Uluslararası İleri Doğa Bilimleri ve Mühendislik Araştırmaları Dergisi
Sayı 8, S. 102-108, 2, 2024
© Telif hakkı IJANSER'e aittir
Araştırma Makalesi

https://as-proceeding.com/index.php/ijanser
ISSN: 2980-0811

International Journal of Advanced Natural Sciences and Engineering Researches
Volume 8, pp. 102-108, 2, 2024
Copyright © 2024 IJANSER
Research Article

Evaluation Of Medical Diagnosis Capabilities Of Three Artificial Intelligence Models – ChatGPT-3.5, Google Gemini, Microsoft Copilot

Yordanka Eneva*, Bora Doğan

Department of Physics and Biophysics, Faculty of Pharmacy, Medical University of Varna "Prof. Dr. Paraskev Stoyanov", Bulgaria

Email of corresponding author: yordanka.eneva@mu-varna.bg

(Received: 05 March 2024, Accepted: 08 March 2024)

(4th International Artificial Intelligence and Data Science Congress ICADA 2024, March 14-15, 2024)

ATIF/REFERENCE: Eneva, Y. & Doğan, B. (2024). Evaluation Of Medical Diagnosis Capabilities Of Three Artificial Intelligence Models-ChatGPT-3.5, Google Gemini, Microsoft Copilot. *International Journal of Advanced Natural Sciences and Engineering Researches, 8(2), 102-108.*

Abstract – The widespread adoption of artificial intelligence (AI) in various domains, including medicine, has prompted extensive research into its diagnostic capabilities. This study conducts a comparative analysis of three prominent AI models – ChatGPT-3.5, Microsoft Copilot, and Google Gemini – to evaluate their performance in medical diagnosis. Clinical vignettes from Texas Tech University Health Sciences Center were utilized to assess the accuracy and precision of the AI models in diagnosing internal medicine cases. Results indicate that ChatGPT-3.5 achieved the highest accuracy rate, correctly diagnosing 70.59% of cases, outperforming Google Gemini and Microsoft Copilot. While all models demonstrated the potential to assist in diagnosis, variations in approach and performance were observed. ChatGPT-3.5 provided concise answers without explicitly stating its lack of medical expertise, while Google Gemini and Microsoft Copilot acknowledged their limitations but offered more detailed explanations and recommendations. Statistical analysis, conducted using the chi-square test for independence revealed significant differences in diagnostic capabilities among the AI models, emphasizing the importance of careful selection in clinical decision-making. This study contributes valuable insights into the application of AI in medical diagnosis and underscores the need for continued refinement of AI models to enhance diagnostic accuracy and support healthcare professionals in delivering optimal patient care.

Keywords – Artificial Intelligence, Medical Diagnosis, ChatGPT-3.5, Microsoft Copilot, Google Gemini, Diagnostic Capabilities, Comparative Analysis

I. INTRODUCTION

The opening of artificial intelligence (AI) to mass use in early 2023 has raised fundamental questions about its application across various human activities. Numerous studies have since been initiated, as many of them are related to the use of AI in various fields of medicine, exploring its potential in assessments and analyses in the health system [1], [2], prevention, screening, prediction of the development of diseases and their treatment, drug development, etc. [3]–[8]. Specialists, including doctors and researchers from various healthcare sectors, are actively engaged in this research.

AI has shown promise in medical diagnostics by expediting the diagnostic process and reducing the time required. It can also provide patients with information about their disease progression, potential drug interactions, side effects, and other pertinent concerns. Moreover, AI can assist in prescribing personalized regimens, diets, exercise routines, work schedules, and rest patterns, while monitoring adherence throughout the therapeutic process.

Research in this area is particularly valuable for determining the validity and accuracy of AI diagnosis. This provoked our team in the present study to determine the diagnostic capabilities of three widely available AI models – ChatGPT-3.5, Microsoft Copilot, and Google Gemini. We selected these models because they are freely accessible and open for mass utilization, thus catering to a broad user base. While AI cannot supplant the role of healthcare professionals, it can significantly complement and streamline their work.

II. MATERIALS AND METHOD

A. *Background*

In November 2022, ChatGPT, a "Generative Pre-trained Transformer" - language model with artificial intelligence developed by OpenAI was launched for free use. It started with the GPT electronic system with a transformer architecture, and a neural network, published in 2017 [9]. The first version, known as GPT-1, was introduced in June 2018. It demonstrated the efficiency when using different data sets and these models have been shown to be able to generate coherent and contextually appropriate user-understandable text. Development was followed by GPT-2 in 2019, which handled 1.5 billion parameters, and then by GPT-3, officially released in June 2020, featuring 175 billion parameters. GPT-3 demonstrates remarkable abilities in solving the tasks of understanding and generating natural language text [10]. Since then, it has been continuously updated and improved. Today, everyone can use the free model ChatGPT-3.5, which is accessed through an account, or ChatGPT-4 for a fee.

Google Gemini is a family of large language models (LLM) developed by Google DeepMind, open-sourced in December 2023. Gemini models can understand and process information from various modalities, including text, code, images, audio, and video [11], and are accessible through a Google account.

Microsoft Copilot is an AI-based productivity tool launched in 2021 that integrates with popular Microsoft 365 apps like Word, Excel, PowerPoint, Outlook, and Teams. Copilot uses advanced language models to provide intelligent suggestions and help. It works with data from Microsoft Graph, improving the user experience in various applications. [12].

Artificial intelligence has found various applications in the daily lives of many people, from daily entertainment to carrying out various professional activities and consulting. Many studies show its potential application in the medical field as well – training of medical students in various specialties, patient charting, diagnostics and consultations related to the treatment of patients, tracking the stages of treatment, etc. [13]. The use of AI in making diagnoses has made a significant contribution. It makes it possible to track the degree of accuracy and reliability of various AIs and the possibility of their use in medical practice by both doctors and patients. These tools can analyze and understand people's conditions and offer clinical solutions [14], [15].

B. *Meta-analysis*

We began the present study with a meta-analysis of the published literature related to AI medical diagnosis following the guidelines for meta-analyses (PRISMA). For this purpose, we used data from Scopus, PubMed, and Web of Science for the period until February 1, 2024. The study started with the keyword “AI”. Due to a large number of articles over 1,000,000, we limited it to the keyword “medicine”. The volume of articles turned out to be too much again. In them, there are various medical studies, a large part of which is in the field of education of various medical professionals and students. One of the objects of our research is ChatGPT and its possibilities for making medical diagnoses. This gave us the reason to use the expression “ChatGPT made medical diagnosis”. After this limit, we set aside 260 titles for consideration. Due to unpublished summaries and different types of diagnostics, some were dropped and our result decreased to 213. After examining the summaries, we discovered that not all of them were related to the study of ChatGPT's capabilities in medical diagnoses. Thus, more titles were dropped and we limited ourselves to examining 52 full-text articles related to the evaluation of the diagnostic abilities of this artificial intelligence. A map (Fig. 1), visualizing them, was made with VOSviewer_1.6.20_exe. A keyword search was conducted for Microsoft Copilot and Google Gemini, but no results were found in the specified databases.

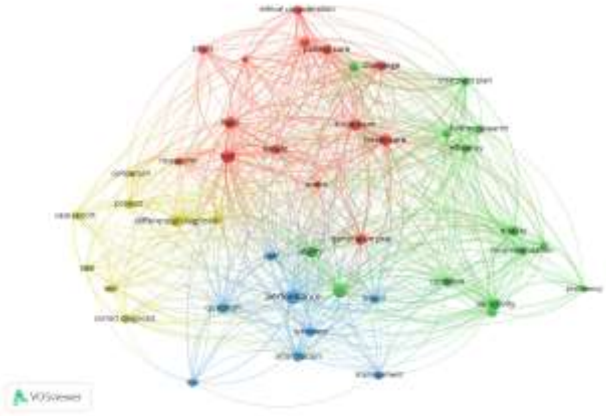

Fig. 1 Mapping of the scientific articles that appeared with keywords “ChatGPT made medical diagnosis” in Scopus, PubMed, and Web of Science for the period until 01.02.2024.

C. *Purpose of the study*

The purpose of the present study is to conduct a comparative analysis of the results of a test study of the medical diagnosis abilities of the three artificial intelligences ChatGPT-3.5, Microsoft Copilot, and Google Gemini. In other words, the aim was to evaluate the performance of the AI models for the given diagnoses and to compare the results in terms of the accuracy and precision of the answers. We utilized 34 clinical vignettes with varying difficulty levels as inputs, asking the AI to determine the most likely diagnosis.

D. *Data Source and Workflow*

The prompts for the AIs were sourced from Texas Tech University Health Sciences Center (TTUHSC), comprising 101 clinical vignettes encompassing various medical inquiries within the realm of Internal Medicine [16]. These vignettes engage readers to discern answers, articulate reasoning, and apply medical expertise across different stages of the diagnostic process. Ranging from inquiries about the most likely diagnosis to recommendations for management strategies, these questions prompt readers to evaluate patient observations, suggest pertinent laboratory tests, and consider optimal treatment options, including antibiotic choices. To assess and compare the performance of the AIs in determining the correct diagnosis based on the provided clinical symptoms and the results from functional and blood tests, we narrowed down our focus to 34 clinical vignettes where a specific diagnosis is requested, with the corresponding correct answer provided in the answer sheet.

The process of creating inputs for the AIs involved the direct transfer of clinical vignettes from TTUHSC into the AI models, adhering strictly to their sequential order. Each clinical scenario was introduced into a fresh chat session to maintain the integrity of the analysis, ensuring that prior cases did not influence the outcomes. This approach aimed to isolate each clinical vignette, allowing the AI models to evaluate and respond to the presented medical challenges independently. The same question was asked in all cases: “What is the most likely diagnosis?”.

To comprehensively account for response variation and enhance the reliability of the AI assessments, each clinical vignette underwent a testing protocol. Specifically, the same clinical scenario was subjected to three consecutive tests conducted in a distinct chat session. Importantly, no modifications were made to the text between these testing sessions, preserving the consistency and integrity of the experimental setup. This process not only provided insights into the performance of the AI models but also served to validate their diagnostic capabilities across multiple iterations of analysis.

Finally, the chi-square test for independence was conducted in R (version 4.3.2; 2023 The R Foundation for Statistical Computing), using the chisq.test() function to determine if there is a significant association between the AI model used and the distribution of responses. This test helped us assess whether there are significant differences in the performance of the AI models in terms of medical diagnosis capabilities.

III. RESULTS

Overall, ChatGPT accurately diagnosed 24 out of 34 clinical vignettes, achieving an accuracy rate of 70.59%. This performance surpassed that of other AI models; Google Gemini achieved 21 correct diagnoses out of 34, resulting in an accuracy rate of 61.76%, while Microsoft Copilot trailed behind with only 12 correct diagnoses out of 34, yielding an accuracy rate of 35.29% (Fig. 2). Google Gemini typically generates three answer variants, referred to as drafts. Through observation, we noted instances where the correct diagnosis appeared in one draft while the remaining two provided incorrect diagnoses or none at all. Consequently, we classified an answer as correct and reflected it to the accuracy rate if the appropriate diagnosis appeared in at least one draft or among the potential diagnoses listed in a differential diagnosis.

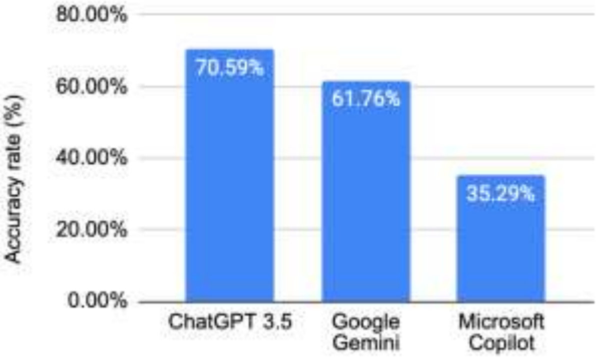

Fig. 2 Accuracy rates (%) of ChatGPT-3.5, Google Gemini (overall of three drafts), and Microsoft Copilot in determining the correct diagnosis out of 34 clinical vignettes

In addition to comparing accuracy rates, we categorized the responses received from the AI models into four distinct groups: Correct, Partially Correct, Wrong, and No Answer. This classification was devised to account for instances where the AI generated a differential diagnosis (listing multiple potential diagnoses instead of a definitive one) or declined to assist, citing its lack of expertise as a chatbot. Accordingly, we classified a response as Correct when it exclusively contained the accurate diagnosis, while Partially Correct encompassed scenarios where the correct diagnosis was present within a differential diagnosis. Responses were labeled as Wrong if they provided an incorrect diagnosis or a differential diagnosis lacking the correct one. Finally, responses that received no diagnosis were categorized as No Answer (Fig. 3).

ChatGPT-3.5 did not generate a differential diagnosis unless directly prompted and provided a diagnosis for all 34 clinical vignettes. In contrast to other AI models, it did not explicitly state that it is not a medical specialist. Instead, it solely focused on the specific question asked, providing its reasoning behind the choice briefly and without offering additional recommendations. Overall, the generated answers were shorter than those generated by the other two AI models in most of the cases.

In most cases, Google Gemini stated that it was not a medical expert and thus refused to assist us, directing us to a medical specialist. However, when a medical diagnosis was provided, it clearly explained the reasoning behind the decision, highlighting the presenting complaints and the test results of the hypothetical patient. Moreover, in many cases, it provided additional treatment recommendations and pointers on how to verify the diagnosis. However, most of the answers provided by the three drafts were counted as Partially Correct since the correct diagnosis was part of a differential diagnosis.

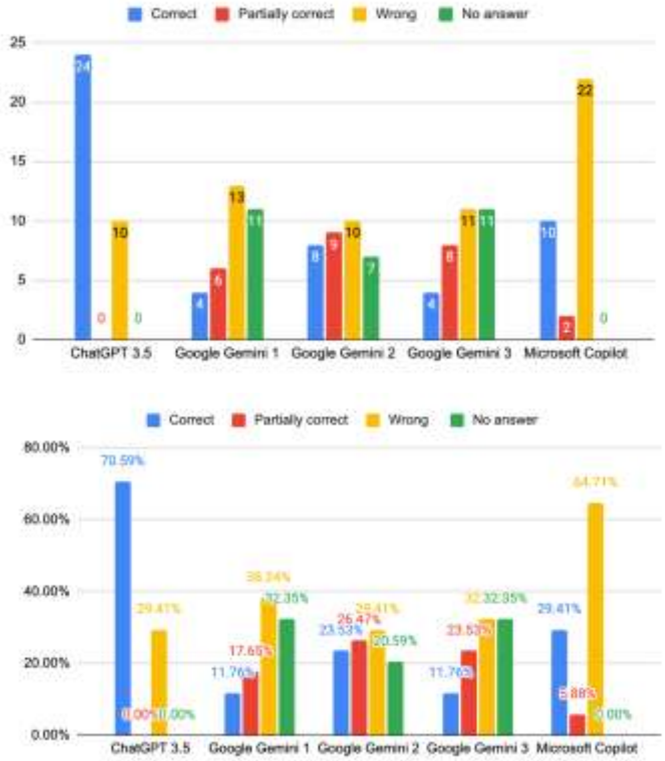

Fig. 3 Performance distribution of ChatGPT-3.5, Google Gemini (each of the three drafts presented separately)), and Microsoft Copilot, showing absolute frequencies (upper figure) and relative frequencies in percentages (lower figure)

Microsoft Copilot, which trails behind in overall accuracy, also underlined in most cases that it is not a medical specialist and stressed the necessity of consulting a real healthcare professional. However, like ChatGPT-3.5 in all cases, it provided a diagnosis and explained the reasoning behind the choice clearly. In some cases, similar to Google Gemini, it provided additional information about management strategies and methods for verifying the diagnosis. Unlike the other AI models, it also included hyperlinks where users could read more about the discussed issues.

The results of the chi-square test (Table 1) revealed a significant association between the choice of AI model and the distribution of responses in diagnosing medical conditions ($\chi^2 = 68.16$, df = 12, $p < 0.001$). This indicates that the performance of the AI models (ChatGPT-3.5, Google Gemini Draft 1, Google Gemini Draft 2, Google Gemini Draft 3, and Microsoft Copilot) varied significantly in providing accurate diagnoses across a set of clinical vignettes. With a p-value of 7.064e-10, well below the conventional significance level of 0.05, we reject the null hypothesis, suggesting that there are notable differences in their diagnostic capabilities. These findings underscore the importance of careful evaluation and selection when incorporating AI models into medical decision-making processes, as certain models may outperform others in providing accurate diagnoses.

Table 1. The results of the chi-squared test of independence, indicating a significant difference between the diagnostic capabilities of the AI models

	X-squared	df	p-value
Pearson's Chi-squared test	68.16	12	7.06E-10

To compare the diagnostic performance of Google Gemini Draft 1, Google Gemini Draft 2, and Google Gemini Draft 3 pairwise chi-square tests were conducted. The rationale behind this analysis was to discern any potential improvements or variations in diagnostic accuracy across the iterations of the Google Gemini. However, the results revealed adjusted p-values of 1 for all three comparisons, indicating that none of the pairwise differences reached statistical significance after adjusting for multiple comparisons using the Bonferroni correction. In other words, we fail to reject the null hypothesis for all three pairwise comparisons. This suggests that there is insufficient evidence to conclude that there are significant differences in diagnostic performance between Google Gemini Draft 1, Google Gemini Draft 2, and Google Gemini Draft 3 based on the given data.

Additionally, we conducted a post-hoc analysis to compare the diagnostic performance of two prominent AI models, ChatGPT-3.5 and Google Gemini By applying a chi-square test with Bonferroni correction, we aimed to determine whether there were significant differences in the proportions of correct diagnoses generated by ChatGPT-3.5 and Google Gemini AI models. The results revealed a statistically significant disparity ($\chi^2 = 8.3251$, df = 1, p-value = 0.00391), indicating that ChatGPT-3.5 exhibited superior diagnostic accuracy compared to Google Gemini.

IV. DISCUSSION

The study is unique in that it is the first to evaluate the diagnostic capabilities of Microsoft Copilot and Google Gemini, as the findings of this study shed light on the diagnostic capabilities of three prominent AI models in the realm of medical diagnosis. Notably, ChatGPT-3.5 demonstrated the highest accuracy rate among the three models, correctly diagnosing 70.59% of the clinical vignettes. This performance surpasses that of Google Gemini and Microsoft Copilot, indicating its potential as a valuable tool in supporting medical professionals in diagnostic processes.

One significant observation is the difference in approach among the AI models. ChatGPT-3.5, while not explicitly stating its lack of medical expertise, focused solely on answering the specific question posed, without providing additional information or recommendations. In contrast, both Google Gemini and Microsoft Copilot acknowledged their limitations as non-medical specialists but offered more detailed explanations, additional treatment recommendations, and methods for verifying diagnoses.

The results of the chi-square test underscore the significant differences in diagnostic capabilities among the AI models. This highlights the importance of careful consideration when selecting AI tools for medical decision-making processes. Healthcare professionals must weigh factors such as accuracy, comprehensiveness, and user-friendliness when integrating AI into clinical practice.

V. CONCLUSION

In conclusion, this study contributes to the growing body of research exploring the application of AI in medical diagnosis. The findings highlight the promising diagnostic capabilities of ChatGPT-3.5, Microsoft Copilot, and Google Gemini, with ChatGPT-3.5 demonstrating the highest accuracy rate among the three models. Future research should focus on further refining AI models to enhance diagnostic accuracy and reliability. Additionally, investigating the impact of AI integration on clinical decision-making processes and patient outcomes would provide valuable insights into the real-world application of AI in healthcare settings. Overall, the continued advancement and evaluation of AI technologies hold promise for improving diagnostic processes and ultimately enhancing patient care.

REFERENCES

[1] Tse Chiang Chen, Emily Kaminski at all, „Chat GPT as a Neuro-Score Calculator: Analysis of a Large Language Model's Performance on Various Neurological Exam Grading Scales", World Neurosurgery, Vol. 179, 2023, Pages e342-e347

[2] Hongyan Wang, WeiZhen Wu, at all, "Performance and exploration of ChatGPT in medical examination, records and education in Chinese: Pave the way for medical AI", International Journal of Medical Informatics, Volume 177, September 2023, 105173

[3] Patel, V.; Shah, M. A comprehensive study on artificial intelligence and machine learning in drug discovery and drug development. Intell. Med. 2021. [Google Scholar] [CrossRef]

[4] Nakamura, T.; Sasano, T. Artificial intelligence and cardiology: Current status and perspective. J. Cardiol. 2022, 79, 326–333. [Google Scholar] [CrossRef]

[5] Muthalaly, R.G.; Evans, R.M. Applications of Machine Learning in Cardiac Electrophysiology. Arrhythm Electrophysiol. Rev. 2020, 9, 71–77. [Google Scholar] [CrossRef] [PubMed]

[6] Asha, P.; Srivani, P.; Ahmed, A.A.A.; Kolhe, A.; Nomani, M.Z.M. Artificial intelligence in medical Imaging: An analysis of innovative technique and its future promise. Mater. Today Proc. 2021, 56, 2236–2239. [Google Scholar] [CrossRef]

[7] Yao, L.; Zhang, H.; Zhang, M.; Chen, X.; Zhang, J.; Huang, J.; Zhang, L. Application of artificial intelligence in renal disease. Clin. Ehealth 2021, 4, 54–61. [Google Scholar] [CrossRef]

[8] Van den Eynde, J.; Lachmann, M.; Laugwitz, K.-L.; Manlhiot, C.; Kutty, S. Successfully Implemented Artificial Intelligence and Machine Learning Applications In Cardiology: State-of-the-Art Review. Trends Cardiovasc. Med. 2022. [Google Scholar] [CrossRef] [PubMed]

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, Illia Polosukhin, „Attention is all you need", „Advances in neural information processing systems", 2017.

[10] "Introducing ChatGPT." Accessed: Feb. 24, 2024. [Online]. Available: https://openai.com/blog/chatgpt

[11] "Gemini - Google DeepMind." Accessed: Feb. 24, 2024. [Online]. Available: https://deepmind.google/technologies/gemini/#introduction

[12] "Your Everyday AI Companion | Microsoft Bing." Accessed: Feb. 24, 2024. [Online]. Available: https://www.microsoft.com/en-us/bing

[13] Khan RA, Jawaid M, Khan AR, Sajjad M. ChatGPT—reshaping medical education and clinical management. Pak J Med Sci. 2023; 39(2): 605-607. . [Google Scholar] [Web of Science] [PubMed]

[14] Amann J at all, "To explain or not to explain?-Artificial intelligence explainability in clinical decision support systems" PLOS Digit Health. 2022 Feb 17;1(2):e0000016. doi: 10.1371/journal.pdig.0000016. PMID: 36812545; PMCID: PMC9931364.

[15] Vasey B et all, "Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI", Nature Medicine, (2022) 28, pages924–933

[16] J. Agrimor et al., Interesting Clinical Vignettes: 101 Ice Breakers for Medical Rounds. Texas Tech University Health Sciences Center (TTUHSC). Accessed: Feb. 09, 2024. [Online]. Available: https://www.ttuhsc.edu/clinical-research/documents/JMD-Cases-of-Interest.pdf