Exploring the Ethical Implications of Using AI-Based Software for MRI Diagnosis in Clinical Settings

Komal Azeem
Queen's University
komal.azeem@queensu.ca

Ricky Leigh
Queen's University
21rjl13@queensu.ca

Rebecca Krichker
Queen's University
20rck4@queensu.ca

Sydney Basil
Queen's University
20sb85@queensu.ca

Cecile Woo
Queen's University
21ysw@queensu.ca

***Abstract*—The increasing usage of artificial intelligence in MRI disease classification and diagnosis presents several ethical implications related to patient privacy, data security, and responsible use. This paper will review some current use cases of AI-based MRI image classification models and propose a framework for ethics policymakers and medical information officers to ensure patient safety and responsible usage of AI in clinical settings.**

I. Introduction

The increasing popularity of Artificial Intelligence (AI) has led to its growing usage in the healthcare industry. The most common use cases range from improving patient interactions to helping physicians in their diagnosis [1]. An area of interest for clinicians has been the upcoming use of image-based detection models that assists in disease diagnosis. These machine learning algorithms use patient data to train and test the model and eventually outline areas of concern [2]. Given this new and emerging application, it is increasingly important for healthcare policy makers as well as private hospitals to understand the ethical implications associated with its use. This paper will delve into some current use cases of AI for MRI classification and disease diagnosis. It will also present a framework that policymakers and medical information officers should consult when assessing the ethical validity of an AI-based service. This paper will address the ethical implications of AI usage in MRI technology for diagnostic purposes through three main phases, outlined in Figure 1. The initial section of this paper will provide a review of the ongoing use cases of AI in MRI technology, providing a background on the datasets, model accuracy, and overt ethical implications. The second phase of this paper will address these specific ethical issues and will serve as a basis for producing a guideline for policy makers and medical information officers who are looking to implement the model for clinical use. The final phase will assess the validity and necessity of such a guideline.

II. Related Works

Existing guidelines have addressed the various ethical implications of using AI in healthcare settings, but the broad nature of these guidelines have made it difficult to apply to varying situations. The guideline developed by Bouderhem [3] is one of the most prominent guideline pertaining to the ethical use of AI in healthcare. Bourderhem [3] outlines a wide range of use cases of AI and analyzes the current ethical challenges and provides various recommendations to address these issues. Unfortunately, due to its broad scope, it fails to apply itself to more specific use cases. As a result, we have chosen to address one category of use cases and cater a comprehensive guideline that outlines major weaknesses and ethical concerns.

III. Reviewing Current Use Cases And Literature

A. *Diagnosing Alzheimer's Disease Through CNN-Based MRI Detection*

Jain et al. [4] explored the use of Convolutional Neural Networks (CNNs) and transfer learning to develop an image classification model that could aid physicians in the early diagnosis of Alzheimer's Disease. Since the training of a CNN typically requires a large database of information, transfer learning is utilized to easily facilitate the process of developing a new model. Transfer learning is the use of one model's output to train another model [5]. Currently, MRI is the most common method of detecting early deterioration as it provides clear anatomical abnormalities, which is strongly linked to the development of Mild Cognitive Impairment [6]. As a result, the classification model categorized MRI results as either Mild Cognitive Impairment (MCI), Cognitively Normal (CN), and Alzheimer's Disease (AD). These classifications were then used to identify any neural degeneration early on and guided the physician's treatment plan.

This paper [4] identifies how a pre-existing CNN, VGG-16, that was originally trained on data from ImageNet was used to facilitate the production of a new 3-way classification CNN for

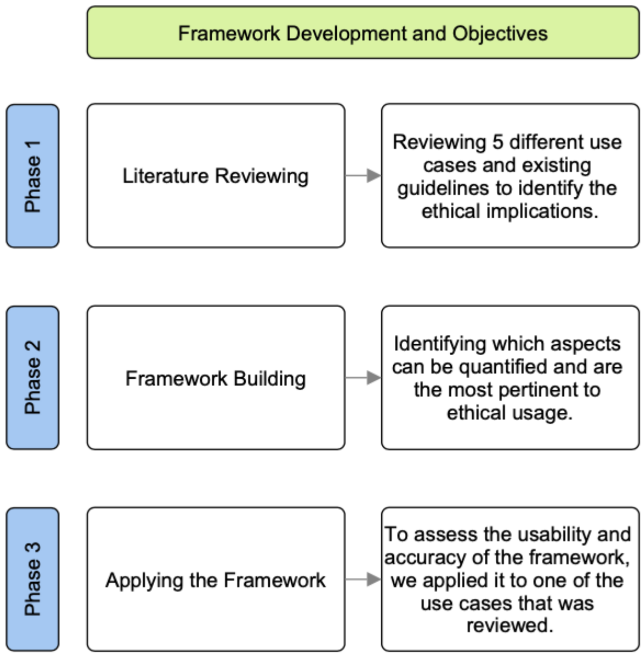

Fig. 1. Flowchart of the phases and objectives of this paper.

MRI images. The resultant precision of their developed validation set was 95.73%. For binary classifications, the model accuracy for AD vs MCI was 99.30%, 99.14% for AD vs CN, and 99.22% for MCI vs CN. Further analysis of the 3-way model using confusion matrices revealed a precision, recall, and F1 score for each condition: the AD condition produced values of 1, 0.91, and 0.95, the CN condition produced values of 0.99, 0.97, 0.98, and the MCI produced values of 0.90, 1, and 0.94, respectively.

Due to the covert nature of the initial dataset and its use of transfer learning, inferring the level of diversity involved in the dataset is very difficult. Any bias present in the initial dataset from ImageNet poses a great risk to the developed CNN, as it will be reflected in the results. A lack of diversity in patient representativeness could result in an overfitting of specific features, leading to inaccurate classification and reported accuracies, which is a large risk for potential clinical usage; racial, sexual, and clinical diversity is an important consideration when assessing the accuracy of the model and the following clinical implications. Also, due to the black box nature of CNN's, the decision-making strategies employed by the CNN are not fully transparent, and can only be altered to a certain extent. Lastly, there is a question of cost of development, implementation, and access; when determining for who and how this technology is implemented, clinics should consider which patients will have access to this technology, and if it provides them with a greater advantage than an individual who is unable to afford it. There remains a question of affordability and insurance coverage for the model usage, as different countries have varying policies on the type of technology and services available to the patient.

B. Using AI in MRI Classification of Liver Tumors

Zhen et al.'s [7] primary objective was to develop a deep learning system (DLS) for accurate liver tumor diagnosis using MRI and clinical data. The model addressed the limitations of current diagnostic methods, which are often subjective and rely on the radiologist's experience. The DLS was created to classify liver tumors into seven categories, differentiate between benign and malignant tumors using unenhanced MRI, and further classify malignant tumors by integrating unenhanced images and clinical data. Ultimately, the researchers aimed to provide a more efficient, accessible, and accurate diagnostic tool that could potentially reduce the need for contrast agents and the associated costs and side effects.

Zhen et al. [7] used convolutional neural networks (CNNs), specifically the Google Inception-ResNet V2 architecture, which was pre-trained on a large image dataset and subsequently fine-tuned using the study's liver tumor MRI data. The dataset included 31,608 MRI images from 1,210 patients for training, and 6,816 images from 201 patients for validation. The seven-way classifier, which utilized six MRI sequences, achieved area under ROC curve (AUC) values ranging from 0.897 to 0.987, with sensitivity between 53.3% and 100% and specificity between 91.6% and 99.5%. The binary classifier, using unenhanced sequences, reached an AUC of 0.946, indicating its ability to distinguish malignant from benign tumors with accuracy comparable to that of a classifier based on enhanced sequences (AUC of 0.951). The three-way malignancy classifier, which integrated unenhanced images and clinical data, demonstrated significantly enhanced AUCs (ranging from 0.963 to 0.998) compared to models that relied solely on enhanced images for similar classifications, closely matching the radiologists' performance.

The study received approval from an independent institutional review board (IRB) at Sir Run Run Shaw Hospital, China, however there were concerns regarding the robustness of Chinese IRBs. Factors such as lack of thorough review processes and insufficient oversight may have compromised the ethical review capacity of the study [8]. Although written informed consent was not mandated in accordance with local laws, this approach is different from other countries' standards. This discrepancy highlights the variability in ethical standards globally, raising issues of patient privacy and the generalizability of research. While the study employed saliency maps to enhance interpretability, it acknowledges the lack of full transparency regarding the AI models' decision-making processes. Lastly, the study relied on patient medical records and MRI images, with limited discussion on specific security measures or anonymization protocols.

C. Utilizing Machine Learning in MRI Technology to Diagnose Schizophrenia

The paper by Sadeghi et al. [9] reviewed and evaluated the applications of artificial intelligence (AI), specifically machine learning (ML) and deep learning (DL), in the diagnosis of schizophrenia using magnetic resonance imaging (MRI). Schizophrenia is a complex psychiatric disorder that poses significant challenges to the accurate diagnosis due to the heterogeneity of the symptoms and the absence of definitive biomarkers [10]. Sadeghi et al. [9] explored AI-based computer-aided diagnostic systems (CADS) designed to automate the process of diagnosis by using structural MRI (sMRI) and functional MRI (fMRI) datasets. The paper emphasizes the integration of AI into clinical workflows to assist healthcare professionals. It also provides a summary of the advancements of AI methods, specifically their performance, challenges, and potential improvements for diagnosing schizophrenia.

The study used various AI models, including conventional ML methods (e.g. Support Vector Machines [SVM], Random Forest) and advanced DL architectures (e.g. CNNs, Autoencoders) [9]. The reported accuracies varied widely depending on the dataset, preprocessing techniques, and feature selection methods. For example, ML models like SVM showed up to 94% accuracy with carefully extracted features [9]. At the same time, DL methods, such as 3d-CNNs, achieved comparable or better results through direct analyses of raw data. The models leveraged features such as gray matter volume, connectivity matrices, and task-based fMRI signal data [9]. However, the performance consistency was impacted due to challenges such as dataset imbalance, small sample sizes, and MRI image noise. The paper concluded that DL's automatic feature extraction was more advantageous than the ML's reliance on manual feature engineering but acknowledged the computational intensity required as a trade-off.

Although the paper did not explicitly discuss the ethical implications of AI-based schizophrenia diagnosis, its findings raise potential concerns. One such implication is that AI models trained on neuroimaging data may overemphasize structural and functional brain abnormalities at the expense of behavioural and functional impairments, which are critical for DSM-5 diagnosis [11]. Hence, this could lead to biases in clinical decision-making if AI tools are not used with comprehensive psychiatric evaluations. Furthermore, while publicly available datasets (e.g. COBRE, SchizConnect) enhance accessibility, limitations in sample diversity and generalizability could lead to the misrepresentation of schizophrenia phenotypes in AI-based diagnostic tools. A final challenge is the lack of standardized evaluation metrics and transparency in model decisions, which can hinder replicability and trust in clinical settings. Future research should focus on integrating multi-modal assessments (e.g., neuroimaging and behavior markers), improving or adding data governance frameworks, and ensuring that AI models are interpretable and clinically relevant.

D. Detecting Brain Tumors Using Convolutional Neural Networks

Research conducted by Rahman et al. [12] leveraged artificial intelligence, specifically the EfficientNetB2 deep learning architecture, to detect patterns indicative of brain tumors in MRI scans. By improving MRI image quality through preprocessing techniques such as cropping, equalization, and homomorphic filtering, the research aimed to enhance the accuracy, efficiency, and consistency of tumor detection. Using publicly available datasets featuring diverse MRI images of individuals with and without brain tumors, the study sought to create a tool that could assist, rather than replace, physicians. This AI-driven methodology is designed to provide rapid, precise, and reliable tumor detection, reducing diagnostic delays and mitigating variability caused by human interpretation. This study used the EfficientNetB2 deep learning architecture, a convolutional neural network model known for its efficiency and scalability in image classification tasks [12]. Fine-tuned for brain tumor detection, the model was trained on three publicly available datasets, achieving high validation accuracies of 99.83% on the BD-BrainTumor dataset, 99.75% on the Brain-tumor-detection dataset, and 99.2% on the Brain-MRI-images-for-brain-tumor-detection dataset. While these results indicate strong performance, real-world clinical settings introduce variability that may impact accuracy, such as differences in MRI protocols, scanner types, and patient demographics. Thus, external validation on diverse, real-world datasets is necessary to assess generalizability. The architecture leverages a balanced scaling approach to optimize depth, width, and resolution, enabling it to capture intricate patterns in MRI scans. Preprocessing techniques, such as cropping, equalization, and homomorphic filtering, further enhanced the input data quality, boosting the model's ability to identify tumor regions accurately. However, despite its strong performance, the model is intended as a decision-support tool rather than a replacement for clinical expertise. Further studies on model interpretability and robustness in real-world environments will be critical for ensuring safe and effective deployment in healthcare settings. Publicly available datasets from Kaggle, including BD-BrainTumor, Brain-tumor-detection, and Brain-MRI-images-for-brain-tumor-detection were used for the study [12]. While these datasets are anonymized and promote accessibility, the lack of information about patient consent during data collection raises ethical concerns. In addition, no explicit security measures, such as encryption or secure storage, were mentioned, which could pose risks in clinical applications requiring compliance with privacy regulations like HIPAA. The datasets may also be biased, potentially limiting the model's generalizability to diverse populations, tumor types, and imaging conditions. The absence of real-world clinical validation further increases this concern. Additionally, the study does not address the explainability of the model, which is critical for building trust in AI-driven diagnoses. While intended to assist physicians, over-reliance on such models without sufficient human oversight could be problematic. Addressing these issues is crucial for the ethical and effective integration of an AI model such as this one in clinical healthcare.

E. Using Deep Learning Models For Early-Stage Breast Cancer Screening

Breast cancer is the second leading cause of cancer-related deaths among women [13]. This has prompted researchers and healthcare systems to implement large-scale mammography screening programs aimed at early detection to improve outcomes [14]. For instance, the USA performs approximately 43 million mammograms annually [5]. However, the effectiveness of these screenings is hindered by variability in cancer detection professionals' accuracy [15]. Additionally, a shortage of mammography specialists worldwide limits the availability and scalability of these programs [14]. To address these challenges, McKinney et al. [16] introduced an artificial intelligence (AI) system, specifically a deep learning model, designed to screen mammograms for early-stage breast cancer, enhancing both the accuracy and scalability of breast cancer detection.

The AI system for breast cancer screening consisted of an ensemble of three deep learning models: lesions, breast and full case. The lesion model identified the ten most specific regions of suspicion in mammogram images of suggestive cancer using a model called RetinaNet2. Each extracted region was passed through a feature detector called MobileNetV23. Malignancy predictions were then produced for each region and combined into a composite score. The breast model processed images of each breast independently. The model used a ResNet-v2-50 network (a type of CNN) as an Image Feature Extractor. Each breast had two views of a mammogram, which were concatenated and passed through an additional neural network to predict a cancer score for each breast. The cancer scores for the right breast and left breast were then compared and the maximum score between the two breasts were taken as the case-level score. Finally, the case model considered the complete set of mammogram views; the model uses a ResNet-v1-50 network as a feature extractor. The complete set of mammogram views contained four images, which were concatenated. The concatenated vector was passed through a hidden layer used for binary classification to determine whether the patient (case) had cancer (yes or no).

All models were trained with data augmentation applied to each image. Each model generated a cancer risk score between 0 and 1, with the final score being the mean of the three models' predictions. The model was trained using datasets from two UK screening centers and one US center, representing both populations. The UK dataset included 25,856 women, while the US dataset contained 3,097 women. McKinney et al. [16] evaluated the model's accuracy through three approaches. First, they compared AI predictions with historical clinical decisions, finding that the AI model demonstrated higher specificity and improved sensitivity compared to both UK first readers and US single readers in radiology practice. The model's performance on the UK dataset had an AUC of 0.996, while on the US dataset, the AUC was 0.883. Second, they conducted cross-cultural testing by applying the UK-trained model to the US dataset, which showed improved specificity and sensitivity compared to radiologists. Here, the AI model achieved an AUC of 0.889 on the US dataset. Finally, they compared the AI system's performance against six US board-certified radiologists interpreting 500 challenging US cases. In this comparison, the AI system significantly outperformed the radiologists' average performance, with the mean radioligst reading AUC being 0.75, whereas the AI system achieved an AUC of 0.871.

One limitation highlighted by McKinney et al. [16] was dataset representativeness. While the UK dataset mirrored the nationwide screening population, the US dataset came from a single screening center. For this AI system to achieve its potential for scalability and accessibility, the datasets need to be truly representative of diverse populations. This issue is reflected in the system performance: AUC values were highest in the UK where it was developed but decreased when trained on US data. Although this decline in performance was minimal, it raises concerns about the system's effectiveness across different populations, particularly given that performance variations were observed even between the demographically similar UK and US.

IV. Developing a Framework

Upon conducting a thorough review of the varying use cases of AI models in MRI technology and diagnosis, the obvious, yet unique, ethical concerns became clear. In order to develop a successful and practical framework for policymakers, it was essential to grasp the common underlying issues in each use case. It became abundantly clear that patient privacy, data collection, interpretation of results, and responsible use were critical to the ethical use of each of the models. As a result, the proposed guideline, provided in the Appendix, presents a graded rubric system that allows developers and future executors to filter their models through an ethical framework to determine the strengths of their models.

A. Criteria Justification

In recent years, significant progress has been made towards regulating AI in healthcare through scientific reviews and policy initiatives [3], [17], [18]. We drew upon these efforts to establish our specific guidelines. In particular, Bouderhem's [3] article provides a comprehensive analysis of the technical, ethical and regulatory challenges related to the application of AI in healthcare. We specifically chose to closely examine Bouderhem's [3] article given its publication recency and thorough examination of AI's prevalence, opportunities, challenges and risks. Bouderhem [3] outlines a broad range of AI applications, including care management, drug discovery, medical imaging analysis and more. Bouderhem's [3] analysis thoroughly examines global policies, such as the United States' General Data Protection Regulation (GDPR), the United States' Health Insurance Portability and Accountability Act (HIPAA) and the European AI Act proposal. This article concludes itself by calling on the WHO to strengthen its regulatory role in AI-driven healthcare, arguing that current legal frameworks are insufficient.

Bourdenhem's message is important and his background research is rigorous, however, the broad scope and general recommendations create ambiguity about implementation [3]. To build upon their work, a more specific guideline for policymakers and healthcare institutions was developed to regulate the AI applications in healthcare. One specific and pertinents application of AI in healthcare was chosen as the basis of this guideline: the role of AI in MRI detection. Bouderhem's [3] article notes that policies like the AI Act proposal struggle with defining AI broadly. This makes it difficult to provide clear ethical and regulatory recommendations. Narrowing the focus to AI in MRI detection specifically will ensure our criteria is specific and practical for future implementation.

The usage of AI in MRI detection was chosen due to its promising potential across a wide range of clinical applications, including cancer detection, cardiac imaging, and musculoskeletal assessment [19]. MRI technology is also highly versatile, benefiting patients across all demographics, including adolescents, children, and older adults [19]. Despite its advantages, traditional MRI technology has notable limitations, such as long scan times and high sensitivity to patient movement, which can lead to blurred images and increased costs [19]. However, AI-driven advancements have demonstrated the ability to accelerate scans, enhance image quality, and reduce expenses [19]. Given the notable potential of AI in MRI, it is crucial to address its challenges and establish clear, rigorous standards. Implementing well-defined technical and ethical guidelines will ensure responsible adoption in healthcare institutions, maximizing its benefits while maintaining patient safety and data integrity.

The goal of this paper is to establish clear guidelines with numbered criteria (1-4) to systematically evaluate whether an AI model meant for MRI technology meets technical and ethical standards. Building on the challenges outlined by Bouderhem [3], three key categories were developed with subsequent criteria for the responsible implementation of AI-based MRI technology in healthcare institutions: patient privacy and security, data collection, and interpretation and responsibility.

First, a section on patient privacy and security addresses the risks posed by patient information being shared or stolen [3]. This criteria emphasizes that patient data used for AI training should be encrypted and anonymized, with strict mechanisms in place to prevent unauthorized access, data breaches, or re-identification risks. In addition, patients should have clear and accessible options to give informed consent for data collection, processing, and sharing, with the ability to opt out if they choose. Furthermore, AI systems should comply with legal and ethical data protection standards such as GDPR and HIPAA to uphold patient privacy.

Second, a section on data collection addresses how biased algorithms can result in discrimination and inaccurate predictions [3]. This section emphasizes the importance of diverse and unbiased datasets to prevent discrimination, particularly against vulnerable populations. AI models should incorporate MRI data from underrepresented cases to improve diagnostic precision across all groups of patients. Healthcare providers and patients should also be fully informed about how AI models are trained, including potential biases or data gaps. In addition, continuous monitoring must be implemented to identify and mitigate biases in real-world applications, ensuring that AI models remain equitable and effective.

Finally, a section on interpretation and responsibility addresses the lack of performance indicators in AI system. These are metrics that healthcare providers need to detect errors and biases that could have legal implications such as medical malpractice liability [3]. Our criteria focus on AI's role as a supportive tool rather than a replacement for physician judgment. The model should enhance, rather than dictate, clinical decisions, with physicians fully trained to interpret and verify AI-generated outputs. Transparency is essential, meaning that AI models should provide explainable results, confidence scores, and clearly defined limitations to prevent over reliance on automated assessments. Finally, AI-generated insights should be integrated into clinical workflows, ensuring that physicians can track and evaluate their influence on final diagnostic decisions.

B. *Utilizing the Scoring System*

The advancement of AI-driven healthcare models presents significant ethical and regulatory challenges, highlighting the need for a structured and objective grading system. A well-defined framework is essential to ensure transparency and consistency while evaluating AI compliance with patient privacy and ethical considerations. Without standardizing the grading system, assessing the ethical nature of the AI models may become subjective, resulting in reduced trust in the AI applications in healthcare. The proposed grading system provides a uniform approach to measure compliance with ethical standards and technical effectiveness, allowing healthcare providers, policy makers, AI developers, and patients to make informed decisions based on measurable criteria rather than arbitrary judgment.

The grading system utilizes a four-tier rubric, assigning numerical values from 1 to 4 based on compliance and performance levels. This structured approach ensures that evaluations remain comparable across AI models, and aligned with industry best practices and regulatory requirements. By adhering to fundamental legal and ethical standards, the grading system minimizes ambiguity and provides a precise evaluation of an AI model's adherence to ethical and technical benchmarks.

1) Considerations for Level Selection: The selection of each level within the grading system was guided by considering compliance, effort, and ethical responsibility. The grading structure follows a hierarchical pattern where Level 1 represents a complete lack of necessary safeguards, making the AI system unreliable. Level 2 reflects basic but inconsistent measures, highlighting the need for further refinement. Level 3 shows a a strong adherence to ethical and technical standards, with only minor area that needed improvement. Finally, Level 4 reflects full compliance, where AI systems not only meet but exceed industry standards by implementing proactive measures for security, fairness, and usability. This progressive framework ensures that AI models are assessed in a manner that acknowledges incremental improvements while maintaining strict ethical requirements. Ethical considerations such as patient privacy, informed consent, and bias mitigation are inherently qualitative, making their assessment challenging. Hence, the grading system translates these principles into measurable indicators. For example, privacy measures are assessed based on encryption and anonymization, while informed consent is evaluated through clarity, accessibility, and comprehensiveness. Bias mitigation efforts are quantified by analyzing demographic representation and the frequency of bias examinations. By transforming these qualitative aspects into metrics, the grading system ensures that AI compliance is systematically evaluated, allowing stakeholders to precisely gauge an AI model's ethical performance.

2) Ethical Trade-offs and Balance: The grading system acknowledges the trade-offs in AI development, balancing model accuracy, transparency, security, and fairness. AI models that emphasize accuracy at the expense of transparency receive lower interpretability scores, while models prioritizing privacy without sufficient compliance measures receive lower regulatory adherence. This balanced approach prevents any single aspect from being overemphasized at the expense of another, allowing a balanced evaluation of the model.

3) Pass/Fail Criteria Thresholds: The grading system classifies AI models into four levels based on their ethical compliance, security, fairness, and transparency, ranging from severely inadequate to fully compliant. Level 1 models pose severe ethical and security risks, failing to meet basic industry standards, making them unsuitable for real-world usage. Level 2 models demonstrate partial compliance, but require significant modifications to mitigate ethical concerns before being considered reliable. Level 3 models are deemed ethically sound and compliant with industry regulations, incorporating strong security, privacy, and fairness measures, though they may need minor refinements to optimize ethical performance. Level 4 models represent full compliance with industry standards, demonstrating proactive strategies to ensure long-term reliability, ethical integrity, and responsible AI implementation. To pass the grading system, an AI model must achieve a minimum threshold of 50% in each section, ensuring it meets basic ethical, security, and transparency requirements. The evaluation framework aligns with GDPR, HIPAA, and industry best practices, emphasizing ethical compliance while considering accessibility for real-world application. The scoring system categorizes models based on performance, where below 50% indicates failure and requires major improvements, scores between 50-75% are satisfactory but need refinement, and scores above 75% demonstrate strong ethical compliance to best practices. Each section of the evaluation contributes to the overall ethical assessment of an AI model. In Section one: Patient Privacy and Security (Total Score: 20), models scoring 0-10 fail to meet ethical security standards and require major improvements, while scores between 11-15 meet baseline requirements but need refinement, and those above 15 are generally compliant with minor adjustments needed. In Section two: Data Collection Practices (Total Score: 16), models below 8 fail ethical standards, scores between 9-12 are satisfactory but could improve, and scores above 12 demonstrate responsible data collection practices. In Section three: Interpretation and Responsibility (Total Score: 28), scores below 14 indicate a need for major improvements, those between 15-21 are satisfactory but require refinements, and those above 21 meet ethical guidelines and demonstrate strong accountability. This structured evaluation ensures that higher-scoring models reflect increasingly proactive and comprehensive ethical measures, rather than simply meeting minimum compliance. By incentivizing models to prioritize security, fairness, and transparency, the system establishes a clear guidance for improvement and aligns with regulatory best practices. This framework not only advances ethical AI development, but also provides a guiding structure for developers and policymakers to create AI solutions that are both innovative and responsible.

V. Application of the Framework

In this section, the proposed ethical guidelines will be applied to the paper ”Advanced AI-driven approach for enhanced brain tumour detection from MRI images utilizing EfficientNetB2” by Rahman et al [12]. This paper was chosen for review as it explores the use of AI in medical imaging, specifically for brain tumour detection, which aligns well with the key areas of the criteria: patient privacy and security, data collection, and AI interpretability. The criteria will help highlight the ethical considerations associated with the three main issues as well as transparency in AI decision-making of this specific paper. Applying these standards will assess how well the study [12] adheres to the ethical principles in AI deployment within healthcare, pinpointing areas that require improvement. This section serves as an example of how policymakers and clinicians who hope to implement AI-based technology should approach future applications and ethical concerns.

A. Application of Section 1 Guidelines

Section One of the guidelines, found in the Appendix A, includes actionable measures such as encryption and anonymization of patient data, mechanisms to prevent unauthorized access, informed consent procedures, opt-out options, and adherence to regulatory frameworks such as GDPR and HIPAA. Applying these guidelines to the article “Advanced AI-driven approach for enhanced brain tumor detection from MRI images utilizing EfficientNetB2” [12] makes it evident that the study lacks critical privacy safeguards. While the research utilizes publicly available datasets such as those from Kaggle, it does not clarify whether these datasets meet standard privacy regulations or whether patients were given a choice to opt out. Another key component of this section is informed consent, which means patients should be aware of how their data is being used, collected, and stored. The study does not discuss the consent mechanism or transparency in using the AI model. This lack of coverage raises ethical concerns, as patient data is central to AI training, and its misuse could lead to privacy violations in patient data. Furthermore, the study does not detail security measures implemented to prevent unauthorized data access, leaving patient information vulnerable. Regulatory compliance is another missing factor, as there is no mention of how the model aligns with data protection laws. Given these shortcomings, the study scores a level 2 for section 1. The reasoning for the score is that they use publicly available datasets. However, they lack mention or action of protection and transparency regarding data security, consent, and patient control over personal information. To improve, researchers should implement encryption, outline informed consent processes, ensure opt-out options, and align with GDPR/HIPAA standards to enhance data protection.

B. Application of Section 2 Guidelines

Section Two of the guidelines [A], highlights the importance of utilizing diverse and representative datasets, along with ensuring transparency in data collection, to mitigate bias and improve the reliability of the results. The study by Rahman et al. [12] uses three publicly available MRI datasets (BD-BrainTumor, Brain-Tumor-Detection, and Brain-MRI-Images-for-Brain-Tumor-detection) to train an EfficientNetB2 deep learning model for brain tumor detection. While this demonstrates an effort toward dataset diversity, the study does not provide demographic breakdowns, fairness assessments, or bias mitigation strategies. Issues with data diversity can raise concerns about how well the model generalizes across different populations, tumor types, and imaging conditions. A key weakness is the lack of inclusion of rare cases, which may impact the model accuracy for underrepresented tumor types or patient demographics. Additionally, no information about dataset bias or continuous monitoring mechanisms is provided, making it difficult to determine if the model adapts to real-world variations. The absence of explainability features and clinician involvement in data interpretation further limits transparency. Trust in AI-driven diagnostics may be reduced without clearly disclosing how patients and healthcare providers are informed about dataset limitations. Hence, the study scores a level 3 in data collection. The study effectively uses multiple datasets and advanced preprocessing techniques but lacks bias analysis, fairness checks, and long-term validation in clinical settings. The study should include rare tumor cases, conduct fairness audits, and implement ongoing bias monitoring to improve its score in section 2. Future AI models should prioritize transparency and real-world validation to ensure equitable healthcare applications.

C. Application of Section 3 Guidelines

Section 3 of the AI ethics assessment matrix [A] involves interpretation and accountability, whether and how the AI models aid doctors in decision-making, are transparent, and yield accountability. It establishes whether the AI is presenting useful help, offering explainability, offering confidence scores, defining boundaries, and embedded in clinical workflows. Section 3 also deals with physicians' training and how accountability is distributed between human clinicians and AI technology. Applying these criteria to Rahman et al.'s [12] work, we see that the proposed EfficientNetB2-based model for brain tumor detection is very accurate and possesses good clinical potential. However, it lacks any significant features of interpretability or transparency. The model's decision-making process is not well explained, and while confidence scores are mentioned, their calibration and reliability are not mentioned. Moreover, education of physicians in AI interpretation is not addressed, and over-reliance or misinterpretation of AI results is feared. Generally, this study falls partially within Section 3 of the matrix, at an estimated level 3. The model is applicable in decision support but requires a high level of physician control. The weakest aspect is the lack of formal training of clinicians, which can be resolved by making AI interpretation courses mandatory and having specific guidelines on AI-assisted MRI evaluation. In addition, policymakers need to ensure that AI algorithms applied in radiology are transparent, have well-documented limitations, and function effectively in clinical practice to enhance trust and reliability in health applications.

D. Overall Performance

The objective of the AI ethics guideline is to ensure that AI-driven healthcare applications align with ethical, legal, and safety standards while promoting fairness, transparency, and accountability. The framework assesses AI models in MRI-based diagnostics across three key areas: patient privacy and security, data collection practices, and interpretation and responsibility. By applying this guideline to Advanced AI-driven approach for enhanced brain tumor detection from MRI images utilizing EfficientNetB2, we can evaluate how well this study adheres to ethical principles in AI deployment within medical imaging. In Section 1, which focuses on patient privacy and security, the study performed poorly, scoring a Level 2. While the dataset used in the study was publicly available, the paper did not address key privacy safeguards such as encryption, informed consent, or patient opt-out mechanisms. There is also a lack of discussion regarding compliance with GDPR or HIPAA standards, raising ethical concerns about data security. To improve, future AI research must prioritize explicit policies on patient consent, transparency in data use, and robust security measures to prevent unauthorized access and misuse of medical data. In Section 2, which evaluates data collection fairness and representativeness, the study performed slightly better, scoring a Level 3. The research used multiple datasets, demonstrating some level of diversity in training data. However, the absence of demographic breakdowns and fairness assessments makes it difficult to determine whether the model generalizes well across different patient groups. The study also lacked bias mitigation strategies, and no measures were in place for continuous monitoring of AI bias in real-world applications. To align better with ethical standards, future research should integrate rare cases, conduct fairness audits, and establish long-term monitoring strategies to prevent AI bias from affecting clinical outcomes. In Section 3, which assesses AI interpretability and physician responsibility, the study scored another Level 3. While the model demonstrated strong accuracy, it lacked crucial elements of explainability and transparency. Confidence scores were not clearly calibrated, and there was no structured training for physicians to interpret AI-generated outputs. Without a clear explanation of how the model reaches its decisions, there is a risk of over-reliance on AI or physician misinterpretation. To improve, AI developers should implement transparent confidence scoring, ensure full documentation of model limitations, and provide mandatory AI training for clinicians to mitigate risks of blind AI reliance.

VI. Conclusion

In conclusion, this paper provided a broad review of the current usage of AI models in MRI technology to help physicians with patient diagnosis. While conducting these reviews, several ethical issues specific to MRI usage became evident. As a result, a comprehensive guideline was developed for policymakers and developers to grade the ethical implications of the model. Looking ahead, the implications of these ethical guidelines extend far beyond this study. As AI continues to shape medical diagnostics, researchers, policymakers, and healthcare providers must work together to establish universal ethical standards for AI in radiology and medical imaging. Future AI models should be designed with privacy-first architectures, fairness-aware algorithms, and clinician-in-the-loop frameworks to ensure ethical and responsible AI implementation in patient care. Transparency, accountability, and human oversight must remain central principles in AI development, ensuring that AI enhances medical decision-making without undermining physician expertise or compromising patient rights. By adopting these ethical standards, the future of AI in healthcare can be both transformational and ethically sound.

VII. Future Work

Future work could improve the level of detail of the proposed guidelines and allow for a more flexible grading system. It may be difficult for model developers and policy makers to objectively assess the nature of the model, so it may be helpful to have an external party assess the ethical nature of their models.

VIII. Limitations

The obvious limitations of this review include the lack of specificity when detailing the original datasets. It is difficult to outline the representativeness and diversity of the dataset as well as the accuracy of the models since th specifics of the training data are undisclosed.

REFERENCES

[1] M. A. Mazurowski, M. Buda, A. Saha, and M. R. Bashir, "Deep learning in radiology: An overview of the concepts and a survey of the state of the art with focus on mri," *Journal of magnetic resonance imaging*, vol. 49, no. 4, pp. 939–954, 2019.

[2] A. Hernandez-Trinidad, B. O. Murillo-Ortiz, R. Guzman-Cabrera, and T. Cordova-Fraga, "Applications of artificial intelligence in the classification of magnetic resonance images: Advances and perspectives," *New Advances in Magnetic Resonance Imaging*, 2023.

[3] R. Bouderhem, "Shaping the future of ai in healthcare through ethics and governance," *Humanities and social sciences communications*, vol. 11, no. 1, pp. 1–12, 2024.

[4] R. Jain, N. Jain, A. Aggarwal, and D. J. Hemanth, "Convolutional neural network based alzheimer's disease classification from magnetic resonance brain images," *Cognitive Systems Research*, vol. 57, pp. 147–159, 2019.

[5] H. Aboutorab, O. K. Hussain, M. Saberi, F. K. Hussain, and E. Chang, "A survey on the suitability of risk identification techniques in the current networked environment," *Journal of Network and Computer Applications*, vol. 178, p. 102984, 2021.

[6] X. Zhao, C. K. E. Ang, U. R. Acharya, and K. H. Cheong, "Application of artificial intelligence techniques for the detection of alzheimer's disease using structural mri images," *Biocybernetics and Biomedical Engineering*, vol. 41, no. 2, pp. 456–473, 2021.

[7] S.-h. Zhen, M. Cheng, Y.-b. Tao, Y.-f. Wang, S. Juengpanich, Z.-y. Jiang, Y.-k. Jiang, Y.-y. Yan, W. Lu, J.-m. Lue, *et al.*, "Deep learning for accurate diagnosis of liver tumor based on magnetic resonance imaging and clinical data," *Frontiers in oncology*, vol. 10, p. 680, 2020.

[8] L. Lu, S. Shi, B. Liu, and C. Liu, "Analysis of factors influencing the organizational capacity of institutional review boards in china: a crisp-set qualitative comparative analysis based on 107 cases," *BMC Medical Ethics*, vol. 24, no. 1, p. 74, 2023.

[9] D. Sadeghi, A. Shoeibi, N. Ghassemi, P. Moridian, A. Khadem, R. Alizadehsani, M. Teshnehlab, J. M. Gorriz, F. Khozeimeh, Y.-D. Zhang, *et al.*, "An overview of artificial intelligence techniques for diagnosis of schizophrenia based on magnetic resonance imaging modalities: Methods, challenges, and future works," *Computers in Biology and Medicine*, vol. 146, p. 105554, 2022.

[10] T. R. Insel, "Rethinking schizophrenia," *Nature*, vol. 468, no. 7321, pp. 187–193, 2010.

[11] D. Bzdok and A. Meyer-Lindenberg, "Machine learning for precision psychiatry: opportunities and challenges," *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, vol. 3, no. 3, pp. 223–230, 2018.

[12] A. Zubair Rahman, M. Gupta, S. Aarathi, T. Mahesh, V. Vinoth Kumar, S. Yogesh Kumaran, and S. Guluwadi, "Advanced ai-driven approach for enhanced brain tumor detection from mri images utilizing efficientnetb2 with equalization and homomorphic filtering," *BMC Medical Informatics and Decision Making*, vol. 24, no. 1, p. 113, 2024.

[13] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians*, vol. 68, no. 6, pp. 394–424, 2018.

[14] L. Tabár, B. Vitak, T. H.-H. Chen, A. M.-F. Yen, A. Cohen, T. Tot, S. Y.-H. Chiu, S. L.-S. Chen, J. C.-Y. Fann, J. Rosell, *et al.*, "Swedish two-county trial: impact of mammographic screening on breast cancer mortality during 3 decades," *Radiology*, vol. 260, no. 3, pp. 658–663, 2011.

[15] J. G. Elmore, S. L. Jackson, L. Abraham, D. L. Miglioretti, P. A. Carney, B. M. Geller, B. C. Yankaskas, K. Kerlikowske, T. Onega, R. D. Rosenberg, *et al.*, "Variability in interpretive performance at screening mammography and radiologists' characteristics associated with accuracy," *Radiology*, vol. 253, no. 3, pp. 641–651, 2009.

[16] S. M. McKinney, M. Sieniek, V. Godbole, J. Godwin, N. Antropova, H. Ashrafian, T. Back, M. Chesus, G. S. Corrado, A. Darzi, *et al.*, "International evaluation of an ai system for breast cancer screening," *Nature*, vol. 577, no. 7788, pp. 89–94, 2020.

[17] N. Stogiannos, R. Malik, A. Kumar, A. Barnes, M. Pogose, H. Harvey, M. F. McEntee, and C. Malamateniou, "Black box no more: a scoping review of ai governance frameworks to guide procurement and adoption of ai in medical imaging and radiotherapy in the uk," *The British Journal of Radiology*, vol. 96, no. 1152, p. 20221157, 2023.

[18] M. Goisauf and M. Cano Abadía, "Ethics of ai in radiology: a review of ethical and societal implications," *Frontiers in big Data*, vol. 5, p. 850383, 2022.

[19] E. Shimron and O. Perlman, "Ai in mri: computational frameworks for a faster, optimized, and automated imaging workflow," 2023.

Appendix

TABLE I: Section one of the guideline outlining patient privacy and security.

Section 1: Patient Privacy and Security	1	2	3	4
The patient data being used to train the model is encrypted and anonymized before training.	There are no policies or security measures in place to protect sensitive information, creating a high risk of exposure and misuse.	Patient data is anonymized, but encryption standards are weak or inconsistently applied. Some measures exist to protect privacy, but gaps remain in ensuring robust security and compliance with data protection regulations.	Patient data is both encrypted and anonymized before training, following standard security protocols. Data protection measures meet most regulatory requirements, but periodic audits and updates are needed.	Patient data undergoes advanced encryption and thorough anonymization before training, meeting industry standards.
There are mechanisms in place to prevent unauthorized access, data breaches, and re-identification risks.	No security measures in place to prevent unauthorized access of data, with a high-risk of data breaches and/or re-identification risks.	Basic security measures exist but are insufficient or inconsistently applied.	Strong security measures, but some challenges remain.	Comprehensive security framework in place to prevent prevent unauthorized access of data, with a low-risk of data breaches and/or re-identification risks.
The AI system provides patients with clear, accessible options to give informed consent for data collection, processing, and sharing.	No informed consent process. Data is collected and shared without patient knowledge. Consent notices do not exist, or they lack any transparency on data collection, frequency, and third-party access.	Some effort toward consent, but the information is vague, difficult to access, or missing critical details, such as what data is collected, how often, and which third parties are involved. Patients may not fully understand how their data is used.	Clear consent process with minor areas of improvement. Most key details on data collection, frequency, and third-party access are provided, but there may be minor gaps or lack of clarity in certain areas.	Fully transparent, easy-to-understand consent process where patients have full control over their data. Consent notices clearly specify what data is collected, how frequently, and which third parties have access, ensuring true informed consent.

Patients can opt-out of the collection of their personal data or the application of the AI model and were explicitly informed of its use in diagnosis, training, and development.	No opt-out option. AI use is mandatory, and patients have no ability to refuse data collection. There is no transparency about AI involvement in diagnosis, training, or development.	Opt-out exists but is difficult to access or poorly explained. Information about opting out is either too vague or too complex, such as being written in overly broad or highly detailed language. Patients may not be aware that opting out is possible, or they may struggle to navigate the process due to a lack of clarity.	Clear opt-out options with minor challenges. Patients are explicitly informed of AI usage and have a way to opt-out but the process may have small barriers, such as requiring multiple steps, unclear instructions, or limited accessibility. Some patients may still find it challenging to opt-out.	Clear opt-out system with full transparency and patient control. Patients are explicitly informed about the use of AI in diagnosis, training, and development. The opt-out process is simple, accessible, and user-friendly, allowing patients to easily withdraw consent at any time.
The AI system ensure compliance with legal and ethical data protection standards (e.g., GDPR, HIPAA) to safeguard patient privacy.	No efforts are made to meet legal and ethical data protection standards. There is no mention of compliance mechanisms. The system poses a high risk of privacy breaches.	Some attempts to address compliance are evident, but significant gaps remain. Key safeguards are missing or insufficient, leaving patient data vulnerable.	The system adheres to most legal and ethical data protection requirements, with only minor weaknesses that need improvement. Privacy risks are minimal but not fully eliminated.	The system adhered to all legal and ethical data protection requirements. Strong safeguards are in place, with compliance being actively monitored and continuously improved to maintain data security and patient privacy.
Total /20				

TABLE II: Section two of the guideline outlining ethical data collection.

Section 2: Data Collection	1	2	3	4
Data collection methods designed to minimize bias and ensure diversity in the dataset, particularly to prevent discrimination against vulnerable populations.	Data collection lacks diversity and includes significant biases, leading to potential discrimination. There are no active efforts to ensure representation, and vulnerable populations are underrepresented or excluded.	Some measures are in place to promote diversity, but gaps remain in ensuring fair representation. The dataset includes different demographic groups, but bias assessments are infrequent, and vulnerable populations may still be underrepresented.	Data collection is designed to minimize bias and includes a representative sample of diverse populations. Regular audits and fairness assessments are conducted, and adjustments are made to address any identified biases.	Data collection follows best practices to ensure diversity and prevent discrimination against vulnerable populations. Proactive strategies, such as targeted data collection, fairness-aware algorithms, and continuous bias monitoring, are implemented to maintain equity and inclusivity.
Patients and healthcare providers are adequately informed about how AI models are trained, including any potential data gaps or biases.	No measures are in place to ensure that patients and healthcare providers understand how AI models are trained. There is minimal or non-existent awareness of data gaps or biases.	Limited efforts are made to provide transparency. Some basic information is available, but it is insufficient for meaningful understanding of data gaps or biases.	Patients and healthcare providers are provided with clear resources to understand AI model training, including potential biases and data gaps. Transparency is achieved through adequate but standard materials, but the depth of information may be limited.	Patients and healthcare providers are provided with well-structured and accessible resources that clearly outline the AI model's training process, data sources, potential biases, and data gaps. Transparency is robust, with materials that provide potential data gaps and biases associated with the AI model.

The AI model incorporates MRI data from patients with rarer cases that are underrepresented to ensure equitable and accurate diagnostic performance.	The AI model does not incorporate MRI data from underrepresented rare cases, leading to biased and inaccurate diagnostics.	Limited efforts are made to include MRI data from rare cases, but representation remains insufficient, impacting diagnostic equity in demographic, geographic, or age diversity.	The AI model actively integrates MRI data from rare cases to improve diagnostic accuracy and equity, addressing some data gaps.	The AI model systematically ensures broad representation of rare cases across demographics, geography, and age groups. Clear documentation highlights data sources, biases, and measures taken to enhance equity and accuracy.
The AI model plans to undergo continuous monitoring to identify and mitigate biases in real-world MRI diagnostics, with mechanisms in place to update training data.	The AI model provides little to no measures for continuous monitoring or bias mitigation. The AI model remains unchanged, lacking a mechanism to address biases in real-world MRI diagnostics.	Limited monitoring exists, but it is irregular or lacks depth. Bias mitigation efforts are minimal, and updates to training data are infrequent.	The AI model undergoes structured monitoring to identify and reduce biases. Mechanisms exist to update training data periodically, improving fairness and accuracy.	A comprehensive and proactive monitoring system is in place, continuously tracking biases in real-world MRI diagnostics. Transparent reporting, bias mitigation strategies, and regular training data updates ensure fairness and accuracy.
Total /16				

TABLE III: Section three of the guideline outlining interpretation and responsible usage.

Section 3: Interpretation and Responsibility	**1**	**2**	**3**	**4**
The model supplements the judgement of the physician's decision and strengthens their diagnosis.	The model provides little to no meaningful assistance in clinical decision-making. It lacks interpretability, produces inconsistent results, and may even introduce errors or biases that hinder accurate diagnoses.	The model offers some assistance but is not fully reliable. It can highlight potential findings, but its outputs require significant physician oversight due to occasional inaccuracies or lack of explainability.	The model consistently enhances physician judgment by providing reliable diagnostic insights. It improves efficiency and accuracy while maintaining transparency, but physicians must still verify results before making final decisions.	The model serves as a highly effective decision-support tool, strengthening physician diagnoses with high accuracy and clear explainability. It integrates seamlessly into clinical workflows, continuously learns from real-world data, and provides interpretable recommendations that align with best medical practices.
Physicians are informed and trained how to interpret and use the results.	No formal training provided for physicians on AI-assisted MRI interpretation.	Some basic training is available, but lacks depth and does not cover AI limitations or biases.	Regular training sessions are provided, covering AI interpretation and some ethical concerns.	Comprehensive training ensures physicians understand AI outputs, biases, and ethical concerns with ongoing learning opportunities.
The model's outputs are explainable and transparent, allowing physicians to verify AI-generated MRI assessments before making a clinical decision.	AI model provides no explanations for its assessments; physicians cannot verify results.	AI provides minimal explanations, but they are unclear or overly complex.	AI outputs include confidence scores and basic explanations, allowing some verification.	AI model outputs are fully transparent, with clear confidence scores, reasoning, and decision-support tools for physicians.

The model serves as a supplementary tool, enabling physicians to enhance their MRI interpretation while retaining full responsibility for the final diagnosis.	AI operates autonomously with little to no physician oversight, posing risks of over-reliance.	AI is supplementary but not clearly defined as a decision-support tool; physicians may over-rely on it.	AI serves as a clear supplementary tool, with physicians retaining final decision-making authority.	Ai is fully integrated into workflows as a decision-support tool, ensuring that physicians enhance their diagnoses while maintaining full responsibility.
The model provides confidence scores, allowing physicians to assess the reliability of AI-generated findings in the context of other clinical data.	The model does not provide confidence scores, making it difficult for physicians to assess the reliability of AI-generated findings	The model provides basic confidence scores, but they are not well-calibrated or interpretable, requiring significant physician intuition to use effectively.	The model provides confidence scores that are generally well-calibrated and useful, helping physicians weigh AI findings in context, though occasional inconsistencies exist.	The model provides highly accurate, well-calibrated confidence scores that seamlessly integrate with clinical data, allowing physicians to assess reliability with high confidence.
The model's limitations are clearly defined, ensuring that physicians remain aware of potential biases and do not over-rely on AI outputs.	The model does not communicate its limitations, leading to potential over-reliance or misinterpretation by physicians.	The model provides some general disclaimers about limitations, but they lack specificity or transparency, requiring physicians to infer potential biases.	The model clearly defines its key limitations, including known biases, and presents this information in an accessible way, though some edge cases may still be unclear.	The model thoroughly documents its limitations, biases, and potential failure cases, ensuring physicians have a comprehensive understanding and do not over-rely on AI outputs.
The model's insights are documented within clinical workflows, allowing physicians to track its influence on their final diagnostic decisions.	The model's insights are not documented within clinical workflows, making it difficult to assess its influence on decision-making.	The model's insights are recorded sporadically but lack structured documentation, limiting the ability to track AI influence effectively.	The model's insights are consistently documented in clinical workflows, allowing physicians to review and assess AI's role in decision-making, though some integration gaps remain.	The model's insights are fully integrated into clinical workflows, with structured documentation that allows physicians to track AI influence seamlessly, supporting transparency and accountability.
Total /28				