{Score}_{\text{group}} = \text{Mean Absolute} + \text{Max Absolute} + \text{Std Dev} \tag{1}$$

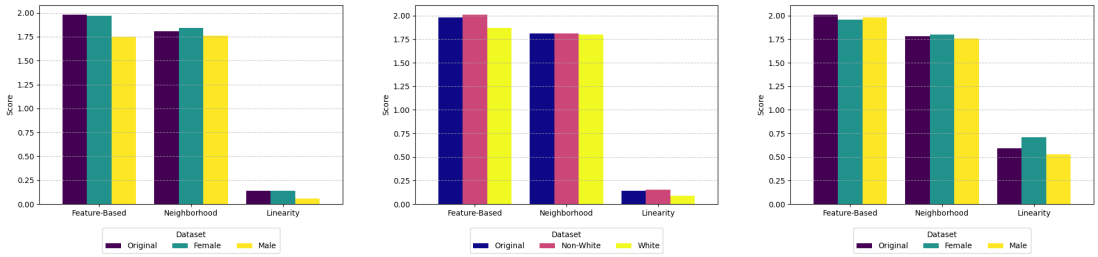
Other results obtained during this work that were not included due to space constraints, such as the figures for each complexity measure and the PCA of the datasets. These plots are available on Github[1].

5.1. Intersectional-Bias Dataset

Figure 1(a) shows the scores for the Intersectional-Bias Dataset and its variations from the attribute *sex*. We highlight that the measures for the disadvantaged group were higher than for the bias-favored group. This difference indicates that the available attributes in the dataset with only women face more challenges separating the instances into their target classes than the male variation. In addition, due to the increase in the linearity group, it is inferred that separating classes using linear computing is more complicated. The PCA analysis can explain the value in the group of neighborhood measures for both datasets, which shows a more significant overlap between female and male instances.

The scores concerning the protected attribute *race* are shown in Figure 1(b). Similarly, the disadvantaged class also obtained higher metrics. The overlap of the instances of the non-white dataset is likewise more significant, according to the PCA analysis for both variations.

Table 1 shows the pre-training bias metrics obtained for the Intersectional-Bias dataset. We point out that this dataset has a more prominent representation of disadvantaged classes in both protected attributes, justifying the negative value for Class Imbalance. All values were generally positive and close to zero, indicating slight bias, favoring the privileged class. The scores in Figures 1(a) and 1(b) also show this behavior, with a not-so-significant difference between the protected attributes groups.



(a) Intersectional-Bias and sex variations (b) Intersectional-Bias and race variations (c) Indian Liver Patient and sex variations

Figure 1. Scores for original datasets and variations

Table 1. Pre-Training Bias Metrics for Intersectional-Bias Dataset

Metric name	Value
Class Imbalance (Sex)	-0.103
KL Divergence (Sex)	0.076
KS (Sex)	0.194
CDDL (Sex)	-0.182
Class Imbalance (Race)	-0.268
KL Divergence (Race)	0.017
KS (Race)	0.092
CDDL (Race)	0.075

Table 2. Pre-Training Bias Metrics for Indian Liver Patient Dataset

Metric name	Value
Class Imbalance (Sex)	0.512
KL Divergence (Sex)	0.017
KS (Sex)	0.086
CDDL (Sex)	0.063

5.2. Indian Liver Patient Dataset

The scores obtained for the Indian Liver Patient Dataset and its variations are illustrated in Figure 1(c). The analysis indicates that the female dataset reported higher values in the Neighborhood and Linearity groups, indicating greater complexity. However, for the Feature-Based group, the male dataset was demonstrated to be more complex, mainly due to the significant increase in the F4 metric, impacting the score. Nevertheless, the score difference for this group was approximately 1%, indicating a more similar complexity compared to the other calculated groups. The PCA analysis obtained was very similar for the two datasets despite the higher score in the neighborhood group obtained by the female dataset.

The pre-training bias metrics for the Indian Liver Patient Dataset are displayed in Table 2. The positive and higher value of the Class Imbalance measure demonstrates a more significant representation of men than women with a considerable difference. Similarly to the Intersectional-Bias Dataset, the other measures were positive and close to zero. As evidenced by the majority of the scores, this result indicates a bias favoring the privileged class. However, the metrics' values were insignificant for the protected attribute considered.

6. Conclusion

Our results indicate that data complexity metrics can consistently reflect group-specific bias in health datasets. In the Intersectional-Bias Dataset, the difference in complexity

between the disadvantaged and advantaged groups was evident from the scores and pre-training metric calculations. Both the female and non-white groups proved to be the most negatively affected in the overall context of the dataset. Nevertheless, in the Indian Liver Patient Dataset, scores for the disadvantaged class did not increase in all complexity groups, and the pre-training bias metrics were not considerably low values. However, as the results show, we point out that the dataset is not free of bias that favors the male group.

The conclusions reached in this work contribute to the current context of machine learning models and data classification and motivate further experiments. By calculating the complexity of subsets of data divided according to a protected attribute and comparing the results with pre-training bias metrics, we suggest that this new approach can pose a new tool to obtain fairer results, especially in health scenarios. For future works, we aim to expand our experiments to new datasets and conduct a systematic analysis of results to derive stronger hypotheses and conclusions.

Acknowledgments

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

References

[Arruda et al. 2020] Arruda, J. L., Prudêncio, R. B., and Lorena, A. C. (2020). Measuring instance hardness using data complexity measures. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part II 9*, pages 483–497. Springer.

[Karamizadeh et al. 2013] Karamizadeh, S., Abdullah, S. M., Manaf, A. A., Zamani, M., and Hooman, A. (2013). An overview of principal component analysis. *Journal of signal and information processing*, 4(3):173–175.

[Lorena et al. 2019] Lorena, A. C., Garcia, L. P. F., Lehmann, J., Souto, M. C. P., and Ho, T. K. (2019). How complex is your classification problem?: A survey on measuring classification complexity. *ACM Computing Surveys (CSUR)*, 52(1):1–34.

[Maslej et al. 2022] Maslej, M. et al. (2022). Intersectional-Bias-Assessment. INCF. Available on internet: https://training.incf.org/lesson/intersectional-approach-model-construction-and-evaluation-mental-healthcare.

[Ramana and Venkateswarlu 2022] Ramana, B. and Venkateswarlu, N. (2022). ILPD (Indian Liver Patient Dataset). UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5D02C.

[Rodrigues 2023] Rodrigues, D. D. (2023). Assessing pre-training bias in health data and estimating its impact on machine learning algorithms. Bachelor's thesis, Ciência da Computação, Instituto de Informática, Universidade Federal do Rio Grande do Sul.

[Sotoca et al. 2005] Sotoca, J. M., Sánchez, J. S., and Mollineda, R. A. (2005). A review of data complexity measures and their applicability to pattern classification problems. *Actas del III Taller Nacional de Mineria de Datos y Aprendizaje. TAMIDA*, 77.