Auditing of AI systems through explainability

Jorge Vindel-Alfageme[1,*], Juan Antonio Recio-Garcia[1] and María Belén Díaz-Agudo[1]

[1]Universidad Complutense de Madrid (UCM), C/ del Profesor José García Santesmases, 9, 28040, Madrid, Spain

Abstract

Artificial intelligence (AI) application has expanded widely. One of the fields that has benefited the most from it is healthcare. The complexity of each biological system is to consider when implementing AI. AI models are susceptible to bias. These errors impact people's lives in healthcare. Explainable AI (XAI) methods allow us to delve deeper into AI models to understand them. Therefore, they become a tool in their bias analysis. On the other hand, Case-Based Reasoning (CBR) is a framework based on empirical evidence that allows deciding the optimal solution from experience gained from real cases. It is worth highlighting the importance these systems can have in healthcare, where each patient represents a case, and new patients' analyses would be facilitated by being fitted within its framework. This PhD project focuses on the development of methodologies for auditing healthcare AI systems, with an emphasis on identifying and mitigating biases. Using XAI, combined with CBR, the aim is to create a reusable framework that assesses the fairness of predictive models in medical diagnosis. The work includes the formalization of an ontology to structure risks and solutions, as well as the implementation of a technological platform that integrates validated use cases in healthcare.

Keywords

Artificial intelligence bias, artificial intelligence in healthcare care, case-based reasoning

1. Introduction

It is often said that biases associated with artificial intelligence (AI) models are the result of the negative legacy left by the data used to train the model [1]. However, there can be algorithmic biases linked to the model's training and the algorithm itself, and not just those due to the input data [2]. These are the two main sources of bias that we can find in AI models. Thus, these sources of bias are present in AI models generally applied in AI implementation domains, including healthcare.

To fully understand the specific biases that can cause AI models to suffer from unexpected error rates, we can associate each type of bias with a data processing stage. In a data analysis workflow, multiple steps in data processing can be identified. Speaking in the same general terms as above, negative legacy would result from altered data recording or error-inducing preprocessing, and algorithmic bias would appear during model training.

Analysis can be considered to begin with the very recording of data using specialized tools. Data recording gives them a characteristic structure that determines their subsequent processing. During data recording, there may be underestimation in the sampling of sample classes, underestimation of sample subgroups depending on the values adopted by some variables (for example, the variable "sex" may cause a sample group to be underrepresented), or an imbalance between the number of samples from different classes and subgroups. These problems are what can induce a negative legacy in the model, since it is the data in their raw form that determines biased results. To overcome these drawbacks, tools such as resampling or synthetic sample generation can be used to help maintain a balance between the sample types in the data set [3]. Evaluating the data set in its raw version can serve to detect possible biases in the model and mitigate their effect through resampling techniques.

To assess the presence of biases associated with the AI model, classification model performance metrics can be used. These metrics are based on measuring the samples classified by an AI model

ICCBR 2025 Doctoral Consortium, Workshops at the 33rd International Conference on Case-Based Reasoning (ICCBR-WS 2025)

*Corresponding author.

jorgevin@ucm.es (J. Vindel-Alfageme); jareciog@ucm.es (J. A. Recio-Garcia); belend@ucm.es (M. B. Díaz-Agudo)

https://www.github.com/JorgeVindelAlfageme (J. Vindel-Alfageme)

0000-0002-1375-6902 (J. Vindel-Alfageme); 0000-0001-8731-6195 (J. A. Recio-Garcia); 0000-0003-2818-027X (M. B. Díaz-Agudo)

according to the following types: true positives (samples correctly classified in a model where there is a class with a condition different from the control or reference class), false positives, true negatives (samples correctly classified in a model where there is a class with the control or reference class), and false negatives. These values are typically represented by the following acronyms, respectively: TP, FP, TN, and FN. Thus, different formulas are defined based on these four values, which represent metrics that speak to the model's effectiveness.

There are many defined metrics and some common ones used for model performance evaluation, such as accuracy (obtained by dividing the number of correctly classified samples by the full sample size), specificity (the probability that a result is negative, conditioned on the sample being truly negative), positive predictive value (PPV) (the probability that a sample with a positive result on a test has the condition being assessed), etc.

The bias present in the model can be assessed by calculating disparities in error rates. Using the formula for one of the previous metrics, rates for these metrics can be established to evaluate the disparity between sample groups, based on the values of these error measures associated with each of the groups. From the data sets, subgroups of samples can be established according to the criterion applied by the user. One criterion for assessing bias is grouping samples according to the value of variables considered sensitive, such as sex (female or male), ethnicity (Caucasian, African, Latino or Hispanic, Middle Eastern Asian, East Asian, Pacific Islander, etc.), or age. Evaluating the disparity rates between the selected groups allows us to verify whether the ratio between performance metrics exceeds a lower (e.g., 0.8) or upper (e.g., 1.25) threshold. This way, you can see if there are subgroups of samples that tend to be classified better or worse by the model based on the above metrics compared to a reference subgroup. There are criteria for defining the reference subgroup. For example, the majority subgroup can be used as the reference group.

The bias assessment of AI models allows us to verify the presence of biases in the models, which are often fed with data volumes where some sample subgroups are larger than others, or where the underestimation of some subgroups leads to the extraction of patterns in the data that are not representative of reality [4, 5, 6].

To date, there are multiple studies that have studied the presence of biases in AI, specifically when applied to the field of healthcare [3, 7, 8, 9, 10]. These studies have attempted to identify biases such as those described above using assessment methods such as disparity calculation, which has revealed biases associated with the AI classifiers produced. However, as this is a novel area, there is no clear consensus among the different studies on how to classify the types of biases and, therefore, what would be the appropriate methods to mitigate them accordingly. There is also general talk of a lack of transparency and standards associated with data recording and publication, to make it easier to detect biases and correct them by considering both the processing and the origin of the data [11].

In general, there is a lack of a clear protocol for bias detection and mitigation in data processing workflows that ultimately produce AI classifiers with applications in healthcare. This protocol would need to establish what type of input data is being handled, how to detect potential types of bias in each case (to achieve this, it is important to consider error evaluation metrics and potential disparities associated with these metrics), and how these can be mitigated after detection. Considering the lack of consensus among researchers, this area reveals that further work is still needed to clarify these problems, something that has a clear impact on people's lives, as it is a healthcare issue.

Given this problem, the first milestone to be achieved in this doctoral project would be to conduct a literature review of the state of the art regarding bias detection and mitigation in AI models in healthcare. After gathering bibliographic information on the screened cases that fit this topic, an ontology could be designed that would gather information on how to process each data set, considering bias assessment, until the classification model is generated. This would establish an organized and consistent protocol for any data set, enabling the design of more accurate and effective AI models.

Case-Based Reasoning (CBR) is a framework based on empirical evidence that allows for determining an optimal solution based on the experience gained from the real cases it was defined with. The above ontology would be represented by a CBR model, where the analyzed datasets define the framework of the model, which invites us to follow guidelines for analyzing and assessing biases, depending on the

nature of the initial dataset and based on the empirical evidence collected after researching the state of the art. To date, there is no record of an ontology applied in this area of healthcare knowledge capable of guiding experts in assessing biases in AI.

It is important to highlight the importance that CBR models can have in healthcare, where each patient represents a case, and the analysis of new patients would be facilitated by fitting within this framework. CBR models would provide a highly suitable system to ensure correct data processing, given the absence of clear standards.

Finally, this ontology could be tested with new datasets to verify whether the system is truly capable of ensuring bias analysis and assessment. This would launch a resource with a significant impact on the medical field, ensuring the ethical and responsible use of AI tools in research.

2. Research plan

Considering the PhD plan with the corresponding funding, the objectives described in the introduction would be carried out over a total of three years. These three years represent a time span that can be divided into phases linked to milestones to be met throughout the PhD.

The first phase would correspond to the first year of the PhD. During this phase, the literature related to the generation of AI models from datasets in the healthcare field for which bias assessment has been considered would be compiled and reviewed. Based on the compilation of all the cases gathered with the literature, an ontology for data analysis would be designed to guide the expert-level bias assessment, implemented according to a CBR model.

During the second phase of the doctorate, which would correspond to years 2 and 3 of the project, the CBR model is expected to be tested with new cases, which would represent new discoveries within the CBR system. Finally, all the findings would be compiled in the final project report.

2.1. Research objectives

The research objectives are related to elucidating the clear bias detection and mitigation mechanisms that affect AI models in healthcare, given that no clear standards currently exist. This can be achieved through empirical evidence reflected in an ontology used to construct a CBR model, which can guide expert-level analysis of new medical datasets.

Surrounding this main objective are several other objectives that must be met. For example, it is necessary to clearly establish the standard structures of the datasets that can be analysed, since, to date, there is still no well-defined standard for datasets for different medical problems. Next, it is important to understand the different stages of data processing and how, throughout them, it is possible to analyse the presence of biases that will affect the final AI model. Among the available evaluation methods, we find the disparity rates of performance metrics between user-defined sample subgroups. Another issue surrounding these performance metrics is their most relevant use, considering the type of dataset being manipulated. There is no literature clarifying the most appropriate use of each performance metric. Finally, tasks such as transforming case evidence into a CBR model and ontology would constitute a paradigm in the field of artificial intelligence and the medical domain, as we have not encountered a similar system to date. Thanks to the CBR architecture, a tool would be created capable of ensuring better use of artificial intelligence in the medical field.

2.2. Approach / Methodology

Considering the research plan and the objectives to be carried out, the first part of the project would consist of preparing a bibliographic review of the state-of-the-art bias assessment in data set analysis in the healthcare field. To do this, it would first be necessary to determine which article search engines are most appropriate for conducting this research based on the topic. An advanced search would then be carried out in the relevant databases to find the raw bibliographic corpus. The articles found would then be manually screened to finally identify the valuable articles for conducting the bibliographic review.

During the writing of the bibliographic review, the data analysis would be adapted to each data set based on the accumulated empirical evidence. This would allow for the establishment of calculations and tools for assessing the bias that may affect each data set and at each stage. With these tools, analyses can be performed to understand their scope and thus better understand the possibilities that the final ontology will cover. Finally, in this first stage, the ontology and fundamental structure of the CBR model would be designed.

During the second phase of the doctoral program, the ontology's principles would be applied to process new medical datasets to validate the ontology's usefulness. Based on these new cases, the ontology may be redesigned to truly represent all possible aspects of dataset evaluation and serve as a means of providing the best possible support during expert analysis of medical datasets.

3. Progress summary

A tool has been developed for measuring bias in classification models using the disparity ratios of different metrics called Aequitas [12]. Aequitas is a tool specifically designed to facilitate the assessment of bias in the form of disparity ratios, using the classification results of an AI model as input. By forming subgroups of samples based on the values they adopt for certain attributes considered sensitive, the user can easily calculate the performance metrics for each subgroup and establish disparity ratios between them, in order to calculate and diagram the presence of bias, as well as calculate the statistical significance associated with these disparity ratios. It is an easy-to-use and versatile tool for AI classification models. Thus, it is considered a valuable method for assessing bias in AI classification models used in the healthcare domain.

Aequitas is a versatile tool that has been studied to understand how it can be applied to medical datasets. Its simplest use relies on the results of a binary classification from an AI model, which also includes attributes with values from which new sample subgroups can be established. This approach allows the user to establish any subgroups and compare the classification rates between them. In the case of bias, there would be a disparity rate far from the quotient equal to 1. This approach allows Aequitas to be used with relatively simple tabular datasets, although it can also be applied in a less straightforward way to evaluate disparity rates, as is the case with image datasets [5, 6].

So far, Aequitas has been tested on simple datasets to learn how to calculate disparity rates between subgroups, how to represent the results in disparity rates and how to calculate their statistical significance, and how to mitigate model bias by reweighting the model.

4. Conclusion and future work

The creation of an ontology and a CBR model that records the instances of datasets and AI models applied in the medical domain and allows for dataset-specific bias assessment is a paradigm that is lacking. The lack of transparency in certain datasets [11] and the absence of a standard for bias detection and mitigation currently makes the issue of bias assessment in AI models applied in the medical field of notable importance. The emergence of calculations such as disparity rates points to some of the consensus that has been established, although it is still an area in which much work remains to be done, among other things due to the high direct impact that changes in the healthcare domain have on the general population. The establishment of a reference framework for bias analysis at the expert level intercedes in the responsible and ethical use of AI, an area in constant expansion due to its applicability, which also occurs at a dizzying speed. This project will help shed light on the most appropriate use of bias assessment in a domain where AI is of enormous importance. The defined ontology can then be constantly tested on new datasets, further refining it and better adapting it to a constantly evolving field of knowledge.

Declaration on Generative AI

During the preparation of this work, the authors used Microsoft Translator Service for the purpose of: assisting translation from Spanish to English. After using this service, the author(s) reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] T. Kamishima, S. Akaho, H. Asoh, J. Sakuma, Fairness-Aware Classifier with Prejudice Remover Regularizer, in: D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, P. A. Flach, T. De Bie, N. Cristianini (Eds.), Machine Learning and Knowledge Discovery in Databases, volume 7524, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 35–50. URL: http://link.springer.com/10.1007/978-3-642-33486-3_3. doi:10.1007/978-3-642-33486-3_3.

[2] S. Hooker, Moving beyond "algorithmic bias is a data problem", Patterns 2 (2021) 100241. URL: https://linkinghub.elsevier.com/retrieve/pii/S2666389921000611. doi:10.1016/j.patter.2021.100241.

[3] F. Chen, L. Wang, J. Hong, J. Jiang, L. Zhou, Unmasking bias in artificial intelligence: a systematic review of bias detection and mitigation strategies in electronic health record-based models, Journal of the American Medical Informatics Association 31 (2024) 1172–1183. URL: https://academic.oup.com/jamia/article/31/5/1172/7634193. doi:10.1093/jamia/ocae060.

[4] I. Straw, H. Wu, Investigating for bias in healthcare algorithms: a sex-stratified analysis of supervised machine learning models in liver disease prediction, BMJ Health & Care Informatics 29 (2022). URL: https://informatics.bmj.com/content/29/1/e100457. doi:10.1136/bmjhci-2021-100457.

[5] L. Seyyed-Kalantari, G. Liu, M. McDermott, I. Y. Chen, M. Ghassemi, CheXclusion: Fairness gaps in deep chest X-ray classifiers, 2020. URL: https://arxiv.org/abs/2003.00827. doi:10.48550/ARXIV.2003.00827.

[6] L. Seyyed-Kalantari, H. Zhang, M. B. A. McDermott, I. Y. Chen, M. Ghassemi, Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations, Nature Medicine 27 (2021) 2176–2182. URL: https://www.nature.com/articles/s41591-021-01595-0. doi:10.1038/s41591-021-01595-0.

[7] M. Nagendran, Y. Chen, C. A. Lovejoy, A. C. Gordon, M. Komorowski, H. Harvey, E. J. Topol, J. P. A. Ioannidis, G. S. Collins, M. Maruthappu, Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies, BMJ (Clinical research ed.) 368 (2020) m689. doi:10.1136/bmj.m689.

[8] M. Sasseville, S. Ouellet, C. Rhéaume, M. Sahlia, V. Couture, P. Després, J.-S. Paquette, D. Darmon, F. Bergeron, M.-P. Gagnon, Bias Mitigation in Primary Health Care Artificial Intelligence Models: Scoping Review, Journal of Medical Internet Research 27 (2025) e60269. doi:10.2196/60269.

[9] O. Perets, E. Stagno, E. B. Yehuda, M. McNichol, L. Anthony Celi, N. Rappoport, M. Dorotic, Inherent Bias in Electronic Health Records: A Scoping Review of Sources of Bias, medRxiv: The Preprint Server for Health Sciences (2024) 2024.04.09.24305594. doi:10.1101/2024.04.09.24305594.

[10] Y. Huang, J. Guo, W.-H. Chen, H.-Y. Lin, H. Tang, F. Wang, H. Xu, J. Bian, A scoping review of fair machine learning techniques when using real-world data, Journal of Biomedical Informatics 151 (2024) 104622. doi:10.1016/j.jbi.2024.104622.

[11] R. Daneshjou, M. P. Smith, M. D. Sun, V. Rotemberg, J. Zou, Lack of Transparency and Potential Bias in Artificial Intelligence Data Sets and Algorithms: A Scoping Review, JAMA dermatology 157 (2021) 1362–1369. doi:10.1001/jamadermatol.2021.3129.

[12] P. Saleiro, B. Kuester, L. Hinkson, J. London, A. Stevens, A. Anisfeld, K. T. Rodolfa, R. Ghani, Aequitas: A Bias and Fairness Audit Toolkit, 2019. URL: http://arxiv.org/abs/1811.05577. doi:10.48550/arXiv.1811.05577, arXiv:1811.05577.