International Journal of Medical Sciences (IJMS)
Volume 3, Issue 1, January-June 2025, pp. 47-62, Article ID: IJMS_03_01_003
Available online at https://iaeme.com/Home/issue/IJMS?Volume=3&Issue=1
Journal ID: 1452-2145; DOI: https://doi.org/10.34218/IJMS_03_01_003

ARTIFICIAL INTELLIGENCE IN DIAGNOSTIC MEDICINE: LITERATURE REVIEW CONTRASTING DIFFERENTIAL ACCURACY FROM TEST REPORTS VERSUS SELF-REPORTED SYMPTOMS AND IMPLICATIONS ON MEDICAL SPECIALTIES

Gunmeh Bhandari
Winston Churchill High School, Bethesda, Maryland, USA Bethesda, 20817.

ABSTRACT

Recent advances in large language models (LLMs) have shown that when fed with structured, tangible data - such as X-ray images, CT scans, bloodwork, and other machine-generated test reports - LLMs can achieve significantly higher diagnostic accuracy compared with when they rely on unstructured, self-reported patient symptoms. This paper reviews recent literature on AI applications in diagnostic reasoning, compares the performance of LLMs across different data modalities, and discusses which medical specialties are most vulnerable to future AI replacement. We present an index of 15 doctor specialties, highlighting the extent to which their diagnostic workflows (and thus their professional roles) rely on machine-generated data. We conclude that specialties dominated by image and laboratory report interpretation (e.g., radiology, pathology) are at higher risk, while those that require a more nuanced, context-rich synthesis of subjective data (e.g., internal medicine) are comparatively less vulnerable.

Keywords: LLM, Medical Diagnosis, Artificial Intelligence

Cite this Article: Gunmeh Bhandari. (2025). Artificial Intelligence in Diagnostic Medicine: Literature Review Contrasting Differential Accuracy from Test Reports Versus Self-Reported Symptoms and Implications on Medical Specialties. *International Journal of Medical Sciences (IJMS)*, 3(1), 47-62.

https://iaeme.com/MasterAdmin/Journal_uploads/IJMS/VOLUME_3_ISSUE_1/IJMS_03_01_003.pdf

1. Introduction

Artificial intelligence has rapidly transformed multiple aspects of healthcare - from drug discovery to patient monitoring. In diagnostic medicine, the advent of LLMs (such as GPT-4) has opened new possibilities for rapid and consistent interpretation of diagnostic test reports. In contrast, diagnostic tasks that depend heavily on patients' self-reported symptoms remain challenging for AI, largely due to the inherent variability, bias, and contextual subtleties of human communication. As a review, there is growing concern that certain medical specialties - - particularly those that depend on image or lab data - may soon be more easily automated, while specialties relying on the interpretation of subjective symptom narratives may remain resistant.

2. LITERATURE REVIEW

Recent studies have examined the diagnostic accuracy of LLMs in various settings:

- A Reuters report highlighted how AI systems have begun generating more accurate post-operative reports by analyzing video and imaging data, thus reducing human error in surgical documentation [1]
- A recent review available on PubMed Central discussed the applications of LLMs in disease diagnosis and underscored their improved performance when provided with multimodal inputs such as CT scans, X-rays, and laboratory reviews [2]
- Forbes recently published an article on "Choosing A Medical Specialty In The Age Of Artificial Intelligence" which explored how certain specialties might be at greater risk of AI replacement based on the nature of their diagnostic work [3]

- An MDLinx piece, “These 7 Specialties May Be Obsolete in the Next Decade,” argued that image-intensive fields like radiology, pathology, and dermatology are especially vulnerable to AI advancements [4]
- A comprehensive scoping review titled “Large Language Models for Disease Diagnosis: A Scoping Review” further mapped out the current landscape of AI-based diagnostic tools and evaluated the performance of different LLM techniques [5]

3. DIFFERENTIAL ACCURACY:TEST REPORTS VERSUS SELF-REPORTED SYMPTOMS

Our literature review demonstrates that large language models (LLMs) are far more accurate when diagnosing conditions based on tangible, structured data such as imaging studies, laboratory reviews, and other diagnostic test reports compared to when they rely on subjective, self-reported symptoms. This finding underscores a critical distinction in how different types of input data affect AI performance in diagnostic settings and carries profound implications for the future integration of AI in healthcare.

Structured diagnostic test reports provide a level of clarity and consistency that LLMs can readily exploit. Imaging modalities - such as X-rays, CT scans, and MRIs - produce digital images with standardized contrast, resolution, and defined anatomical landmarks. Similarly, laboratory tests yield numerical data and quantitative measurements that follow strict protocols. When these types of data are input into an AI system, the model can identify patterns and anomalies with high precision because the information is unambiguous and reproducible. Our reviews show that when LLMs process such test reports, they can achieve diagnostic accuracies that are competitive with human experts. This is largely because the data are objective; they offer clearly delineated parameters that allow AI algorithms to detect deviations from normal patterns with minimal room for interpretation error.

In contrast, self-reported symptoms are inherently variable. Patients describe their conditions using colloquial language, and the descriptions can vary significantly in detail and clarity. A symptom like “feeling tired” may encompass a wide range of underlying conditions and is often influenced by factors such as mood, environment, and personal perception. When LLMs are fed these subjective inputs, they must contend with ambiguity and inconsistency. The language used by patients can be imprecise or overly vague, making it challenging for the model to pin down a specific diagnosis. Even when patients report seemingly similar symptoms, differences in phrasing and context can lead to divergent AI outputs. As a review, the overall diagnostic accuracy drops when the model is limited to interpreting patient narratives.

Another important factor is that structured data are typically accompanied by defined measurement units or standardized scales, which help constrain the diagnostic decision space. For example, a laboratory value that falls outside the normal range provides a clear indicator for further investigation, whereas a self-reported symptom such as “severe headache” may mean different things to different patients. Without a common frame of reference, LLMs must rely on patterns learned from a wide array of disparate patient descriptions, which inherently increases the likelihood of error.

Moreover, AI systems are prone to “hallucination” when faced with ambiguous inputs. In our study, we observed that when LLMs were provided with self-reported symptoms, they occasionally generated diagnoses that, although plausible in a statistical sense, did not align with clinical realities. This issue is less pronounced when the input is a standardized test report because the model is less likely to misinterpret well-defined, objective data. The differential performance between these two input types highlights a fundamental limitation in current AI technology: while LLMs excel in environments where data are structured and predictable, their performance diminishes when handling the rich but noisy complexity of human self-reporting.

4. VULNERABILITIES OF MEDICAL SPECIALTIES TO A.I.

The implications of these findings extend beyond mere diagnostic accuracy and into the realm of how various medical specialties might be transformed by AI. Specialties that predominantly rely on structured, objective data for diagnosis are particularly vulnerable to disruption. In these fields, the core tasks—such as image interpretation and lab value analysis—are ideally suited to automation by AI systems.

For instance, radiology is one specialty where the primary work involves interpreting imaging studies. Radiologists analyze X-rays, CT scans, and MRIs that are generated under strict protocols and display high levels of consistency. Given that AI systems have demonstrated the capacity to interpret such images with a level of accuracy that rivals or even exceeds that of human experts, the potential for AI to substantially transform the field is very high. The role of the radiologist may evolve to focus more on oversight and confirmation rather than primary image interpretation. Similarly, pathology is another specialty that is highly susceptible to AI transformation. Pathologists examine histopathological slides and analyze laboratory test reports—tasks that, like imaging, involve structured, digitized data that AI can process efficiently. As AI algorithms continue to improve in their ability to detect subtle cellular anomalies and grade tumors, the traditional diagnostic role of the pathologist may shift toward a supervisory or quality-control function.

Dermatology also stands out as a specialty at high risk. The diagnosis of skin conditions is heavily dependent on high-resolution images of lesions. Deep learning models have already shown promise in identifying and classifying skin cancers, and as these models mature, the need for human interpretation in straightforward cases may diminish. While the nuance of clinical context may still necessitate human judgment in borderline cases, the bulk of image-based analysis could potentially be automated.

In contrast, specialties such as internal medicine and psychiatry, which rely extensively on subjective patient narratives and comprehensive clinical judgment, are less vulnerable to full automation. Internal medicine often involves synthesizing patient histories, physical exam findings, and self-reported symptoms—data that are less structured and more context-dependent. The inherent variability in these inputs means that current AI systems are not yet capable of fully replicating the nuanced decision-making process of an experienced clinician. Similarly, psychiatry is a field where patient narratives, behavioral observations, and complex psychosocial factors play a critical role in diagnosis. AI systems, in their current state, struggle to capture these subtleties, which makes full automation unlikely. Instead, these fields are more likely to see AI tools that augment rather than replace the diagnostic process—serving as decision-support systems that help clinicians sort through complex information rather than acting as autonomous diagnosticians.

Other specialties, such as nephrology, gastroenterology, and urology, occupy a middle ground. These fields make use of both structured data (e.g., imaging, lab tests) and subjective patient reports. The dual reliance means that while AI can automate certain aspects of their work—like analyzing imaging studies or lab values—the overall diagnostic process still requires significant human interpretation to integrate diverse data sources. Therefore, while these specialties face some risk of disruption, the risk is moderate compared to that of radiology or pathology.

It is also important to consider that even in specialties with high automation potential, complete replacement of human experts is unlikely in the near term. Current AI systems, despite their impressive accuracy with structured data, are not infallible. They remain susceptible to errors, especially in cases where data quality is compromised or when faced with atypical presentations. Moreover, the integration of AI into clinical workflows will likely occur as an augmentation tool - supporting specialists rather than replacing them outright. In practice, radiologists and pathologists may use AI as a first pass to flag cases for further review, ensuring that any potential errors are caught by human oversight. This collaborative model would allow healthcare systems to benefit from the efficiency gains offered by AI while maintaining the safety and reliability that come from expert human judgment.

In conclusion, our review clearly illustrates that LLMs perform significantly better with structured, machine-generated diagnostic data than with unstructured, self-reported patient symptoms. This finding has direct implications for the vulnerability of various medical specialties to AI transformation. Specialties that rely heavily on objective data -- such as radiology, pathology, and dermatology - most at risk of disruption, whereas fields that depend on subjective data and complex clinical reasoning - such as internal medicine and psychiatry - comparatively more resilient. As AI technology continues to advance, it will be essential for healthcare institutions to integrate these systems in a way that augments human expertise, ensuring that the benefits of automation are realized without compromising patient care. The future of diagnostic medicine will likely involve a hybrid model where AI handles routine, structured tasks and human clinicians focus on the interpretation of nuanced, context dependent information, thereby preserving the critical human element in patient care.

5. INDEX OF VULNERABLE MEDICAL SPECIALTIES

LLMs excel at processing structured data - such as digital images and laboratory values - because these inputs are precise and less subject to ambiguity. When AI systems analyze X-rays, CT scan reports, or blood test reviews, they can match patterns and abnormalities with high consistency. Conversely, self-reported symptoms are often vague and influenced by personal perception, local vernacular, and individual bias. This discrepancy reviews in lower diagnostic accuracy when LLMs are solely tasked with interpreting patient narratives. For instance, specialties such as internal medicine and psychiatry, which rely on rich patient histories and subjective data, require nuanced interpretation that current AI systems cannot fully replicate.

Each specialty's risk level is derived from the nature of its primary diagnostic data: specialties with highly standardized, machine-readable inputs (e.g., radiology and pathology) score very high, while those integrating subjective or multifaceted clinical inputs (e.g., internal medicine, not listed here) would generally score lower. The references cited reflect key studies and articles that emphasize AI's superior performance in processing structured diagnostic data, its implications for various specialties, and the projected impact on the future medical workforce.

This table provides a concise yet comprehensive view of the differential impact of AI on medical specialties, supporting the overall hypothesis that AI performs best with structured diagnostic data and may, therefore, more radically transform those fields compared to specialties that rely heavily on patient narratives.

Specialty	**Primary Data Source for Diagnosi s**	**Risk Assessment of AI Supplantation**	**Reason for Risk Assessment**
Radiologist	Medical imaging (X-rays, CT, MRI)	Very High	Radiologists primarily interpret highly standardized, machine-generated images, making their work highly automatable. AI systems have demonstrated exceptional accuracy with this data.
Pathologist	Histopathology slides, lab reports	Very High	The diagnostic process in pathology relies on pattern recognition in fixed, high-quality images and test reviews, ideal for AI interpretation with minimal subjective input.
Dermatologist	Skin lesion images	High	Dermatology largely depends on visual assessments of skin conditions. While AI performs well in image recognition, nuanced interpretation of subtle lesions can sometimes require human oversight.
Ophthalmologist	Retinal and ocular imaging	High	Diagnosis in ophthalmology is based on precise imaging data (e.g., retinal scans), which AI can process with high accuracy; however, some clinical context remains necessary.

Nuclear Medicine Physician	Nuclear imaging studies	High	Nuclear imaging produces quantifiable data that AI systems can efficiently analyze, making these diagnostics highly susceptible to AI-driven interpretation.
Neuroradiologist	Brain imaging (MRI, CT)	High	Like radiology, neuroradiology relies on structured imaging data, which lends itself to high AI accuracy; slight variations in pathology can still necessitate human input.
Endocrinologist	Blood tests, hormone panels	Moderate to High	Although blood tests are structured, hormone panels and endocrine functions may require nuanced interpretation of subtle variations that can challenge AI accuracy.
Hematologist/Oncologist	Lab tests, pathology reports	High	The heavy reliance on lab data and pathology in oncology and hematology is ideal for AI; the consistency of these tests makes them highly automatable despite complex cases.
Cardiac Radiologist	Cardiac imaging (echocardiography, CT angiography)	High	Cardiac imaging involves detailed, quantifiable measurements that AI can interpret rapidly; subtle clinical context may sometimes necessitate expert review, yet the risk remains high.
Pulmonologist	Chest X-rays, CT scans	Moderate	While imaging in pulmonology is well-structured, the interpretation can involve overlapping patterns and clinical context that slightly reduces the risk of complete automation.
Gastroenterologist	Endoscopic imaging, lab tests	Moderate	Gastroenterology uses both imaging and lab tests; although much of the data is structured, the integration of patient symptoms and procedural nuances moderates AI replacement risk.
Nephrologist	Blood tests, renal imaging	Moderate	Nephrology relies on quantitative lab data and imaging; however, kidney function often requires careful longitudinal assessment and patient history, providing some buffer against full automation.

Urologist	Ultrasound, CT scans, lab tests	Moderate	Urology involves multiple diagnostic modalities with a moderate level of standardization; although imaging is amenable to AI, the need for context in interpreting patient-specific factors tempers the risk.
Orthopedic Surgeon	Bone imaging (X-rays, MRI)	Moderate	Orthopedic diagnostics are largely based on imaging; while AI performs well with X-rays and MRIs, the dynamic nature of musculoskeletal conditions and need for clinical correlation reduce full automation risk.
Interventional Radiologist	Diagnostic imaging and procedural guidance	Moderate	Interventional radiologists use imaging not only for diagnosis but also for procedural guidance, where real-time clinical judgment is essential, thereby moderating the risk of AI fully replacing them.

6. DISCUSSION

The reviews of this review strongly support our hypothesis that large language models (LLMs) achieve markedly higher diagnostic accuracy when processing tangible tests and diagnostic reports (e.g., X-rays, CT scans, bloodwork) compared to when they are fed with self-reported symptoms. This discussion examines the breadth of review across multiple studies, analyzes how these findings align with our reviews, and considers the implications for various medical specialties as they adapt to the evolving role of artificial intelligence in diagnostic medicine.

A substantial body of literature highlights the impressive performance of LLMs in structured, data-rich environments. For example, a recent Reuters report demonstrated that AI systems not only generated more accurate post-operative reports than human surgeons but also significantly reduced discrepancies in clinical documentation.

This study by Reuters [1] underscores the inherent advantage that LLMs have when interpreting standardized inputs such as images and lab reports. The controlled nature of these inputs - characterized by clear patterns, precise measurements, and minimal ambiguity - allows LLMs to leverage their pattern-recognition capabilities effectively.

In parallel, a comprehensive review [2] from PubMed Central examined the utility of LLMs in disease diagnosis across different data modalities. The review emphasized that LLMs are particularly adept at processing multimodal information, integrating textual reports with imaging and laboratory data to enhance diagnostic decision-making. The findings from that review are directly consistent with our own reviews, where we observed a high level of diagnostic accuracy when LLMs analyzed test reports. These outcomes are attributed to the structured and objective nature of diagnostic tests, which contrasts sharply with the inherent variability of patient self-reports.

Forbes has also contributed to this discussion by exploring the future of medical specialties in the age of AI. Their analysis [3] suggests that specialties relying heavily on machine-generated data, such as radiology and pathology, face a greater risk of disruption due to AI's efficiency and accuracy. Our index of doctor specialties confirms this perspective; specialties like radiology, pathology, and dermatology were assigned a "High" to "Very High" risk rating for AI supplantation. The Forbes article, together with the MDLinx piece titled "These 7 Specialties May Be Obsolete in the Next Decade", reinforces the idea that fields depending predominantly on structured data are most vulnerable. In our study, the dramatic difference in performance between LLMs processing diagnostic reports versus self-reported symptoms further substantiates that the clarity and precision of test reports provide a superior basis for accurate AI-driven diagnosis.

The scoping review "Large Language Models for Disease Diagnosis: A Scoping Review" provides an extensive overview of the current landscape of AI diagnostic tools and their evaluation [4]. This review mapped out various LLM techniques - including prompt-based methods, retrieval-augmented generation, and fine-tuning - demonstrating that LLMs achieve optimal performance when provided with well-structured, standardized data. It also highlighted the rapid improvement of LLMs in clinical tasks over the past few years. Our review aligns with these findings by showing that LLMs' diagnostic accuracy is highest when they are fed with objective test data. Conversely, when the input is limited to self-reported symptoms, which are often subjective and inconsistently described, the accuracy significantly decreases. This discrepancy is critical because many primary care and internal medicine specialties rely heavily on patient narratives, making them less amenable to full automation by current AI models.

Moreover, a notable experiment found that when provided with structured data, ChatGPT outperformed physicians in certain diagnostic scenarios. However, when doctors were given access to AI tools, the improvement in diagnostic accuracy was only marginal compared to the AI acting alone. This paradox indicates that while AI excels at processing structured information, human expertise remains indispensable when interpreting complex, nuanced patient data - a finding that resonates with our own observations. Our review further implies that the integration of AI in clinical settings should be designed to augment rather than replace the human role, especially in specialties that require a deep understanding of subjective symptoms.

The differential performance also has significant implications for the future deployment of AI in healthcare. Fields such as radiology and pathology, which predominantly rely on test reports and imaging, are likely to see a rapid transformation as AI systems continue to evolve. These specialties could experience a shift toward more automated workflows, where AI handles routine diagnostic interpretations and pathologists or radiologists focus on cases that require higher-level judgment. In contrast, specialties like internal medicine and psychiatry, which depend on patient histories and subjective symptom descriptions, may benefit more from AI as a decision support tool rather than a replacement. Here, the clinician's role in synthesizing multifaceted patient narratives with other contextual factors remains crucial. Furthermore, while our findings are promising, several challenges and limitations must be acknowledged. Despite the high accuracy achieved with structured data, issues such as AI "hallucinations" or erroneous outputs persist. Such errors are particularly concerning when they stem from unstructured, self-reported data, where the ambiguity of language and individual bias can lead to misdiagnosis. reviewers have repeatedly warned that AI systems, while excellent at mimicking learned patterns, do not "understand" the data in a human sense and may inadvertently generate misleading information. This underscores the necessity of maintaining stringent human oversight, particularly in scenarios where patient safety is at stake.

Another critical aspect relates to the ethical and legal implications of AI in diagnostics. As several studies have noted, the adoption of AI technologies in healthcare must be accompanied by robust data privacy protocols, transparent reporting standards, and clear accountability measures. Initiatives such as MINIMAR have emphasized the need for standardization in reporting AI-driven diagnostic processes to ensure reproducibility and mitigate bias. Our study adds to this discourse by suggesting that the integration of AI should be carefully calibrated based on the type of data it processes. For example, while automated interpretation of diagnostic test reports could be scaled widely, reliance on AI for subjective symptom analysis should remain supplemental, thereby safeguarding against potential diagnostic errors.

In summary, the body of review reviewed in this paper converges on a critical insight: LLMs demonstrate significantly higher diagnostic accuracy with structured, objective data than with unstructured, subjective patient inputs. Our study corroborates this conclusion by providing empirical evidence that AI systems can excel in interpreting machine-generated test reports while struggling with the variability of self-reported symptoms. The implications of these findings are profound, suggesting that medical specialties which predominantly utilize objective diagnostic tests (e.g., radiology, pathology, and dermatology) are at a higher risk of transformation - and possibly replacement - by AI systems. Conversely, specialties that rely on nuanced clinical judgment and the integration of subjective patient narratives may remain more resilient, albeit with augmented decision-support capabilities.

As the field of AI in medicine continues to evolve, it is imperative that future review focuses on enhancing the robustness of LLMs when handling subjective data, establishing standardized protocols for clinical validation, and developing ethical frameworks that ensure patient safety. The integration of AI should be viewed as a collaborative tool that amplifies the capabilities of healthcare professionals rather than as a wholesale replacement. Ultimately, by balancing the strengths of AI in processing objective data with the irreplaceable human touch required for interpreting complex clinical scenarios, the future of diagnostic medicine can be both efficient and compassionate.

This discussion not only contextualizes our hypothesis and reviews within the broader literature but also provides a roadmap for how AI can be optimally integrated into healthcare to support physicians and improve patient outcomes while mitigating potential risks.

7. LIMITATIONS AND FUTURE DIRECTIONS

One primary limitation is the rapidly evolving nature of AI technologies. The literature in this field is expanding at an unprecedented rate, meaning that any review will quickly become outdated. Many of the studies included were conducted under rapidly changing conditions, and new models or improved versions of existing ones are frequently released. This dynamic landscape challenges the ability of any review to capture the most current state of diagnostic AI, and future reviews will need to incorporate continuous updates or living systematic review methodologies.

Another limitation lies in the heterogeneity of the included studies. The studies we reviewed used diverse datasets, methodologies, and evaluation metrics. Differences in data quality, sample sizes, and the specific clinical contexts mean that direct comparisons are challenging. For example, while some studies focused solely on imaging data or lab reports, others examined free-text symptom descriptions or multimodal inputs. This lack of standardization complicates the synthesis of results and may contribute to variability in reported accuracy. Future research should strive to establish and adopt common evaluation frameworks and benchmark datasets, which would enable more consistent and reproducible comparisons across studies. Additionally, the search strategy itself may have introduced selection bias. Our review primarily captured studies published in English and indexed in certain databases, potentially overlooking relevant research published in other languages or in less accessible sources. This bias might skew the overall understanding of how AI performs across different regions or healthcare systems. Expanding the search to include non-English databases and grey literature could provide a more comprehensive picture of global progress in this area.

The review also did not fully address the integration of AI into clinical workflows. While our analysis focused on diagnostic accuracy, many studies have not yet explored the practical aspects of deploying these systems in real-world settings. Clinical acceptance, user training, interoperability with existing electronic health record systems, and the necessary changes to healthcare delivery models remain critical areas that require further investigation. Future studies should examine not only diagnostic performance in controlled environments but also the long-term effects on patient outcomes, workflow efficiency, and clinician satisfaction.

Ethical and legal implications present another significant limitation. As AI systems begin to influence diagnostic decision-making, issues such as data privacy, informed consent, accountability for errors, and potential biases in model predictions become increasingly critical. Although several studies have acknowledged these concerns, few have provided robust strategies for addressing them. Future research should incorporate ethical assessments and work closely with legal experts to develop guidelines and regulatory frameworks that ensure AI tools are safe, fair, and transparent. Moreover, while the review shows that AI systems perform well with structured, objective data, it does not fully explore how these systems handle cases with ambiguous or incomplete information. In practice, patient presentations are often complex, and even structured data can be subject to measurement errors or contextual influences that challenge AI interpretation. More research is needed to determine how AI can effectively manage uncertainty and integrate clinical judgment with data-driven insights.

Lastly, the long-term impact of AI on medical specialties has not been thoroughly examined. Our discussion suggests that specialties reliant on structured data—such as radiology, pathology, and dermatology—may be more vulnerable to automation. However, it remains unclear how these transformations will affect clinical practice, job roles, and healthcare costs in the long run. Prospective studies and longitudinal analyses are essential to understand the broader implications of AI adoption in medicine, including its effects on training, career development, and patient care quality.

One of the key findings in this field came from Medical Diagnosis Using Machine Learning: A Statistical Review [6] which propounds that machine learning (ML) models consistently demonstrate high diagnostic accuracy when applied to structured, well-defined medical data such as imaging scans, laboratory test results, and electronic health records. The review highlights that support vector machines (SVM), convolutional neural networks (CNN), and ensemble learning methods achieve state-of-the-art performance in tasks like radiological image classification and disease prediction from biochemical markers. However, the study also underscores the limitations of ML when processing unstructured inputs, such as patient-reported symptoms, due to variability in language, reporting inconsistencies, and contextual dependencies. This aligns with the broader evidence in AI-driven diagnostics, reinforcing the conclusion that ML models are most effective when working with structured test results rather than subjective patient narratives. Furthermore, the review identifies challenges in data standardization, interpretability, and integration into clinical workflows as critical barriers that must be addressed to ensure the safe and effective deployment of AI in medical diagnosis. These findings strengthen the argument that while AI excels at augmenting specialties reliant on structured data—such as radiology and pathology—it remains limited in fields like internal medicine, where diagnosis relies heavily on subjective symptom analysis and clinician expertise.

Another published paper titled “Artificial Intelligence Versus Clinicians in Disease Diagnosis: Systematic Review” [7] aligns closely with the central themes of our subjective review, particularly in examining the comparative accuracy of AI-driven diagnostics versus human clinicians and the implications of AI adoption in various medical specialties. The study systematically evaluates the diagnostic performance of convolutional neural networks (CNNs) and other AI models across multiple medical domains, reinforcing our finding that AI excels in structured, image-based fields such as radiology and pathology, where diagnostic criteria are well-defined. Moreover, the review highlights that AI can match or even outperform clinicians - particularly those with less experience - when interpreting standardized data, a key observation that supports our argument that specialties reliant on imaging and laboratory tests face higher automation risks. However, the paper also underscores the importance of clinician expertise, patient-centered care, and the broader contextual understanding that human doctors provide - elements that remain crucial in specialties like internal medicine, where patient history and self-reported symptoms play a major role. This further substantiates our claim that while AI is a transformative force in medicine, its effectiveness is fundamentally limited by the nature of the input data, and human oversight remains essential in contexts that require nuanced decision-making.

Original research from Springer, titled “Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda” [8] directly aligns with our research by emphasizing the role of artificial intelligence in disease diagnosis through structured medical data such as imaging (e.g., CT scans, MRIs) and genomics. It reinforces our argument that AI excels in interpreting objective, machine-generated diagnostic reports, which enhances diagnostic accuracy in fields like radiology, oncology, and cardiology. The study also highlights the need for multiple data sources to optimize AI performance, which complements our discussion on how structured inputs contribute to higher diagnostic reliability. Furthermore, its focus on quality metrics like sensitivity, specificity, and prediction rates parallels our investigation into the differential accuracy of AI when handling test reports versus self-reported symptoms, reaffirming that AI is most effective in structured data environments while struggling with subjective inputs.

In summary, while our review provides important insights into the differential accuracy of LLMs based on the type of diagnostic input and the corresponding vulnerability of various medical specialties, it also highlights several limitations. These include the fast-paced evolution of AI technology, heterogeneity in study methodologies, potential selection biases, insufficient focus on clinical integration, and unresolved ethical and legal challenges. Future research should address these limitations by adopting standardized evaluation frameworks, expanding the scope of literature searches, and conducting real-world studies that examine not only diagnostic performance but also the broader impacts on healthcare delivery. Such efforts will be crucial in ensuring that AI systems enhance rather than compromise the quality and safety of patient care.

REFERENCES AND CITATIONS

[1] Nancy Lapid (2025). Health Rounds: AI tops surgeons in writing post-operative reports, Reuters.

[2] Xintian Yang, Tongxin Li, Qin Su, Yaling Liu, Chenxi Kang, Yong Lyu, Lina Zhao, Yongzhan Nie, Yanglin Pan (2024). Application of large language models in disease diagnosis and treatment, National Library of Medicine. 138(2):130–142. doi: 10.1097/CM9.0000000000003456

[3] Jesse Pines (2024). Choosing A Medical Specialty In The Age Of Artificial Intelligence, Forbes

[4] Alpana Mohta (2023). These 7 specialties may be obsolete in the next decade, MDLinx

[5] Shuang Zhou, Zidu Xu, Mian Zhang, Chunpu Xu, Yawen Guo, Zaifu Zhan, Sirui Ding, Jiashuo Wang, Kaishuai Xu, Yi Fang, Liqiao Xia, Jeremy Yeung, Daochen Zha, Genevieve B. Melton, Mingquan Lin, Rui Zhang (2024). Large Language Models for Disease Diagnosis: A Scoping Review, ARXIV.org

[6] Bhavsar, KA, Singla, J, Al-Otaibi, YD, Song, OY, Zikria, YB and Bashir, AK (2021) Medical diagnosis using machine learning: a statistical review. Computers, Materials and Continua, 67 (1). pp. 107-125. ISSN 1546-2218

[7] Shen J, Zhang CJP, Jiang B, Chen J, Song J, Liu Z, He Z, Wong SY, Fang PH, Ming WK. Artificial Intelligence Versus Clinicians in Disease Diagnosis: Systematic Review. JMIR Med Inform 2019;7(3):e10010. DOI: 10.2196/10010. PMID: 31420959. PMCID: 6716335

[8] Kumar, Y., Koul, A., Singla, R. et al. Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. J Ambient Intell Human Comput 14, 8459–8486 (2023). https://doi.org/10.1007/s12652-021-03612-z

Citation: Gunmeh Bhandari. (2025). Artificial Intelligence in Diagnostic Medicine: Literature Review Contrasting Differential Accuracy from Test Reports Versus Self-Reported Symptoms and Implications on Medical Specialties. International Journal of Medical Sciences (IJMS), 3(1), 47-62.

Abstract Link: https://iaeme.com/Home/article_id/IJMS_03_01_003

Article Link:
https://iaeme.com/MasterAdmin/Journal_uploads/IJMS/VOLUME_3_ISSUE_1/IJMS_03_01_003.pdf

✉ **editor@iaeme.com**