PhD proposal: Improving Algorithmic Fairness with Missing Values

Context. In the past three decades, Machine Learning (ML) methods have been broadly used in medicine, notably for diagnosis purposes [Kononenko, 2001, Erickson et al., 2017, Shehab et al., 2022]. As ML algorithms work by extracting information from large data sets, they can be sensitive to hidden biases. Applying such algorithms without further guarantees can lead to reproduce and even reinforce existing bias in the data, which can be dramatic in sensitive domains such as medical diagnosis [see Gianfrancesco et al., 2018]. Algorithmic fairness, which has become increasingly popular in the last decade, aims at reducing the influence of sensitive attributes on a prediction and can thus be used to mitigate biases in diagnosis Rajkomar et al. [2018]. However, algorithmic fairness is must often studied in the context of complete observations, whereas many real data sets contain missing values. This PhD project aims at (i) studying theoretically the impact of missing values on fairness and (ii) creating and analyzing algorithms trained on missing data with good fairness properties and predictive accuracy.

Research topic. There are two main ways of producing fair classifiers: the in-processing approach, which modifies the existing learning procedure (via a change of loss, or equivalently a fairness constraint), or the post-processing approach which first learns a classifier and then applies post-processing operations to finally produce a fair classifier. While second approaches are much more generic (as the post-processing step can be applied to wide range of algorithms), in-processing approaches may outperform them as they directly target the correct quantity. Accordingly, this PhD project is decomposed into two axes: studying first generic post-processing approaches in presence of missing data and then in-processing approaches based on random forests. Our final aim is to compare these approaches and provide guidance about practical use of fairness techniques in medical diagnosis. We will test these approaches on the Traumabase Dataset[1], a real-world data set composed of clinical data related to emergency [see, e.g., Jiang et al., 2020].

Link with the theme (health). In ICU settings, such as the Traumabase example, machine learning models are developed to predict patient outcomes. For example, using real-time data collected in ambulances, we aim to predict the risk of hemorrhagic shock, the need for neurosurgery, and the availability of specialized resources in trauma centers. These predictive models improve patient triage, ensuring they are directed to the most appropriate hospitals when necessary. This has significant economic and human implications, as misdirection can either unnecessarily mobilize resources or lead to critical consequences for patients. The Traumabase dataset includes patients with major trauma from car accidents, falls, and stab wounds, with 80% being men. As a result, there is concern that predictive models may not perform equitably for women. Another source of non equity stems from missing data themselves. If certain features are not collected for some patients, there is a risk that they may receive suboptimal care simply due to the missing values. Finally, it is well known that a patient's socio-economic status can influence the quality of care they receive, which may lead to undesirable disparities.

1 First axis - Post-processing methods

Since the seminal work of Rubin [1976], missing data have been categorized as Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR). MCAR values are easier to study since missingness is independent of both inputs X and output Y (in a supervised setting). Thus, MCAR scenarios imply a loss of information, but do not introduce additional biases. Conversely, in MNAR scenarios, missingness may depend on unobserved values, introducing unrecoverable biases. While much of the literature focuses on parameter estimation with missing values, recent research explores its impact on prediction [Le Morvan et al., 2021, Ayme et al., 2022, Josse et al., 2024].

In this first axis, we place ourselves in a binary classification setting, and assume that we may have access to a (binary) sensitive attribute (for example, gender). We want to study the impact of MCAR missing values on

[1] `https://www.traumabase.eu/fr_FR`

fairness properties, when missing values may occur on S only, on the other input variables only, or on both. In particular, we want to know whether artificially introducing missing values in a data set can help to create fair predictors (while maintaining good predictive performances). Besides, is it possible to create fair predictors that do not use S at prediction in the three above scenarios? We will assess the algorithmic fairness via the well-studied *Demographic Parity* metric [Calders et al., 2009, Denis et al., 2021].

Existing literature. Some recent works consider the settings in which either the sensitive attribute or the other input variables are missing [Zhang and Long, 2021, Kallus et al., 2022, Feng et al., 2024]. Kallus et al. [2022] circumvent this problem by using an additional data set in order to reconstruct the sensitive attribute, given other variables. Such imputation approaches may induce a bias, whose magnitude depends on the tuning parameters of imputation procedures [Chen et al., 2019]. Very few finite-sample analyses of fairness in presence of missing data exist, with the notable exception of Zhang and Long [2021].

2 Second axis - In-processing methods with random forests

Random forests [Breiman, 2001] are among the state-of-the-art methods to solve supervised learning problems with tabular data sets [Fernández-Delgado et al., 2014]. Each tree of a random forest recursively splits the input space based on a data-dependent criterion. This splitting criterion is chosen with respect to the final metric to optimize. In this second axis, we turn to in-processing methods and design random forests tailored for achieving fairness by modifying the splitting criterion accordingly.

Several propositions of fair random forests have already been made [see, e.g., Raff et al., 2018, Zhang et al., 2021]. Raff et al. [2018] change the splitting criterion to discourage splitting along variables that are correlated with both the sensitive attribute and the output. Based on the framework developed in Zhang and Ntoutsi [2019] for single trees, Zhang et al. [2021] propose a new splitting criterion and design FARF, an online and fair random forest. Very few theoretical results on fair random forests exist, as the RF analysis is already challenging [Biau and Scornet, 2016]. Given the expertise of the supervisors, we plan to fill in the gap, at least for some stylized random forests models as the centered forests [Biau, 2012].

As tree-based methods, random forests are naturally well-equipped to deal with inputs with missing values in both training and test sets [Twala et al., 2008, see, e.g.,]. Therefore, one can wonder what theoretical guarantees are available for fair random forests in presence of missing data, in terms of fairness and accuracy. We are particularly interested in comparing the in-processing approach of this axis to the post-processing approach of the first axis. We are also interested in studying the behavior of variable importance measures [MDI and MDA, see Breiman, 2001, 2002] in presence of missing data, to see how they can be used to detect or encourage fairness, in the light of recent works [Bénard et al., 2022].

3 PhD supervisors

Erwan Scornet (PhD director, LPSM, Sorbonne Université; Research theme: Missing values, random forests) is a researcher at LPSM at Sorbonne University since 2023. He is a specialist in random forest algorithms. Author of 9 papers on missing data and 19 on tree-based methods, he has co-authored a literature review on the subject reference in the field (cited 4000 times). He has co-supervised five theses, all of which have been defended, and is currently co-director of two other theses.

Julie Josse (co-supervisor, PreMedicaL, Inria Montpellier; Research themes: Missing values, medical applications) is a research director at Inria, leading the PreMeDICaL team. She leads the Traumatrix program, developing AI-driven ambulance decision support tools. She also advances reproducible research through open-source R software. Julie has a strong international network, participating in prestigious events like Berkeley's causality semester and the Rousseeuw Prize. A former visiting researcher at Stanford and Google Brain Paris, she has received the Prix Jeunes Chercheurs Inria-Académie des Sciences and a Marie Curie grant. She has published 35 papers on missing values and supervised 14 PhD students.

Christophe Denis (co-supervisor, SAMM, Université Paris 1; Research theme: Fairness) is professor at SAMM at Panthéon-Sorbonne University since 2024. His research mainly focuses on supervised learning. He is an expert in Algorithmic fairness. Author of 4 papers on fairness. He has already supervised two theses. He is currently co-director of two other theses.

References

Igor Kononenko. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in medicine*, 23(1):89–109, 2001.

Bradley J Erickson, Panagiotis Korfiatis, Zeynettin Akkus, and Timothy L Kline. Machine learning for medical imaging. *radiographics*, 37(2):505–515, 2017.

Mohammad Shehab, Laith Abualigah, Qusai Shambour, Muhannad A Abu-Hashem, Mohd Khaled Yousef Shambour, Ahmed Izzat Alsalibi, and Amir H Gandomi. Machine learning in medical applications: A review of state-of-the-art methods. *Computers in Biology and Medicine*, 145:105458, 2022.

Milena A Gianfrancesco, Suzanne Tamang, Jinoos Yazdany, and Gabriela Schmajuk. Potential biases in machine learning algorithms using electronic health record data. *JAMA internal medicine*, 178(11):1544–1547, 2018.

Alvin Rajkomar, Michaela Hardt, Michael D Howell, Greg Corrado, and Marshall H Chin. Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine*, 169(12):866–872, 2018.

Wei Jiang, Julie Josse, Marc Lavielle, TraumaBase Group, et al. Logistic regression with missing covariates—parameter estimation, model selection and prediction within a joint-modeling framework. *Computational Statistics & Data Analysis*, 145:106907, 2020.

Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

Marine Le Morvan, Julie Josse, Erwan Scornet, and Gaël Varoquaux. What'sa good imputation to predict with missing values? *Advances in Neural Information Processing Systems*, 34:11530–11540, 2021.

Alexis Ayme, Claire Boyer, Aymeric Dieuleveut, and Erwan Scornet. Near-optimal rate of consistency for linear models with missing values. In *International Conference on Machine Learning*, pages 1211–1243. PMLR, 2022.

Julie Josse, Jacob M Chen, Nicolas Prost, Gaël Varoquaux, and Erwan Scornet. On the consistency of supervised learning with missing values. *Statistical Papers*, 65(9):5447–5479, 2024.

Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE international conference on data mining workshops*, pages 13–18. IEEE, 2009.

Christophe Denis, Romuald Elie, Mohamed Hebiri, and François Hu. Fairness guarantee in multi-class classification. *arXiv preprint arXiv:2109.13642*, 2021.

Yiliang Zhang and Qi Long. Assessing fairness in the presence of missing data. *Advances in neural information processing systems*, 34:16007–16019, 2021.

Nathan Kallus, Xiaojie Mao, and Angela Zhou. Assessing algorithmic fairness with unobserved protected class using data combination. *Management Science*, 68(3):1959–1981, 2022.

Raymond Feng, Flavio Calmon, and Hao Wang. Adapting fairness interventions to missing values. *Advances in Neural Information Processing Systems*, 36, 2024.

Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 339–348, 2019.

Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. Do we need hundreds of classifiers to solve real world classification problems? *The journal of machine learning research*, 15(1):3133–3181, 2014.

Edward Raff, Jared Sylvester, and Steven Mills. Fair forests: Regularized tree induction to minimize model bias. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 243–250, 2018.

Wenbin Zhang, Albert Bifet, Xiangliang Zhang, Jeremy C Weiss, and Wolfgang Nejdl. Farf: A fair and adaptive random forests classifier. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 245–256. Springer, 2021.

Wenbin Zhang and Eirini Ntoutsi. Faht: an adaptive fairness-aware decision tree classifier. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 1480–1486, 2019.

Gérard Biau and Erwan Scornet. A random forest guided tour. *Test*, 25:197–227, 2016.

Gérard Biau. Analysis of a random forests model. *The Journal of Machine Learning Research*, 13(1):1063–1095, 2012.

Bheki ETH Twala, MC Jones, and David J Hand. Good methods for coping with missing data in decision trees. *Pattern Recognition Letters*, 29(7):950–956, 2008.

Leo Breiman. Manual on setting up, using, and understanding random forests v3. 1. *Statistics Department University of California Berkeley, CA, USA*, 1(58):3–42, 2002.

Clément Bénard, Sébastien Da Veiga, and Erwan Scornet. Mean decrease accuracy for random forests: inconsistency, and a practical solution via the sobol-mda. *Biometrika*, 109(4):881–900, 2022.