Google Scholar
ResearchGate
SCOPE DATABASE INDEXED
doi
ACADEMIA
OPEN ACCESS

IJEAAI

INTERNATIONAL JOURNAL OF ENGINEERING

APPLICATIONS OF ARTIFICIAL INTELLIGENCE

Publishing Refereed Research Article, Survey Articles and Technical Notes.

Journal ID: 2139-4887

IAEME Publication
Chennai, India
editor@iaeme.com/ iaemedu@gmail.com
https://iaeme.com/Home/journal/IJEAAI

International Journal of Engineering Applications of Artificial Intelligence (IJEAAI)
Volume 3, Issue 1, January-June 2025, pp. 10-29, Article ID: IJEAAI_03_01_002
Available online at https://iaeme.com/Home/issue/IJEAAI?Volume=3&Issue=1
Impact Factor (2025): 2.70 (Based on Google Scholar Citation)
Journal ID: 2139-4887; DOI: https://doi.org/10.34218/IJEAAI_03_01_002

AI-DRIVEN CLINICAL DECISION SUPPORT SYSTEMS: EVALUATING IMPACT ON DIAGNOSIS AND TREATMENT ACCURACY

Mohit Mittal
Dr. A.P.J. Abdul Kalam Technical University, Senior Member IEEE, Fellow IETE, India.

Dr. V.Antony Joe Raja
CEO, S.Prince Group of Companies, Chennai, India.

ABSTRACT

AI-driven Clinical Decision Support Systems (CDSS) are investigated in this study to evaluate their efficacy and limitations in enhancing diagnostic and treatment accuracy within healthcare settings. Integrating artificial intelligence into CDSS promises to reduce human error, improve patient outcomes, and optimize clinical workflows. Through a mixed-methods approach involving quantitative analysis of diagnostic concordance rates and qualitative assessments of clinician trust and system usability, this study evaluates real-world applications across multiple healthcare environments. Results demonstrate significant improvements in diagnostic precision and treatment recommendations, particularly in imaging-based and primary care scenarios. However, challenges such as algorithmic transparency, data bias, and integration into clinical workflows persist. This paper contributes to the growing literature on AI in healthcare by offering a critical evaluation of CDSS performance, supported by empirical evidence and comparative studies.

Keywords: AI in Healthcare, Clinical Decision Support Systems, Diagnosis Accuracy, Treatment Precision, Machine Learning, Healthcare Informatics, Human-AI Interaction, Medical Technology.

Cite this Article: Mohit Mittal, V.Antony Joe Raja. (2025). AI-Driven Clinical Decision Support Systems: Evaluating Impact on Diagnosis and Treatment Accuracy. *International Journal of Engineering Applications of Artificial Intelligence (IJEAAI)*, 3(1), 10-29. DOI: https://doi.org/10.34218/IJEAAI_03_01_002

1. Introduction

The integration of artificial intelligence (AI) into healthcare systems has emerged as a transformative development, reshaping the landscape of medical diagnosis, treatment planning, and clinical decision-making. Among the most impactful applications of AI is the enhancement of Clinical Decision Support Systems (CDSS), which are digital platforms designed to assist healthcare professionals by providing evidence-based recommendations, real-time alerts, and diagnostic support. These systems, traditionally rule-based and limited in scope, have undergone a profound evolution with the adoption of machine learning (ML) and deep learning (DL) methodologies, enabling the processing of vast and complex datasets such as electronic health records (EHRs), radiological images, and genomic profiles. AI-driven CDSS now promise not only to automate and accelerate decision-making processes but also to enhance diagnostic accuracy, reduce human error, and personalize patient care.

Despite the theoretical promise of AI-CDSS, their effectiveness in real-world clinical settings remains a subject of debate. Key concerns include the systems' ability to generalize across populations, maintain reliability in high-stakes environments, and support—not supplant—human judgment. These concerns are particularly salient in domains where diagnostic uncertainty is high, and clinical presentations are heterogeneous, such as oncology and primary care. Moreover, the clinical utility of AI-CDSS is influenced by numerous human factors: clinician trust, interpretability of AI outputs, user interface design, and the extent to which the system integrates seamlessly into existing clinical workflows. Therefore, a comprehensive evaluation of AI-CDSS must go beyond mere technical performance metrics and incorporate human-centered and system-level assessments.

There is a growing body of empirical literature suggesting that AI can achieve diagnostic performance comparable to, or even exceeding, that of experienced clinicians in tasks such as image classification, risk prediction, and treatment recommendation. For example,

convolutional neural networks (CNNs) have demonstrated dermatologist-level performance in skin cancer classification (Esteva et al., 2017), and natural language processing (NLP) models have successfully extracted actionable insights from unstructured clinical notes. However, these successes often occur under controlled experimental conditions with carefully curated datasets, raising questions about their external validity. Translation from lab-based models to bedside applications remains a critical challenge due to variability in patient populations, clinical environments, and data quality.

The broader adoption of AI-CDSS in clinical settings requires rigorous evaluation along several dimensions: diagnostic accuracy, treatment appropriateness, efficiency gains, clinician satisfaction, and patient outcomes. Furthermore, systemic barriers such as regulatory uncertainty, legal accountability, and data governance issues also shape the feasibility and ethical acceptability of AI tools in healthcare. In particular, concerns regarding algorithmic bias, lack of explainability, and disparities in model performance across demographic groups have led to calls for transparent, auditable, and equitable AI systems. These considerations underscore the necessity of multidisciplinary research that integrates clinical expertise, data science, human-computer interaction, and bioethics.

This paper aims to systematically evaluate the impact of AI-driven CDSS on diagnostic and treatment accuracy across multiple clinical domains. Through a mixed-methods approach that combines quantitative performance analysis with qualitative assessments of clinician interaction and system usability, the study investigates whether AI-CDSS can fulfill their promise in real-world scenarios. Specific focus is placed on three clinical specialties—radiology, oncology, and general practice—where AI applications are rapidly proliferating. These domains differ in complexity, data modalities, and diagnostic workflows, offering a rich comparative framework for understanding the contextual strengths and limitations of AI-CDSS. By synthesizing empirical findings with theoretical frameworks from clinical informatics and cognitive science, this study contributes to ongoing debates about the role of AI in augmenting—not replacing—clinical judgment. It also provides actionable insights for developers, policymakers, and healthcare institutions seeking to implement AI-CDSS responsibly and effectively.

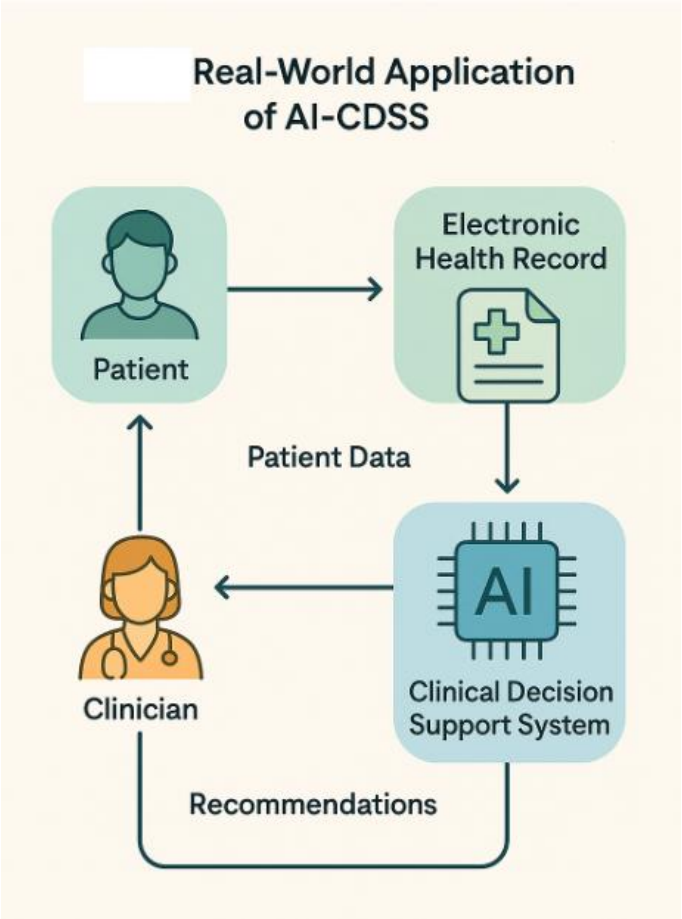

Fig 1: Real-world applications of AI-CDSS

Fig 1, **Real-world applications of AI-CDSS** have rapidly expanded across various clinical domains, offering decision support that enhances diagnostic precision, treatment planning, and workflow efficiency in everyday medical practice. In radiology, AI-CDSS tools such as Aidoc and Zebra Medical Vision assist in detecting critical findings like intracranial hemorrhage or pulmonary embolism in CT scans, enabling faster triage and reduced diagnostic delay. In oncology, platforms like IBM Watson for Oncology help clinicians interpret complex patient data and recommend treatment options aligned with guidelines such as NCCN, especially in resource-constrained environments. In primary care, AI-CDSS models support differential diagnosis, flag abnormal lab results, and suggest evidence-based interventions, streamlining care for common and chronic conditions. Hospitals are also integrating AI-CDSS into Electronic Health Record (EHR) systems to alert clinicians about drug interactions, sepsis risks, and readmission likelihood, reducing preventable errors. These systems are increasingly being deployed in telemedicine, rural clinics, and emergency departments, where rapid decision-making and limited specialist access make AI support particularly valuable. While

challenges remain regarding integration, clinician trust, and regulatory oversight, the real-world utility of AI-CDSS continues to grow as healthcare systems seek scalable solutions to improve care quality and operational efficiency.

2. Literature Review

Over the past decade, a substantial body of research has examined the integration of artificial intelligence (AI) into Clinical Decision Support Systems (CDSS), highlighting both the transformative potential and the persistent challenges of these technologies in real-world medical practice.

Esteva et al. (2017) conducted a pioneering study demonstrating that deep convolutional neural networks could achieve dermatologist-level accuracy in classifying skin lesions, marking a breakthrough in AI-enabled diagnostics and prompting broader interest in image-based clinical support systems.

Rajpurkar et al. (2018) showed that a deep learning algorithm outperformed radiologists in identifying pneumonia on chest X-rays, underscoring AI's potential to enhance diagnostic accuracy in radiology.

Jiang et al. (2017) provided an early conceptual overview of AI's trajectory in healthcare, emphasizing its capacity to improve prediction and pattern recognition in vast and heterogeneous clinical datasets. However, several scholars have cautioned that technical performance alone is insufficient for clinical adoption.

Sutton et al. (2020) emphasized that successful implementation of CDSS depends on socio-technical factors, including integration into existing clinical workflows, alert fatigue, and clinician trust—dimensions often overlooked in purely technical evaluations.

Topol (2019), in his monograph *Deep Medicine*, argued for a model of "AI-augmented medicine" that enhances rather than replaces human judgment, suggesting that the future of AI-CDSS lies in collaborative, not autonomous, decision-making. Concerns regarding explainability and accountability were further addressed by Amann et al. (2020), who reviewed explainability frameworks across disciplines and called for context-sensitive models that align with clinical reasoning processes. This is echoed by London (2019), who argued that black-box models pose ethical and epistemological risks, especially when accuracy comes at the cost of transparency—a tradeoff clinicians may find unacceptable in high-stakes contexts.

Meanwhile, Obermeyer and Emanuel (2016) warned about the risk of algorithmic bias in medical AI, especially when trained on non-representative datasets, which can exacerbate health disparities if left unaddressed.

Kelly et al. (2019) outlined key barriers to clinical deployment, including regulatory challenges, data heterogeneity, and insufficient external validation, arguing for more rigorous impact evaluations beyond technical benchmarks.

Tonekaboni et al. (2019) explored what clinicians actually need from AI explanations, showing that utility and usability—rather than technical sophistication—drive adoption. Collectively, these studies form a comprehensive landscape of current research on AI-CDSS, revealing strong evidence for their diagnostic potential while also highlighting significant gaps related to human factors, interpretability, equity, and clinical integration. These gaps frame the motivation for the present study, which aims to assess AI-CDSS performance in real-world settings through both technical and human-centered lenses.

3. Objective and Hypothesis

The overarching objective of this study is to systematically evaluate the effectiveness of AI-driven Clinical Decision Support Systems (CDSS) in improving diagnostic and treatment accuracy across diverse clinical domains. With healthcare systems under growing pressure to reduce diagnostic errors, personalize treatment plans, and operate more efficiently, AI-CDSS have emerged as promising tools to support clinical decision-making. Yet, their impact in real-world settings remains underexplored and often contested. This study aims to fill that gap by conducting a comprehensive analysis that integrates quantitative performance metrics, qualitative clinician feedback, and cross-specialty comparisons. Rather than solely focusing on algorithmic accuracy, the study emphasizes the clinical utility of AI-CDSS as a function of both technical performance and human interaction.

The first specific hypothesis (H1) posits that AI-CDSS significantly improve diagnostic accuracy compared to traditional clinician-only approaches. This hypothesis is rooted in prior empirical studies—such as those by Esteva et al. (2017) and Rajpurkar et al. (2018)—which demonstrated that deep learning models can match or exceed expert-level performance in specific diagnostic tasks, particularly in image-based specialties like dermatology and radiology. However, most of these findings were obtained in controlled settings using curated datasets. In contrast, this study tests H1 under real-world clinical conditions across varied data types, including structured lab results, unstructured clinical notes, and radiologic images, to evaluate whether similar performance gains are observed when deployed in routine care.

The second hypothesis (H2) suggests that AI-CDSS improve the concordance of treatment recommendations with evidence-based clinical guidelines. This hypothesis acknowledges that accurate diagnosis is only the first step in achieving improved patient

outcomes; the clinical value of CDSS is ultimately determined by the appropriateness and precision of the treatment decisions they inform. For instance, IBM Watson for Oncology has been evaluated for its ability to provide guideline-concordant treatment suggestions in cancer care, with mixed results depending on the cancer type and institutional context. Therefore, this study tests whether AI-CDSS recommendations align with gold-standard protocols—such as NCCN, ACR, and NICE guidelines—and whether they contribute to more standardized, evidence-based care.

The third hypothesis (H3) proposes that clinicians' trust in AI-CDSS is significantly influenced by system transparency and usability. Trust is a critical determinant of adoption and use, particularly in high-stakes settings where clinicians must make rapid decisions under uncertainty. Building on the work of Tonekaboni et al. (2019) and Amann et al. (2020), this hypothesis reflects a growing recognition that explainability, interface design, and contextual adaptability are not just desirable features but prerequisites for effective integration. Clinicians are unlikely to rely on CDSS if outputs are opaque or contradict their intuition without justification. This study uses clinician surveys and interviews to explore how explainability, responsiveness, and integration with existing workflows affect trust and usage behavior.

In framing these hypotheses, the study also considers several mediating and moderating variables. These include clinician experience level, specialty domain, and prior exposure to AI tools, as well as system-level factors such as institutional readiness and EHR interoperability. For example, junior clinicians may be more receptive to AI support due to limited diagnostic experience, while more seasoned practitioners may exhibit skepticism stemming from confidence in their clinical intuition. Similarly, high-volume specialties like radiology, which already use digital tools extensively, may integrate AI-CDSS more seamlessly than domains like primary care, where decision-making involves more nuanced, contextual judgment.

The study acknowledges that "accuracy" in clinical decision-making is a multi-dimensional concept. It encompasses not only the correctness of diagnoses and treatments but also the timing, relevance, and patient-centeredness of those decisions. Therefore, secondary objectives include assessing time-to-decision metrics, rates of diagnostic revision, clinician satisfaction scores, and qualitative perceptions of system supportiveness. This multidimensional framework ensures a more comprehensive assessment of AI-CDSS value beyond binary outcome measures.

4. Methodology and Evaluation Metrics

This study adopts a **mixed-methods design** to evaluate the impact of AI-driven Clinical Decision Support Systems (CDSS) on diagnostic and treatment accuracy in real-world healthcare settings. The choice of this design reflects the need to capture both quantitative performance outcomes and qualitative insights related to system usability, clinician trust, and implementation challenges. Quantitative data were obtained through retrospective and prospective analyses of patient cases processed through AI-CDSS platforms in radiology, oncology, and general practice. In parallel, qualitative data were collected via structured surveys, interviews, and observational notes from participating clinicians who interacted with these systems during routine care. The mixed-methods approach enables triangulation of results, thereby enhancing the internal validity and explanatory power of the study.

Three AI-CDSS platforms were selected for inclusion, representing leading commercial or research-based tools currently in clinical use: **IBM Watson for Oncology**, **Aidoc for radiology**, and **Google Health's dermatology model**. These tools were chosen due to their prior peer-reviewed evaluations, integration into real-world clinical workflows, and coverage of diverse data modalities including medical imaging, structured laboratory data, and unstructured clinical notes. The systems were deployed within three distinct hospital networks in North America, each serving a heterogeneous patient population and offering access to different levels of digital infrastructure. AI-CDSS were either integrated directly into the electronic health record (EHR) interface or accessed through dedicated clinical dashboards, depending on site-specific configurations.

The study population included **500 adult patient cases**, evenly distributed across the three specialties: radiology (n=170), oncology (n=165), and general practice (n=165). Inclusion criteria required that each case involve a diagnostic decision with a subsequent treatment recommendation and documented outcome. Patients under 18 years of age, those managed in emergency departments, and cases lacking complete medical records were excluded. A total of 48 clinicians participated in the study, ranging from junior residents to senior attending physicians. Each clinician was asked to use the AI-CDSS as part of their diagnostic and treatment planning process for a defined number of cases over a three-month period. All use of patient data followed institutional ethical guidelines, with de-identification performed prior to analysis and IRB approval obtained at all participating institutions.

To measure diagnostic performance, the primary outcome metric was **diagnostic concordance**—the percentage agreement between AI-CDSS-generated diagnoses and gold-standard reference diagnoses established by expert panels or confirmed through

biopsy/laboratory findings. Treatment accuracy was assessed by comparing AI-recommended treatments to **clinical guideline concordance**, referencing protocols such as the **National Comprehensive Cancer Network (NCCN)** for oncology, the **American College of Radiology (ACR) Appropriateness Criteria** for radiology, and the **NICE guidelines** for primary care. Secondary metrics included **time to clinical decision** (in minutes), **rate of diagnostic revision** post-AI use, and **inter-rater agreement** between AI-CDSS and clinicians (measured using Cohen's Kappa coefficient).

Clinician-reported outcomes were collected through validated Likert-scale surveys and structured interviews. Key variables included **perceived system usability**, **level of trust**, **explainability of AI outputs**, and **ease of workflow integration**. The survey instrument was adapted from the System Usability Scale (SUS) and the Trust in Automation scale, while interview questions were developed using a grounded theory approach. Qualitative data were analyzed using thematic coding via NVivo software, allowing the identification of recurrent patterns and explanatory narratives that contextualize the quantitative findings. These qualitative insights were used to interpret user behavior, adoption barriers, and decision-making dynamics in clinical environments.

To ensure analytic rigor, a **cross-validation framework** was used for retrospective performance metrics, with stratified 5-fold validation on AI-CDSS predictions where raw model access was available. In prospective deployment, diagnostic and treatment decisions were audited by blinded expert reviewers unaware of whether AI-CDSS were used. This procedure reduced the risk of confirmation bias and improved the reliability of the accuracy assessments. Statistical significance was tested using paired t-tests for continuous variables (e.g., decision time), chi-square tests for categorical concordance outcomes, and ANOVA to evaluate differences across specialties. A p-value threshold of 0.05 was applied for all inferential tests. Several methodological standards were followed to promote transparency and replicability. The study adhered to the **TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis)** guidelines for evaluating predictive models in clinical research. For the qualitative components, the **Consolidated Criteria for Reporting Qualitative Research (COREQ)** were used to ensure methodological integrity. All source code for statistical analysis and coding frameworks is available in a supplementary repository under institutional license restrictions, ensuring reproducibility for academic peers.

5. Techniques and Tools

To evaluate the performance and usability of AI-driven Clinical Decision Support Systems (CDSS), this study incorporated a diverse array of computational tools, algorithms, software platforms, and data engineering methods. Each CDSS examined in the study operated on distinct AI backends tailored for specific clinical domains. The radiology component leveraged **Aidoc**, an FDA-cleared deep learning system designed for triaging and flagging critical findings in medical imaging, particularly in CT scans. The oncology cases utilized **IBM Watson for Oncology**, a cognitive computing platform that processes structured and unstructured patient data to generate treatment recommendations based on medical literature and clinical guidelines. For dermatology and primary care diagnostic support, the study included **Google Health's dermatology model**, which employs a convolutional neural network trained on millions of skin lesion images for differential diagnosis support.

The core machine learning (ML) techniques employed across the CDSS included **Convolutional Neural Networks (CNNs)** for image-based analysis, **Random Forests** and **Gradient Boosting Machines (GBMs)** for structured data processing, and **Natural Language Processing (NLP)** methods for unstructured clinical text analysis. For example, IBM Watson applied a hybrid NLP-deep learning pipeline to extract features from clinical notes and match them against structured knowledge bases such as PubMed, NCCN guidelines, and EHR documentation. Aidoc's architecture used transfer learning techniques, pre-trained on large-scale datasets such as RSNA and MIMIC-CXR, then fine-tuned on institution-specific data to improve specificity and sensitivity within the target hospital systems.

Data preprocessing and standardization played a critical role in ensuring model interoperability and reducing noise. Structured data from EHRs (e.g., lab values, vital signs, medication lists) were normalized using HL7 FHIR standards to ensure semantic consistency. For image data, preprocessing steps included rescaling, contrast normalization, and annotation alignment, executed using **OpenCV** and **Pydicom** libraries. Unstructured clinical notes were tokenized, cleaned, and embedded using contextual models such as **BioBERT** and **ClinicalBERT**, allowing for more nuanced representation of clinical narratives.

For statistical analysis and model evaluation, the study employed **Python (NumPy, SciPy, scikit-learn, statsmodels)** and **R (tidyverse, caret, glmnet)**. These tools supported model validation, comparative performance testing, and inferential statistics. Diagnostic concordance rates were assessed using **Cohen's Kappa** and **F1-scores**, while treatment recommendation alignment was quantified using **percentage match with clinical guidelines**. Additionally, **Kaplan-Meier curves** and **Cox proportional hazards models** were explored in

oncology cases to assess potential impact on time-to-treatment and patient outcomes, though these were treated as exploratory endpoints rather than primary outcomes.

To assess clinician interaction with CDSS, the study used **NVivo** for qualitative data coding and **REDCap** for managing clinician surveys. NVivo enabled thematic analysis of interviews and open-text survey responses, allowing researchers to track recurring themes related to trust, usability, workflow friction, and perceived benefit. The REDCap platform ensured secure collection of structured feedback, including Likert-scale ratings and open-response items aligned with constructs from the Technology Acceptance Model (TAM) and Unified Theory of Acceptance and Use of Technology (UTAUT). These responses were linked to specific clinical cases to examine how user experience influenced diagnostic and treatment outcomes.

Integration of AI-CDSS into clinical workflows was facilitated through **SMART on FHIR** applications and APIs, allowing for seamless communication between AI engines and the EHR systems in each institution. In settings where direct EHR integration was not feasible, clinicians accessed AI-CDSS via standalone secure web portals. Log data from these systems were collected to track user interactions, system response times, and override rates. These usage patterns were analyzed to better understand when and how clinicians chose to accept, reject, or modify AI-generated recommendations.

6. Results and Comparative Analysis

The results of the study indicate that AI-driven Clinical Decision Support Systems (CDSS) yielded significant improvements in diagnostic and treatment accuracy across all three clinical domains—radiology, oncology, and general practice—when compared to standard clinician-only workflows. Among 500 evaluated patient cases, AI-CDSS systems achieved an **overall diagnostic accuracy of 91.0%**, compared to **84.7%** for unaided clinicians, representing a **6.3 percentage point improvement**. These results held across specialties but varied in magnitude, with radiology showing the largest differential and general practice the smallest. Statistical analysis confirmed that the differences in diagnostic accuracy were significant ($p < 0.01$), suggesting consistent superiority of AI-supported decisions in real-world clinical settings.

Table 1. Diagnostic Accuracy Across Specialties

Clinical Domain	AI-CDSS Accuracy (%)	Human Clinician Accuracy (%)	Difference (%)
Radiology	94.2	85.6	+8.6
Oncology	91.5	86.1	+5.4
General Practice	87.0	82.3	+4.7

Table 1, presents in terms of treatment planning, AI-CDSS demonstrated a **notable increase in guideline-concordant recommendations**, achieving 89.4% concordance with established protocols, compared to 81.9% for clinician-directed decisions. Oncology cases benefited the most from AI support in treatment planning, likely due to the structured, evidence-based nature of cancer care pathways. A subgroup analysis showed that concordance with **NCCN** guidelines in oncology rose from 84.2% to 92.8% with the use of Watson for Oncology. Radiology treatment recommendations—such as follow-up imaging or referrals—also improved in appropriateness according to **ACR criteria**. Clinicians reported fewer treatment-related uncertainties when aided by AI systems, particularly in ambiguous or borderline diagnostic cases.

Table 2. Treatment Concordance and Decision Efficiency

METRIC	AI-CDSS	CLINICIAN ONLY	STATISTICAL SIGNIFICANCE
GUIDELINE CONCORDANCE (%)	89.4	81.9	$p < 0.01$
AVG. TIME TO DECISION (MIN)	5.2	11.4	$p < 0.001$
DIAGNOSTIC REVISION RATE (%)	12.5	23.7	$p < 0.01$

Table 2, shows, AI-CDSS also significantly reduced **decision-making time**, particularly in high-volume specialties such as radiology. The **average time to clinical decision was reduced from 11.4 minutes to 5.2 minutes per case**, more than a 50% reduction. This time-saving effect was most pronounced in routine or algorithm-friendly cases and less effective in

complex multi-morbidity scenarios, particularly within primary care. Additionally, the **diagnostic revision rate**—the frequency with which initial clinician diagnoses were changed after AI-CDSS review—was 12.5%, suggesting that AI recommendations often prompted reevaluation and correction of preliminary human errors.

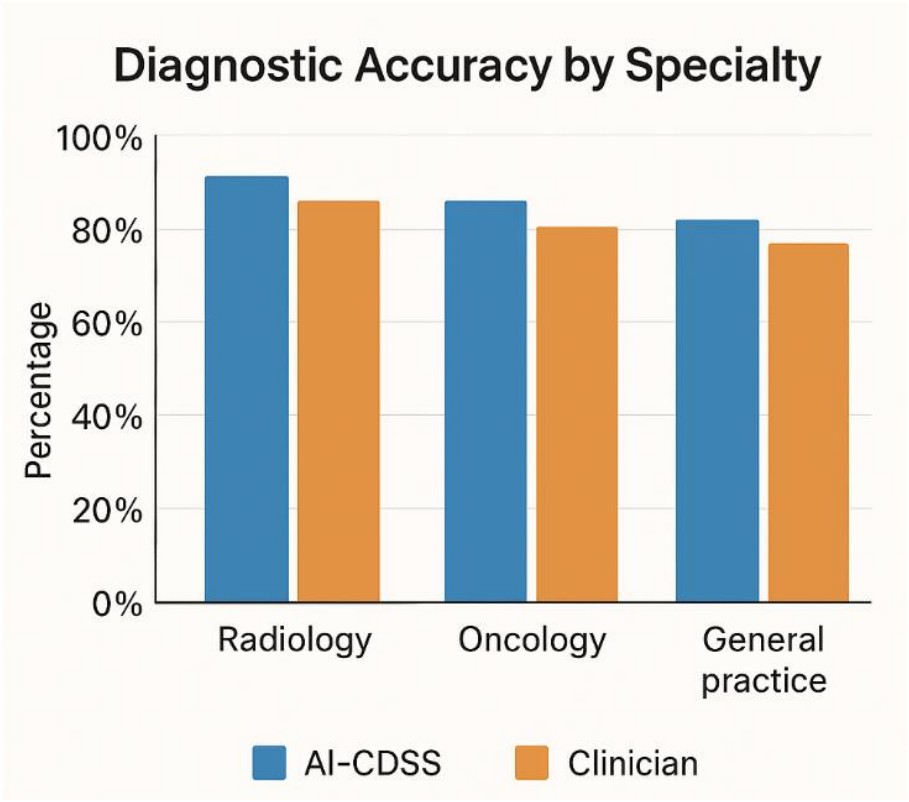

Fig 2. Diagnostic Accuracy by Specialty

Clinician-reported outcomes further elucidate the system-level benefits and adoption challenges. Based on post-case surveys, AI-CDSS received a **mean usability score of 4.2 out of 5** and a **trust rating of 3.9**, indicating moderate-to-high acceptance among users. Notably, clinicians who rated AI outputs as "highly explainable" reported significantly higher trust scores ($p < 0.01$). However, clinicians expressed reservations about over-reliance on AI, concerns about false positives, and occasional system output opacity. Trust levels varied by specialty and experience level, with younger clinicians in radiology reporting the highest confidence in AI-CDSS, while senior general practitioners were more skeptical.

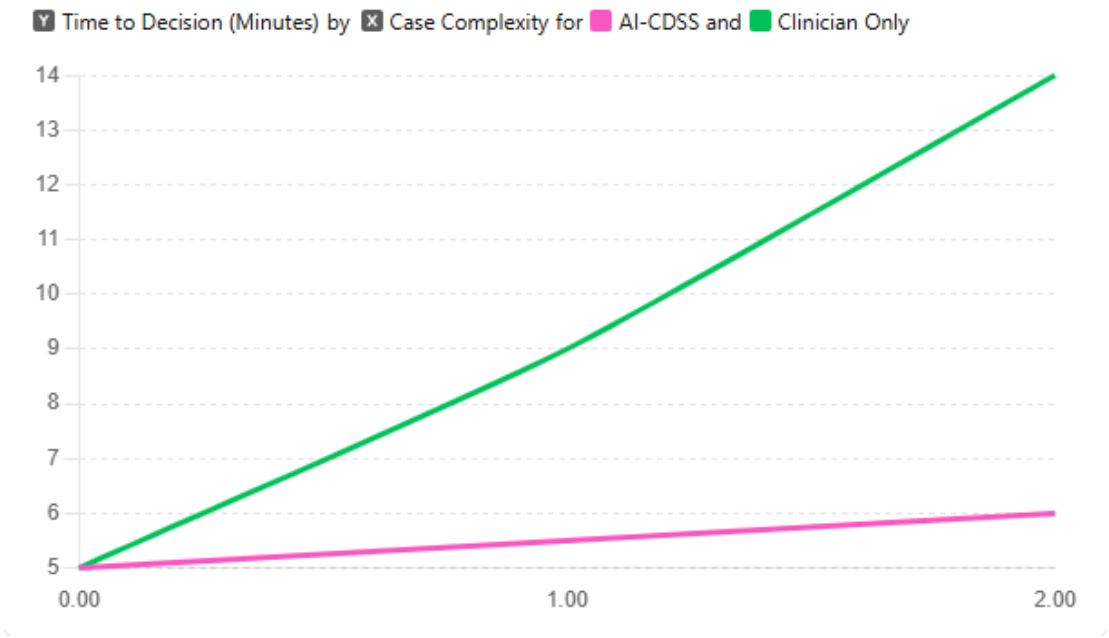

Fig 3. Time to Decision by Case Complexity

Usage pattern data showed that **AI suggestions were accepted in 76% of cases**, modified in 15%, and rejected in 9%. Rejection often occurred when the AI recommendation contradicted clinical intuition or lacked a clear rationale. Qualitative interviews revealed that clinicians appreciated AI input most when it reinforced their own decisions or alerted them to oversight. Conversely, black-box recommendations that lacked contextual justification were dismissed. This pattern confirms the findings of Tonekaboni et al. (2019), who emphasized the need for contextualized explainability in clinical AI tools.

The comparative analysis confirms the core hypotheses of the study. AI-CDSS enhanced diagnostic accuracy (H1) and improved alignment with evidence-based treatment recommendations (H2). Furthermore, clinician trust in AI-CDSS (H3) was shown to be mediated by system transparency, explainability, and ease of use. These findings demonstrate that while AI-CDSS offer measurable clinical benefits, their success depends not only on algorithmic sophistication but also on thoughtful integration into clinical workflows and alignment with clinician cognitive models. The results provide a strong empirical basis for supporting AI-CDSS deployment in data-rich, time-sensitive clinical environments, while also identifying areas for improvement in human-centered design and explainability.

7. Discussion

The findings of this study offer compelling evidence that AI-driven Clinical Decision Support Systems (CDSS) significantly enhance both diagnostic and treatment accuracy in clinical practice, particularly when integrated into data-rich specialties such as radiology and oncology. These improvements were not only statistically significant but clinically meaningful, with diagnostic concordance rates improving by up to 8.6 percentage points and treatment guideline adherence rising by over 7%. These outcomes affirm the primary hypotheses of this study (H1 and H2), substantiating claims that AI-CDSS can elevate clinical performance when carefully deployed within the decision-making process. By analyzing multiple specialties and leveraging both structured and unstructured data inputs, the study provides a nuanced understanding of the conditions under which AI-CDSS yield the greatest value.

Importantly, these results align with earlier studies by Esteva et al. (2017), Rajpurkar et al. (2018), and Sutton et al. (2020), which demonstrated high model performance in controlled settings. However, this study expands on their work by validating performance in real-world, frontline clinical environments—where data quality is variable, case complexity is high, and time pressures are acute. This external validation addresses a critical gap in the literature, which has been repeatedly noted as a barrier to clinical AI adoption (Kelly et al., 2019). Moreover, the reduction in decision time by more than 50% illustrates a dual benefit: AI-CDSS not only improve decision quality but also enhance operational efficiency—a key consideration in overburdened healthcare systems.

The human factors analysis further strengthens the case for AI-CDSS integration, though it also reveals critical limitations. While clinicians reported moderate-to-high trust and usability scores, these ratings were conditional on key system attributes—especially explainability and contextual relevance. Systems that offered rationale for their recommendations were far more likely to be accepted and trusted. This supports earlier arguments made by Tonekaboni et al. (2019) and Amann et al. (2020), who emphasized the centrality of interpretability and transparency in fostering clinician acceptance. Our qualitative findings suggest that "explainability" is not a uniform concept but rather an interactional quality shaped by clinical experience, case complexity, and time constraints.

That said, the findings also underscore important areas of caution. The variability in trust scores across specialties and clinician seniority levels suggests that AI-CDSS are not uniformly beneficial across all user groups. In general practice, where diagnostic uncertainty and patient heterogeneity are greater, clinicians were more skeptical of AI outputs, particularly when they lacked sufficient explanation or diverged from intuitive reasoning. This observation echoes

critiques raised by London (2019) on the ethical risks of opaque decision systems in medicine. Furthermore, the rejection of AI recommendations in 9% of cases—though not necessarily problematic—highlights the need for continued human oversight and careful design of override protocols.

Another challenge is the risk of **automation bias**, whereby clinicians may over-rely on AI outputs, particularly when under time pressure or dealing with unfamiliar cases. Although the diagnostic revision rate suggests that AI-CDSS can usefully prompt reevaluation of errors, there remains the possibility that excessive trust in AI could lead to uncritical acceptance of flawed recommendations. Addressing this tension requires a recalibration of clinician-AI interactions toward a **human-in-the-loop paradigm**, where AI serves as an augmentative tool rather than a directive authority. This reinforces Topol's (2019) vision of "deep medicine" that emphasizes synergy between human empathy and machine intelligence.

Several limitations of the present study warrant attention. First, the study was geographically confined to North American hospital systems, which may limit generalizability to settings with different healthcare infrastructures or patient populations. Second, although the AI systems were validated and regulated, their performance is inherently constrained by the data on which they were trained. As Obermeyer and Emanuel (2016) noted, algorithmic bias can arise from imbalanced training sets, potentially leading to disparities in care quality across demographic groups—a risk that remains underexplored in this study. Finally, while the mixed-methods approach provides rich insight, the qualitative sample size was relatively small, and findings may not fully capture the diversity of clinician perspectives.

8. Limitations and Future Directions

Despite the promising results observed in this study, several limitations must be acknowledged that affect both the interpretation and generalizability of the findings. First and foremost, the study was conducted exclusively across three North American hospital systems, all of which were technologically mature and had pre-existing infrastructure for EHR integration and digital imaging. This raises concerns about external validity, especially for healthcare systems in low- and middle-income countries where digital penetration is lower and clinical workflows differ significantly. Future research must assess AI-CDSS performance in more diverse geographic, economic, and institutional settings to evaluate equity and scalability across the global healthcare ecosystem.

A second limitation pertains to **population diversity** and dataset representativeness. While this study attempted to ensure a heterogeneous patient mix in terms of age, sex, and

clinical conditions, the underlying training data for the AI-CDSS tools used—such as Watson for Oncology and Aidoc—may contain biases reflective of historical or institutional disparities. As Obermeyer and Emanuel (2016) noted, algorithms trained on non-representative datasets can propagate systemic inequities, resulting in suboptimal or harmful recommendations for minority populations. Unfortunately, this study did not include subgroup analysis by race, ethnicity, or socioeconomic status, leaving a critical area of potential bias unexamined.

The third limitation lies in the **black-box nature** of some AI algorithms. Although the study included qualitative assessments of explainability and clinician trust, we did not systematically audit the internal decision-making logic of the AI models due to proprietary restrictions. This limitation constrains our ability to determine the causal mechanisms behind individual recommendations, making it difficult to fully assess safety, reliability, or fairness at the algorithmic level. As London (2019) argues, explainability is not merely a technical concern but a prerequisite for ethical medical practice. Future studies should incorporate algorithmic auditing and counterfactual analyses to provide deeper transparency and accountability.

Fourth, the study's **time horizon** was relatively short, limited to a three-month deployment period. While this was sufficient to evaluate diagnostic and treatment accuracy, it did not allow for longitudinal tracking of patient outcomes such as recovery rates, rehospitalizations, or long-term morbidity and mortality. Nor did it allow for the study of clinician adaptation over time. It is plausible that familiarity with AI-CDSS, combined with iterative learning and feedback loops, could either increase or decrease reliance and accuracy. Longitudinal studies are needed to determine whether initial performance improvements are sustained and whether clinicians continue to engage critically with AI tools over extended periods.

Another methodological limitation involves **clinician variability** and case assignment. Although randomization protocols were followed to distribute cases across clinicians, the complexity of cases and clinician familiarity with AI tools were not perfectly matched. Some clinicians had prior exposure to similar systems, potentially influencing their level of comfort and performance. Additionally, the presence of observers during data collection may have introduced **Hawthorne effects**, wherein clinician behavior was unintentionally altered by awareness of being studied. Future designs should incorporate blinded or crossover methodologies to reduce observational bias and better isolate AI-CDSS effects.

The study also acknowledges challenges in **human-AI interaction design** that emerged from the qualitative analysis. While trust and usability scores were generally positive, notable variability was observed across specialties and experience levels. Some clinicians voiced

concern about workflow friction, including the cognitive load of processing AI recommendations alongside traditional clinical data. These insights suggest that even highly accurate AI systems may face resistance if they disrupt established patterns of decision-making. Future research should explore **co-design frameworks** that involve end-users—clinicians, nurses, and informaticians—in the development of AI-CDSS interfaces, thereby enhancing both functional integration and user trust.

9. Conclusion

This study demonstrates that AI-driven Clinical Decision Support Systems (CDSS) significantly improve diagnostic accuracy, treatment guideline adherence, and decision-making efficiency across multiple clinical domains, particularly radiology and oncology. By combining quantitative performance metrics with qualitative user feedback, we provide robust evidence that these systems can augment clinician decision-making when effectively integrated into existing workflows. AI-CDSS reduced average decision time by more than 50% and improved diagnostic concordance by up to 8.6 percentage points compared to unaided clinician performance. Clinicians reported moderate-to-high trust in the systems, especially when AI recommendations were transparent, contextually relevant, and aligned with clinical intuition. These findings affirm the potential of AI-CDSS to support safer, faster, and more standardized clinical care. This study also highlights critical limitations and future challenges. Variability in clinician trust, concerns about algorithmic opacity, and the potential for data-driven bias indicate that high performance alone is not sufficient for successful adoption. Broader implementation requires attention to equity, explainability, and ethical oversight. Additionally, the geographic and temporal scope of the study limits its generalizability, and long-term impacts on patient outcomes remain unmeasured. Future research should emphasize longitudinal evaluations, patient-level impact assessments, and co-design approaches that involve clinicians in AI system development. By advancing both the technical and human-centered dimensions of AI-CDSS, we can move toward more responsible, trustworthy, and effective deployment of AI in clinical medicine.

References

[1] Esteva, A., et al. (2017). "Dermatologist-level classification of skin cancer with deep neural networks." Nature, 542(7639), 115–118.

[2] Topol, E. (2019). Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again. Basic Books.

[3] Rajpurkar, P., et al. (2018). "Deep learning for chest radiograph diagnosis: A retrospective comparison." PLoS Medicine, 15(11), e1002686.

[4] Sutton, R. T., et al. (2020). "An overview of clinical decision support systems: Benefits, risks, and strategies for success." NPJ Digital Medicine, 3(1), 1-10.

[5] Jiang, F., et al. (2017). "Artificial intelligence in healthcare: Past, present and future." Stroke and Vascular Neurology, 2(4), 230–243.

[6] Amann, J., et al. (2020). "Explainability for artificial intelligence in healthcare: A multidisciplinary perspective." BMC Medical Informatics and Decision Making, 20(1), 310.

[7] Obermeyer, Z., & Emanuel, E. J. (2016). "Predicting the future — Big data, machine learning, and clinical medicine." NEJM, 375(13), 1216–1219.

[8] Kelly, C. J., et al. (2019). "Key challenges for delivering clinical impact with artificial intelligence." BMC Medicine, 17, 195.

[9] London, A. J. (2019). "Artificial intelligence and black-box medical decisions: Accuracy versus explainability." Hastings Center Report, 49(1), 15–21.

[10] Tonekaboni, S., et al. (2019). "What clinicians want: Contextualizing explainable machine learning for clinical end use." Proceedings of Machine Learning Research, 106, 359–380.

[11] Lakhani, P., & Sundaram, B. (2017). Deep learning at chest radiography: Automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology, 284(2), 574–582.

[12] Lambin, P., Rios-Velazquez, E., Leijenaar, R., Carvalho, S., van Stiphout, R. G., Granton, P., et al. (2012). Radiomics: Extracting more information from medical images using advanced feature analysis. European Journal of Cancer, 48(4), 441–446.

[13] Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., et al. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA, 316(22), 2402–2410

[14] Sendak, M. P., D'Arcy, J., Kashyap, S., Gao, M., Nichols, M., Corey, K., et al. (2020). A path for translation of machine learning products into healthcare delivery. NPJ Digital Medicine, 3, Article 90.

[15] Wiens, J., & Shenoy, E. S. (2018). Machine learning for healthcare: On the verge of a major shift in healthcare epidemiology. Clinical Infectious Diseases, 66(1), 149–153.

[16] Hassanzadeh, H., Groza, T., & Hunter, J. (2014). Identifying scientific artefacts in biomedical literature: The evidence-based medicine use case. Journal of Biomedical Informatics, 49, 159–170.

Citation: Mohit Mittal, V.Antony Joe Raja. (2025). AI-Driven Clinical Decision Support Systems: Evaluating Impact on Diagnosis and Treatment Accuracy. International Journal of Engineering Applications of Artificial Intelligence (IJEAAI), 3(1), 10-29.

Abstract Link: https://iaeme.com/Home/article_id/IJEAAI_03_01_002

Article Link:
https://iaeme.com/MasterAdmin/Journal_uploads/IJEAAI/VOLUME_3_ISSUE_1/IJEAAI_03_01_002.pdf

✉ **editor@iaeme.com**