Innovation in Applied Nursing Informatics
G. Strudwick et al. (Eds.)

doi:10.3233/SHTI240171

Fairness in Classifying and Grouping Health Equity Information

Ruinan JIN,[a] Xiaoxiao LI,[a] Lorraine J. BLOCK[b], Ivan BESCHASTNIKH[a], Leanne M. CURRIE[a], and Charlene E. RONQUILLO [a]
[a] *The University of British Columbia, Vancouver and Okanagan, BC, Canada*
[b] *Vancouver Coastal Health Authority*

ORCiD ID: Ruinan Jin https://orcid.org/0009-0000-1421-2555;
Xiaoxiao Li https://orcid.org/0000-0002-8833-0244;
Lorraine Block https://orcid.org/0000-0002-9496-3208;
Ivan Beschastnikh https://orcid.org/0000-0003-1676-8834;
Leanne M. Currie https://orcid.org/0000-0002-8232-2809;
Charlene E. Ronquillo https://orcid.org/0000-0002-6520-1765

Abstract. This paper explores the balance between fairness and performance in machine learning classification, predicting the likelihood of a patient receiving anti-microbial treatment using structured data in community nursing wound care electronic health records. The data includes two important predictors (gender and language) of the social determinants of health, which we used to evaluate the fairness of the classifiers. At the same time, the impact of various groupings of language codes on classifiers' performance and fairness is analyzed. Most common statistical learning-based classifiers are evaluated. The findings indicate that while K-Nearest Neighbors offers the best fairness metrics among different grouping settings, the performance of all classifiers is generally consistent across different language code groupings. Also, grouping more variables tends to improve the fairness metrics over all classifiers while maintaining their performance.

Keywords. Fairness and Bias, Electronic Health Record, Feature Engineering

1. Introduction

Machine Learning (ML) has gained substantial popularity in the field of high-dimensional data classification [1], particularly within the context of electronic health records (EHR) [2]. Despite remarkable advancements in performance, there remains a critical research gap regarding the fairness and potential biases of ML algorithms. The concern is that, without meticulous design and oversight, ML algorithms could continue or even worsen existing health disparities [3].

This paper aims to address this gap by exploring the impact of classification fairness and performance in classifying social determinants of health (SDoH), focusing on wound care in community nursing EHRs. The paper consists of two goals: 1) investigate the tradeoff between performance and fairness among different classification algorithms, and 2) explore the effect of grouping predictor variables during pre-processing on classifiers.

2. Methods

2.1 Community Nursing Wound Care EHR

This study used data acquired from a large health region in British Columbia, Canada, serving a diverse multicultural population. The data from Jan 1, 2019 to Dec, 31 2021, were extracted from two distinct community EHRs: 1) general community care with detailed patient info, and 2) for nursing wound care management, with extensive data on wound assessments and treatments. Over the three-year span, the patient count was 4,843, 4,952, and 5,196, respectively. All patients were present in both community EHRs and analysis was based on patient ID.

We aimed to predict the likelihood of a patient receiving anti-microbial treatment based on various factors: SDoH (gender, language spoken, marital status, additional contact listed, age), and care provision (diagnosis, body part under assessment, type of assessment, and the wound being assessed). All data were included in the classification algorithms. The fairness metrics were calculated on patient language and gender, two areas of potential bias. In the data preprocessing stage, we employed one-hot encoding for all variables, where each class is assigned a binary value of 0 or 1, except birth dates [4].

The data contained 238 language codes. This vast dimensionality, particularly evident in the one-hot encoding process, posed a significant challenge. To address this, we explored the strategy of grouping language codes based on their frequency of occurrence. Specifically, we considered encoding only the top k language codes, while aggregating the remaining codes into one category. In this study, we experimented with different values of k, namely 1 (English or Other), 2 (English and Cantonese, or Other), 119 (half of the total language codes), and the full set of 238.

2.2 Classification algorithms

Existing research on high-dimensional structured data classification investigates the effect of feature selection and classification performance on different algorithms. We selected classifiers aligned with existing studies on bacteremia prediction [2] and feature selection [1]. Our chosen classifiers are grouped as linear and non-linear:

Linear classifiers:

- **Logistic Regression (LR)** [5] is used for binary classification. It models the probability of a default class and is recognized for its simplicity and interpretability. Two LR algorithms with L1 and L2 regularization were run.
- **Linear Discriminant Analysis (LDA)** [6] distinguishes itself through its foundations in Bayes' rule. It presumes Gaussian class-conditional densities, with each class having its own mean but sharing a common covariance matrix.

Non-linear classifiers:

- **Random Forest (RF)** [7] is a prominent non-linear method and is particularly valued for its feature selection capabilities, as demonstrated through RF's feature importance metric.
- **K-Nearest Neighbors (KNN)** [8] classifies samples based on the closest feature space neighbors. This method is particularly useful for scenarios where the relationship between features and classes is not linearly separable.

- **Support Vector Machines (SVM)** [9] stands out in classification for its effectiveness in high-dimensional spaces. This method, particularly adept at binary classification, creates optimal hyperplanes in multidimensional space to distinguish between classes.

2.3 Fairness and Bias in Health Data

A common strategy for promoting fairness in ML involves enforcing equal treatment for various groups, guided by certain methods. For example, one method would focus on addressing disparate treatment, ensuring procedural fairness, and providing equal opportunities. Another method would emphasize reducing disparate impacts and inequalities in outcomes, fostering distributive justice.

2.3.1 Fairness Metrics

We evaluate the fairness of classifiers on two sensitive predictors, patient gender and language code, using the three metrics below. Let $D \leftarrow (X, Y, C)$ be the dataset with X be protected variables, Y to be everything else, and C to be classes.

- **Disparate Impact (DI)** evaluates how much more (or less) likely a model is to predict that a data point of sensitive group 0 would recidivate vs sensitive group 1 [10]. Mathematically, the classification algorithm has a DI of τ where $\frac{Pr(f(Y)=1|X=0)}{Pr(f(Y)=1|X=1)} \leq \tau$. A fair classifier has a DI close to 1.
- **True Positive Parity (TPP)** refers to the situation where different groups have an equal probability of being correctly identified as positive by the classifier [11]. Similar to DI, TPP assumes the label C to be binary. A classification algorithm is said to have TPP if for groups $x \in X$, $|P_{X=x}(f(Y) = 1|C = 1) - P_{X=x'}(f(Y) = 1|C = 1)|$. We evaluate the difference between the two terms, where a smaller difference indicates a more fair classifier.
- **Predictive Value Parity (PVP)** measures the accuracy of positive predictions across different groups in binary classification [11]. The PVP for a given class is defined as $|P_{X=x}(C = 1|f(Y) = 1) - P_{X=x'}(C = 1|f(Y) = 1)|$. A smaller difference between the two terms indicates a more fair classifier.

In our experiments, we implemented the grouping strategies based on language code as outlined in section 2.1, followed by constructing classifiers as described in section 2.2. We then assessed DI, TPP, and PVP for each classifier, considering *gender* and *language* separately. Our paper presents the average values of DI, TPP, and PVP across *gender* and *language* for each classifier.

3. Results

The classification performance and fairness metrics associated with each grouping strategy are depicted in the 3D plots below. Each plot illustrates the classification performance via a color bar and shown in the dots showing the different classification algorithms in each plot, where lighter shades indicate superior classifier performance. The fairness metrics are represented along the three axes of these plots. For TPP and PPV, a smaller value indicates a more fair classification. For DI, a value close to 1 is considered ideal, reflecting equitable treatment across groups.

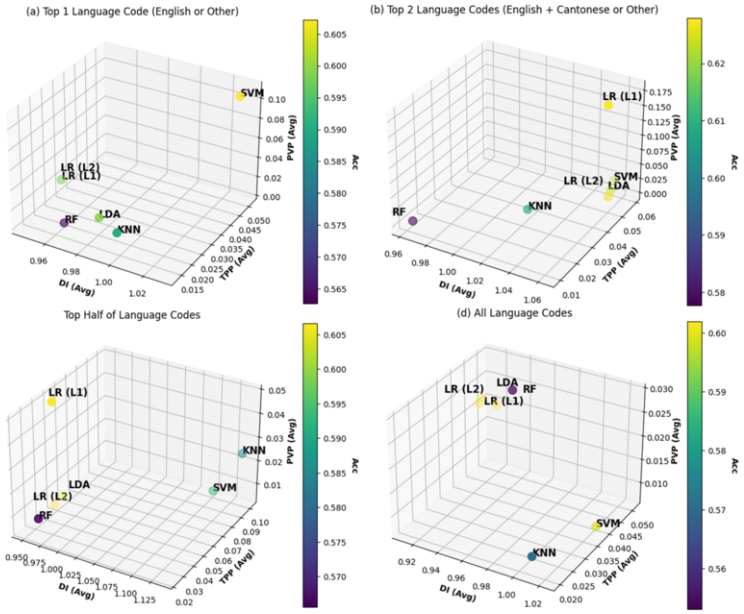

Figure 1. Visualization of classification and fairness performance for different language groupings. The DI is on the x-axis, TPP on the y-axis, and PVP on the z-axis.

4. Discussion

4.1 Fairness among Different Classifiers

Among the classifiers we evaluated, the KNN algorithm consistently achieves the best averaged fairness metrics for *gender* and *language*. Across all four plots, KNN maintains a DI score very close to 1, demonstrating its effectiveness in fair classification. In scenarios (a), (b), and (d), the KNN classifier is positioned towards the bottom of the figure, indicating a minimal difference in TPP and PPV across different groups. hile KNN excels in fairness, its performance is lower than the best-performing classifiers, though not by much. For instance, in scenario (a), KNN achieves an accuracy of 59.05%, which is only slightly lower than the 60.72% accuracy achieved by SVM. This highlights KNN's balanced approach between maintaining fair classification and ensuring competent performance.

4.2 Effect of Grouping on Language Code

As discussed in Section 2, our experiments involved varying the grouping of language codes, with the outcomes illustrated in the subplots of Figure 1. In subplot (d), where all language codes are retained, the highest accuracy recorded is 60.19% for LDA. The peak accuracy of 62.79% is observed in subplot (b) with the L1-regularized LR, where only the top 2 language codes (English, Cantonese) are preserved. Across all four subplots, the accuracies hover around 60%, suggesting that grouping language codes has a negligible impact on classifier performance. In terms of fairness, we note a deterioration in fairness metrics with an increase in k. As k grows from plot (a) to (d), the data points are progressively shifted toward the front top of the plot. This shift indicates an increase

in both TPP and PPV. Consequently, this suggests that classifiers tend to yield varying results across different sensitive groups, particularly as the number of language codes included in the analysis increases.

5. Conclusions

This study shows that while grouping language codes in EHRs impacts fairness metrics, it has a minimal effect on classifier performance. KNN outperforms other classifiers in fairness but does not lead in accuracy. Grouping more variables tends to improve the fairness of classifiers, highlighting the trade-offs between fairness and performance in ML classifications on different groupings. As interest grows in adding SDoH to ML algorithms, our findings provide methodological advances in enhancing fairness without significantly compromising classifier performance.

References

[1] R.-C. Chen, C. Dewi, S.-W. Huang, and R.E.J.J.o.B.D. Caraka, Selecting critical features for data classification based on machine learning methods, 7 (2020), 52.
[2] O. Garnica, D. Gómez, V. Ramos, J.I. Hidalgo, and J.M.J.E.J. Ruiz-Giardín, Diagnosing hospital bacteraemia in the framework of predictive, preventive and personalised medicine using electronic health records and machine learning classifiers, 12 (2021), 365-381.
[3] E. Ooghe, E. Schokkaert, D.J.S.C. Van de Gaer, and Welfare, Equality of opportunity versus equality of opportunity sets, 28 (2007), 209-230.
[4] J.T. Hancock and T.M.J.J.o.B.D. Khoshgoftaar, Survey on categorical data for neural networks, 7 (2020), 1-41.
[5] J.L. Alzen, L.S. Langdon, and V.K.J.I.j.o.S.e. Otero, A logistic regression investigation of the relationship between the Learning Assistant model and failure rates in introductory STEM courses, 5 (2018), 1-12.
[6] A. Tharwat, T. Gaber, A. Ibrahim, and A.E.J.A.c. Hassanien, Linear discriminant analysis: A detailed tutorial, 30 (2017), 169-190.
[7] L.J.M.l. Breiman, Random forests, 45 (2001), 5-32.
[8] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, KNN model-based approach in classification, in: On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings, Springer, 2003, pp. 986-996.
[9] M.A. Hearst, S.T. Dumais, E. Osuna, J. Platt, B.J.I.I.S. Scholkopf, and t. applications, Support vector machines, 13 (1998), 18-28.
[10] M. Feldman, S.A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, Certifying and removing disparate impact, in: proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, 2015, pp. 259-268.
[11] P. Garg, J. Villasenor, and V. Foggo, Fairness metrics: A comparative analysis, in: 2020 IEEE International Conference on Big Data (Big Data), IEEE, 2020, pp. 3662-3666.