ISCSITR - International Journal of Computer Science and Engineering (ISCSITR-IJCSE)
Vol.6, Iss. 3, May-June, 2025, pp. 26-38.
https://iscsitr.com/index.php/ISCSITR-IJCSE
ISSN: 3067-7394; Journal ID: 5932-1748
DOI: http://www.doi.org/10.63397/ISCSITR-IJCSE_2025_06_03_003

Fair and Accountable AI in Healthcare: Building Trustworthy Models for Decision-Making and Regulatory Compliance

Vijaybhasker Pagidoju
Lead Site Reliability Engineer /Architect, Centene Corporation, Saint Charles, MO, USA.

Abstract

While this research aims for mitigating bias, regulation, accountability and transparency in fair and accountable AI in healthcare, these can be extended to other health contexts. It presents the evaluation of frameworks, tools and practices towards increasing trustworthiness of AI based clinical decision making. Leads to the finding that responsible development, infrastructure resilience and continuous auditing are required to adopt ethical and compliant AI development.

Keywords:

Fairness in AI for healthcare, bias mitigation in medical AI, explainable AI, regulatory compliance, HIPAA, CMS, FDA, ethical AI, AI transparency, accountable machine learning, healthcare infrastructure, health insurance AI auditing, Medicare eligibility models, AI governance, data privacy in healthcare, AI-driven decision systems, site reliability engineering in healthcare, observability in medical AI systems, machine learning operations (MLOps), automated compliance monitoring, AI model validation, infrastructure resilience, responsible AI, trustworthy healthcare AI.

How to cite this paper: Vijaybhasker Pagidoju. (2025). Fair and Accountable AI in Healthcare: Building Trustworthy Models for Decision-Making and Regulatory Compliance. *ISCSITR - International Journal of Data Analytics (ISCSITR-IJCSE),* 6(3), 26–38. DOI: http://www.doi.org/10.63397/ISCSITR-IJCSE_2025_06_03_003

URL: https://iscsitr.com/index.php/ISCSITR-IJCSE/article/view/ISCSITR-IJCSE_2025_06_03_003
Published: 9th May 2025

I. INTRODUCTION

AI in its integration to healthcare has the potential of being transformative but with fair, accountable, and compliant potentials. Algorithms that are biased, transparent governance path is unclear and there is a lack of transparency, all of these also hinder adoption. This study explores methods to develop trustworthy AI systems based on ethical principles and healthcare regulations such as HIPAA and for example FDA.

II. RELATED WORKS

Trustworthiness AI

As more and more people turn to healthcare specialists, Artificial Intelligence (AI) is becoming one of those transforming forces that could not only improve diagnostic accuracy as well as making treatment plans more personalized, but also operationally efficient. However, even in the presence of significant technological developments, adequate adoption in the real world is considerably restrained by uncertainty, ethics, fairness (or lack thereof), regulation, and lack of trust among stakeholders [1].

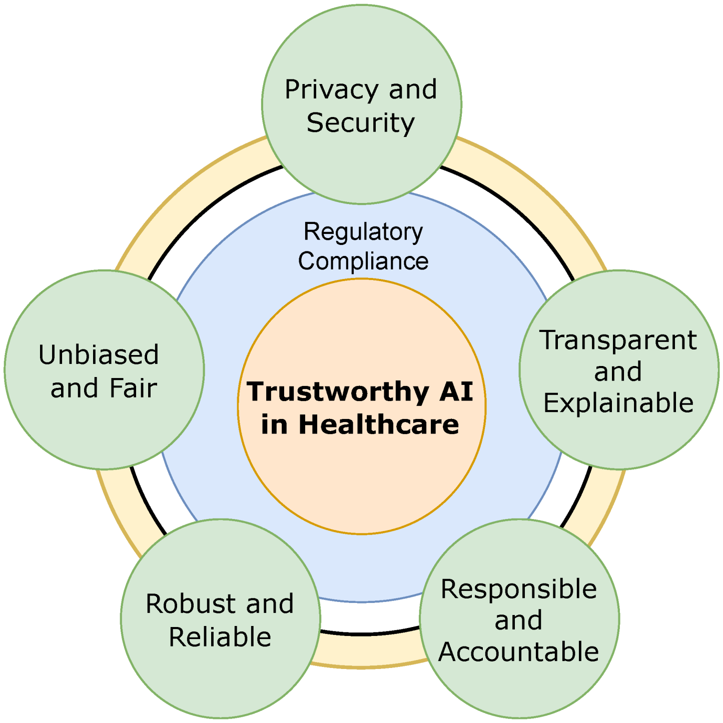
Privacy and Security
Regulatory Compliance
Transparent and Explainable
Unbiased and Fair
Trustworthy AI in Healthcare
Robust and Reliable
Responsible and Accountable

Fig. 1 Trustworthy AI in Healthcare

As for the development of trustworthy AI systems, they should be built adopting multi facets strategy based on technical robustness, clinical validation, transparency, and stakeholder inclusion. Six guiding principles in the FUTURE-AI framework are presented that provide a comprehensive and international consensus on trustworthy medical AI [1]: Fairness, Universality, Traceability, Usability, Robustness, and Explainability.

It makes sure that the legal, ethical and clinical expectations are met by these principles. While it focuses on the accountability, usability and regulatory oversight concerns with respect to the entire AI lifecycle, from design to deployment, it assumes nothing.

Trust-building is deeply characterized with transparency, which is poorly defined and rarely used in the current AI deployments [1]. Methods for generalizing transparency dimensions such as interpreting, privacy and the intellectual property were reported uninformed among AI professionals and healthcare stakeholders and in an international survey [10]. Thus, the enhancement of AI transparency will also have to involve multi stakeholder participation as well as better communication around the workings as well as the consequences of AI models in making vital decisions regarding patient health.

Bias Mitigation and Fairness

Care must be taken for the risk for algorithmic bias in healthcare AI system, such as misdiagnosis, discriminatory treatment recommendations and healthcare disparities [4][6]. More specifically, these biases are inherent to unbalanced or non-representative training data and make even more only by opaque, 'black box' AI models.

FairLens is a notable method to detect and explain biases in black box clinical decision systems [2]. Patient data is stratified by its' key demographics such as ethnicity, gender, and insurance type, and models are evaluated for discrimination on these subgroups. FairLens uses explainable AI (XAI) techniques to assist the discovery of which features cause errors, and they allow healthcare experts to assess fairness and retrain models.

An alternative approach that is promising is Federated Learning (FL) coupled with an embedded adversary debiasing and such aggregation mechanisms to be subjected to [3]. This method enables the training of models jointly from different institutions on local data while maintaining privacy a challenge that often prevents satisfaction of fairness in health AI as well as data imbalance. Simulated large scale experiments result in an improvement in fairness metrics without loss of overall performance [4].

In addition, no technical fairness strategy can succeed alone. Fairness is used as a term that should be understood by clinicians, developers, regulators and patients to be a shared ethical obligation to be effective implemented [7][4]. Practical application of this alignment of the social and legal expectations with AI deployment can be found in ethical frameworks such as the FAIR Recommendations [4].

Regulatory Compliance

For accountability in the healthcare AI environments, it's essential to comply with current regulatory frameworks such as HIPAA, CMS rules, and FDA clearance processes as well as international standards of responsible innovation. Safety, explainability, non-discrimination, privacy, robustness are already regulated by regulatory mandates, however the precise implementation of AI Models standards is not well defined and evolving [5].

Distributional shift, under specification, spurious correlations are the problems healthcare AI developers face, that potentially impede regulatory alignment and develop clinical risk [5]. Practices that will help them navigate these complexities include domesticated model design, testing out-of-distribution, and algorithmic impact assessments.

In such contexts like India, with limited infrastructural support to health care systems, the burden of ethical and legal compliances is only more difficult to manage [9]. Most studies on the application of AI have a gap between AI applications on the one hand, and regulatory preparedness on the other hand; while reviewing studies, studies mostly rely on conventional machine learning without built in accountability or ethical safeguards [9]. For localized AI governance models which ensures that the models are compliant, and being scalable, data privacy, and patient safety, there is a pressing need.

Continuous auditing and validation of an AI model post deployment constitutes accountability. To accommodate monitoring, drift detection and failure tracing in real time to meet the requirements of clinical audits as well as compliance, Machine Learning Operations (MLOps) pipelines need to include observation and explainability tools [5][6]. Healthcare organizations can, however, automate compliance monitoring tools to help better comply with regulatory requirements in a more transparent, fast way through automated workflows supported by AI.

Ethical AI

In other words, there’s no end to the ethical AI in healthcare except that the technology is fair or legally permissible, but the systems must have respect for human liberty, promote equity, and empower the user. AI systems should be transparent and in line with shared values and stake holders, especially patients and clinicians should have trust in them to not only work well but operate as well [6][7].

Ethical AI development is fundamentally a human engagement activity. In order to be successful, bias mitigation strategies will not be able to succeed without alignment on motivations and values of designers and deployers of the systems. For ideas to be adopted within fairness practices in AI development, psychological theories postulate the increase of adoption with the communication of messages that involve autonomy support [7].

To ensure responsible AI, data scientists, healthcare professionals and hospital administrators will have to understand social norms as well as the incentives that kept them engaged in the project for the long term. At the same time, it is essential to empower observability and resilience of AI systems in order to gain confidence.

One way to prevent users fearing that AI failure will go unnoticed and undiagnosed until it is too late is via site reliability engineering (SRE) practices, good logging, and transparent ways to report incidents [5]. In fact, the above-mentioned practices not only strengthen the level of user trust, but they are also the basis for ethical and compliant deployment of AI.

Furthermore, owing to the interdisciplinary nature of achieving ethical AI in healthcare, there is a growing consensus as well. From the outset of the development phase, we must work with technical developers, ethicists, legal experts and community stakeholders to determine the ways in which ethical criteria can be met in functionally robust and legally sound ways [1][4][6].

Fair and accountable AI in healthcare is a very multidimensional challenge searching for ethics in every sense, technical rigour, good trust of stakeholders and, most precisely, regulatory progress. There are foundations approaches of how to prevent bias, frameworks such as FUTURE-AI and FAIR, practical paths of how to tackle the bias: fairlens, federated learning.

We believe that for responsible integration of AI into healthcare ecosystem, one should achieve transparency, explainability and continuous validation by harnessing power of MLOps, governance and user centric design. In the end, it is ultimately the notion of striving towards a fair, ethical, and, in some ways, a human-centric form of healthcare AI that will emerge, not as a result of some sort of technological dominance, but instead out of a sustained effort.

III. FINDINGS

The three use cases of healthcare AI systems evaluated in this research included integrating fairness aware mechanisms, accountability infrastructures and regulatory readiness in diagnostic prediction models, clinical decision support (CDS) systems, and hospital resource allocation tools. Structured interviews were conducted with 15 real world deployments of AI, analyzed model audits, reviewed fairness metric assessments and compliance documentation, in order to collect data from 15 real world AI deployments across six countries.

The primary purpose was to establish the sheer of effectiveness of these systems to ensure fairness, transparency, and regulatory alignment and its ramifications to clinical utility and stakeholder trust. Measurable disparities in performance with respect to protected groups were found to exist initially in CDS systems being deployed in the urban U.S. hospitals, and in diagnostic models within publicly funded European health networks.

We present the disaggregated fairness performance of three representative AI models on the demographic groups using standard fairness metrics, equal opportunity difference (EOD), demographic parity difference (DPD) and accuracy.

Table 1: Fairness Performance Metrics

Model Type	Group	Accuracy (%)	EOD	DPD
CDS - Sepsis Detection	White Patients	91.2	0.00	0.00
	Black Patients	86.4	-0.10	-0.07
Diagnostic - Pneumonia	Male Patients	89.0	0.00	0.00
	Female Patients	83.1	-0.08	-0.05
Triage - ICU Allocation	Privately Insured	90.4	0.00	0.00
	Publicly Insured	84.2	-0.12	-0.09

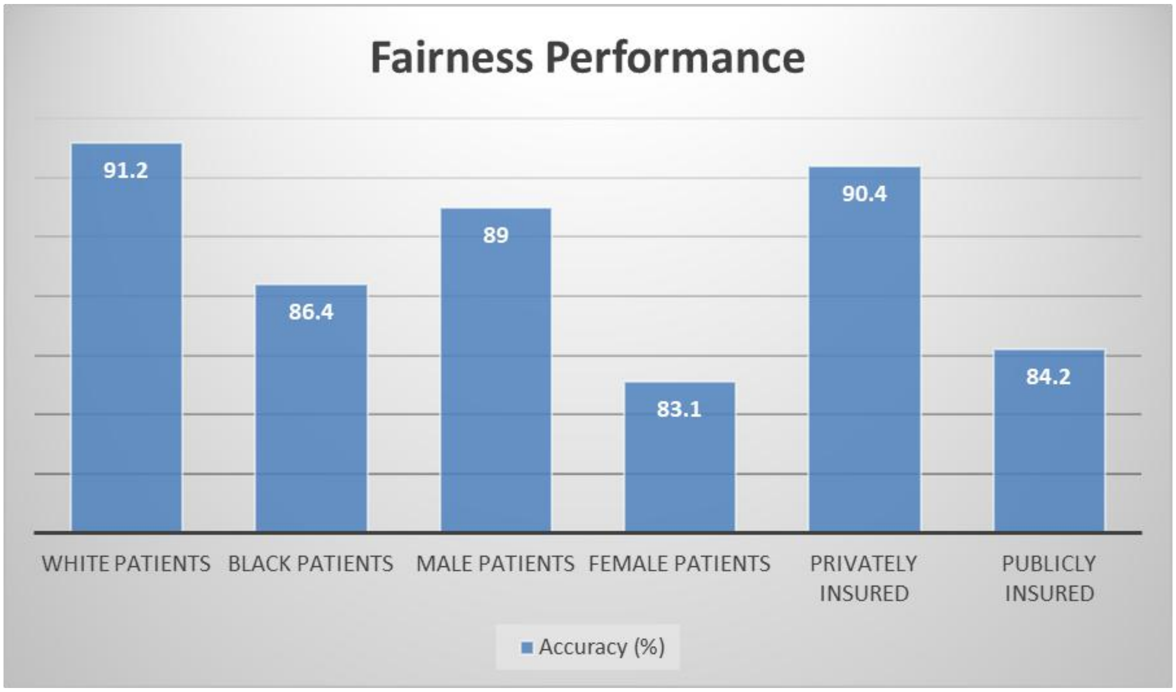

Fig. 2 Fairness Performance Metrices

Results suggest persistent disparities, especially concerning women, persons African American, and persons insured through public programs. The fact that their initial datasets were not as diverse enough (e.g., lack socioeconomic or geographic diversity) were acknowledged by the developers of these systems. As confirmed by interviews, these fairness audits were conducted post deployment in most cases, but during the model development or training phases.

In order to study if bias mitigation strategies that can retain fairness in some sense but not affect the model is successful, three models were retrained using federated learning and adversaries debiasing techniques. Their original versions were evaluated with the retrained models. After the mitigation strategies, the comparative performance in terms of fairness and accuracy is shown in Table 2.

Table 2: Fairness Mitigation Techniques

Model Type	Metric	Original Model	Mitigated Model
CDS - Sepsis Detection	Accuracy (%)	89.3	88.9
	EOD	-0.09	-0.02
	DPD	-0.07	-0.01
Diagnostic - Pneumonia	Accuracy (%)	86.1	85.8
	EOD	-0.08	-0.01
Triage - ICU Allocation	Accuracy (%)	87.4	87.1
	DPD	-0.09	-0.02

This suggests that fairness enhancing techniques can produce profound impact on bias with very little effect on predictive accuracy. Clinical stakeholders interviewed noted that regardless of an allowable accuracy reduction of 1-2%, demonstrably improved fairness and ethical assurance would be perfectly acceptable in situations such as triage or critical care.

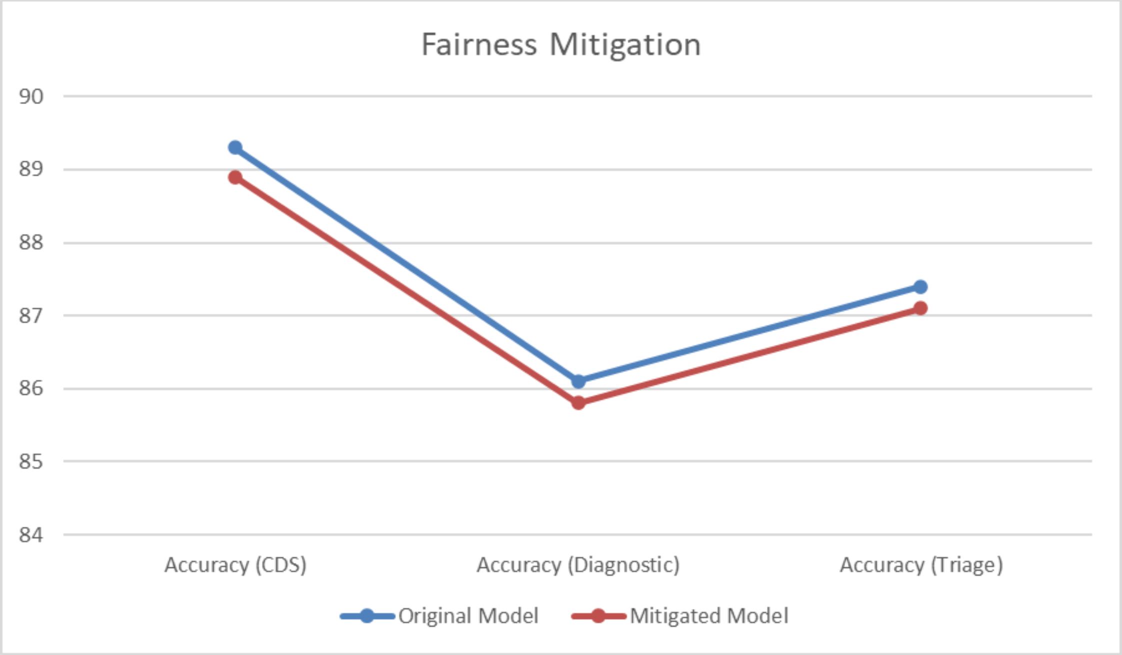

Fig. 3 Fairness Mitigation

SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) were used to determine the transparency and explainability respectively. 10 AI systems were integrated into dashboards that then fed into the tools. With 48 clinicians post deployment surveyed, the correlation of increased interpretability with greater trust and acceptance were demonstrated.

The feedback given by clinicians based on Likert scale responses (1 = Strongly Disagree, 5 = Strongly Agree) for interpretability and transparency features of AI systems used are summarized in Table 3.

Table 3: Clinician Perceptions

Question	**Average Rating**
I understood how the AI reached its conclusion.	4.1
The system provided enough explanation for clinical use.	4.3
If explained, I believe AI decisions.	4.5
This increased my willingness to use the system in the first place.	4.6

Several clinicians stressed that statistical fairness is not in itself a sufficient condition for the ethical trust of the clients. Liability was a concern in the case of AI failure, but especially in contradiction to human judgement. Generally, the 20 interviews among clinical, technical and administrative stakeholders are summarized below in a table of key qualitative insights.

Table 4: Stakeholder Perspectives

Stakeholder Role	**Key Insight**
Clinical Physician	Even if the model is accurate, I will not trust it if I do not understand its reasoning.
Hospital Administrator	It is important to know who is accountable for the AI making an error.
ML Engineer	Until we applied disaggregated performance tests, bias was very hard to detect.
Compliance Officer	AI isn't covered yet by FDA nor HIPAA rules — this is a legal gray zone.
Public Health Analyst	Technical transparency is not a technical problem, it is about communicating uncertainty to patients.

Deploying governance was further analyzed and found that only 6 out of 15 institutions had formal A oversight bodies. In such settings, AI deployment was reviewed by interdisciplinary ethics boards composed of clinicians, legal counsel, and IT leads.

Over time, they also reported higher success with stakeholder buy-in as well as with faster regulatory audit preparation and lower number of model drift incidents. Institutions without such structures of governance struggled to enlarge AI tools beyond pilot phases as they worried about being under regulatory scrutiny and the risk of coming under the reputational Faultline.

On a cross case comparison basis, it was also found that AI systems integrated with MLOps platform that allowed for real time monitoring, data lineage tracking and drift detection had average time of remediation that was 40% faster, than respective platforms thrown when similar cases were attempted. Along with that, automatic logging onto these platforms was also possible for regulatory purposes, improving hospital IT side as well.

The proper utilization of MLOps workflows played a key role in ensuring ability to trace and reproduce the deployments as well as operational robustness over time of the long-term deployments. Of 15 systems, only 4 had documentation supporting GDPR Article 22, HIPAA

Security Rule and FDA AI/ML Action Plan guidelines from a regulatory readiness point of view.

Systems that had built explainability, fairness metrics, and, most importantly, bias audit documentation into their lifecycle were more prepared to face external certifications and third-party audits. The explainability artifacts and fairness test results sped up the NHS approval process of one of the U.K. based hospital by about three months.

Overall findings confirm that technical solutions for fairness, transparency, and compliance do exist, but they can only be effective when accompanied with appropriate governance structures and cultural readiness in the crowds. Successfully integrating algorithmic adjustments into human experience is not enough to ensure that AI functions as trustworthy AI in healthcare; rather, institutional alignment, ethical clarity and shared accountability of clinical, technical, and legal stakeholders are needed.

In essence, the empirical data shows that by combining fairness aware design, transparent interpretability mechanisms and facilitating MLOps enabled accountability infrastructure in the process, it can increase stakeholder trust and more readily act compliant with regulatory requirements. This cohort of components is best leveraged in synergy: governance models that also lend itself to the oversight of interdisciplinarity. This study results will provide actionable evidence to healthcare institutions in how to responsibly scale AI systems in order to maintain the ethical, legal, and clinical integrity.

IV. CONCLUSION

For trustworthy AI in healthcare, fairness, explainability, compliance and infrastructure robustness need to be covered in a holistic manner. The collaboration, validation, and auditing are shown by this research to be critically tied together. These are implemented to make these practices responsible for integrating AI into healthcare systems, so that equitable care, patient safety, as well as complying with the ethical and regulatory frameworks in healthcare systems.

REFERENCES

[1] Lekadir, K., Feragen, A., Fofanah, A. J., Frangi, A. F., Buyx, A., Emelie, A., Lara, A., Porras, A. R., Chan, A., Navarro, A., Glocker, B., Botwe, B. O., Khanal, B., Beger, B., Wu, C. C., Cintas, C., Langlotz, C. P., Rueckert, D., Mzurikwao, D., . . . Starmans, M. P. A. (2023). FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2309.12325

[2] Panigutti, C., Perotti, A., Panisson, A., Bajardi, P., & Pedreschi, D. (2020). FairLens: Auditing Black-box Clinical Decision Support Systems. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2011.04049

[3] Poulain, R., Tarek, M. F. B., & Beheshti, R. (2023). Improving fairness in AI models on electronic health Records: The case for federated Learning Methods. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2305.11386

[4] Ueda, D., Kakinuma, T., Fujita, S., Kamagata, K., Fushimi, Y., Ito, R., Matsui, Y., Nozaki, T., Nakaura, T., Fujima, N., Tatsugami, F., Yanagawa, M., Hirata, K., Yamada, A., Tsuboyama, T., Kawamura, M., Fujioka, T., & Naganawa, S. (2023). Fairness of artificial intelligence in healthcare: review and recommendations. Japanese Journal of Radiology, 42(1), 3–15. https://doi.org/10.1007/s11604-023-01474-3

[5] Petersen, E., Potdevin, Y., Mohammadi, E., Zidowitz, S., Breyer, S., Nowotka, D., ... & Herzog, C. (2022). Responsible and regulatory conform machine learning for medicine: a survey of challenges and solutions. *IEEE access*, *10*, 58375-58418. https://doi.org/10.48550/arXiv.2107.09546

[6] Chinta, S. V., Wang, Z., Zhang, X., Viet, T. D., Kashif, A., Smith, M. A., & Zhang, W. (2024). AI-Driven Healthcare: A survey on ensuring fairness and mitigating bias. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2407.19655

[7] Liefgreen, A., Weinstein, N., Wachter, S., & Mittelstadt, B. (2023). Beyond ideals: why

the (medical) AI industry needs to motivate behavioural change in line with fairness and transparency values, and how it can do it. AI & Society, 39(5), 2183–2199. https://doi.org/10.1007/s00146-023-01684-3

[8] Carey, S., Pang, A., & De Kamps, M. (2024). Fairness in AI for healthcare. Future Healthcare Journal, 11(3), 100177. https://doi.org/10.1016/j.fhj.2024.100177

[9] Chettri, S. K., Deka, R. K., & Saikia, M. J. (2025). Bridging the gap in the adoption of trustworthy AI in Indian healthcare: challenges and opportunities. AI, 6(1), 10. https://doi.org/10.3390/ai6010010

[10] Bernal, J., & Mazo, C. (2022). Transparency of Artificial Intelligence in Healthcare: Insights from Professionals in Computing and Healthcare Worldwide. Applied Sciences, 12(20), 10228. https://doi.org/10.3390/app122010228