A practical framework for auditing fairness in medical AI

Andreea M. Oprescu, Jorge Vindel-Alfageme, Erik Campos-Espinosa, Marta Caro-Martínez, Belén Díaz-Agudo, M. Carmen Romero-Ternero, and Juan A. Recio-Garcia

Department of Software Engineering and Artificial Intelligence
Instituto de Tecnologías del Conocimiento
Universidad Complutense de Madrid, Spain
{jorgevin,erikcamp,martcaro,belend,jareciog}@ucm.es
Departamento de Tecnología Electrónica
Instituto de Ingeniería Informática (I3US)
Universidad de Sevilla, Spain
{aoprescu,mcromerot}@us.es

Abstract. The increasing deployment of Artificial Intelligence (AI) systems in healthcare has raised significant concerns about bias, fairness, and ethical implications of automated decision-making. While medical AI systems offer capabilities for diagnosis, treatment planning, and patient care, they also inherit and potentially amplify biases present in training data, particularly affecting underrepresented populations. This paper presents a comprehensive framework for auditing AI systems in medical domains through functional audit processes that systematically evaluate bias and fairness. Our approach integrates three key components: data quality assessment, fairness analysis, and application of explainable AI (XAI) techniques. We demonstrate the practical application through the auditing of a machine learning model to predict COVID-19 patient mortality. The results reveal disparities in model performance across different demographic groups.

Keywords: Fairness · Bias auditing · Explainable AI · Healthcare ethics

1 Introduction

AI technologies in healthcare enable diagnostic tools, personalized treatment recommendations, and predictive analytics for patient outcomes. However, AI models present critical challenges related to algorithmic fairness, bias detection, and ethical compliance in medical decision-making systems.

Medical AI systems are particularly susceptible to bias due to historical inequities in healthcare data collection and treatment practices. These biases manifest from underrepresentation of demographic groups in training datasets, systematic exclusion from clinical trials, and perpetuation of discriminatory practices embedded in medical records. Such biases can lead to suboptimal or harmful

outcomes for minority populations, women, elderly patients, and other vulnerable groups. AI auditing can be characterized by its purpose and methodological characteristics. According to Mokander [1], AI auditing is a governance mechanism utilized by different actors: (i) regulators assessing legal compliance, (ii) technology providers mitigating risks, and (iii) stakeholders making informed engagement decisions [2].

The literature on auditing AI systems includes academic articles and books, auditing tools developed by private companies, industry standards, and policy guidance documents. AI system audits can be approached from different perspectives: *functional audits* verify if systems perform tasks accurately and reliably; *code audits* review source code; and *impact audits* investigate outcome effects [3, 4]. Audit strategies vary from generic governance-focused approaches to specific bias evidence collection methods based on domain regulations. This paper follows the latter approach.

Recent advances in eXplainable AI (XAI) have developed methods providing transparency and interpretability for AI systems. We propose our functional audit processes based on the synergy between explainability and auditing concepts. Research demonstrates XAI's utility for auditing: [5] in insurance, [6] using SHAP [7] in finance, [8] in health governance, and others developing "Accountable eXplainable AI" (AXAI) frameworks [9].

While our functional audit framework provides quantitative bias evidence through technical metrics, result interpretation requires interdisciplinary collaboration between AI experts, healthcare professionals, and social scientists. Identified disparities reflect complex social realities shaping healthcare delivery, not merely statistical anomalies. Effective healthcare AI auditing must address underlying social dynamics influencing data collection and model deployment. Furthermore, AI bias interpretation requires considering the social context of data collection and deployment. The COVID-19 pandemic exemplifies how social inequalities can be reflected and amplified through healthcare AI systems. Socio-economic factors, healthcare access, and historical treatment patterns influence both data collection and model predictions. Therefore, effective AI auditing must incorporate sociotechnological perspectives considering how social realities shape technical artifacts and their interpretation [10].

This paper runs as follows: Section 2 describes our three-stage functional audit process for medical domains, including data quality assessment, fairness analysis using Aequitas, and XAI-enhanced auditing. Section 3 presents a COVID-19 mortality prediction case study demonstrating our framework's practical application. Section 4 discusses results from data quality assessment, fairness analysis, and explainable AI methods. Finally, Section 5 concludes with key contributions and outlines future work directions.

2 Functional audit process for medical domains

The medical domains presents unique challenges for AI fairness due to the complex interplay between biological, social, and historical factors that influence

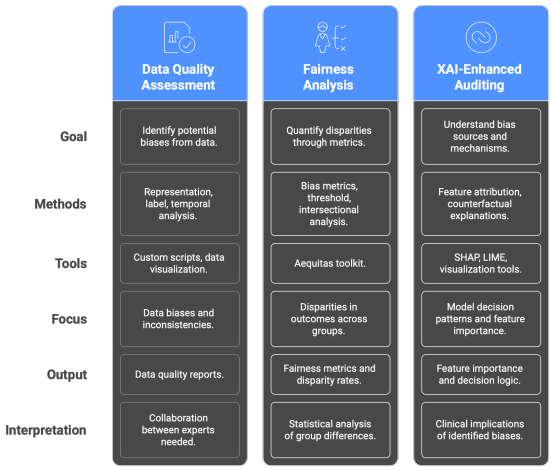

Fig. 1: Description of the proposed auditing process

health outcomes and data collection practices. Understanding these biases is needed for developing effective auditing methodologies. Several key sources of data-centric ethical bias are presented: *Sampling or data representation bias*: biomedical datasets reflect the limitations of the healthcare systems they are extracted from, as, for example, systematic exclusion or undersampling of certain demographic groups. *Labeling bias*: Inconsistent or biased annotation practices that reflect historical prejudices. *Outcome availability bias*: adverse outcomes (e.g., death) can have a higher prevalence in specific demographics, limiting the diversity of positive cases available for model learning. *Temporal bias*: Changes in medical practices and diagnostic criteria over time that are not properly accounted for in historical datasets.

To address these biases, we propose a three-stage audit process: *(1) data quality assessment, (2) fairness analysis,* and *(3) eXplainable AI (XAI) integration* (see Figure 1). This holistic approach identifies potential biases from data to model evaluation, requiring collaboration between domain experts, AI specialists, and social scientists for proper contextual interpretation.

2.1 Data Quality Assessment

Comprehensive data quality assessment forms a crucial component of our auditing procedures: *Representation Analysis*: Systematic evaluation of demographic representation throughout the different splits of the dataset (i.e training and testing), identifying underrepresented groups and potential sampling biases. *Label Quality Assessment*: Analysis of annotation consistency and potential bias in labeling practices, particularly important in medical contexts where diagnostic labels may reflect historical biases. *Temporal Consistency*: Evaluation of how medical practices and diagnostic criteria have evolved over time and their impact on model training and performance.

2.2 Fairness analysis

There are several toolkits for fairness analysis that can be applied to perform this stage. In our case, we have chosen Aequitas, an open-source bias and fairness assessment toolkit to perform functional bias audit processes [11]. Aequitas offers a comprehensive bias audit for medical AI, using fairness metrics such as demographic parity, equalized odds, calibration, and false positive/negative rate balance, with particular emphasis on metrics capturing disparities in diagnostic accuracy across demographic groups.

The tool receives as input the test set containing true outcome labels, model predictions, and sensitive, multivalued demographic attributes that serve as reference variables for organizing sample subgroups. While our current implementation focuses on basic bias metrics assessment, the full Aequitas framework enables more advanced analyses including threshold optimization and intersectional bias detection for comprehensive fairness evaluation.

The following steps describe a fairness audit with Aequitas: *(1) Define subgroups* using multivalue attributes to investigate for bias. *(2) Select fairness metrics*, maintaining consistency with model optimization metrics. *(3) Calculate group-specific metrics* based on selected fairness metrics. *(4) Establish reference group* using majority criterion (selecting the subgroup with the highest number of patients for each multivalue attribute). *(5) Quantify disparity* by comparing subgroup metrics to reference group through disparity rates.

2.3 XAI-Enhanced Auditing

Our framework integrates XAI techniques as a complementary component to Aequitas fairness metrics, creating a comprehensive approach to bias detection in medical AI systems. While Aequitas quantifies disparities through statistical metrics, XAI methods provide insights into underlying mechanisms causing biased outcomes.

XAI integration serves multiple purposes: Feature attribution methods like SHAP [7] and LIME [12] analyze how features contribute to decisions across demographic groups, revealing inappropriate reliance on sensitive attributes or proxies [13]. Beyond individual analysis, XAI identifies systematic bias patterns not apparent from aggregate metrics [14], such as visualization of feature importance distributions showing different decision logic across populations [15]. Counterfactual explanations [16] demonstrate how sensitive attribute changes affect predictions, distinguishing legitimate medical factors from discriminatory patterns. This synergistic approach enables auditors to detect bias presence and understand its sources, facilitating targeted mitigation strategies [17]. XAI interpretability is valuable for communicating results to healthcare professionals who need to understand clinical implications of identified biases [18].

3 Use case: Audit of a COVID-19 mortality Machine Learning (ML) model

To demonstrate our functional audit process, we present a biomedical case study using a real-world COVID-19 clinical dataset to train a ML model to predict patient mortality.

3.1 Dataset

The dataset used is *COVID Data for Shared Learning (CDSL)*, a multimodal biomedical dataset available on Physionet [19], containing 4,479 hospitalized patients in Spanish hospitals HM Hospitales from 2019-12-26 to 2021-02-13. We focus exclusively on tabular data for binary mortality prediction to investigate sex and age biases. We adopt the data preprocessing methodology described in [20]. After preprocessing, features included laboratory values and vital signs from the first 48 hours, admission duration, comorbidities, mechanical ventilation, ICU admission, age and sex. Exclusion criteria was applied: unknown patient destination after discharge (n=159), patients with a hospitalization stay of more than 30 day (n=159), patients without laboratory data within the first 48 hours (n=2016). Following exclusion criteria, 2,145 patients were included with a class imbalance ratio of 5.96 (14.36% mortality class); 1,273 male and 872 female patients. Missing values were handled using Multiple Imputation by Chained Equations for numerical features and zero-filling for categorical ones. Categorical features were one-hot encoded and numerical features were Z-score normalized after stratified train-test split (80%-20%) to prevent data leakage. The split was performed using stratified sampling to preserve the proportion of the target classes (mortality vs. survival) in both subsets.

3.2 Classification Model

We trained and optimized an Extreme Gradient Boosting (XGBoost) model implementing cost sensitive learning to address class imbalance by assigning a higher weight to the positive class (mortality). To evaluate the model's ability to generalize, Repeated Stratified K-Fold cross-validation scheme with 10 splits and 5 repeats (n_splits=10, n_repeats=5) was employed. The metric used to optimize the model is F1-Score, to stress the importance of reducing the number of false negatives (FN) and trying to increase the number of true positives (TP).

The confusion matrix values for the test set were: 353 true negatives (TN), 15 false positives (FP), 16 FN, and 46 TP. The classifier achieved a precision of 0.757 (95% CI: 0.638–0.871), recall of 0.739 (95% CI: 0.627–0.844), and an F1-Score of 0.748 (95% CI: 0.648–0.829).

3.3 Results of the data quality asessment

Our comprehensive data quality assessment revealed important characteristics of the data subsets that inform the subsequent bias analysis.

Table 1: Number and (%) of patients by hospital outcome for train and test sets

Set	Group	Category	Frequency		Percentage (%)	
			Survival	Death	Survival	Death
Train	Age	Adult (27-59 years old)	474	8	98.3	1.7
		Aged adult (60-79 years old)	666	78	89.5	10.5
		Oldest adults (80-98 years old)	312	156	66.7	33.3
	Sex	Male	858	166	83.8	16.2
		Female	612	80	88.4	11.6
Test	Age	Adult (27-59 years old)	144	0	100.0	0.0
		Aged adult (60-79 years old)	155	22	87.6	12.4
		Oldest adults (80-98 years old)	67	39	63.2	36.8
	Sex	Male	210	39	84.3	15.7
		Female	157	23	87.2	12.8

Representation Analysis: To identify sampling bias from underrepresented groups, we analyzed the training and testing sets, which were created by stratifying patients based on their outcome. The sensitive variables available in the dataset are "sex" and "age". Table 1 details the patient demographics for the training and testing sets, which were stratified by outcome. In both the training and testing sets, survival rates decline significantly with increasing age. Regarding sex, the dataset includes a larger number of male patients than female patients across both splits. **Label Quality Assessment**: patient outcome (mortality or survival) is derived from the administrative variables. Patients were labeled with a mortality outcome (1) if their discharge status was recorded as "Death" within 30 days of admission. All other patients were labeled with a survival outcome (0). For this retrospective dataset, no further assessment of the discharge status can be performed. **Temporal Consistency**: Given that all data were collected during a specific pandemic period, temporal bias analysis was not relevant for this dataset.

3.4 Results of the fairness analysis

After evaluating the model's performance, we assessed for disparities in predictive outcomes across the available multivalue attributes: sex and age.

For the analysis, we have considered samples from the "Male" subgroup (253 individuals) and from the "Female" subgroup (177 individuals) regarding the variable sex. Also, we have created the following categories to stratify patients by age span with respect to the variable age: "Adult (27-59 years old)" (138 individuals), "Aged adult (60-79 years old)" (189 individuals), and "Oldest adults (80-98 years old)" (103 individuals). Table 2 shows how samples from the test set were classified regarding the different subgroups they belong to.

Although Aequitas does not use the F1 score, it does support its core components, sensitivity and precision. Therefore, we can analyze disparities in the

Table 2: Classification of the samples in the test set

Multivalue attribute	Subgroup tag	FP	FN	TN	TP
"sex" (Sex)	"Female"	7	5	147	18
	"Male"	8	11	206	28
"age_cat" (Age category)	"Adult (27-59 years old)"	0	0	138	0
	"Aged adult (60-79 years old)"	4	8	162	15
	"Oldest adults (80-98 years old)"	11	8	53	31

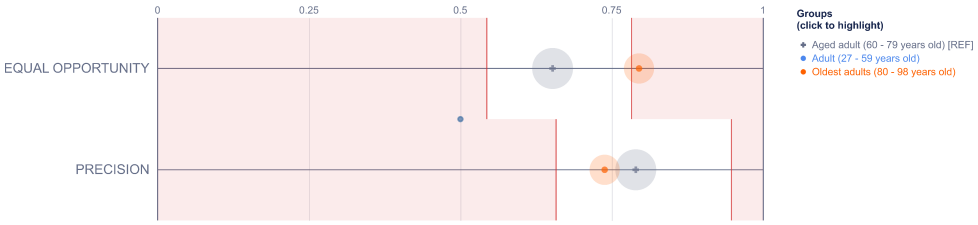

Fig. 2: Equal Opportunity and Precision results, grouped by age category

F1 score by examining the fairness metrics for these two underlying performance measures. In Aequitas, sensitivity is equivalent to the True Positive Rate (TPR), and its corresponding fairness goal is Equal Opportunity. In parallel, Precision measures the accuracy of the model's positive predictions and its fairness metric is Predictive Parity.

Figure 2 shows Equal Opportunity and Precision metrics for age groups on a 0-1 scale. Aged adults score approximately 0.68 (Equal Opportunity) and 0.8 (Precision), while Adults show lower Equal Opportunity (0.6) but higher Precision (0.85). Oldest adults achieve the highest Equal Opportunity (0.75) with moderate Precision (0.78). Figure 3 presents the same analysis by sex. Females show higher Equal Opportunity than Males but lower Precision, resulting in more FP and reduced Precision.

Figure 4 shows the absolute performance of a model using two different metrics: TPR and Precision. The model's performance varies across the different age groups. The Aged Adults group achieved a TPR of 0.65 and a Precision of 0.79, which indicates moderate sensitivity and relatively high accuracy in positive predictions. For the Adult group, although the model correctly identified all 138 non-positive cases, the TPR and Precision metrics are undefined, as there were no positive instances in the test set for this subgroup. The Oldest Adults group yielded the highest TPR of 0.79, with a Precision of 0.74. The model is moderately precise in detecting TP within this older population.

Overall, these findings suggest that the model performs best in the Oldest Adults group and moderately in the Aged Adults group. For the Adult group, the model correctly classified all negative cases, altough the performance on positive cases could not be evaluated, as the test set for the Adult group contained no positive instances.

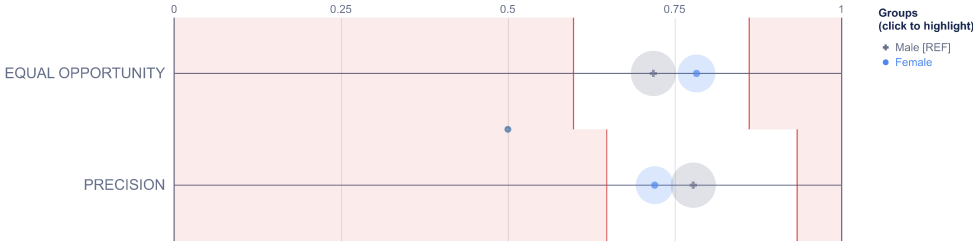

Fig. 3: Equal Opportunity and Precision results, grouped by sex

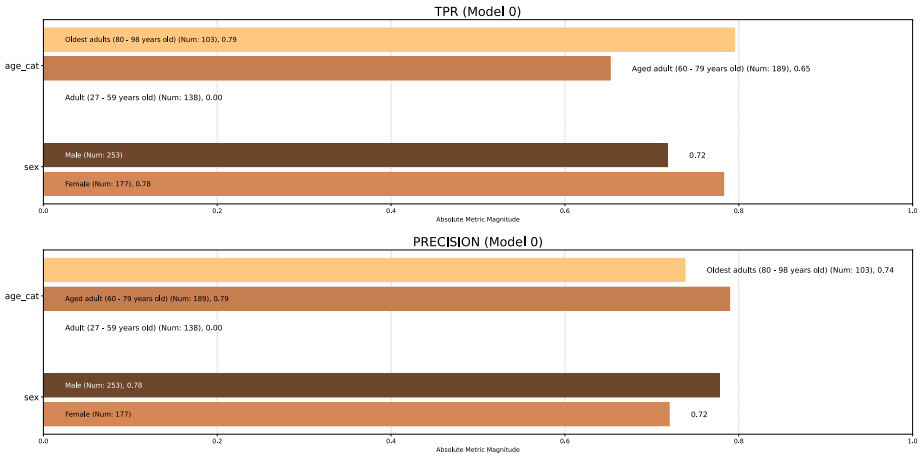

Fig. 4: Model performance (TPR and Precision) across demographic groups

3.5 Results of the explainable AI methods

Following our three-stage audit methodology, after data quality assessment and fairness evaluation with Aequitas, we implemented the third component: XAI techniques (ALE for global explanations, LIME, SHAP, and DiCE for local explanations) to complement bias detection results and understand underlying model decision patterns.

ALE analysis identified four key mortality predictors: *lab_creatinine*, *lab_til*, *sex*, and *mechvent*. Figure 5 shows serum creatinine increases mortality risk above 2.0, total bilirubin exhibits a U-shaped relationship (decreasing until 0.5, then increasing), while female sex and mechanical ventilation associate with lower mortality predictions.

Analyzing the local explanations obtained with LIME, SHAP, and DiCE, we found that the same four attributes are the most important ones when explaining a single instance (chosen randomly) from the testset. Therefore, global and local explanations supported each other. Figure 6 provides a SHAP local explanation, showing that the four most influential features are the same as those identified by ALE (Figure 5) and LIME.

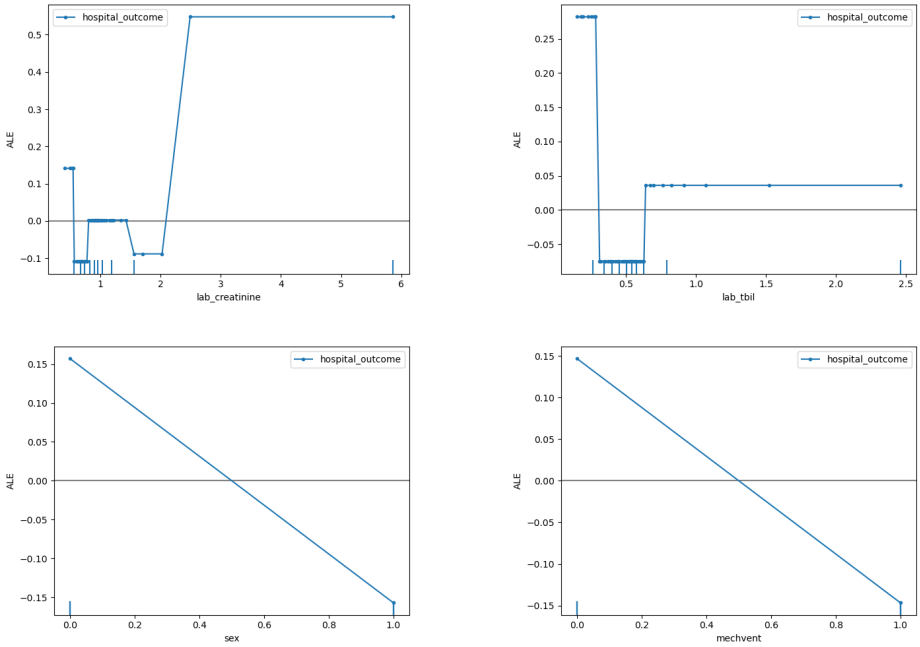

Fig. 5: Global explanation obtained with ALE.

According to the SHAP analysis, sex has little influence, while the other three attributes have a similar impact. Only total bilirubin (*lab_til*) affects hospital outcome positively (as it increases, *hospital_outcome* also increases), while the rest have a negative effect. In the case of the result of DiCE, we obtain a slightly different explanation. DiCE obtains a counterfactual, which means an example of a different instance with a different outcome where there is at least one attribute value that has a wide variation comparing it with the instance to explain. Here, the only attribute that showed a variation was *lab_creatinine*, meaning that for this specific instance, changing only the *lab_creatinine* value is enough to return a different *hospital_outcome*.

This XAI analysis complements the Aequitas results regarding age and sex biases. While Aequitas detected disparities for both attributes, XAI reveals that age has minimal influence on model predictions, suggesting the detected bias stems from data imbalance rather than learned behavior. However, sex significantly affects *hospital_outcome* predictions, with the model predicting higher mortality for male patients. This confirms genuine algorithmic bias that requires mitigation strategies beyond simple data rebalancing.

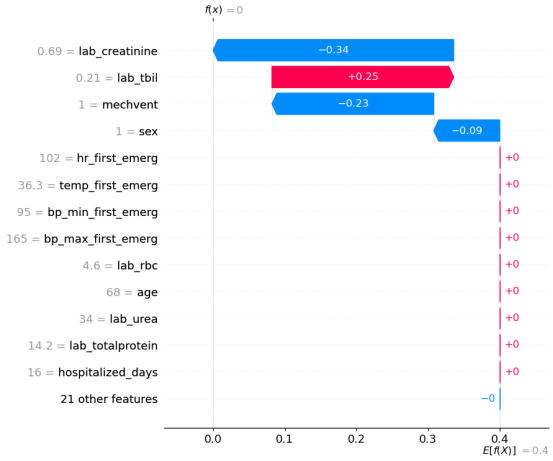

Fig. 6: Local explanation obtained with SHAP.

4 Conclusions and Future Work

This paper presents a comprehensive framework for auditing AI systems in medical domains through functional audit processes that systematically evaluate bias and fairness. Our approach integrates three key components: data quality assessment, application of fairness toolkits such as Aequitas, and implementation of explainable AI techniques. The practical application through a COVID-19 case study demonstrates the framework's effectiveness in identifying significant disparities in model performance across demographic groups. The key contributions of our approach include: (1) Integration of various forms of discrimination, including intersectional bias that may affect vulnerable populations; (2) A systematic methodology that combines quantitative bias assessment with qualitative XAI insights; (3) Practical tools for healthcare practitioners and AI developers to identify and address ethical concerns in medical AI systems.

The technical analysis presented through our functional audit framework provides quantitative evidence of bias patterns, but understanding their full implications requires incorporating sociological interpretation. The disparities identified through Aequitas metrics and XAI explanations are not merely statistical anomalies but reflect complex social realities that shape healthcare outcomes. Our COVID-19 case study exemplifies this complexity: the choice of variables (sex and age) was constrained by data availability, yet missing socio-economic or ethnicity variables could provide crucial insights into health disparities. Traditional ethics models based on general principles are insufficient for healthcare AI; domain-specific ethical frameworks must consider how social determinants of health influence both data collection and model interpretation [21]. The intersection of technical bias detection with social context analysis is essential for developing comprehensive understanding of how AI systems may perpetuate or amplify existing healthcare inequalities.

Looking ahead, we plan to implement a Case-Based Reasoning (CBR) methodology based on our theoretical framework. CBR provides a problem-solving approach that learns from past experiences, well-suited for complex, context-dependent auditing scenarios where universal rules are insufficient. The CBR approach will transform AI auditing from ad-hoc practices into a structured, reusable methodology that captures collaborative insights from domain experts, AI specialists, and social scientists about bias patterns, appropriate fairness metrics, and effective XAI techniques for specific medical contexts.

Furthermore, our framework aligns directly with the principles of the new European Health Data Space (EHDS) regulation. The EHDS creates a legal and infrastructural environment for the trustworthy use of health data, mandating the traceability, transparency, and interoperability that are foundational to auditable AI. The intersection of the EHDS and AI regulation (particularly the AI Act) creates a dual compliance environment where both the ethical design of ML models and the lawful handling of biomedical data must be auditable and demonstrable. Additionally, our framework will incorporate regulatory and ethical compliance assessment, including GDPR data protection requirements, medical device regulations for AI diagnostics, and established medical ethics principles. Success requires ongoing collaboration between AI researchers, medical professionals, ethicists, and regulatory bodies to maintain relevant and effective auditing procedures aligned with evolving medical AI standards.

Acknowledgements Supported by AUDITIA-X project PID2023-150566OB-I00 (MCIN/AEI/10.13039/501100011033), SHOW-X project (Comunidad de Madrid/UCM), and ARTIFACTS project PID2022-141045OB-C4X (MCIN/AEI/10.13039/501100011033/FEDER).

Bibliography

[1] J. Mökander. Auditing of ai: Legal, ethical and technical approaches. *Digital Society*, 2(3):49, 2023.

[2] S. Brown, J. Davidovic, and A. Hasan. The algorithm audit: Scoring the algorithms that score us. *Big Data & Society*, 8(1), 2021.

[3] E. Berghout, R. Fijneman, L. Hendriks, M. de Boer, and B. J. Butijn. *Advanced Digital Auditing: Theory and Practice of Auditing Complex Information Systems and Technologies*. Springer, 2023.

[4] Brent D. Mittelstadt. Auditing for transparency in content personalization systems. *International Journal of Communication*, 10:4991–5002, October 2016.

[5] C. A. Zhang, S. Cho, and M. Vasarhelyi. Explainable artificial intelligence (xai) in auditing. *International Journal of Accounting Information Systems*, 46, 2022.

[6] N. Bussmann, P. Giudici, D. Marinelli, and J. Papenbrock. Explainable ai in fintech risk management. *Frontiers in Artificial Intelligence*, 3:26, 2020.
[7] S.M. Lundberg and S. Lee. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*, 2017.
[8] S. Khanna and S. Srivastava. Ai governance in healthcare: Explainability standards, safety protocols, and human-ai interactions dynamics in contemporary medical ai systems. *Empirical Quests for Management Essences*, 1(1):130–143, 2021.
[9] M. M. Khan and J. Vice. Toward accountable and explainable ai part 1: Theory and examples. *IEEE Access*, 10:99686–99701, 2022.
[10] A. Cortina Orts. *¿Ética o ideología de la IA? : el eclipse de la razón comunicativa en una sociedad tecnologizada*. Paidós, Barcelona, 2024.
[11] Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T. Rodolfa, and Rayid Ghani. Aequitas: A Bias and Fairness Audit Toolkit, April 2019. arXiv:1811.05577.
[12] M.T. Ribeiro, S. Singh, and C. Guestrin. "why should i trust you?" explaining the predictions of any classifier. In *Proc. 22nd ACM SIGKDD Conf.*, pages 1135–1144, 2016.
[13] J. Dai, S. Upadhyay, et al. Fairness via explanation quality: Evaluating disparities in the quality of post hoc explanations. In *Proc. 2022 AAAI/ACM Conf. on AI, Ethics, and Society*, pages 203–214, 2022.
[14] Jianlong Zhou, Amir H Gandomi, Fang Chen, and Andreas Holzinger. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, 10(5):593, 2021.
[15] Wojciech Samek, Alexander Binder, and Grégoire ... Montavon. Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, 28(11):2660–2673, 2016.
[16] Riccardo Guidotti. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, pages 1–55, 2022.
[17] Marta Caro-Martínez, Juan A. Recio-García, Belén Díaz-Agudo, and Jesus M. Darias et al. isee: A case-based reasoning platform for the design of explanation experiences. *Knowledge-Based Systems*, 302:112305, 2024.
[18] Andrew Silva and Mariah et al. Schrum. Explainable artificial intelligence: Evaluating the objective and subjective impacts of xai on human-agent interaction. *Int. Journal of Human–Computer Interaction*, 39(7):1390–1404, 2023.
[19] Álvaro Ritoré, Andreea M Oprescu, Alberto Estirado Bronchalo, and Miguel Ángel Armengol de la Hoz. COVID data for shared learning (CDSL): A comprehensive, multimodal COVID-19 dataset from HM hospitales, 2024.
[20] Joy Tzung-Yu Wu, Miguel Ángel Armengol de la Hoz, and Po-Chih et al Kuo. Developing and validating multi-modal models for mortality prediction in COVID-19 patients: A multi-center retrospective study. *J. Digit. Imaging*, 35(6):1514–1529, December 2022.
[21] E. Ratti, M. Morrison, and I. Jakab. Ethical and social considerations of applying artificial intelligence in healthcare—a two-pronged scoping review. *BMC Med Ethics*, 26(68), 2025.