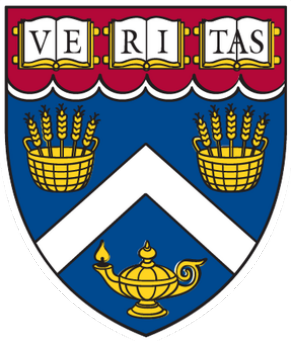

HARVARD EXTENSION SCHOOL

CSCI E-184: Data Science and Artificial Intelligence: Ethics, Governance, and Laws

Algorithmic Decision-Making In Healthcare

Exploring the ethical considerations of using algorithms for healthcare decisions, such as patient diagnosis, treatment recommendations, or resource allocation

FINAL PROJECT

REPORT

Group Name: The Midnight Crew

Jamal O'Garro | Seymur Hasanov | Seraphim Eilken | Daniella Olivas | Xueying Prawat | Ricardo Villarreal Canales

May 13, 2025

Table of Contents

Abstract

Artificial intelligence (AI) is increasingly used in healthcare to support decision-making in diagnosis, treatment planning, and resource allocation. While these tools offer opportunities to improve patient outcomes and system efficiency, they also introduce significant ethical challenges. This report explores the implications of algorithmic decision-making in clinical settings, focusing on concerns such as bias, lack of transparency, data privacy, and accountability.

Through case studies—such as IBM Watson for Oncology and AI-based tools for skin cancer detection—we examine how poorly validated models, limited training data, and opaque algorithms can lead to inaccurate recommendations and unequal care. These examples highlight the need for careful oversight and inclusive design.

We apply ethical theories, including utilitarianism and deontology, to frame the moral responsibilities involved in deploying AI in medicine. These perspectives help balance the goals of improving overall outcomes while safeguarding individual rights.

To address the identified challenges, we propose key recommendations: improving data diversity and quality, ensuring model explainability, mandating independent audits, and integrating human oversight through agent-based system design. Our findings stress that ethical implementation of AI in healthcare is essential to building trustworthy and effective systems that serve all patients fairly and responsibly.

1. Introduction

With the rapid advancement of medical knowledge and the emergence of new diseases, healthcare professionals face increasing pressure to stay up to date. Research suggests that clinical physicians need to spend approximately 4 to 5 hours per week just to keep pace with the evolving medical literature to provide effective treatment plans [1]. To address this challenge, AI has been introduced into healthcare to support clinical decision-making. AI algorithms can assist with tasks such as data interpretation, diagnosis, reasoning, and treatment recommendations, helping to extend the capabilities of human practitioners and reduce cognitive burden. Despite these advancements, the deployment of AI in healthcare raises critical ethical questions. Concerns about fairness, bias in training datasets, patient safety, and transparency [2]. Without rigorous oversight and clearly defined ethical standards, AI-based tools could negatively impact healthcare.

This report aims to explore the role of AI algorithmic decision-making in healthcare, focusing on the ethical implications of using AI in clinical environments. This report will provide an overview of AI applications in healthcare, highlight ethical concerns, and analyze case studies of AI-based tools for oncology purposes.

2. Background and Context

In today's healthcare landscape, AI and algorithmic decision-making are playing an increasingly influential role. These technologies now assist clinicians with a range of tasks, including diagnosing illnesses, offering treatment recommendations, prioritizing patient care, and managing medical resources. By processing vast quantities of medical data, these systems can help improve diagnostic accuracy, enhance operational efficiency, and tailor care to individual patients. Tools such as clinical decision support systems, imaging analysis software, and predictive models are gradually becoming essential components of modern medical practice. They allow healthcare professionals to better manage the growing complexity of both patient needs and treatment options.

Yet, while the adoption of these technologies brings undeniable benefits, it also introduces serious challenges. Algorithms trained on data that is biased or incomplete can unintentionally reinforce inequalities already present in healthcare systems. In addition, many AI models operate in ways that are difficult for both clinicians and patients to interpret, creating a lack of transparency that can erode trust. The collection and use of sensitive patient data also raise significant privacy concerns. Another unresolved issue is accountability. When an AI system contributes to a medical error, it is often unclear who should be held responsible. Even algorithms that perform reliably in controlled settings may produce inconsistent or unreliable results when applied to diverse patient populations or in real-world clinical environments.

This report examines both the potential and the pitfalls of healthcare AI. Through selected case studies, it highlights the ethical risks that can arise when using these tools and proposes practical recommendations for ensuring their responsible and equitable use.

3. Case Studies

The following case studies are examples of how AI technology has been involved in healthcare.

IBM Watson for Oncology (Failure in Cancer Recommendations)

IBM Watson for Oncology (WFO) was developed as a cutting-edge AI tool to support cancer treatment decisions. This new technology combines vast medical knowledge with patient records to recommend optimal cancer therapies. Despite the hype and partnerships with prestigious hospitals (like Memorial Sloan Kettering Cancer Center), the system ultimately failed to deliver reliable clinical value.

The key points to analyze when reviewing this case study are that Watson was heavily marketed as a revolutionary AI in cancer care, but it failed in practice. This new AI technology gave unsafe or incorrect treatment recommendations. The main issue was that the training data was based on a limited set of handpicked cases and clinical input, but not on broad, real-world patient data. Some ethical considerations in this case study are: Overpromising AI capabilities, lack of transparency in model training (accountability gap). [3]

The WFO model was tested to compare the consistency between WFO and clinicians in different countries and regions. The clinicians were a team of oncologists, surgeons, pathologists, and radiologists who would help determine the treatment scheme. The final report would be compared with the WFO model. However, the result showed that the clinical decision was not consistent with the WFO model. In some scenarios, Watson recommended non-standard treatments. The results showed that the consistency between WFO and MDT for cancer patients is not completely consistent, especially in patients with advanced cancer. This is also noticeable when the model was applied in other countries. [1]

Despite the system's innovative promise, WFO struggled to deliver clinically useful outcomes in real-world settings. Its failure was due to a mix of limited training data, unrealistic marketing claims, and lack of transparency in decision-making. Feedback from clinical users highlighted that the system was often not aligned with actual practice standards, especially for complex or late-stage cancer users. [4] This case shows how algorithmic tools, if not rigorously validated and responsibly implemented, can pose risks to patient care and professional trust.

Figure 1 shows the main issues that led to the failure of IBM WFO, along with their estimated impact level (on a 1–10 scale). The impact level reflects how much each issue contributed to the system's failure, based on public reports, case studies, and peer-reviewed articles.

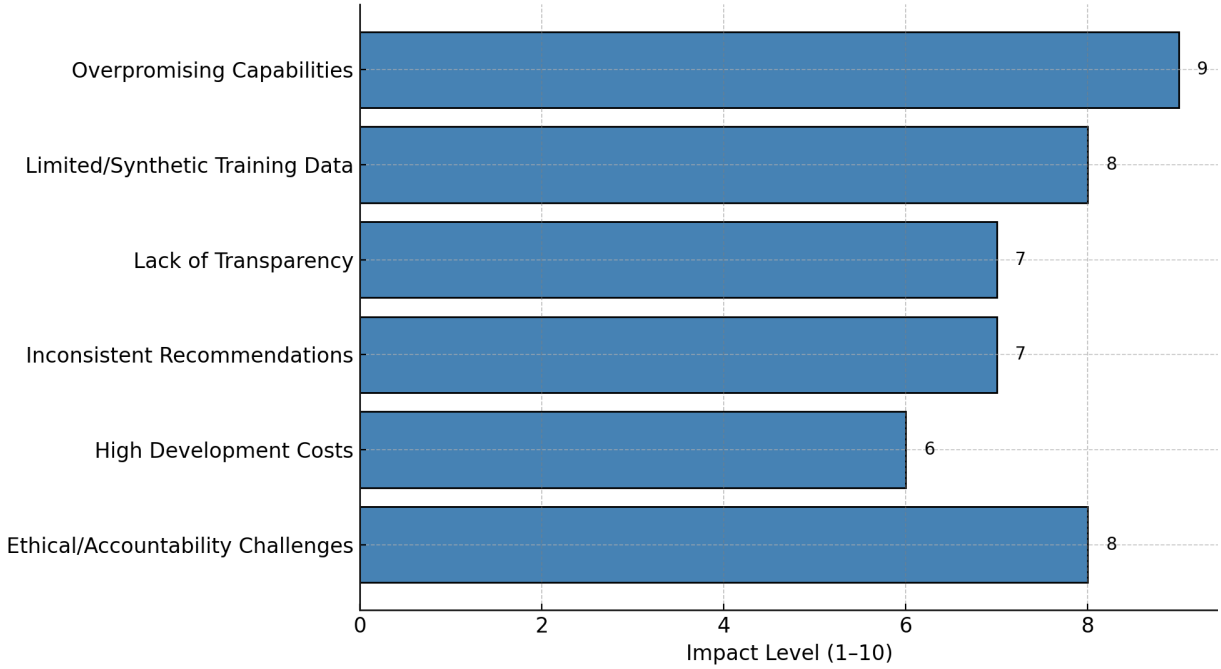

Figure 1. Key factors leading to IBM Watson for oncology failure

Skin Cancer and AI

Diagnostic Performance of Augmented Intelligence with 2D and 3D photography in Melanoma Detection

The main goal of this study was to evaluate the performance of 2D and 3D convolutional neural networks (CNNs) alongside dermatologists in detecting melanoma among high-risk patients. In this study, the AI scores of 3D and 2D divides were measured twice in a subgroup to provide robust evidence for the management and interpretation of AI scores in clinical practices. In this study, the sensitivity, specificity, positive predictive value, and negative predictive value were calculated. When looking at the result and the comparison between melanoma, the 2D CNN model agreed in 58.33% of cases, and the 3D CNN model in 66.67% of cases. The overall findings were that 3D CNN had a sensitivity of 90% and ROC-AUC of 0.92 (strong diagnostic power, but specificity was lower, at about 64.6%. Dermatologists performed similarly in sensitivity but had higher specificity (92.3%). In conclusion, in this study, it was found that 2D CNN underperformed while 3D AI was very promising for melanoma detection. However, it is still better to have an expert oversight [5].

Figure 2 compares the diagnostic performance of 2D CNN, 3D CNN, dermatologists, and dermatologists assisted by AI. The graph shows that while 3D CNN and expert dermatologists achieved equally high sensitivity (90%) in detecting melanoma, dermatologists significantly outperformed AI models in specificity.

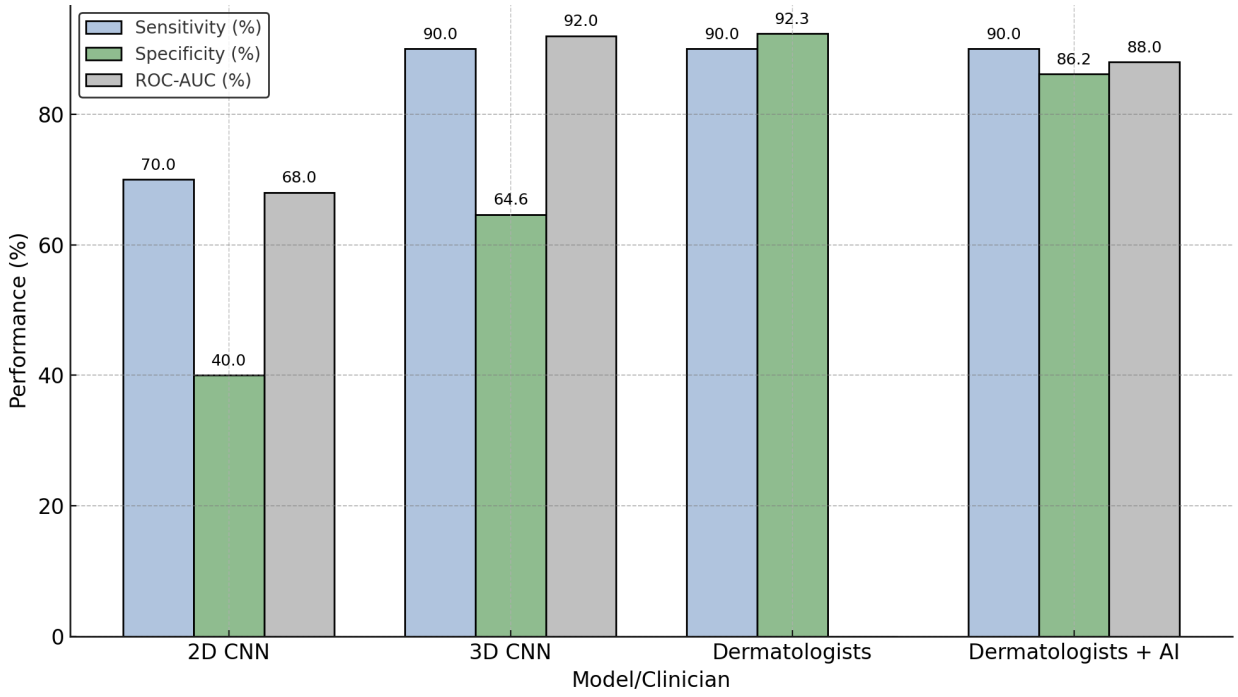

Figure 2. Diagnostic performance in melanoma detection [5] [prepared based on Table 2 of the referenced citation]

The 2D CNN lagged in all key performance measures. The ROC-AUC values further confirm the superior diagnostic accuracy of 3D CNN over 2D CNN, but they also emphasize that human oversight remains critical, particularly to reduce false positives and avoid unnecessary interventions.

Artificial Intelligence in Skin Cancer Diagnosis - Reality Check

This study shows a broad review of AI tools used in skin cancer diagnosis. The paper examined AI tools used by consumers, general practitioners, and dermatologists, and it emphasizes that many AI systems lack clinical validation and struggle in real-world settings, especially with diverse skin types and uncontrolled imaging conditions. Table 1 summarizes the challenges and opportunities of implementing AI in skin cancer diagnosis [6].

Table 1. Challenges and opportunities of implementing AI in Skin Cancer diagnosis

Feature	Challenges	Opportunities
Patient Selection	Bias toward tech-savvy or well-connected patients	Faster, more accessible screening for underserved populations
Image Acquisition	Non-standard imaging protocols affect accuracy	AI + tech advances can improve image quality and consistency
Diversity	Limited skin types in training data can lead to bias	Diverse datasets improve model fairness and accuracy
Privacy	Risk of misuse or breach of sensitive medical data	AI can support secure, anonymized data handling
Transparency	Hard to interpret AI decisions, limiting trust	Explainable AI boosts clinician and patient confidence
Routine Integration	Workflow and system incompatibilities slow adoption	AI can automate tasks, freeing clinicians for patient-focused care
Fairness	Risk of discrimination based on social or demographic factors	Fairness-aware models support health equity
Accuracy	Real-world performance needs more validation	AI can outperform experts in controlled settings
Regulations	Evolving rules create uncertainty and slowdowns	Strong frameworks build trust and protect patient safety

Together, these studies highlight that while AI holds considerable promise in skin cancer diagnosis, significant ethical and practical challenges remain, which are shown on Table 2. Across tools used by dermatologists, general practitioners, and consumers, the recurring issues of dataset bias, lack of clinical validation, limited interpretability, and inconsistent real-world performance are clear barriers to safe and equitable implementation. The evidence underscores that AI should not be deployed in isolation but rather as an assistive tool, with strong human oversight, regulatory standards, and ongoing validation. Future development must prioritize fairness, transparency, and diverse data representation to ensure AI serves all patient populations effectively.

Table 2: Ethical Issues Across Case Studies in Healthcare AI

Case Study	Ethical Concern	Root Cause	Real-World Impact
IBM Watson for Oncology	Unsafe or inaccurate recommendations	Narrow training data, lack of clinical validation, opacity	Patient risk, overhyped expectations, damaged institutional trust
2D CNN for Skin Cancer Detection	Poor diagnostic performance, underfit	Limited accuracy, less robust modeling, lack of validation	Missed diagnoses, particularly for high-risk patients
3D CNN for Skin Cancer Detection	Low specificity, limited transparency	Technical complexity, hard-to-interpret AI scores	Difficult to use without expert oversight
Consumer AI Tools for Skin Cancer	Bias against darker skin tones	Lack of diverse image datasets	Discriminatory performance, unequal care access

4. Ethical Considerations Identified

AI in healthcare has brought multiple ethical considerations, including bias, fairness, transparency, privacy, and accountability. One of the most urgent issues is the presence of bias in training data sets. Sometimes, the datasets used to train AI systems do not reflect the full range of diversity in real-world patients, like the example from IBM's Watson for Oncology [3]. As a result, these models may exhibit inaccurate diagnoses for populations that are underrepresented. When an algorithm performs better for some patients than others, it becomes hard to argue that it is offering fair or consistent care.

Transparency is another big concern. It is important to understand how the results are drawn, what variables are taken into consideration, and how they correlate. Having a transparent AI system will build trust between patients, doctors, and AI systems. In addition, privacy and security of patient data should be taken into consideration when discussing AI systems. If patient data is not handled carefully, personal information may be vulnerable to breaches or misuse. Patients have the right to know how their data is collected, used, and protected. A lack of transparency can result in a loss of confidence in AI systems in healthcare institutions [7].

Accountability is also a central concern. When an AI system makes an error that contributes to patient harm, it is often unclear who bears the responsibility: the software developer, the healthcare provider, the institution, or the data scientist who trained the model. It is important to establish clear legal and ethical frameworks to address possible issues[2].

While AI promises efficiency and cost savings, those benefits should be focused primarily on patient care. Ethical deployment of AI in healthcare requires attention not just to individual interactions but also to systemic consequences, such as the commodification of healthcare, increased surveillance, and the marginalization of populations that fall outside the algorithm's design parameters.

5. Application of Ethical Theories and Frameworks

To better understand the ethical issues that arise when using AI in healthcare, it is useful to apply two key ethical theories: utilitarianism and deontology. These frameworks help highlight different ways of thinking about fairness, responsibility, and what counts as a "good" outcome.

Utilitarianism focuses on maximizing overall benefit, essentially doing the most good for the most people [19]. From this perspective, using AI in healthcare is promising. Algorithms that speed up diagnoses, reduce errors, or help allocate resources more efficiently can potentially save lives and improve care for many. For example, 3D convolutional neural networks used in skin cancer detection show strong diagnostic power, which could lead to earlier interventions and better outcomes on a broad scale [5].

But utilitarianism also has blind spots. If an AI system is trained on data that underrepresents certain groups (e.g., patients with darker skin tones), it might not work as well for those individuals. Even if the tool benefits most people, it could still harm others by providing less accurate care. This tradeoff is especially concerning in healthcare, where equity matters just as much as efficiency. As discussed in the article Utilitarian Principlism as a Framework for Crisis Healthcare Ethics, even in crisis scenarios where resource optimization is critical, it is still necessary to consider fairness and prevent harm to marginalized populations [19].

Deontology, on the other hand, emphasizes duties, rights, and moral rules, regardless of the outcomes [19]. In the context of AI in medicine, this means focusing on things like respecting patient autonomy, ensuring informed consent, and protecting privacy. A deontologist would argue that even a highly accurate AI tool is unethical if patients don’t understand how it works or if it is used without clear accountability. For instance, the failure of IBM Watson for Oncology wasn’t just about poor predictions, it also raised questions about transparency and who was ultimately responsible for its recommendations [3].

Taken together, these two frameworks highlight the tension between doing good at scale and protecting individual rights. A balanced approach would aim to achieve both: use AI to improve outcomes broadly (utilitarianism) while also building systems that are explainable, fair, and accountable (deontology). Ethical implementation of healthcare AI means not just asking “does it work?” but also “for whom?” and “under what conditions?”

6. Recommendations

This report proposes the following recommendations to address the ethical challenges and responsibly leverage the full potential of algorithmic decision-making in healthcare. These recommendations foster trust, ensure equity, and maintain patient safety as Artificial Intelligence healthcare solutions become more integrated into medical practice.

Improving Data Diversity and Quality

A primary concern identified in the case studies examined in this report is the impact of narrow or biased training data. Such biased data can negatively impact the training of the algorithms that power Artificial Intelligence solutions used in algorithmic decision-making, particularly medical diagnosis [15]. To mitigate this bias, healthcare organizations and AI engineers should mandate the creation and use of diverse and representative datasets when developing and validating Artificial Intelligence systems designed for healthcare. They should ensure the data is fairly representative of various demographic groups (e.g., age, gender, ethnicity, socioeconomic status) and is accurate to ensure models perform equitably across the majority of patient populations. It is also essential to invest in robust data governance frameworks that ensure data quality, integrity, and accuracy, which include documentation of the data sources and preprocessing steps used to create datasets.

Ensuring Explainability and Transparency in Critical Healthcare Artificial Intelligence

The "black box" nature of some Artificial Intelligence systems, as highlighted in the examination of transparency and the IBM Watson case, diminishes trust between the developers, users, and subjects of AI systems used in healthcare. As a result, it is important to prioritize developing and deploying explainable Artificial Intelligence models, especially for critical applications like diagnosis and treatment planning, so medical practitioners and patients can understand the underlying factors influencing an AI system's medical recommendations. In addition, Artificial Intelligence systems should be delivered with clear documentation detailing their intended use, performance metrics, limitations, and characteristics of the data they were trained on. These practices establish trust by ensuring that end users understand the mechanics of the system and how they influence the outputs produced by the AI model, which will eventually increase its adoption in healthcare [16].

Mandatory Third-Party Audits of Healthcare Algorithms

To ensure objectivity and rigor in the development, deployment, and usage of Artificial Intelligence systems used to produce healthcare algorithms, a system for independent, third-party audits should be introduced and enforced by lawmakers and regulatory bodies [17]. These audits should provide visibility into the accuracy, fairness, bias, privacy, and other tradeoffs considered when producing data for and developing Artificial Intelligence solutions before they are approved for use on human patients. The results of these audits should be transparent to healthcare providers, patients, and regulators and presented in the form of written reports and other useful media with the intent to inform interested parties of how these solutions performed during the auditing process. Information about the auditors, their credentials, and prior experience auditing similar algorithms, as well as disclosure about the performance of algorithms they have previously audited, should be disclosed to create greater trust in the auditing process.

Policy and Regulatory Recommendations

Artificial Intelligence's rapidly evolving nature requires an approach of active governance to ensure that existing regulations do not become antiquated over time. For example, healthcare regulations, such as HIPAA in the United States, should be regularly amended to address concerns arising from AI use, including algorithm bias, data privacy, and security [18]. Additionally, emphasis should be placed on the importance of developing AI governance proposals at the national and international levels to standardize best practices, safety protocols, and ethical guidelines for healthcare AI. In addition, regulatory bodies should be put in place to perform audits of AI model efficacy, enforce ethical standards, and impose penalties for any violations found during investigations. Any process created to extend existing healthcare regulations regarding AI should involve direct input from legal experts, ethicists, medical practitioners, software engineers, Artificial Intelligence experts, and regulatory bodies to introduce a fair, equitable, and effective regulatory environment.

Leveraging Advanced Architectures for Safer Artificial Intelligence Implementation

While the previous subsections describe policy-driven recommendations that can be introduced and enforced by healthcare practitioners, AI software developers, lawmakers, and regulators, this section focuses on a potential technical implementation of a systematic approach to developing AI software for the medical field. We propose the need for an agent-based architecture, coupled with a human-in-the-loop system, to offer a structured way to manage Artificial Intelligence in algorithmic medical decision-making workflows. This recommendation is based on the efficacy of taking this approach in complex medical areas such as clinical trials, mimicking specialization and collaboration, decomposition of complex medical problems, and producing specialized agents that focus on a particular problem domain, as demonstrated by existing research [11]. This approach has led to improved performance when compared to using a general-purpose large

language model by producing better results on performance benchmarks and providing more actionable and explainable outcomes [11].

Healthcare providers should explore and implement agent-based architectures where individual AI components can review existing decision-making processes, provide confidence scores for their outputs, and flag uncertain cases for human review. The "human-in-the-loop" approach ensures that a qualified professional can approve, override, or modify recommendations produced by Artificial Intelligence systems, acting as a critical guardrail and arguably the most crucial part of the algorithmic healthcare decision-making process. Human in the Loop (HITL) is a system design where humans are actively involved in an AI system's decision-making process. This is particularly important in the most critical stages of the software's involvement, where high stakes or ethical considerations must be accounted for before acting on a decision. It aims to couple the efficiency of using artificial intelligence to evaluate complex problems and perform recommendations while requiring human intuition and additional context that may not be available to the system to ensure that the results and outputs generated by the AI align with human values and societal norms [8].

The agentic AI process should promote the identification of potential bias in datasets and algorithmic outputs, not only flagging concerns but also proposing potential fixes or mitigation strategies that require human approval before implementation. This model should be extended to include agents that can perform a broader ethical review process that reviews existing research proposals, clinical trial designs, or new Artificial Intelligence deployments by identifying potential ethical concerns and generating and flagging them in a formal review document. Ultimately, any healthcare-related decision influenced by Artificial Intelligence software should clearly articulate potential ethical concerns and technical tradeoffs made when developing the system or producing an algorithmic decision, with final approval resting with a human expert. This aligns with research that emphasizes the opportunities presented by using AI as a tool to augment existing medical practices while preserving the role of human experts to make the final decision in high-stakes medical situations [9].

By implementing these recommendations, stakeholders can work towards an ecosystem where the use of Artificial Intelligence in healthcare is developed and deployed ethically, maximizing its benefits while minimizing patient risks and upholding the core principles of medical ethics. We see a trend in the direction of using Artificial Intelligence in the nursing science and healthcare fields, where the benefits of AI can be maximized by improving diagnostics and treatment, making research more efficient if developed and deployed ethically [10]. We believe safer and more robust Artificial Intelligence will be introduced in healthcare by focusing on implementing the solutions presented in this section.

An Example Agentic System Design

The system described in the previous section should be designed to perform bias analysis on a given dataset, train a model make diagnostic medical predictions, evaluate the performance of this model, prescribe a treatment plan, and generate a report detailing potential biases, mitigation

strategies, and ethical considerations, and produce a disclaimer that summarizes caveats that should be considered when making medical decisions based on the use of the algorithm. In this section, we explore the architecture for a system capable of diagnosing a patient with diabetes.

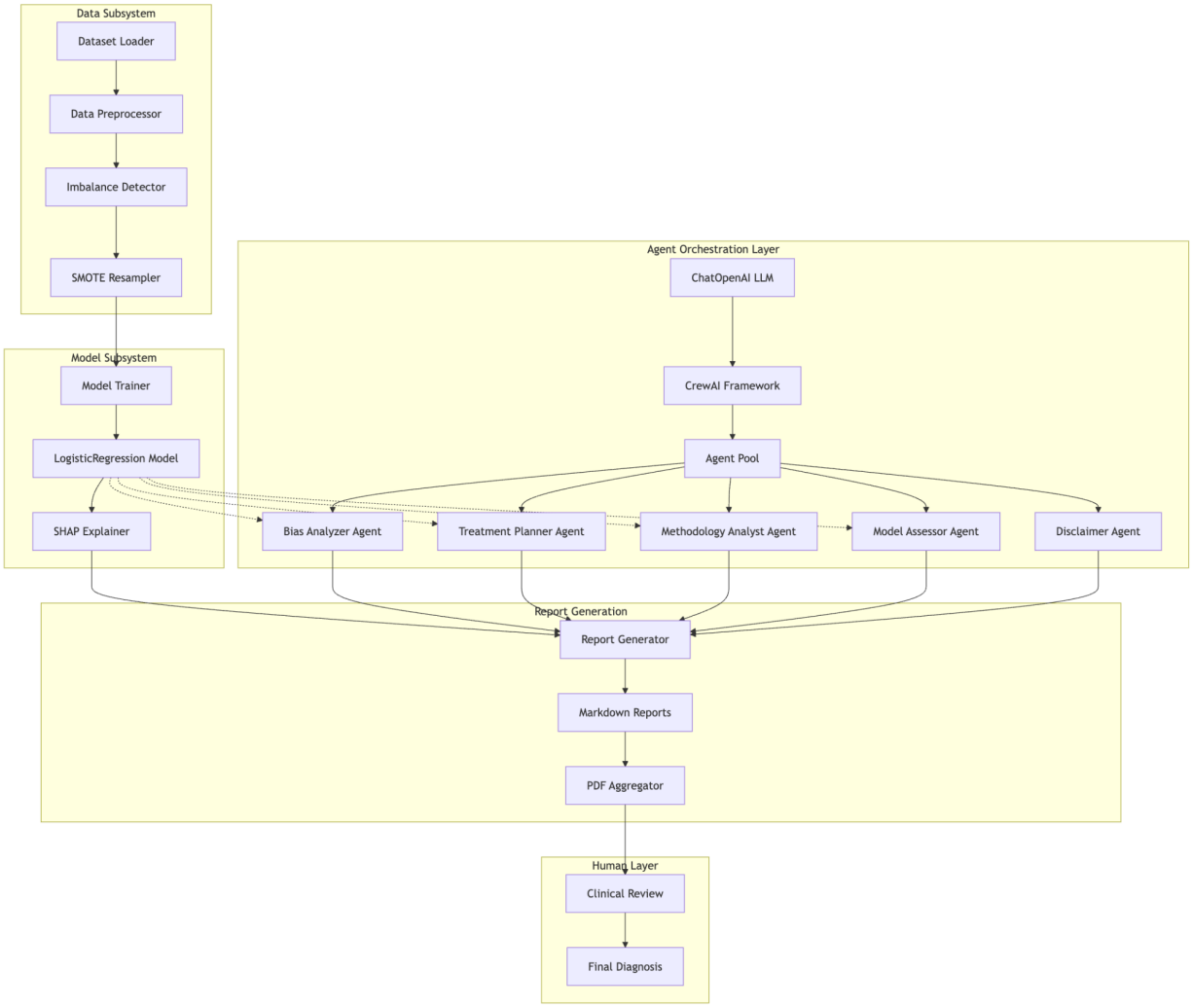

Figure 3: Architecture of proposed agent-based system

This proposed architecture handles data ingestion, data processing, model training, model interpretation, and automated reporting capabilities that summarize the assumptions made during data collection, data processing, model development, model usage, in addition to technical tradeoffs and ethical concerns. The primary steps taken by the system are broken down into Data Ingestion, Class Imbalance Detection, SMOTE Application [20], Model training, SHAP Analysis, Agent-based Report Generation, Report Aggregation, Document Output, and a formal Clinical Review performed by a human.

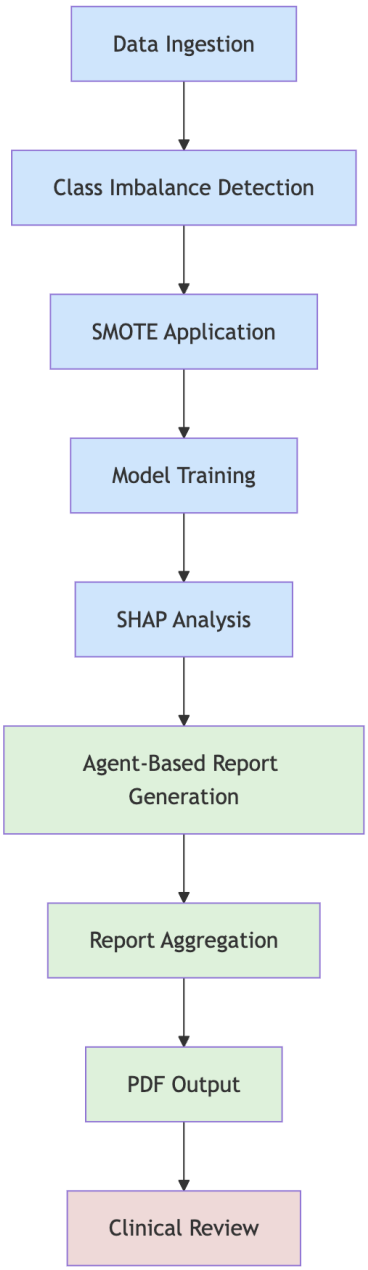

Figure 4: AI agent system flowchart

The flowchart in Figure 4 details how the system behaves from start to finish. First, data is ingested, then checks for class imbalance are performed. If necessary, SMOTE techniques are applied to the data to adjust for any detected class imbalances. Next, the model is trained with the ingested and formatted data. AI agents are then leveraged to create individual reports on their area of expertise, while another agent aggregates each report into a final report. Last, the final analysis is saved in a PDF file or other format that a human-in-the-loop will review to inform a critical medical decision.

The steps of this pipeline are summarized in more detail below.

1. **Data Ingestion** - This step involves preparing the data, cleaning it, and formatting it to ensure it is suitable for analysis.
2. **Class Imbalance Detection** - Evaluates the data to determine if class imbalances are present.
3. **SMOTE Application** - If a class imbalance is detected, SMOTE is applied to make the data more balanced.
4. **Model Training and Evaluation** - The system trains and evaluates a model that is used for making predictions and recommending a treatment.
5. **SHAP Analysis** - SHAP (Shapley Additive Explanations) analysis is performed on the trained model. SHAP values help to understand the contribution of each feature to the model's predictions, providing insights into the model's decision-making process [12]. As part of this process, fairness comparison plots may be generated to help assess the fairness of the model's prediction across different demographic groups and to help ensure that the model does not discriminate against any [13].
6. **Agent-based Report Generation** - AI agents are orchestrated to handle generating detailed reports on the bias, statistical methodology, model efficacy, code quality, and generate disclaimers on the system's intent, tradeoffs made, and ethical concerns of the system. Agents should be created using the CrewAI framework [14].
7. **Report Aggregation** - The system then aggregates each individual report into a comprehensive document that is output from the system.
8. **Report Output (PDF or other format)** - This is the final report produced by the AI system that details the recommended medical treatment as prescribed by the agents performing the system analysis.
9. **Clinical Review (Human-in-the-loop)** - In this final step in the system pipeline, a human healthcare provider uses the produced document to inform a formal diagnosis and treatment, and acts as the final system guardrail.

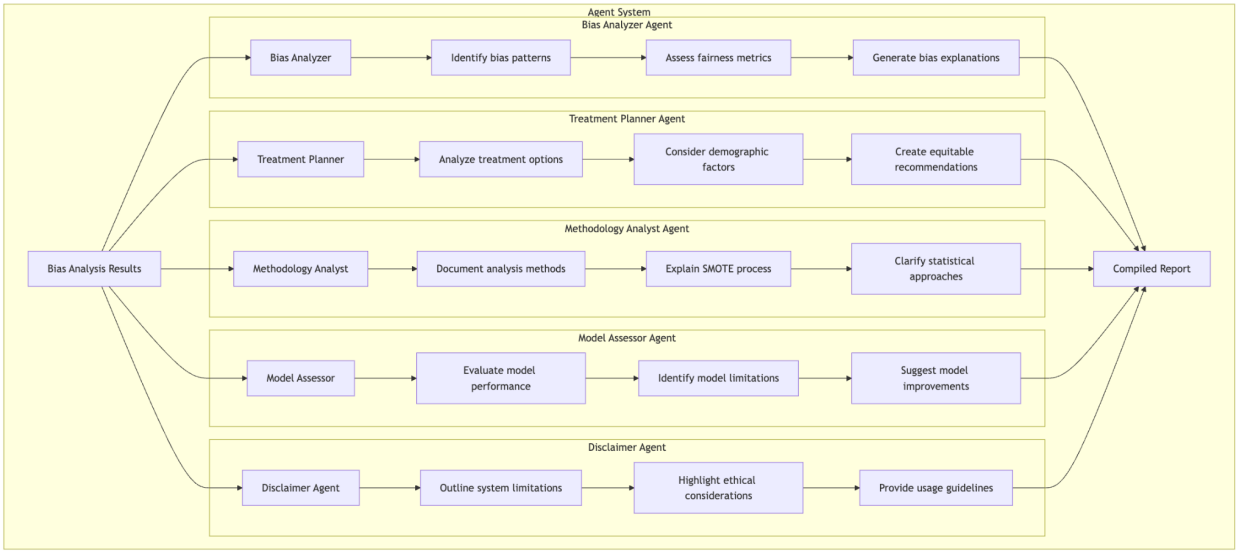

Figure 5: System agent flowchart

Figure 5 further breaks down the steps performed by each AI agent in the "Agent-Based Report Generation" step in the pipeline. Each agent and its responsibilities are documented below.

- **Bias Analyzer Agent** - Generates a bias report that discusses potential bias of the model based on feature influence, ethical considerations, and suggests potential mitigation strategies.
- **Treatment Planner Agent** - Generates a recommended treatment plan based on the prediction made by the trained model.
- **Methodology Analyst Agent** - Generates an explanation of the statistical methodologies employed in the program with a focus on how the data was produced, processed, and used to train the model. It also performs an assessment of the model's prediction capabilities and biases using various statistical techniques.
- **Model Assessor Agent** - Generates a written assessment of the model used to perform a medical diagnosis prediction. The assessment considers model characteristics, its suitability for making predictions, and any potential limitations.
- **Disclaimer Agent** - Generates a disclaimer regarding the intended use of this AI system, highlighting potential limitations arising from the system's design, the data it processes, statistical methodologies, and the AI technology used.

In summary, the final report generated by this proposed agentic system provides a detailed analysis of bias detected in data sources and trained models, highlights potential mitigation strategies, and summarizes ethical considerations, ensuring the model is fair and transparent and raises concerns when it is not. Additionally, the system produces diagnoses and treatment recommendations based on the model's predictions and suggests methods for mitigating any bias identified while detailing technical tradeoffs and concerns by evaluating the architecture of the system by analyzing the data, model, and code used to produce results.

This simple system's modular architecture allows for the introduction of additional features and improvements. To further extend the system, additional agents could be added to dynamically decide which statistical models and machine learning techniques to use based on the provided dataset and problem space. This can be accomplished with machine learning libraries, which can perform cross-validation and recommend models based on the characteristics of the data and model performance during the model selection process.

7. Conclusion

As artificial intelligence becomes more embedded in healthcare decision-making, it is crucial to examine these tools' ethical implications. Through our case studies, we've seen that while AI offers clear benefits, like improving diagnostic accuracy and streamlining care, it also brings real risks. Tools like IBM Watson for Oncology[3] and AI systems for skin cancer[5] detection show that without diverse, representative data and proper validation, AI can mislead clinicians or fail to serve all patients equally.

Key concerns like bias, lack of transparency, data privacy, and unclear accountability aren't just technical problems –they're ethical ones. These issues can erode trust in the healthcare system and lead to unequal treatment. Ethical frameworks such as utilitarianism, which emphasizes the greatest good for the greatest number, and deontology, which stresses duties and rights, offer valuable perspectives. These theories remind us that improving aggregate outcomes is important, but not at the expense of fairness, autonomy, or individual well-being.

Our recommendations aim to move the field in a more responsible direction. That includes building AI systems that are transparent, regularly audited, and inclusive of the diverse populations they're meant to serve. It also means putting clear safeguards in place so that human professionals, not algorithms, remain accountable for clinical decisions.

AI has the potential to support better healthcare, but only if it is implemented thoughtfully and is guided by strong ethical oversight. Ethical oversight isn't a barrier to innovation –it is what makes innovation worth trusting.

8. References

[1] Z. Jie, Z. Zhiying, and L. Li, “A meta-analysis of Watson for Oncology in clinical application,” *Scientific Reports*, vol. 11, no. 1, p. 5792, Mar. 2021. [Online]. Available: https://doi.org/10.1038/s41598-021-84973-5

[2] D. Wen, A. Soltan, E. Trucco, and R. N. Matin, “From data to diagnosis: skin cancer image datasets for artificial intelligence,” *Clinical and Experimental Dermatology*, vol. 49, no. 7, pp. 675–685, 2024. [Online]. Available: https://doi.org/10.1093/ced/llae112

[3] H. Dolfing, “Case Study 20: The $4 Billion AI Failure of IBM Watson for Oncology,” Dec. 2024. [Online]. Available: https://www.henricodolfing.com/2024/12/case-study-ibm-watson-for-oncology-failure.html

[4] C. Ross and I. Swetlitz, “IBM pitched its Watson supercomputer as a revolution in cancer care. It’s nowhere close,” *STAT News*, Sep. 5, 2017. [Online]. Available: https://www.statnews.com/2017/09/05/watson-ibm-cancer/

[5] S. E. Cerminara, P. Cheng, L. Kostner, S. Huber, M. Kunz, J. T. Maul, *et al.*, “Diagnostic performance of augmented intelligence with 2D and 3D total body photography and convolutional neural networks in a high-risk population for melanoma under real-world conditions: A new era of skin cancer screening?,” *European Journal of Cancer*, vol. 190, p. 112954, 2023. [Online]. Available: https://www.ejcancer.com/article/S0959-8049(23)00306-4/fulltext

[6] G. Brancaccio, A. Balato, J. Malvehy, S. Puig, G. Argenziano, and H. Kittler, “Artificial intelligence in skin cancer diagnosis: a reality check,” *Journal of Investigative Dermatology*, vol. 144, no. 3, pp. 492–499, 2024. [Online]. Available: https://doi.org/10.1016/j.jid.2023.10.004

[7] Y. Singh, H. Patel, D. V. Vera-Garcia, Q. A. Hathaway, D. Sarkar, and E. Quaia, “Beyond the hype: Navigating bias in AI-driven cancer detection,” *Oncotarget*, vol. 15, pp. 28665–28674, 2024. [Online]. Available: 10.18632/oncotarget.28665

[8] D. Ligot, “Human in the Loop: A Crucial Safeguard in the Age of AI,” 2024. [Online]. Available: https://hackernoon.com/human-in-the-loop-a-crucial-safeguard-in-the-age-of-ai

[9] T. H. Davenport and R. Kalakota, “The potential for artificial intelligence in healthcare,” *Future Healthcare Journal*, vol. 6, no. 2, p. 94, 2019. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC6616181/

[10] S. Yelne, M. Chaudhary, K. Dod, A. Sayyad, and R. Sharma, “Harnessing the power of AI: a comprehensive review of its impact and challenges in nursing science and healthcare,” *Cureus*, vol. 15, no. 11, 2023. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC10744168/

[11] L. Yue and T. Fu, “Ct-agent: Clinical trial multi-agent with large language model-based reasoning,” *arXiv e-prints*, arXiv:2404.14777, 2024. [Online]. Available: https://arxiv.org/abs/2404.14777

[12] D. B. Acharya, B. Divya, and K. Kuppan, “Explainable and Fair AI: Balancing Performance in Financial and Real Estate Machine Learning Models,” *IEEE Access*, 2024. doi: 10.1109/ACCESS.2024.3484409.

[13] SHAP Developers, “Violin Plot,” *SHAP Documentation*, 2024. [Online]. Available: https://shap.readthedocs.io/en/latest/example_notebooks/api_examples/plots/violin.html. Accessed on: May 11, 2025.

[14] crewAIInc, “crewAI,” *GitHub*, 2024. [Online]. Available: https://github.com/crewAIInc/crewAI. Accessed on: May 11, 2025.
[15] W. Sun, O. Nasraoui, and P. Shafto, “Evolution and impact of bias in human and machine learning algorithm interaction,” *PLOS ONE*, vol. 15, no. 8, p. e0235502, 2020. doi: 10.1371/journal.pone.0235502.
[16] A. A. Biswas, “A Comprehensive Review of Explainable AI for Disease Diagnosis,” *Array*, vol. 100345, 2024. [Online]. Available: https://doi.org/10.1016/j.array.2024.100345
[17] L. Oala, A. G. Murchison, P. Balachandran, S. Choudhary, J. Fehr, A. W. Leite, *et al.*, “Machine learning for health: algorithm auditing & quality control,” *Journal of Medical Systems*, vol. 45, pp. 1–8, 2021. doi: 10.1007/s10916-021-01783-y
[18] C. Mennella, U. Maniscalco, G. De Pietro, and M. Esposito, “Ethical and regulatory challenges of AI technologies in healthcare: A narrative review,” *Heliyon*, vol. 10, no. 4, 2024. doi: 10.1016/j.heliyon.2024.e26297
[19] L. Vearrier and C. M. Henderson, “Utilitarian Principlism as a Framework for Crisis Healthcare Ethics,” *HEC Forum*, vol. 33, no. 1–2, pp. 45–60, 2021. doi: 10.1007/s10730-020-09431-7
[20] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: synthetic minority over-sampling technique,” *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002. doi: 10.1613/jair.953

Appendix

- **Group presentation slides:** AI Ethics Final Project Slides
- **Supplemental demo of the recommended approach: Demo video**