

# Artificial intelligence (ChatGPT) ready to evaluate ECG in real life? Not yet!

Volkan Çamkıran , Hüseyin Tunç, Batool Achmar , Tuğçe Simay Ürker, İlhan Kutlu and Akin Torun

## Abstract

**Objective:** This study aims at evaluating if ChatGPT-based artificial intelligence (AI) models are effective in interpreting electrocardiograms (ECGs) and determine their accuracy as compared to those of cardiologists. The purpose is therefore to explore if ChatGPT can be employed for clinical setting, particularly where there are no available cardiologists.

**Methods:** A total of 107 ECG cases classified according to difficulty (simple, intermediate, complex) were analyzed using three AI models (GPT-ECGReader, GPT-ECGAnalyzer, GPT-ECGInterpreter) and compared with the performance of two cardiologists. The statistical analysis was conducted using chi-square and Fisher exact tests using scikit-learn library in Python 3.8.

**Results:** Cardiologists demonstrated superior accuracy (92.52%) compared to ChatGPT-based models (GPT-ECGReader: 57.94%, GPT-ECGInterpreter: 62.62%, GPT-ECGAnalyzer: 62.62%). Statistically significant differences were observed between cardiologists and AI models ( $p < 0.05$ ). ChatGPT models exhibited enhanced performance with female patients; however, the differences found were not statistically significant. Cardiologists significantly outperformed AI models across all difficulty levels. When it comes to diagnosing patients with arrhythmia (A) and cardiac structural disease ECG patterns, cardiologists gave the best results though there was no statistical difference between them and AI models in diagnosing people with normal (N) ECG patterns.

**Conclusions:** ChatGPT-based models have potential in ECG interpretation; however, they currently lack adequate reliability beyond oversight from a doctor. Additionally, further studies that would improve the accuracy of these models, especially in intricate diagnoses are needed.

## Keywords

ChatGPT, artificial intelligence, electrocardiogram (ECG), diagnostic accuracy, machine learning general

Submission date: 14 November 2024; Acceptance date: 11 February 2025

## Introduction

Over the past few years, artificial intelligence (AI) technology has been applied in the health sector, and as such it has garnered a lot of interest, particularly in the field of cardiology.<sup>1</sup> The interpretation of electrocardiograms (ECGs) is of paramount importance in the diagnosis and treatment of cardiovascular disease. It is anticipated that AI-based systems will have the capacity to automate and accelerate this assessment.<sup>2</sup> For this reason, there is an urgent need to investigate the effectiveness of state-of-the-art AI

models like ChatGPT by OpenAI for translating ECGs from academic and clinical perspectives.

The objective of this study is to conduct a comprehensive analysis of the efficacy of ChatGPT in ECG evaluation

Bahçeşehir Üniversitesi Hastanesi Medical Park Göztepe, İstanbul, Turkey

### Corresponding author:

Volkan Çamkıran, Bahçeşehir Üniversitesi Hastanesi Medical Park Göztepe, Merdivenköy, G-118 Sk., 34732, Kadıköy/İstanbul, Turkey 34732.  
Email: wolkan36@gmail.com



and to assess its potential for success in this domain. Details regarding its training set can be found in “Methods” section. Moreover, the limitations of this study are addressed in the conclusion section. We are motivated by our personal experiences in conducting this research. Previous trials demonstrated that the accuracy of ECG results interpreted by ChatGPT was inconsistent, with significant variability across different readings, leading to fluctuations between high and low accuracy.<sup>3–6</sup> The present situation demands immediate interrogation on whether AI can be trusted in reading ECGs as well as methods through which it may be enhanced further.

The application of AI technologies holds considerable promise for advancements in the medical field, including cardiology. In recent years, there has been a notable increase in the utilization of AI and machine learning (ML) technologies in medicine. It is of the utmost importance that these technologies provide reliable and accurate results in clinical applications.

One significant area of research within medicine is the automatic analysis and interpretation of ECG data in cardiology. Rapid and accurate diagnosis of life-threatening conditions such as ECG evaluation should be a subject of intense scrutiny regarding AI success. A literature review reveals a multitude of studies examining the performance of AI in ECG evaluation. A substantial body of evidence indicates that ML algorithms can achieve high levels of accuracy in the interpretation of ECGs. For example, Hannun et al. reported that deep learning models perform at a comparable level to human cardiologists in recognizing arrhythmias in ECG data.<sup>7</sup> Similarly, Rajpurkar et al. reported high accuracy and precision in the analysis of ECG data.<sup>8</sup> Moreover, Attia et al. showed that AI could identify heart failure using ECG data, implying that AI has various uses in ECG analysis.<sup>9</sup>

However, the majority of these investigations are based on outcomes of models trained on large datasets. This limits their ability to evaluate a variety of ECG data observed in real-world clinical practice. In particular, there is a scarcity of literature that investigates the effectiveness of Large Language Models (LLMs) such as ChatGPT when it comes to medical domain specificities. In the meantime, some studies also point out that AI faces struggle when it comes to assessing ECGs. For example, Ribeiro et al. remarked that AI models may yield erroneous results due to shortcomings in the quality of the training data and the lack of diversity in the data set.<sup>10</sup> Thus, AI has important implications, especially for clinical applications.

The efficacy of ChatGPT in ECG assessment is of significant clinical importance, particularly in instances where cardiologists are unavailable. Emergency room doctors and other physicians working in remote locations can make crucial decisions with the assistance of an AI tool that is capable of accurate and expedient ECG assessment. However, the consequences may be grave if

ChatGPT fails to interpret ECGs accurately. Incorrect ECG assessments may result in misdiagnosis and inappropriate treatment, which could have adverse effects on patient outcomes. Hence, it is crucial to assess how well ChatGPT performs currently to find out whether it can be employed in hospitals or not as suggested earlier.

The hypothesis for this research is that the current version of ChatGPT does not have enough accuracy to assess ECG data. If this supposition is confirmed, the use of ChatGPT by emergency physicians or other physicians working in the periphery where cardiologists are not available will not be beneficial in its current form. Therefore, the AI performance in this area needs to be enhanced. In the future, an AI tool that would make more precise assessments and become more appropriate for clinical may play a significant role in clinical decision support systems.

## Methods

This study was approved by the Human Research Ethics Committee of the Istinye University Medical Faculty on 14 June 2024, protocol number: 24-132. A total of 107 ECG cases from the textbook by Hampton et al. were selected for analysis, with particular attention paid to ensuring that similar examples were not overrepresented.<sup>11</sup> Consent was deemed unnecessary because the ECG pictures were taken from a published book and are already publicly available. They also do not contain identifying personal data. The selected ECGs were classified according to the difficulty levels as simple, intermediate, and complex, and these levels were recorded in the study. The subsequent classification explains the distribution of the 107 ECG cases as mentioned in the 150 ECG Cases book.<sup>11</sup> A medical student should be able to handle simple scenarios, and anybody who has read *The ECG Made Easy* can correctly respond to them. Simple patterns such as usual arrhythmias, normal sinus rhythm, or abnormalities fall under this category. House officers, specialist nurses, or paramedics are expected to be able to respond to moderate cases, especially if they have read *The ECG in Practice*. This involves an intermediate level of competence for the interpretation of the ECG, like recognizing any structural defect, irregularity in the rhythm or ischemic changes.

Finally, complex scenarios are designed to put applicants preparing for the Membership of the Royal Colleges of Physicians exam to the test. They require advanced ECG interpretation skills as well as the ability to manage rare illnesses, overlapping anomalies, or subtle patterns. Using a tiered approach ensures that the difficulty levels correlate with the target audience’s expected level of ability.

The ECG reading engines mentioned in this article are housed on OpenAI’s platform and may be accessed through the Explore GPTs function of the ChatGPT application. Users must go to the ChatGPT interface (available through OpenAI’s platform or the ChatGPT app), choose

the “Explore GPTs” option, and then search for ECG-related utilities such as “ECG Reader.”

To access these models, the user may need to have a membership to OpenAI’s premium service (e.g. ChatGPT Plus for GPT-4 capabilities).

In order to train the ECG reading engines, OpenAI created models that utilized a variety of easily accessible publicly or semi-publicly available information, including guideline documents, textbooks, medical literature, and other sources that may be relevant to the healthcare domain. The training included aspects related to arrhythmias, cardiovascular medicine, and electrocardiograms’ (ECG) reading. This includes the ability to recognize patterns of arrhythmia, conduction abnormalities, ischemia, or even infarction.

As part of OpenAI’s training process, datasets containing medical advice, research papers, and textbooks from European Society of Cardiology, American Heart Association (AHA), and other institutions, were compiled together and training included such datasets. It has to be remembered that there was no supervision of the training conducted by medical boards or physicians. In the absence of such active clinical oversight, various data patterns were analyzed with powerful machine-learning algorithms.

The relevant clinical data in respect to the age and sex of the ECG sample were properly noted on the appropriate fields. After which, chatbot (ChatGPT-40) underwent analysis of the selected ECG samples with a view of providing a clinical interpretation (open-ended question). On ChatGPT-40 platform, we chose three AI models that are “GPT-ECGReader” “GPT-ECGAnalyzer” and “GPT-ECGInterpreter” since they have high usage rates. The images of the 12-lead ECG, accompanied by the corresponding questions, were captured using the iPhone 14 Pro Max and subsequently uploaded into each model. The results of the ECG assessments were recorded for each application. The clarity of the responses was assessed by one cardiologist who examined the pictures and responses based on visual interpretation and clinical experience. The cardiologist used the book 150 ECG PROBLEMS<sup>11</sup> as a standard to determine the correctness of the answers. Figure 1 shows how the pictures were uploaded in addition to the response of each model. The figure also reveals how organized and well-structured the responses were which facilitated evaluation and aided in comprehension. Two groups each containing half of the ECG samples were randomized, where two different cardiologists independently and blindly assessed each group, respectively. The two cardiologists who conducted the evaluation were both professors with 20 years of experience, one specializing in electrophysiology and the other in interventional cardiology. The evaluated ECG results were categorized into five distinct groups. The classification system used was as follows: N=normal, A=arrhythmia, I=ischemia, NI=arrhythmia + ischemia, CSD=cardiac structural disease.

Based on the leads of the ECG, ST-elevation myocardial infarction or non-ST-elevation MI was categorized among those that showed evidence for myocardial infarction. Atrial fibrillation was found to be an abnormal rhythm present in some ECGs as well as atrioventricular block, ventricular tachycardia—V tachyarrhythmias include ventricular fibrillation, supraventricular tachycardia—SVT with wide complex tachycardias or PSVT that conducts over an accessory pathway or associates with preexcitation patterns, complete left bundle branch block, and complete right bundle branch block. Another group was identified as having structural changes where left ventricular hypertrophy and right ventricular hypertrophy cases were found on ECG. The control group was selected based on the criteria of the study and then compared with AI application’s responses as well as cardiologists’ answers.

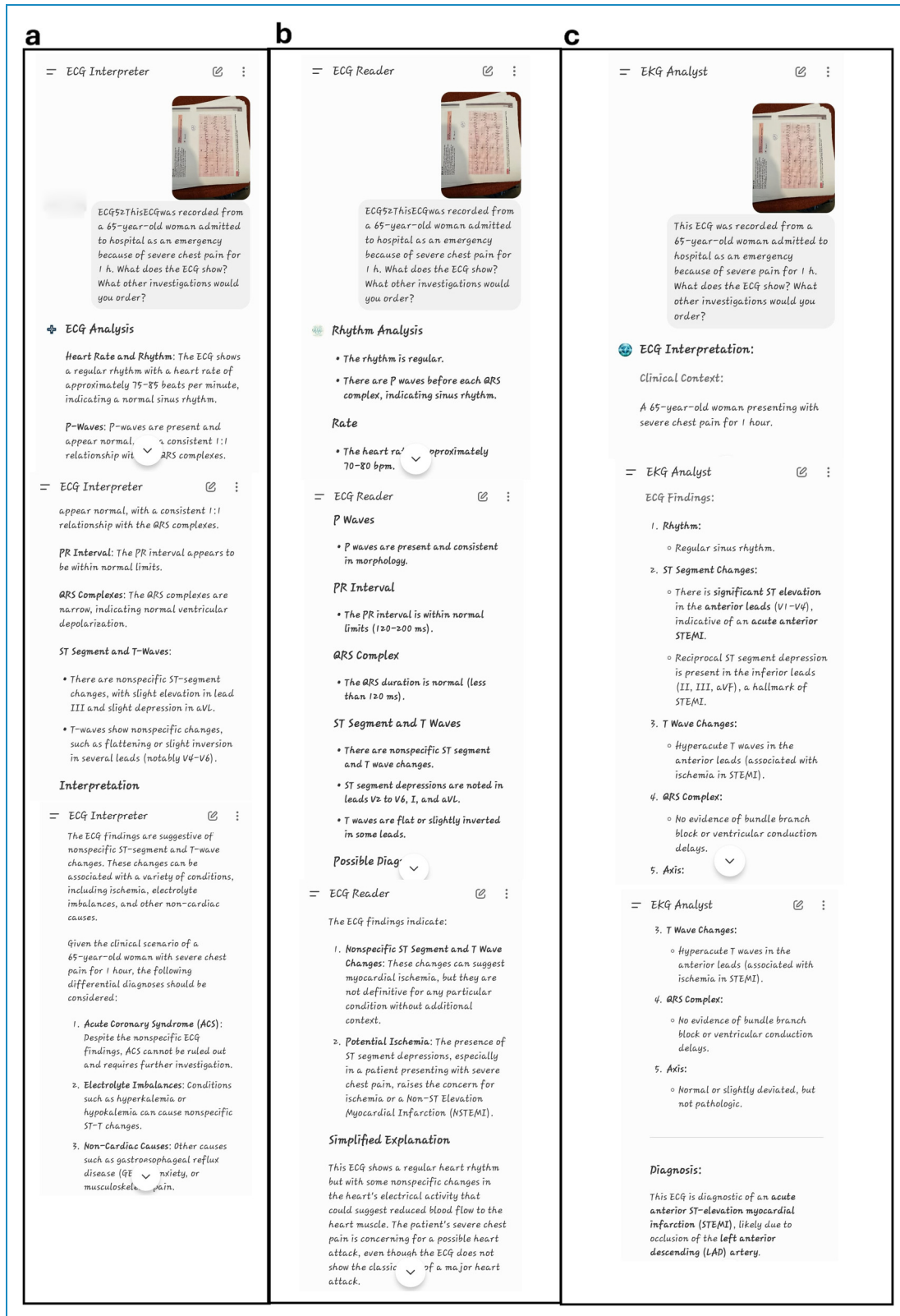
The STAGER checklist<sup>12</sup> was followed to standardize the structure and reporting of this study.

### Statistical analysis

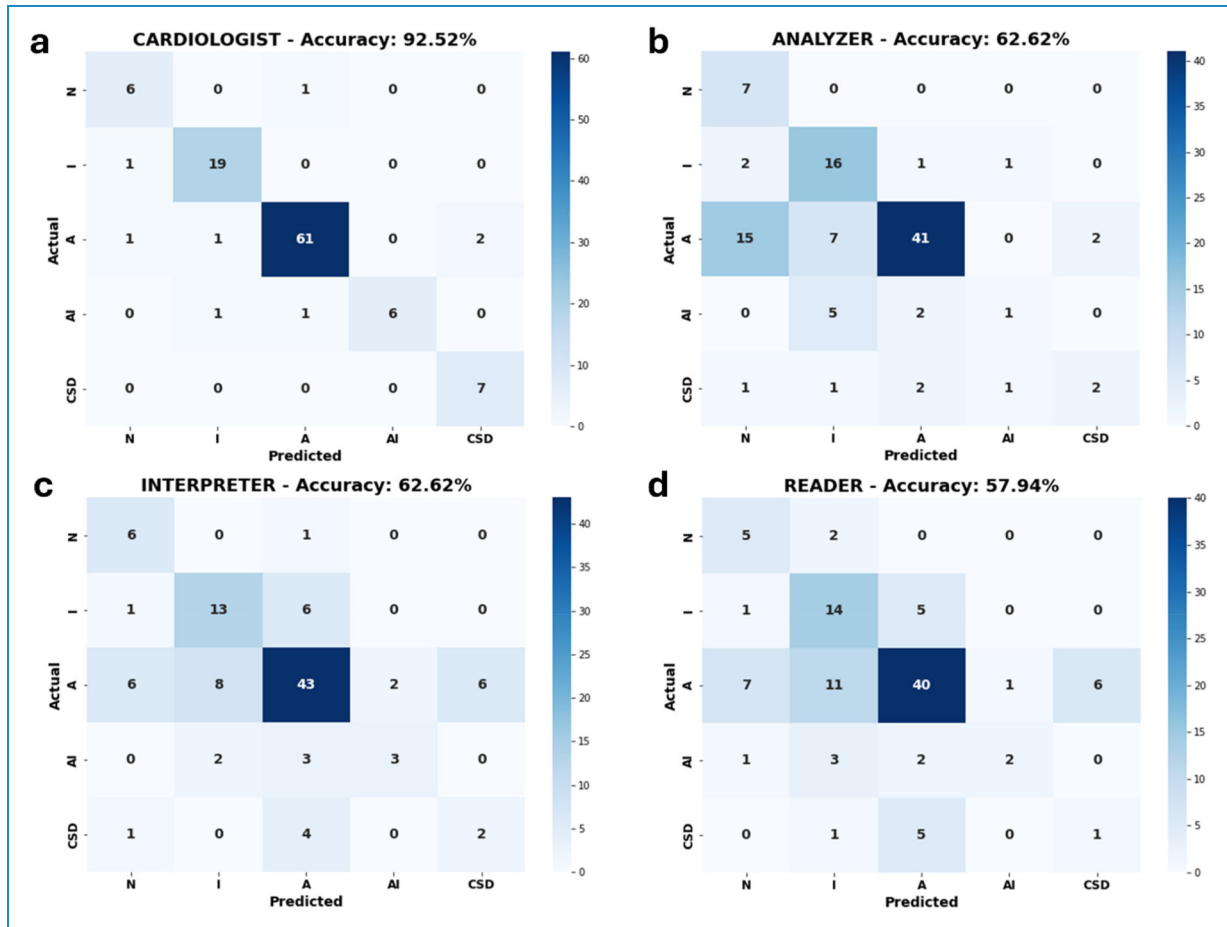
Statistical data analysis was conducted using the scikit-learn library in Python 3.8. Contingency tables were created for each method pair to compare their correct diagnosis success rates. To compare cardiologists’ success rates with LLMs for each diagnosis type, corresponding contingency table pairs were generated. Similarly, contingency tables were produced to observe the effects of difficulty level and patient gender on the accuracy of the methods. After generating the contingency tables, chi-square or Fisher’s exact tests were performed based on the conditions required for the chi-square test. A critical  $p$ -value ( $p$ ) of 0.05 was chosen, and contingency tables yielding  $p < 0.05$  were considered to indicate statistically significant differences between classification approaches. We employed a two-proportion post hoc power analysis to examine the adequacy of our sample size for detecting the difference in accuracy between cardiologists and ChatGPT-based models. Effect sizes were computed using Cohen’s  $h$ , and the power values were estimated using a normal-approximation method for two proportions.

### Results

Figure 2 illustrates the accuracy and misclassification rates of the diagnostic methods for each diagnostic category. In the context of cardiology, the high degree of accuracy observed in the classification of ECG images was accompanied by the identification of some erroneous diagnoses made by ChatGPT-based models. To illustrate, when cardiologists correctly identified the A class, the ChatGPT-based GPT-ECGAnalyzer model incorrectly classified it as the N class, resulting in 15 erroneous diagnoses. Similarly, the GPT-ECGInterpreter model also misclassified the A class



**Figure 1.** Shows how 12-lead ECG photos were entered to ChatGPT model. The right diagnosis according to 150 ECG PROBLEMS book is (acute anterolateral myocardial infarction). (a) GPT-ECGInterpreter diagnosis is nonspecific ST-segment and T-wave changes. (b) GPT-ECGReader diagnosis is nonspecific ST-segment and T-wave changes and potential ischemia. (c) GPT-ECG Analyzer diagnosis is acute anterior ST-elevation myocardial infarction (STEMI).



**Figure 2.** Confusion matrices including number of actual and predicted images for each diagnosis class related to the classification approaches: (a) CARDIOLOGIST, (b) GPT-ECGAnalyzer, (c) GPT-ECGInterpreter, and (d) GPT-ECGReader. Accuracy metric indicates the correctly classified ECG images within all images.

Note: N = normal, A = arrhythmia, I = ischemia, AI = arrhythmia + ischemia, CSD = cardiac structural disease.

as the I class, leading to six incorrect diagnoses. The figure demonstrates that AI models particularly misclassify classes A and CSD, while performing best in predicting class N. For instance, out of seven CSD patterns, the GPT-ECGReader model correctly identified only one and mislabeled five of them as class A.

Table 1 presents a comparison of diagnostic methods for ECG images in pairs. Significant differences were observed between cardiologists (92.52%) and ChatGPT-based GPT-ECGReader (57.94%), GPT-ECGInterpreter (62.62%), and GPT-ECGAnalyzer (62.62%) models. In all these comparisons,  $p$ -values were found to be below 0.05, indicating that cardiologists were significantly more accurate in their diagnoses.

Table 2 presents the post hoc power and effect size analysis results. Based on our post hoc power analysis, the difference between the cardiologist and each of the ChatGPT-based models (GPT-ECGReader, GPT-ECGInterpreter, and GPT-ECGAnalyzer) was statistically substantial, with a difference in diagnostic accuracy that could significantly

**Table 1.** Pairwise comparison of the ECG image diagnosis methods.

Method-1 (Accuracy %)	Method-2 (Accuracy %)	$p$ -value
GPT-ECGReader (57.94%)	CARDIOLOGIST (92.52%)	<b>&lt;0.001</b>
GPT-ECGInterpreter (62.62%)	CARDIOLOGIST (92.52%)	<b>&lt;0.001</b>
GPT-ECGAnalyzer (62.62%)	CARDIOLOGIST (92.52%)	<b>&lt;0.001</b>
GPT-ECGReader (57.94%)	GPT-ECGInterpreter (62.62%)	0.576
GPT-ECGReader (57.94%)	GPT-ECGAnalyzer (62.62%)	0.576
GPT-ECGInterpreter (62.62%)	GPT-ECGAnalyzer (62.62%)	1.00

Note: Chi-square test was performed to quantify the  $p$ -values. The bold values represent  $p$ -values lower than 0.05.



**Table 2.** Power and effect size comparisons between AI models and cardiologists.

Comparison	Sample size ( <i>n</i> )	Cohen's <i>h</i> (effect size)	Post hoc power
GPT-ECGReadervs. Cardiologist	107	0.857	1.000
GPT-ECGInterpretersvs. Cardiologist	107	0.762	1.000
GPT-ECGAnalyzersvs. Cardiologist	107	0.762	1.000

impact patient care. The cardiologist's diagnostic accuracy (92.5%) was significantly higher than that of the three AI models, which ranged from 57.9% to 62.6%. Cohen's *h* values comparing the cardiologist to each ChatGPT-based model ranged from 0.762 to 0.857, indicating a large effect size (with the positive sign reflecting the cardiologist's superior performance). Furthermore, the post hoc power of all comparisons was 1.000, demonstrating that the sample size of 107 ECG cases was more than sufficient to detect these significant gaps in diagnostic accuracy. These results highlight the current limitations of ChatGPT-based models in accurately interpreting ECGs and emphasize the continued importance of expert human interpretation in clinical practice.

Table 3 presents a comparison of the accuracy rates of the methods used for each diagnostic category. It is evident that cardiologists exhibited superior performance in categories A ( $p \leq 0.001$ ) and CSD ( $p < 0.05$  for all three comparisons). It should be noted, however, that models based on ChatGPT have also achieved acceptable levels of accuracy in the N and I categories. Nevertheless, these results indicate that cardiologists generally perform better.

Table 4 compares the accuracy rates of the methods in question according to the difficulty level of the ECG images. In all difficulty levels (simple, intermediate, and complex), cardiologists demonstrated significantly superior performance compared to ChatGPT-based models. For instance, at the simple level, cardiologists achieved a 95.56% accuracy rate, while the GPT-ECGReader model attained a 62.22% accuracy rate ( $p \leq 0.001$ ). The sole exception is the comparison between the GPT-ECGAnalyzer and the cardiologist in the complex group ( $p = 0.137$ ), which indicates a difference but not to a statistically significant extent. Moreover, notable differences were not observed between different ChatGPT methods.

Table 5 presents a comparison of the accuracy rates of the methods according to the patients' gender. The results indicate that cardiologists have a 95.71% accuracy rate when diagnosing male patients, which is statistically

significant ( $p < 0.05$  for all comparisons). In the case of female patients, the discrepancy is less pronounced. Cardiologists achieved an accuracy rate of 88.57%, while the GPT-ECGReader model attained a rate of 71.43% ( $p = 0.133$ ). These results demonstrate that cardiologists achieve high accuracy rates for both sexes. Nevertheless, models based on ChatGPT have demonstrated inferior performance in male patients. As illustrated in the lower table, the ChatGPT model performs better when classifying the female group (although this is not meaningful), and therefore, when competing with the cardiologist for the female group, ChatGPT performs poorly and is unable to compete with the cardiologist. Moreover, significant differences in performance between different ChatGPT methods are not observed across the two sexes.

The objective of Table 6 is to demonstrate whether the performance differences between the methods are significant when stratified by gender. The GPT-ECGReader and GPT-ECGInterpreter models demonstrate higher accuracy rates for female patients compared to male patients; however, these differences are not statistically significant. In addition, cardiologists have achieved high accuracy rates in both sexes. The results demonstrate that there may be performance differences between male and female patients in ECG evaluations, yet these differences are not statistically significant.

The data show that ChatGPT-based models excellently perform in some diagnostic categories such as N and I but have less than desirable accuracy for others such as NI and CSD. Conversely, cardiologists scored high on all diagnostic categories. This indicates that the current form of ChatGPT is not reliable for complex and critical diagnoses of NI and CSD. Thus, ChatGPT-based models should be used carefully in clinical situations.

In general, women showed higher accuracy scores compared to men at all difficulty levels analyzed. Nonetheless, this was not much different statistically. In relation to females' ECGs analysis by the use of ChatGPT-based models, there was a slight improvement in performance, whereas in the case of male patients, the opposite trend has been observed, with a decline in performance.

Considering these limitations and the superior performance of cardiologists, it is evident that there is a need to enhance the capabilities of AI in this domain. Future research may enhance the accuracy of ChatGPT's ECG assessments, thereby facilitating the reliable and convenient use of this tool by emergency physicians and other health-care professionals in diverse settings.

## Discussion

This study examined the performance of ChatGPT-based AI models in ECG evaluation. The findings revealed that cardiologists demonstrated high accuracy rates across all diagnostic categories. However, ChatGPT-based models

**Table 3.** Pairwise comparison of the ECG image diagnosis methods according to each diagnosis class.

Diagnosis	Method-1 (Accuracy %)	Method-2 (Accuracy %)	p-value
N	GPT-ECGReader (71.43%)	CARDIOLOGIST (85.71%)	1.000
	GPT-ECGInterpreter (85.71%)	CARDIOLOGIST (85.71%)	1.000
	GPT-ECGAnalyzer (100.00%)	CARDIOLOGIST (85.71%)	1.000
	GPT-ECGReader (71.43%)	GPT-ECGInterpreter (85.71%)	1.000
	GPT-ECGReader (71.43%)	GPT-ECGAnalyzer (100.00%)	0.462
	GPT-ECGInterpreter (85.71%)	GPT-ECGAnalyzer (100.00%)	1.000
I	GPT-ECGReader (70.00%)	CARDIOLOGIST (95.00%)	0.091
	GPT-ECGInterpreter (65.00%)	CARDIOLOGIST (95.00%)	<b>0.043</b>
	GPT-ECGAnalyzer (80.00%)	CARDIOLOGIST (95.00%)	0.342
	GPT-ECGReader (70.00%)	GPT-ECGInterpreter (65.00%)	1.000
	GPT-ECGReader (70.00%)	GPT-ECGAnalyzer (80.00%)	0.716
	GPT-ECGInterpreter (65.00%)	GPT-ECGAnalyzer (80.00%)	0.480
A	GPT-ECGReader (61.54%)	CARDIOLOGIST (93.85%)	<b>&lt;0.001</b>
	GPT-ECGInterpreter (66.15%)	CARDIOLOGIST (93.85%)	<b>&lt;0.001</b>
	GPT-ECGAnalyzer (63.08%)	CARDIOLOGIST (93.85%)	<b>&lt;0.001</b>
	GPT-ECGReader (61.54%)	GPT-ECGInterpreter (66.15%)	0.715
	GPT-ECGReader (61.54%)	GPT-ECGAnalyzer (63.08%)	1.000
	GPT-ECGInterpreter (66.15%)	GPT-ECGAnalyzer (63.08%)	0.855
NI	GPT-ECGReader (25.00%)	CARDIOLOGIST (75.00%)	0.132
	GPT-ECGInterpreter (37.50%)	CARDIOLOGIST (75.00%)	0.315
	GPT-ECGAnalyzer (12.50%)	CARDIOLOGIST (75.00%)	<b>0.041</b>
	GPT-ECGReader (25.00%)	GPT-ECGInterpreter (37.50%)	1.000
	GPT-ECGReader (25.00%)	GPT-ECGAnalyzer (12.50%)	1.000
	GPT-ECGInterpreter (37.50%)	GPT-ECGAnalyzer (12.50%)	0.569
CSD	GPT-ECGReader (14.29%)	CARDIOLOGIST (100.00%)	<b>0.004</b>
	GPT-ECGInterpreter (28.57%)	CARDIOLOGIST (100.00%)	<b>0.021</b>
	GPT-ECGAnalyzer (28.57%)	CARDIOLOGIST (100.00%)	<b>0.021</b>

(continued)

Table 3. Continued.

Diagnosis	Method-1 (Accuracy %)	Method-2 (Accuracy %)	<i>p</i> -value
	GPT-ECGReader (14.29%)	GPT-ECGInterpreter (28.57%)	1.000
	GPT-ECGReader (14.29%)	GPT-ECGAnalyzer (28.57%)	1.000
	GPT-ECGInterpreter (28.57%)	GPT-ECGAnalyzer (28.57%)	1.000

Note: Fisher exact test was performed to quantify the *p*-values. The bold values represent *p*-values lower than 0.05.

Table 4. Pairwise comparison of the ECG image diagnosis methods according to difficulty levels of the ECG images.

Difficulty Level	Method-1 (Accuracy %)	Method-2 (Accuracy %)	<i>p</i> -value
Simple	GPT-ECGReader (62.22%)	GPT-ECGInterpreter (66.67%)	0.826**
	GPT-ECGReader (62.22%)	GPT-ECGAnalyzer (62.22%)	1.000**
	GPT-ECGReader (62.22%)	CARDIOLOGIST (95.56%)	<b>&lt;0.001*</b>
	GPT-ECGInterpreter (66.67%)	GPT-ECGAnalyzer (62.22%)	0.826**
	GPT-ECGInterpreter (66.67%)	CARDIOLOGIST (95.56%)	<b>&lt;0.001*</b>
	GPT-ECGAnalyzer (62.22%)	CARDIOLOGIST (95.56%)	<b>&lt;0.001*</b>
Intermediate	GPT-ECGReader (57.89%)	GPT-ECGInterpreter (63.16%)	0.814**
	GPT-ECGReader (57.89%)	GPT-ECGAnalyzer (57.89%)	1.000**
	GPT-ECGReader (57.89%)	CARDIOLOGIST (89.47%)	<b>0.004*</b>
	GPT-ECGInterpreter (63.16%)	GPT-ECGAnalyzer (57.89%)	0.814**
	GPT-ECGInterpreter (63.16%)	CARDIOLOGIST (89.47%)	<b>0.014*</b>
	GPT-ECGAnalyzer (57.89%)	CARDIOLOGIST (89.47%)	<b>0.004*</b>
Complex	GPT-ECGReader (50.00%)	GPT-ECGInterpreter (54.17%)	1.000**
	GPT-ECGReader (50.00%)	GPT-ECGAnalyzer (70.83%)	0.238**
	GPT-ECGReader (50.00%)	CARDIOLOGIST (91.67%)	<b>0.003*</b>
	GPT-ECGInterpreter (54.17%)	GPT-ECGAnalyzer (70.83%)	0.371**
	GPT-ECGInterpreter (54.17%)	CARDIOLOGIST (91.67%)	<b>0.008*</b>
	GPT-ECGAnalyzer (70.83%)	CARDIOLOGIST (91.67%)	0.137*

Note: Fisher exact (\*) and chi-square (\*\*) tests were performed to quantify the *p*-values. The bold values represent *p*-values lower than 0.05.

exhibited lower accuracy, particularly in complex and critical diagnoses (NI and CSD). Additionally, it was observed that female patients exhibited higher accuracy rates in ECG evaluations than male patients, though these differences were not statistically significant.

In order to maintain alignment and consistency with previous investigations, textbook examples were purposefully chosen over clinical cases. Since we use examples from the same book (150 ECG Problems) which has been used in several previous studies, results and conclusions can be



**Table 5.** Pairwise comparison of the ECG image diagnosis methods according to genders of the patients.

Gender	Method-1 (Accuracy %)	Method-2 (Accuracy %)	p-value
Female	GPT-ECGReader (71.43%)	GPT-ECGInterpreter (71.43%)	1.000**
	GPT-ECGReader (71.43%)	GPT-ECGAnalyzer (65.71%)	0.797**
	GPT-ECGReader (71.43%)	CARDIOLOGIST (88.57%)	0.133*
	GPT-ECGInterpreter (71.43%)	GPT-ECGAnalyzer (65.71%)	0.797**
	GPT-ECGInterpreter (71.43%)	CARDIOLOGIST (88.57%)	0.133*
	GPT-ECGAnalyzer (65.71%)	CARDIOLOGIST (88.57%)	<b>0.044*</b>
Male	GPT-ECGReader (51.43%)	GPT-ECGInterpreter (58.57%)	0.497**
	GPT-ECGReader (51.43%)	GPT-ECGAnalyzer (61.43%)	0.306**
	GPT-ECGReader (51.43%)	CARDIOLOGIST (95.71%)	<b>&lt;0.001*</b>
	GPT-ECGInterpreter (58.57%)	GPT-ECGAnalyzer (61.43%)	0.863**
	GPT-ECGInterpreter (58.57%)	CARDIOLOGIST (95.71%)	<b>&lt;0.001*</b>
	GPT-ECGAnalyzer (61.43%)	CARDIOLOGIST (95.71%)	<b>&lt;0.001*</b>

Note: Fisher exact (\*) and chi-square (\*\*) tests were performed to quantify the p-values. The bold values represent p-values lower than 0.05.

**Table 6.** Accuracies of the ECG image diagnosis methods according to genders of the patients.

Method	Male	Female	p-value
GPT-ECGReader	51.43%	71.43%	0.080**
GPT-ECGInterpreter	58.57%	71.43%	0.284**
GPT-ECGAnalyzer	61.43%	65.71%	0.830**
CARDIOLOGIST	95.71%	88.57%	0.218*

Note: Fisher exact (\*) and chi-square (\*\*) tests were performed to quantify the p-values.

easily compared across studies. In addition, textbook cases minimize ambiguity and create uniformity in the readers understanding by providing specific clearly conceptualized illustrations of ECG shapes. If real clinical cases are used, assessment of inter-rater reliability might be more difficult due to the variability caused by missing clinical information, poor recording of ECG, and complexity of the case.

Moreover, textbook examples are ideal in testing theoretic knowledge as well as diagnosis in terms of accuracy because they do not contain any confounding factors that could influence the bias of interpretation. Hence, we consider the reasons for inclusion of textbook scenarios in

our investigation satisfying, although we acknowledge the potential disadvantage of being unrepresentative of the reality.

A review of the literature reveals the potential and limitations of AI in ECG interpretation. In a study conducted by Hannun et al., it was demonstrated that deep learning models can achieve high levels of accuracy in the evaluation of ECGs.<sup>7</sup> Similarly, Rajpurkar et al. demonstrated that deep learning models can accurately classify 14 distinct ECG rhythm disorders.<sup>8</sup> Further advancements in the role of AI in ECG interpretation have been made in research conducted in 2023 and 2024. For example, the HeartBEiT model developed by Mount Sinai was trained on 8.5 million ECGs and demonstrated high performance at low sample sizes.<sup>13</sup> In a study conducted in 2023, a new AI tool demonstrated superior performance in detecting myocardial infarction compared to conventional methods.<sup>14</sup>

A further study examining the potential of AI-assisted ECG analysis demonstrated that AI can identify cardiac disease risks at an earlier stage than current risk calculation methods.<sup>15</sup> This is of particular importance for conditions such as coronary artery disease, which can be insidious and result in significant health complications.

There are some published studies that have already investigated ChatGPT's effectiveness. These publications, which give different perspectives and conclusions, present a thorough examination of ChatGPT's capacity to decipher

ECG pictures and respond to pertinent multiple-choice questions.<sup>3-6</sup> The researchers used multiple-choice questions and ECG pictures from the AHA Advanced Cardiovascular Life Support tests in a study published in “Resuscitation” (2023). The accuracy of ChatGPT-4 Plus, Bing Chat Enterprise, and Google Bard in understanding these photos was assessed. With 63.0% of the ECG pictures properly interpreted, ChatGPT-4 Plus showed the best level of accuracy among the chatbots. By comparison, the accuracy rates of Bing Chat Enterprise and Google Bard were found to be lower, at 22.2% and 48.2%, respectively.<sup>4</sup>

In the “JMIR Preprints” (2023) research, 62 ECG-related multiple-choice questions from credible medical tests were used to assess the recently launched ChatGPT-4V, which has visual recognition capabilities. According to the study, ChatGPT-4V obtained an astounding 83.87% total accuracy. But the way it performed differed greatly depending on the kind of inquiry. With an accuracy of 86.21%, it performed exceptionally well in questions recommending treatments. In contrast, diagnostic questions had a lower accuracy of 65.38%, and the lowest accuracy was shown in counting-based questions such as QT interval computations, at 28.57%. The study ascribed ChatGPT-4V’s inadequate ability to integrate different ECG characteristics and its limitations in obtaining exact quantitative measurements to the inferior performance in counting tasks, which resulted in vague diagnoses when multiple-choice choices were not supplied.<sup>5</sup>

The third research compared the GPT-4’s performance to that of cardiologists and emergency care experts and was published in the “American Journal of Emergency Medicine” (2024). 40 multiple-choice questions, broken down into daily and more challenging ECG cases, were utilized in the study. The questions were based on ECG examples from the book “150 ECG Cases.” In routine ECG questions, GPT-4 performed similarly to cardiologists, but it outscored both emergency medicine experts and cardiologists in more difficult questions. In all, GPT-4 answered 36.33 out of 40 questions correctly on average. The results of the study showed that the GPT-4’s diagnostic accuracy was greater for common questions than for more difficult ones, indicating that the test’s performance may be impacted by complexity and the requirement for nuanced knowledge.<sup>3</sup>

Another recent retrospective observational study done in a single center included a cohort of 128 patients presenting to the emergency department with cardiovascular complaints. In this study, researchers trained Chat GPT 4.0 using appropriate cardiology material, after which it was asked to respond to particular ECG interpretation questions, such as yes/no questions concerning rhythm, the PR interval, the QRS complex, and so on. Unlike our study, which assessed ChatGPT’s open-ended diagnostic capabilities, this technique is expected to result in improved ChatGPT performance since closed questions are straightforward and do not include numerous difficulties. While

the results indicate that the ECG segments were mostly correctly interpreted, they also highlight the limits of ChatGPT for practical usage, particularly in detecting crucial conditions such as major adverse cardiac events.<sup>6</sup>

Overall, ChatGPT models’ performance varies according to the kind and difficulty of the queries, even if they have demonstrated encouraging promise in the interpretation of ECG pictures. Significant proficiency in ECG interpretation was shown by ChatGPT-4 Plus and ChatGPT-4V, whose accuracy rates in some situations even surpassed those of human professionals. There are still restrictions, nevertheless, especially with regard to combining different ECG parameters and quantitative measures. Notably, ChatGPT’s inferior capacity to directly analyze ECG leads was shown by the fact that it only outperformed human professionals in the research where ECG images were not included.<sup>3-6</sup>

This research has several notable strengths such as thorough analysis and gender comparisons as well as diagnostic classifications. Additionally, this trial is one of the first few to assess the effectiveness of clinical chat models based on ChatGPT. However, due to limited data collection methods and the dissimilar data used in training the AI model, this finding has some limitations. Also, such a comparison could not be made because only one AI model was evaluated among other available AI models. This study did not fully consider the source and diversity of data which are important in determining model performance.

From these results, it can be concluded that ChatGPT-based models for now should be applied with caution at medical facilities. Low accuracy percentages in complex and serious diagnoses (NI and CSD) indicate that depending on these models in clinical practice might not be necessary. Consequently, cardiologists should use AI models especially LLM as a tool to support rather than replace them. It is crucial to note, however, that the findings of ML models are far superior to LLM’s and should be taken into account as a first choice when introducing AI into the medical sector.

Future studies need to focus on improving ChatGPT-based models’ performance by using larger amounts of diverse datasets for training purposes. Besides, there is a need for comparative evaluation among different AI models so as to identify which of them best suits clinical applications. More research is needed to optimize AI model performance in challenging and emergency diagnoses. Specifically, it is important to test the models with real-world clinical data and to model with these data.

## Conclusions

This study shows how limited ChatGPT-based AI models are when it comes to ECG assessment compared to cardiologists. More research and advances need to be made so that AI models can be safe and effective when used in clinical

practices. Therefore, this study provides valuable information on potential integration of AI into healthcare systems.

## Limitations

When interpreting the results, it is important to keep in mind the limitations of this study as the research was performed on a restricted and homogenized dataset. Lack of diversity within the data set may limit both generalizability and performance of the model. This could be improved by using a larger, more varied dataset.

The current study only addressed the effectiveness of ChatGPT-based AI models. There were no comparisons made with other AI models. One possible area for further research would be to compare different AI models so that we can know which one is better suited for clinical applications.

The model has never been tested in real-life medical settings. Thus, there is no direct information about how effective or trustable it can be in medical applications. In order to evaluate its efficiency and applicability in the medical environment, it should be tried on true-life data samples.

Long-term follow-up data were not included in the study. A comprehensive understanding of the clinical utility of these kinds of models can be achieved through assessing long-term outcomes for such AI models.

To address these limitations, future works should use more varied and bigger data sets, different AI models as well as real-world clinical data to cope with the above-mentioned problems and assess the efficiency of the model in a more comprehensive way.


**Acknowledgments::** The authors would like to thank all who supported this research.

**Contributorship:** All authors contributed to: (1) substantial contributions to conception and design, or acquisition of data, or analysis and interpretation of data, (2) drafting the article or revising it critically for important intellectual content, and (3) final approval of the version to be published.

**Declaration of conflicting interests:** The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**Funding:** The authors received no financial support for the research, authorship, and/or publication of this article.

**ORCID iDs:** Volkan Çamkiran  <https://orcid.org/0000-0003-1908-0648>

Batool Achmar  <https://orcid.org/0009-0003-3918-8792>

**Supplemental material:** Supplemental material for this article is available online.

## References

1. Itchhaporia D. Artificial intelligence in cardiology. *Trends Cardiovasc Med* 2022; 32: 34–41.
2. Muzammil MA, Javid S, Afridi AK, et al. Artificial intelligence-enhanced electrocardiography for accurate diagnosis and management of cardiovascular diseases. *J Electrocardiol* 2024; 83: 30–40.
3. Günay S, Öztürk A, Özerol H, et al. Comparison of emergency medicine specialist, cardiologist, and chat-GPT in electrocardiography assessment. *Am J Emerg Med* 2024; 80: 51–60.
4. Fijačko N, Prosen G, Abella BS, et al. Can novel multimodal chatbots such as Bing Chat Enterprise, ChatGPT-4 Pro, and Google Bard correctly interpret electrocardiogram images? *Resuscitation* 2023; 193: 110009.
5. Zhu L, Mou W, Wu K, et al. Multimodal ChatGPT-4V for electrocardiogram interpretation: Promise and limitations. *JMIR J Med Internet Res* 2024; 26: e54607.
6. Zaboli A, Brigo F, Ziller M, et al. Exploring ChatGPT's potential in ECG interpretation and outcome prediction in emergency department. *Am J Emerg Med* 2024; 88: 7–11.
7. Hannun AY, Rajpurkar P, Haghpanahi M, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat Med* 2019; 25: 65–69.
8. Rajpurkar P, Hannun AY, Haghpanahi M, et al. Cardiologist-level arrhythmia detection with convolutional neural networks. *arXiv.org*. 2017, July 6. <https://arxiv.org/abs/1707.01836>
9. Attia ZI, Friedman PA, Noseworthy PA, et al. Age and sex estimation using artificial intelligence from standard 12-lead ECGs. *Circ Arrhythm Electrophysiol* 2019; 12. DOI: 10.1161/circep.119.007284.
10. Ribeiro AH, Ribeiro MH, Paixão GMM, et al. Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nat Commun* 2020; 11: 1760.
11. Hampton J, Adlam D and Hampton J. 150 ECG Cases. Elsevier Health Sciences. 2019. [http://books.google.ie/books?id=ZA6IDwAAQBAJ&printsec=frontcover&dq=150 ECG Cases Book by David Adlam, Joanna Hampton, and John R Hampton&hl=&cd=1&source=gbs\\_api](http://books.google.ie/books?id=ZA6IDwAAQBAJ&printsec=frontcover&dq=150+ECG+Cases+Book+by+David+Adlam,+Joanna+Hampton,+and+John+R+Hampton&hl=&cd=1&source=gbs_api)
12. Chen J, Zhu L, Mou W, et al. STAGER checklist: Standardized testing and assessment guidelines for evaluating generative artificial intelligence reliability. *iMetaOmics* 2024; 1. DOI: 10.1002/imo2.7.
13. Vaid A, Jiang J, Sawant A, et al. A foundational vision transformer improves diagnostic performance for electrocardiograms. *NPJ Digit Med* 2023; 6: 1–8.
14. New AI tool beats standard approaches for detecting heart attacks. ScienceDaily. [https://www.sciencedaily.com/releases/2023/06/230629125710.htm#:~:text=New AI tool beats standard approaches for detecting heart attacks, Date A June 29&text=Summary A,according to a new study \(2023, June 23\).](https://www.sciencedaily.com/releases/2023/06/230629125710.htm#:~:text=New+AI+tool+beats+standard+approaches+for+detecting+heart+attacks,+Date+June+29&text=Summary+A,according+to+a+new+study+(2023,+June+23).)
15. Artificial Intelligence (AI) in Cardiovascular Medicine - Overview - Mayo Clinic. Mayo Clinic. <https://www.mayoclinic.org/departments-centers/ai-cardiology/overview/ovc-20486648> (2024b, March 16).