

Navigating AI Integration in Breast Cancer Diagnosis: A Mixed-Method Study on Perceptions, Trust, and Adoption

Elia Rezaeian - Ph.D. Candidate
Onur Asan - Advisor

*Department of Systems and Enterprises
Stevens Institute of Technology
Study Date : Sep 2022 - Feb 2025*

1. Study Objectives

In recent years, Artificial Intelligence (AI) has emerged as a transformative force across various domains, with healthcare becoming a particularly promising area of focus. Given the vast amounts of complex data generated in clinical settings, AI offers valuable opportunities to improve processes and enhance patient care. AI systems have shown significant potential in improving diagnostic accuracy [15, 16, 13, 17], supporting treatment planning [14, 5], streamlining workflows, and reducing human error [25, 8, 21].

One area where AI is making noticeable progress is in breast cancer detection. In this context, AI-powered Clinical Decision Support Systems (CDSSs) assist radiologists and oncologists by analyzing medical images, identifying abnormalities, and providing diagnostic suggestions [15]. However, despite these technological advancements, the value of AI-based CDSSs ultimately depends on clinician acceptance and adoption [3, 18, 9, 6]. Trust plays a critical role in this process, as it directly shapes clinicians' willingness to rely on AI systems during diagnosis and treatment [5, 10].

Among the factors influencing trust, explainability has emerged as a particularly important element [22]. AI systems are often seen as "black boxes" that deliver outcomes without showing how or why a decision was made. This lack of transparency can create hesitation or skepticism, especially in sensitive and high-stakes fields like healthcare. When AI systems provide meaningful explanations for their recommendations, clinicians are more likely to trust and integrate them into their workflows. However, the specific ways in which explainability impacts clinician behavior are still not well understood.

Explainability in clinical settings needs to go beyond abstract technical descriptions—it should be intuitive, relevant, and aligned with the informational needs of healthcare professionals. In our study, we introduce multiple levels of explanation for AI-generated breast cancer classifications, aiming to make the system's outputs increasingly interpretable. While several techniques exist to improve transparency, their usefulness depends on how, when, and why the information is delivered. Too much detail, or poorly timed explanations, can overwhelm users and negatively affect their decision-making. Our goal is to identify the level and form of explanation that best supports clinicians without distracting or confusing them.

To explore these dynamics, we designed a mixed-method study that examines how varying levels of AI explainability affect clinicians' interaction with an AI-based CDSS for breast cancer diagnosis. The quantitative component involves a human-subject experiment using a custom-built web application that provides diagnostic recommendations under multiple explainability conditions. We evaluate the effect of these varying levels on clinicians' diagnostic accuracy, self-reported and behavioral trust, and cognitive load. To complement and deepen our understanding of the quantitative findings, we conducted a series of semi-structured interviews with participants after the experimental phase. This qualitative component aims to explore clinicians' experiences, expectations, and concerns regarding the use of AI systems in their clinical workflow.

By combining experimental results with qualitative insights, our study offers a comprehensive view of how AI explainability impacts clinician behavior. It highlights the importance of designing explainability features that are not only technically sound but also meaningful and usable in real-world clinical settings. This work contributes to the development of AI systems that are more transparent, trustworthy, and aligned with the practical needs and values of healthcare professionals.

2. Methodology

2.1. Participants

A total of 28 participants were recruited between January and August 2024 through medical associations, social platforms, and professional networks. All participants were U.S.-based, fluent in English, and over the age of 18. They were compensated \$80 for their participation. The sample comprised primarily radiologists ($\approx 60\%$), with the remaining participants identified as oncologists ($\approx 18\%$) or other healthcare professionals involved in breast cancer care ($\approx 22\%$).

From this group, 11 participants voluntarily participated in the qualitative interview process, providing additional insights into their experiences and perceptions of AI-based CDSS. These participants received \$50 for compensation.

2.2. Experimental Design

The target audience for this experiment is clinicians, including oncologists and radiologists, who traditionally process breast tissue scans to make cancer diagnosis decisions. We follow an interrupted time series (ITS) experiment

design [11] where all participants use all versions of CDSSs (treated as multiple interventions) in a certain sequence. We choose the ITS design to be able to examine how clinicians' trust and diagnostic performance change over time as they interact with AI systems that provide increasing levels of explanation. This design is appropriate for our study as it allows observing both immediate and cumulative effects of each explanation type, while also comparing changes to the baseline condition. By using a fixed sequence of interventions with the same participants, we are able to better understand the specific impact of each explanation level on clinicians' behavior.

Initially, the CDSS presents diagnostic suggestions based on the analysis of breast cancer tissue images with a machine learning model without any accompanying explanations. These suggestions categorize the findings as "healthy", "benign tumor" or "malignant tumor". Subsequently, we introduce variations in the diagnostic process to understand how the level of explanation influences clinicians' trust and decision-making within the CDSS. All the clinical decisions recorded during the experiment are compared with the traditional process where the participants make diagnosis decisions in the absence of any decision support (treated as baseline). The experiments were conducted entirely online, without the need for direct oversight by the research team.

The ITS experiment process follows the interventions below in the given order. We designed the experiment such that the participants were exposed to decision support with an increasing level of explanations. *Table 1* shows a detailed overview of the experimental conditions and their differences. Each condition involves diagnosing a series of ten breast tissue images, followed by post-experiment survey.

- **Baseline (Stand-alone):** Clinicians are not presented with any diagnostic suggestions and are asked to make diagnosis decisions on the breast cancer tissue images based on their own judgment.
- **Intervention I (No Explanation):** Clinicians are presented with diagnostic suggestions (healthy, benign tumor, malignant tumor) without accompanying explanations.
- **Intervention II (AI Confidence):** Building upon the first, this intervention introduces probability estimates for each diagnostic class (healthy, benign tumor, malignant tumor).
- **Intervention III (Tumor Localization):** In addition to the information provided in the second intervention, the CDSS advances by estimating the precise location of the tumor within the breast tissue images. No location information was shown when the prediction was "healthy".
- **Intervention IV (Enhanced Tumor Localization with Confidence Levels):** Compared to the third intervention, clinicians in this scenario receive tumor location information including both low and high confidence

estimates upon tumor detection. No location information was shown when the prediction was "healthy".

We provided the descriptions of each type of CDSS in the tutorial and provided reminders in the experiment interface as shown in *Table 1* (e.g, the interface for the 3rd intervention shows "The image below shows the potential cancerous area").

The AI system behind the CDSS builds upon our previous work [19]. This system integrates a U-Net architecture for image segmentation with a Convolutional Neural Network (CNN) for cancer prediction [20]. The U-Net architecture was designed to depict the boundaries of cancerous areas in breast tissue, helping to localize potential tumors. The CNN then classifies the images into three categories: healthy, benign, and malignant. The system was trained on a publicly available breast cancer dataset [1] consisting of 780 ultrasound images, achieving an 81% diagnostic accuracy. This accuracy result suggests that the tool may be useful for decision support if users exercise some healthy skepticism when following the recommendations provided by this tool.

2.3. Experiment Procedure

2.3.1. Web Application

The experiment platform we developed based on the Python-based Dash framework allows collecting quantitative data representing the clinical users' interaction and collaboration with an AI recommender system, coupled with both pre-and post-experiment surveys, all integrated into a single, user-friendly web application. This platform enables us to gain a comprehensive understanding of how medical professionals engage with AI-based CDSS in a virtual environment. The experiment process, starting with the submission of consent forms and going down to data collection, takes place online within this integrated web application.

2.3.2. Recruitment

We recruited clinicians aged 18 and older, including oncologists and radiologists experienced in interpreting breast tissue scans for cancer diagnosis. All participants were fluent in English. Recruitment efforts focused on medical associations, social media platforms, and professional networks, primarily within the USA.

2.3.3. Tutorial Video and Preparation

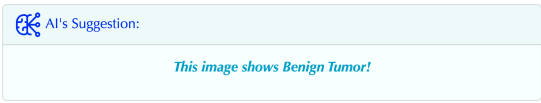
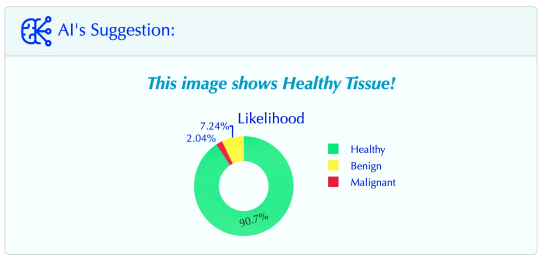
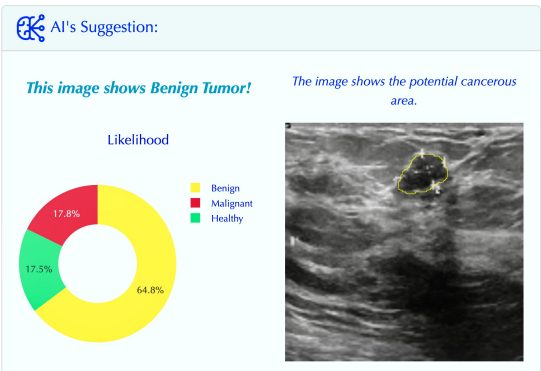
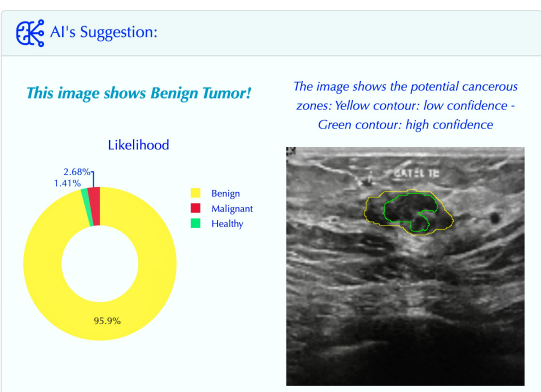
Interested participants contacted us via email to express their interest in the study. We sent them both a comprehensive video tutorial and the access link to the experiment via email. This tutorial explains the goals of the study, how to use the web application, and the tasks they need to complete. We emailed the link exclusively to those who showed interest, ensuring the authenticity of participants.

2.3.4. Consent Form and Pre-Experiment Survey

Participants signed a consent form electronically through the web app, followed by a pre-experiment survey that collected their demographic information and baseline data

Table 1

The experimental conditions involving AI suggestions and their descriptions

AI Suggestion	Description
 <p>AI's Suggestion:</p> <p><i>This image shows Benign Tumor!</i></p>	<p>No explanation (1st) intervention</p> <p>The AI system provides diagnosis suggestions without any accompanying explanations, categorizing them into three distinct categories: healthy, benign tumor, and malignant tumor.</p>
 <p>AI's Suggestion:</p> <p><i>This image shows Healthy Tissue!</i></p> <p>Likelihood</p> <p>90.3% Healthy, 7.24% Benign, 2.04% Malignant</p>	<p>AI Confidence (2nd) intervention</p> <p>The AI system presents diagnosis suggestions in three distinct categories—healthy, benign tumor, and malignant tumor—accompanied by corresponding confidence scores for each category.</p>
 <p>AI's Suggestion:</p> <p><i>This image shows Benign Tumor!</i></p> <p>Likelihood</p> <p>64.8% Benign, 17.3% Healthy, 17.8% Malignant</p> <p>The image shows the potential cancerous area.</p>	<p>Tumor localization (3rd) intervention</p> <p>The AI system provides diagnosis suggestions categorized into three distinct categories—healthy, benign tumor, and malignant tumor. Additionally, it offers corresponding probability scores for each category and estimates the location of the tumor, if detected.</p>
 <p>AI's Suggestion:</p> <p><i>This image shows Benign Tumor!</i></p> <p>Likelihood</p> <p>95.9% Benign, 2.68% Malignant, 1.41% Healthy</p> <p>The image shows the potential cancerous zones: Yellow contour: low confidence - Green contour: high confidence</p>	<p>Enhanced tumor localization (4th) intervention</p> <p>The AI system presents diagnosis suggestions in three categories—healthy, benign tumor, and malignant tumor—along with associated confidence scores. Furthermore, it estimates the location of the tumor, providing both low confidence and high confidence if a tumor is detected.</p>

on their experience and comfort level with AI and clinical decision support systems in general.

2.3.5. Experimental Sessions

After the pre-experiment survey, participants engaged in a series of experimental sessions discussed in *Section 2.2*. While each experiment condition differs in terms of the amount of information presented to participants, in all these scenarios, participants were presented with a series of ten mammogram images one at a time. Their primary task in the baseline condition was to examine each image and provide one of three diagnoses: Healthy, Benign, or Malignant.

This baseline condition aimed to evaluate their diagnostic performance without the assistance of AI.

The conditions that involve CDSS presented the AI suggestions and their corresponding descriptions. In those conditions, participants were also instructed to assess their agreement with and trust the AI's suggestions for each image they viewed. They rated their agreement and trust levels on a Likert scale using sliders (see ??). If their agreement level fell at or below neutral, they were asked to make their own decision, as shown in *Figure 1*. This approach allowed us to capture their true decision for images where they did not agree with the AI. We used the agreement scale to

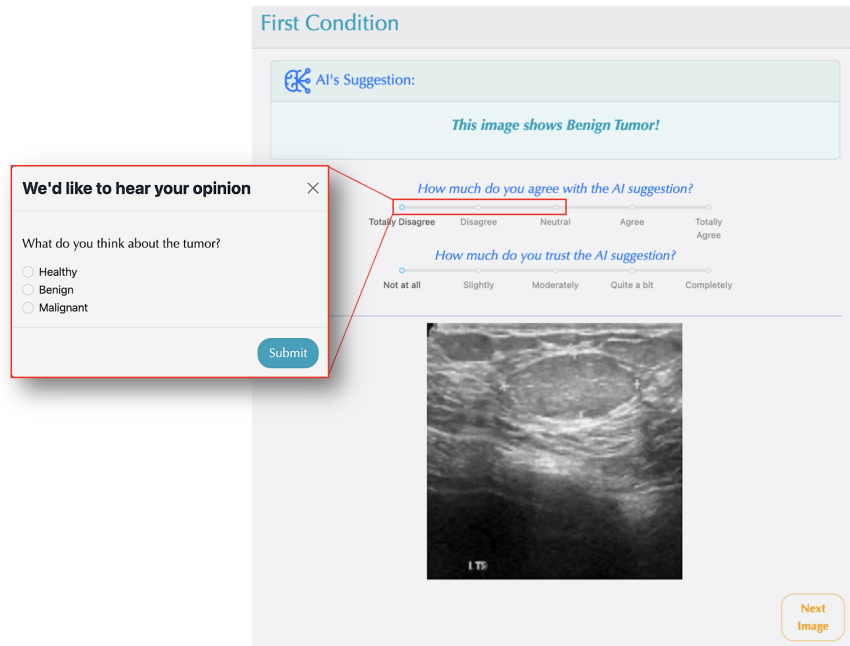


Figure 1: Experiment web interface with an interactive assessment pop-up window, where participants’ opinions are recorded when their diagnostic judgments differ from those of the AI.

capture the degree to which participants aligned with the AI’s suggestions under different levels of explainability. Our aim with using this scale was to observe how their agreement levels varied across conditions in finer granularity than the case where participants’ responses were limited to a binary decision.

2.3.6. Post-Experiment Survey

Upon completion of all experimental sessions, participants proceeded to the post-experiment survey. The questions in this survey are designed to gather specific insights into their experience and assess the differences in each condition in terms of their understanding, trust, and workload. Participants were asked to respond to questions using a 5-point Likert scale

2.3.7. Interview

For the qualitative component, interview scheduling began after participants confirmed their interest and provided informed consent. The interviews followed a pre-determined guide designed to explore clinicians’ perceptions, experiences, and decision-making processes when using AI-based CDSS.

Between June 2024 and September 2024, We conducted the interviews via Zoom and recorded the audio to ensure accuracy in transcription and analysis. The interviews ranged in duration from 12 to 35 minutes, with an average length of 19 minutes.

We transcribed the interview recordings using Zoom’s automated transcription feature, and our research reviewed the accuracy by cross-checking with audio recordings. Thematic analysis followed an inductive approach [4], with themes and sub-themes systematically organized in Excel.

Two researchers, OR and EG, independently coded the transcripts, while OA facilitated weekly discussions to refine coding, resolve discrepancies, and identify emerging patterns. The codebook and operational definitions were iteratively refined until a consensus was reached and applied across all transcripts.

2.4. Measurements

2.4.1. Self-Reported Trust Measures

These metrics capture participants’ perceptions and beliefs about the AI system through surveys. These measures are often used to capture subjective aspects by asking participants to rate their experiences on Likert scales.

- **Trust:** We assessed participants’ trust in AI systems using a 5-point Likert scale from 0 (No trust at all) to 5 (Complete trust). During the experiment, participants rated their trust in AI suggestions for each image on a similar 5-point scale, responding to the question, “How much do you trust the AI system?”, providing a finer measurement.
- **Familiarity:** In the demographic survey, we assessed participants’ familiarity with AI systems using a 5-point Likert scale. Participants were asked, “How familiar are you with Artificial Intelligence systems?” with response options ranging from “Not at all” to “Completely familiar”.
- **Understandability:** After each intervention in the post-experiment survey, participants were asked to evaluate the understandability of the AI suggestions. The statement, “I found the AI’s suggested breast cancer classification to be intuitively understandable”,

was rated on a 5-point Likert scale, ranging from “Strongly disagree” to “Strongly agree”.

- **Perceived Accuracy:** After each intervention, in the post-experiment survey, we assessed participants’ perceptions of the AI’s accuracy by asking them to respond to the statement, “I believe that the AI answers were accurate.” Participants rated the perceived accuracy of the AI suggestions on a scale ranging from “Not accurate” to “Completely accurate.”

2.4.2. Behavioral Trust Measures

These metrics assess observable actions that indicate a participant’s trust in the AI system.

- **Performance:** Participants’ performance was assessed based on their decisions about each image during the experiment. To determine their performance, we considered two factors. First, if the participant’s agreement level with the AI suggestion was higher than neutral, we considered their decision to align with the AI’s suggestion. For cases where the agreement level was neutral or below, we asked participants to independently indicate their decision by choosing among “Healthy”, “Benign”, or “Malignant”, as illustrated in *Figure 1*. We calculated their performance for each intervention by comparing their decisions with the ground truth over ten images as a percentage, reflecting the accuracy of their decisions relative to the correct diagnosis.
- **Agreement:** During the experiment, participants rated their agreement with AI suggestions for each image on a 5-point scale, responding to the question, “How much do you agree with AI suggestion?”
- **Decision Time:** We tracked the duration of participants’ decision-making process for each image throughout the experiment.

2.4.3. AI-related Measures

- **AI Role:** To understand participants’ perspectives on AI’s potential in healthcare, we asked them to indicate their level of agreement with the statement, “Artificial Intelligence will play an important role in the future of medicine.” Responses were captured on a 5-point Likert scale, from Totally Disagree to Totally Agree.
- **AI Usefulness:** This measure examined how relevant participants perceived AI to be for their specific roles. Participants rated their agreement with the statement, “AI would be useful in my job.” providing insight into AI’s perceived practical benefits in their professional contexts.
- **Complexity Perception:** To assess perceived barriers to AI adoption, participants responded to the statement, “There are too many complexities and barriers in medicine for AI to help in clinical settings.” This measure, also rated on a 5-point Likert scale, helped

identify potential challenges related to AI integration in clinical practice.

2.4.4. Cognitive load Measures

- **Mental Demand:** After each intervention, we assessed mental demand by asking participants, “How mentally demanding was the task?” Responses were given on a 5-point Likert scale, from Very Low to Very High. This measure helped us understand the cognitive load experienced by participants at each stage.
- **Stress:** After each intervention, we measured Stress by asking participants, “How stressed were you?” Responses were collected on a 5-point Likert scale, ranging from Very Low to Very High. This measure provided insights into participants’ stress levels throughout the experiment.

3. Results

3.1. The effect of Explainability Level

Our analysis using a mixed-effects models (*Table 2 and Table 3*) provides insight into how different levels of AI explainability influence clinicians’ self-reported and behavioral trust and cognitive load outcomes.

Trust

No significant differences were observed across interventions; however, the AI confidence (2nd) and enhanced tumor localization (4th) interventions showed slight decreases in trust compared to the 1st, while the 3rd intervention showed a small positive effect. When using baseline trust as a reference, trust increased significantly after interacting with any AI system, with the 3rd intervention showing the highest boost.

Understandability

The enhanced tumor localization (4th) intervention significantly reduced understandability ($p = 0.013$), suggesting that too much explanation may decrease clarity.

Perceived Accuracy

Participants rated perceived accuracy significantly lower in the 2nd and 4th interventions compared to the 1st, indicating that added confidence scores and extra detail might reduce perceived system reliability.

Performance

As shown in *Figure 2*, performance was generally better with AI than without, with the 1st and 2nd interventions yielding the highest median accuracy. However, the 3rd and 4th interventions resulted in small but significant decreases in performance. Diagnostic accuracy was highest for healthy cases and lower for benign and malignant.

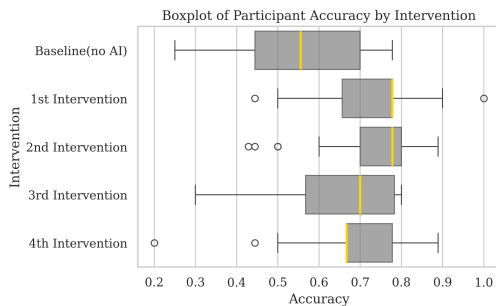
Agreement

Agreement with AI suggestions significantly decreased in the 4th intervention compared to the 1st. A strong correlation was found between trust and agreement ($r_s = 0.85$,

Table 2

Mixed linear model results: Effect of AI explainability on self-reported and behavioral measures

Variable	Self-reported						Behavioral					
	Trust	<i>p</i>	Understandability	<i>p</i>	Perceived Accuracy	<i>p</i>	Agreement	<i>p</i>	Diagnosis Duration	<i>p</i>	Performance	<i>p</i>
Intercept	3.031	0.000	3.179	0.000	3.036	0.000	3.224	0.000	0.260	0.000	0.736	0.000
2nd Intervention	-0.049	0.543	-0.286	0.126	-0.321	0.011	-0.053	0.482	0.054	0.338	0.005	0.835
3rd Intervention	0.059	0.465	-0.107	0.566	-0.036	0.778	0.035	0.640	0.009	0.869	-0.073	0.005
4th Intervention	-0.145	0.073	-0.464	0.013	-0.250	0.048	-0.149	0.048	0.144	0.011	-0.068	0.009
Group Var	0.294	-	0.201	-	0.233	-	0.240	-	0.094	-	0.008	-

**Figure 2:** Participant performance over different interventions**Table 3**

Mixed linear model results: Effect of AI explainability on Cognitive Load:

Variable	Cognitive Load		Mental Demand		Stress	
	β	<i>p</i>	β	<i>p</i>	β	<i>p</i>
Intercept	1.714	0.000	1.071	0.000		
2nd Intervention	-0.036	0.822	0.214	0.185		
3rd Intervention	-0.179	0.262	0.393	0.015		
4th Intervention	-0.107	0.501	0.214	0.185		
Group Var	-0.881		0.785			

$p < 0.05$), showing that participants trusted AI more when their own decisions aligned with its output.

Diagnosis Duration

Decision time was significantly longer in the 4th intervention, indicating increased cognitive load with high levels of explanation.

Mental Demand and Stress

As shown in Table 3, none of the interventions significantly changed mental demand. However, stress significantly increased in the 3rd intervention ($p = 0.015$), possibly due to the introduction of tumor localization and probabilities.

3.2. Impact of AI Confidence Score on User Behavior

Using a separate mixed-effects model (Table 4), we found that low AI confidence significantly reduced trust and agreement, and increased diagnosis duration. High confidence, however, did not significantly improve trust or agreement and even slightly reduced performance, suggesting a risk of over-reliance on AI when confidence is high.

3.3. Thematic Analysis

Thematic analysis of the semi-structured interviews revealed several key themes regarding clinicians' perceptions, expectations, and concerns surrounding AI-based Clinical Decision Support Systems (CDSSs). A total of eight major themes were identified.

Clinicians emphasized the importance of “*Understanding and Perception of AI*”, expressing uncertainties around how AI differs from existing technologies and its role within clinical workflows. While AI was generally viewed as helpful, it was seen as a tool that should support, rather than replace, clinical judgment.

The theme “*AI as a Decision Support, Not a Replacement*” highlighted clinicians' perspectives on the need for AI to complement human expertise. Participants acknowledged AI's potential to enhance decision-making while warning against overdependence that could undermine clinical practice.

In “*AI Impact on the Radiology Workforce*”, concerns were raised about potential job displacement and the erosion of radiologists' diagnostic skills, emphasizing the need to consider AI's long-term implications for professional roles.

The theme “*AI Helpfulness*” reflected positive views on AI's utility in improving diagnostic confidence and supporting less experienced clinicians. Participants appreciated AI's ability to offer guidance and reassurance during decision-making.

“*AI in Clinical Practice*” captured the practical applications of AI, including its role in early cancer detection, screening workflows, clinical prioritization, and disease monitoring. Participants saw AI as a tool to enhance clinical efficiency and accuracy.

In “*Usability and Accessibility*”, clinicians emphasized the importance of integrating AI into existing systems, designing user-friendly interfaces, and addressing resource constraints to support adoption in real-world clinical environments.

The theme “*Impact on Workload*” revealed both opportunities and concerns. While some participants reported gains in efficiency, others noted increased cognitive load or time penalties depending on the system's complexity.

Finally, “*Factors Influencing Trust in AI*” emerged as a critical theme. Clinicians emphasized the importance of clinical validation, transparency, peer endorsements, and explainability in fostering trust. Trust was seen as dynamic—something that builds over time through experience, usability, and clarity in AI outputs.

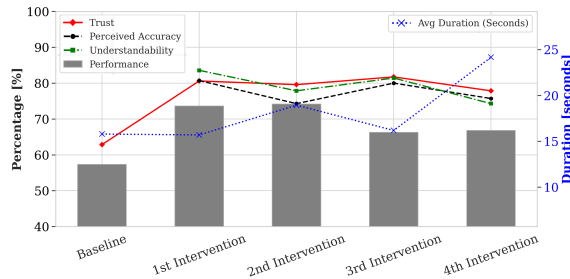
4. Discussion

This mixed-methods study investigated clinicians' experiences with AI and explainable AI (XAI) in diagnostic decision-making. By combining survey data, mixed-effects modeling, and interview analysis, we examined how clinicians develop trust, interpret AI explanations, and integrate

Table 4

Mixed linear model results: Effect of AI confidence score on trust, agreement, performance, and diagnosis duration

AI Confidence Score	Trust		Agreement		Performance		Diagnosis Duration	
Variable	β	p	β	p	β	p	β	p
Intercept	3.031	0.000	3.224	0.000	0.738	0.000	0.260	0.000
Low Confidence	-0.163	0.023	-0.186	0.005	-0.087	0.675	0.131	0.009
High Confidence	0.103	0.169	0.108	0.121	-0.015	0.020	-0.012	0.818
Group Var	0.292		0.238		0.005		0.094	

**Figure 3:** Summary of mean values of all key variables across interventions compared to the baseline with no AI

AI-based Clinical Decision Support Systems (CDSSs) into their workflows.

4.1. Impact of AI Explainability on Trust and Performance

Our quantitative analysis showed that using a CDSS, regardless of explanation level, led to improved trust and diagnostic performance compared to the baseline (no-AI) condition. However, the average clinician accuracy across all interventions remained below the standalone AI model's 81% accuracy, indicating room for enhancing collaborative performance through better system design. Interestingly, clinicians' trust was higher when their decisions aligned with AI recommendations, even when those recommendations were incorrect, highlighting a risk of automation bias, a finding consistent with prior work [23].

These quantitative trends were echoed in the qualitative data. Clinicians viewed AI as a valuable support tool, especially for tasks like early cancer detection, screening, and prioritization. Many emphasized that AI should serve as an assistant, not a replacement, for clinical judgment, reinforcing the theme *"AI as a Decision Support, Not a Replacement"*. Furthermore, the theme *"Trust Development Over Time"* reflected how repeated interactions with the system helped build confidence, aligning with our finding that trust increased across all interventions compared to the baseline and aligns with Hoff's model of learned trust through repeated interaction [12].

However, increasing explanation complexity did not consistently enhance trust or performance. The enhanced tumor localization (4th) intervention, which had the highest level of explainability, resulted in the lowest scores in trust, understandability, and accuracy. Qualitative data provides a meaningful explanation for this trend: participants frequently cited the risk of *"information overload"* and stressed the importance of *"clarity and transparency"* in AI outputs [7]. They reported that overly detailed explanations made the interface harder to interpret and added unnecessary

complexity, mirroring the drop in quantitative scores in the 4th intervention.

4.2. Impact of AI Explainability on Cognitive Load

Quantitative findings showed no significant change in mental demand across interventions, though the 4th intervention did slightly increase it. On the other hand, stress levels significantly increased in the 3rd intervention, which introduced probabilistic tumor localization. These results suggest that detailed explanations can increase cognitive strain, especially if not presented intuitively.

Interestingly, the qualitative results aligned closely with this pattern. Participants noted that while AI could help reduce workload when well-integrated, complex explanation-heavy interfaces could create a *"time penalty"* and additional cognitive steps. The theme *"Balancing Simplicity and Information Overload"* captured this tension, clinicians wanted enough information to trust the system, but not so much that it became distracting. This reinforces the idea that explanation design must be purposeful and user-centered.

4.3. Effect of AI Confidence Scores on User Behavior

The addition of AI confidence scores revealed an important behavioral shift. When confidence was low, participants reported significantly lower trust and agreement, but took more time on their decisions, suggesting greater caution [2, 24]. In contrast, high confidence slightly increased trust but decreased performance, indicating potential overreliance on AI.

This quantitative result directly supports the qualitative theme *"Uncertainty and Confidence in AI Predictions"*. Several clinicians emphasized the need to understand how confident the AI is in its output and wanted this uncertainty to be communicated clearly. However, they also cautioned against blind trust in high-confidence outputs, reinforcing the importance of avoiding automation bias through calibrated, meaningful confidence displays.

4.4. Quantitative and Qualitative Insights

Together, our findings highlight that while clinicians generally trust and value AI systems, trust is not solely based on explanation complexity. In fact, both our data sources show that simpler, cleaner interfaces, as seen in the 1st intervention, foster higher trust, usability, and satisfaction. This supports themes like *"Clarity and Transparency in AI Outputs"* and *"Customization and Flexibility in Explanations"*, which emphasize the need for adaptive and intuitive systems.

Clinicians expressed that trust grows not only through explanation but through repeated, reliable system performance, as reflected in the strong alignment between the

“Trust Development Over Time” theme and quantitative measures. Moreover, trust emerged as a key driver of performance, forming a feedback loop in which higher trust led to higher engagement and accuracy—another point supported by interview statements and survey correlations.

4.5. Limitations

This study has several limitations that should be acknowledged. First, the controlled experimental setting does not fully reflect the complexity of real-world clinical environments. Participants were not interacting with real patients, nor were they working in a hospital setting, which means that crucial factors like emotional involvement, stress, and high-stakes decision-making were absent. Also, important issues such as workflow integration were outside the scope of this paper. The clinicians were interacting with a completely new system for a short duration during the experiment and implications of long-term use might be significantly different. Furthermore, we used a simplified prototype to explore initial clinician behavior and trust formation, which may not fully reflect the design of a mature, operational CDSS. Our user interface was designed for short-term use in controlled experiments targeting new users who are not familiar with such a system which may have introduced additional cognitive load or slowed decision-making. Operational systems should adopt more streamlined and user-friendly interfaces considering long-term use. Further research is necessary to determine the generalizability of our findings to actual clinical practice.

5. Impact of the Study

This study provides important insights into the design and implementation of explainable AI systems in clinical decision-making. By combining quantitative performance measures with qualitative feedback from clinicians, it offers a comprehensive evaluation of how varying levels of AI explainability influence trust, cognitive load, diagnostic accuracy, and system usability.

- **Evidence-based design guidance:** The study identifies key trade-offs between explanation complexity and user trust, highlighting that overly detailed explanations can hinder rather than help clinical decision-making. These findings offer actionable guidance for developers to design XAI systems that prioritize clarity and usability over technical verbosity.
- **Human-centered validation:** By integrating clinicians’ voices through thematic analysis, the study moves beyond system performance metrics to understand real-world expectations, concerns, and cognitive demands. This reinforces the importance of clinician-centered design in AI tool development.
- **Enhancing trust calibration:** The results demonstrate that clinicians’ trust in AI is influenced more by system reliability and clarity than by the volume of information provided. This has direct implications for

improving trust calibration, reducing automation bias, and supporting safe, effective collaboration between humans and AI.

- **Advancing mixed-methods evaluation:** Methodologically, the study contributes a robust framework for evaluating AI systems using both experimental metrics and qualitative insights. This dual approach can be adopted by future studies seeking to understand the nuanced effects of XAI in other high-stakes domains.
- **Informing policy and integration:** Insights about usability, cognitive load, and integration into clinical workflows can inform institutional policies on how AI tools should be introduced and regulated to ensure adoption without compromising professional roles or patient care.

References

- [1] Al-Dhabyani, W., Gomaa, M., Khaled, H., and Fahmy, A. (2020). Dataset of breast ultrasound images. *Data in Brief*, 28:104863.
- [2] Antifakos, S., Kern, N., Schiele, B., and Schwaninger, A. (2005). Towards improving trust in context-aware systems by displaying system confidence. pages 9–14.
- [3] Asan, O., Bayrak, A. E., and Choudhury, A. (2020). Artificial Intelligence and Human Trust in Healthcare: Focus on Clinicians. *Journal of Medical Internet Research*, 22(6):e15154. Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.
- [4] Braun, V. and Clarke, V. (2022). Conceptual and design thinking for thematic analysis. *Qualitative psychology*, (1):3. Publisher: Educational Publishing Foundation.
- [5] Choudhury, A., Asan, O., and Medow, J. E. (2022). Effect of risk, expectancy, and trust on clinicians’ intent to use an artificial intelligence system – Blood Utilization Calculator. *Applied Ergonomics*, 101:103708.
- [6] Davis, F. D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly*, pages 319–340. Publisher: Management Information Systems Research Center, University of Minnesota.
- [7] Derksen, M. E., van Beek, M., de Bruijn, T., Stuit, F., Blankers, M., and Goudriaan, A. E. (2025). Ethical aspects and user preferences in applying machine learning to adjust eHealth addressing substance use: A mixed-methods study. *International Journal of Medical Informatics*, page 105897.
- [8] Gaube, S., Suresh, H., Raue, M., Merritt, A., Berkowitz, S. J., Lerner, E., Coughlin, J. F., Gutttag, J. V., Colak, E., and Ghassemi, M. (2021). Do as AI say: susceptibility in deployment of clinical decision-aids. *npj Digital Medicine*, 4(1):1–8. Number: 1 Publisher: Nature Publishing Group.
- [9] Ghazizadeh, M., Lee, J. D., and Boyle, L. N. (2012). Extending the technology acceptance model to assess automation. *Cognition, Technology & Work*, pages 39–49. Publisher: Springer.
- [10] Glikson, E. and Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, pages 627–660. Publisher: Briarcliff Manor, NY.
- [11] Hartmann, D. P., Gottman, J. M., Jones, R. R., Gardner, W., Kazdin, A. E., and Vaught, R. S. (1980). Interrupted time-series analysis and its application to behavioral data. *Journal of Applied Behavior Analysis*, 13(4):543–559.
- [12] Hoff, K. A. and Bashir, M. (2015). Trust in automation: integrating empirical evidence on factors that influence trust. *Human Factors*, 57(3):407–434.

- [13] Janowczyk, A. and Madabhushi, A. (2016). Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of pathology informatics*, 7(1):29. Publisher: Elsevier.
- [14] McIntosh, C. and Purdie, T. G. (2016). Voxel-based dose prediction with multi-patient atlas selection for automated radiotherapy treatment planning. *Physics in Medicine & Biology*, 62(2):415. Publisher: IOP Publishing.
- [15] McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., and Darzi, A. (2020). International evaluation of an AI system for breast cancer screening. *Nature*, 577(7788):89–94. Publisher: Nature Publishing Group UK London.
- [16] Micocci, M., Borsci, S., Thakerar, V., Walne, S., Manshadi, Y., Edridge, F., Mullarkey, D., Buckle, P., and Hanna, G. B. (2021). Attitudes towards trusting artificial intelligence insights and factors to prevent the passive adherence of GPs: a pilot study. *Journal of Clinical Medicine*, 10(14):3101. Publisher: MDPI.
- [17] Nahata, H. and Singh, S. P. (2020). Deep learning solutions for skin cancer detection and diagnosis. *Machine Learning with Health Care Perspective: Machine Learning and Healthcare*, pages 159–182. Publisher: Springer.
- [18] Parasuraman, R. and Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human factors*, 39(2):230–253. Publisher: SAGE Publications Sage CA: Los Angeles, CA.
- [19] Rezaeian, O., Bayrak, A. E., and Asan, O. (2024). An architecture to support graduated levels of trust for cancer diagnosis with ai. In *International Conference on Human-Computer Interaction*, pages 344–351. Springer.
- [20] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III* 18, pages 234–241. Springer.
- [21] Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1):44–56. Publisher: Nature Publishing Group US New York.
- [22] Tucci, V., Saary, J., and Doyle, T. E. (2022). Factors influencing trust in medical artificial intelligence for healthcare professionals: A narrative review. *J. Med. Artif. Intell.*, 5(4).
- [23] Vicente, L. and Matute, H. (2023). Humans inherit artificial intelligence biases. *Scientific Reports*, (1):15737.
- [24] Zhang, Y., Liao, Q. V., and Bellamy, R. K. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 295–305.
- [25] Čartolovni, A., Tomičić, A., and Mosler, E. L. (2022). Ethical, legal, and social considerations of AI-based medical decision-support tools: A scoping review. *International Journal of Medical Informatics*, 161:104738. Publisher: Elsevier.