International Journal of Paramedicine

Published by the National EMS Management Association (USA)

RESEARCH REPORTS

COMPARATIVE DIAGNOSTIC ACCURACY OF LARGE LANGUAGE MODELS FOR STEMI CLASSIFICATION: AN EXPLORATORY STUDY USING DE-IDENTIFIED EMERGENCY DEPARTMENT CASES

Catherine Royzman*[1]; Jonathan Spagnola, MD[2]; Ruben Kandov, MD[2]

Author Affiliations: 1. Feinstein Institutes for Medical Research, Northwell Health, Manhasset, NY, USA; 2. Staten Island University Hospital, Department of Cardiology, Northwell Health, Staten Island, NY, USA.

Recommended Citation: Royzman, C., Spagnola, J., & Kandov, R. (2026). Comparative diagnostic accuracy of large language models for STEMI classification: An exploratory study using de-identified emergency department cases. *International Journal of Paramedicine*. (15). 103-112. https://doi.org/10.56068/IQHK7751

Keywords: STEMI, artificial intelligence, ECG interpretation, prehospital diagnosis, machine learning, cardiac care, emergency medical services, EMS, paramedicine

Disclosures: The authors declare no conflicts of interest or disclosures. Artificial intelligence tools were evaluated as the subject of this retrospective study and were used only to classify de-identified cases as “STEMI” or “not STEMI” for research purposes; no AI output was used for patient care, diagnosis, treatment, or clinical decision-making. The authors reviewed all AI outputs and performed the analysis and interpretation.

Funding: This study received no external funding. The AI tools were publicly available, and no financial relationships with the developers of these models influenced the study design or analysis.

Received: January 23, 2026
Revised: May 20, 2026
Accepted: May 21, 2026
Published: July 15, 2026

**Corresponding Author:*
croyzman@northwell.edu

ABSTRACT

Background: Early and accurate identification of ST-Elevation Myocardial Infarction (STEMI) in the prehospital setting would reduce morbidity and mortality (Rao et al. 2025). Artificial Intelligence (AI) tools may assist emergency medical personnel by providing rapid electrocardiogram (ECG) interpretation and differential diagnosis (Chen et al. 2022; Nallamothu et al., 2015).

Objective: To evaluate the diagnostic accuracy of ChatGPT-4.0, Gemini 2.5 Pro, and a hybrid model combining ECG-GPT + ChatGPT-4.0 in classifying STEMI versus non-STEMI cases based on ECG and clinical data inputs.

Methods: Fifty-six consecutive de-identified cases (28 STEMI, 28 non-STEMI) from Staten Island University Hospital EMR (1/2025-6/2025) were analyzed. Each case included demographics, vital signs, chief complaints, and a representative 12-lead ECG. The three AI models were independently asked to classify each case as STEMI or not STEMI.

Results: For the STEMI cohort (n=28), detection rates were: ChatGPT-4.0 21.4%, Gemini 2.5 Pro 67.9%, and ECG-GPT + ChatGPT hybrid 71.4%. For the non-STEMI cohort (n=28), correct classification rates were: ChatGPT-4.0 92.9%, Gemini 2.5 Pro 28.6%, and ECG-GPT + ChatGPT hybrid: 96.4%. The ECG-GPT + ChatGPT hybrid model demonstrated the highest diagnostic accuracy at 83.9%.

Conclusion: While ChatGPT-4.0 showed high specificity, it lacked sensitivity for STEMI detection. Gemini 2.5 Pro improved sensitivity but yielded higher false positives. The hybrid ECG-GPT + ChatGPT model outperformed both standalone models, suggesting that multimodal AI integration may enhance triage accuracy for prehospital care. Despite the small population size and limitations associated with this, these results propose a viable means to hypothesize that AI may have potential in assisting responders with early STEMI identification. Future improvements in model precision could support the implementation of AI-based triage systems in emergency response settings, reducing time to treatment and improving outcomes by aiding direct cardiac catheterization lab transfers without unnecessary delays (Garvey et al., 2012; Squire et al., 2014).

INTRODUCTION

Timely identification and treatment of ST-Elevation Myocardial Infarctions (STEMIs) are imperative in improving cardiac patient outcomes (Nallamothu et al., 2015; Rao et al., 2025). In this study, we explore the integration of artificial intelligence (AI) into prehospital triage protocols to enhance early STEMI detection and expedite contact-to-balloon time (Nallamothu et al., 2015). Recognizing the limitations faced by paramedics and emergency medical technicians in the field and physicians in the emergency room, we propose AI as a decision-support tool to assess prehospital ECGs and clinical data.

Importantly, not all AI systems are equivalent in clinical use. Purpose-built medical AI devices, including software as a medical device, are typically developed for a defined intended use, trained and validated on clinical datasets, evaluated with prespecified performance metrics, and deployed within regulated quality, safety, and risk-management frameworks. By contrast, consumer-facing large language models (LLMs) such as ChatGPT and Gemini are general-purpose conversational systems. They are not designed, cleared, or validated as standalone diagnostic devices for STEMI interpretation, and their outputs may be influenced by prompt structure, model updates, incomplete clinical context, and probabilistic language generation. Therefore, this study does not evaluate ChatGPT or Gemini as deployable medical devices. Rather, it examines whether general-purpose multimodal LLMs can perform a constrained experimental classification task and whether pairing an ECG-focused model with a language model improves performance compared with consumer LLMs alone.

Despite its recognized benefits, accurate ECG interpretation by emergency medical technicians (EMTs) in prehospital settings remains challenging. Conventional devices require meticulous electrode placement and skilled interpretation—often compromised by the complexities of prehospital environments (Gregory et al., 2019). Rule-based algorithms used in systems like those in Hong Kong deliver high sensitivity (~94.6%) but only fair specificity (~87.9%), leading to notable false-positive rates and potential over-triage (Lai et al., 2024).

The current state of prehospital STEMI diagnosis varies across EMS systems. In paramedic-led interpretation models, diagnostic performance may be limited by STEMI mimics and variation in training. One study done at Pennsylvania State University found that paramedics recognized STEMI with imperfect sensitivity and specificity when faced with ECG mimics such as left ventricular hypertrophy, bundle branch block, paced rhythms, and supraventricular tachycardia (Mencl et al., 2013). In systems using computerized ECG interpretation, performance can be highly specific but not always sufficiently sensitive, and incorrect computer interpretations may influence paramedic decision-making. The UK RESPECT feasibility study showed that computer interpretation messages affected paramedic STEMI decisions, improving accuracy when correct but worsening accuracy when incorrect (Pillberry et al., 2016). In systems using ECG transmission, physician overread may improve positive predictive value, although studies show variable impact on false-positive activations and workflow time. Therefore, the potential role of AI is not to replace existing EMS protocols, but to serve as adjunctive decision support in systems where rapid interpretation must occur before definitive cardiology review.

AI and deep learning algorithms have the potential to enhance prehospital ECG interpretation (Chen et al., 2022). These methods can automate analysis, reduce human error, and potentially reduce diagnostic delays. For instance, models leveraging convolutional neural networks (CNNs) and long short-term memory (LSTM) architectures have demonstrated impressive real-world performance: an AI model deployed in ambulances in Central Taiwan, using a mini-portable 12-lead ECG device achieved accuracy of 99.2%, sensitivity (recall) of 94.1%, specificity of 99.4%, and an impressively low response time (37 s versus >113 s for physicians). Moreover, this system helped initiate PPCI in ten identified STEMI patients, with a median contact-to-door time of 18.5 minutes (IQR: 16–20.8) (Chen et al., 2022).

To address these gaps, our investigation evaluated three AI-driven approaches—ChatGPT-4.0, Gemini 2.5 Pro, and a hybrid model combining ECG-GPT with ChatGPT—for their ability to discriminate STEMI versus non-STEMI cases based on prehospital ECG findings, vital signs, and chief complaints. Using a balanced dataset of 28 confirmed STEMI and 28 confirmed non-STEMI cases, we measured each model's accuracy, sensitivity, and specificity in order to (1) assess the standalone performance of language-based models on ECG interpretation tasks, (2) determine whether multimodal integration (ECG plus clinical data) enhances diagnostic performance, and (3) explore the feasibility of deploying AI-based triage tools to accelerate in-field identification of STEMI.

METHODS

Study Design and Data Source

This was a retrospective diagnostic accuracy study using de-identified patient data extracted from the electronic medical records (EMR) of Staten Island University Hospital (Northwell Health). Cases included patients transported to the emergency department (ED) with a variety of acute complaints in 2025. Although this study was motivated by prehospital triage, the dataset consisted of retrospective ED-arrival cases rather than prospectively collected ambulance ECGs; therefore, results should be interpreted as an exploratory simulation of early triage decision support rather than validation in an operational EMS system. The study population comprised two cohorts: patients with confirmed ST-Elevation Myocardial Infarction (STEMI) and patients without STEMI. All data were fully de-identified prior to analysis in accordance with institutional review protocols.

Case Selection

A total of 56 cases were randomly selected from the EMR system, consisting of 28 patients with confirmed STEMI (22 male, 6 female) and 28 patients without STEMI. STEMI confirmation was based on final cardiology and interventional reports, while non-STEMI cases were confirmed by discharge diagnoses and absence of acute coronary intervention. Cases were stratified by age to ensure even distribution across both cohorts. Chief presenting complaints in the STEMI cohort included chest pain, abdominal pain, syncope, weakness, or dyspnea, consistent with known STEMI presentations. Non-STEMI cases represented a mixture of chest pain and alternative complaints such as dizziness, confusion, seizures, abdominal pain, or viral syndromes.

Data Preparation

For each patient, the following clinical variables were extracted:

- Demographics: Age, sex;
- Physiological data: Systolic and diastolic blood pressure, heart rate, respiratory rate, and oxygen saturation;
- Presenting complaint: Chief symptom(s) at ED arrival;
- Electrocardiogram (ECG): Representative 12-lead image at presentation.

All data was compiled into a standardized case vignette format. Each vignette was stripped of identifying information to maintain anonymity.

AI Model Evaluation

Three AI models were evaluated:

- ChatGPT-4.0 (OpenAI, 2024)—a general-purpose large language model
- Gemini 2.5 Pro (Google DeepMind, 2025)—an advanced multimodal language model
- ECG-GPT + ChatGPT model – a two-step system combining an ECG-specific model (ECG-GPT, fine-tuned on electrocardiographic data) with ChatGPT for clinical reasoning integration. No prompt was required for ECG-GPT; just an image was attached. The output is a reading of the ECG, which was paired with the prompt (as seen in Appendix) and input into ChatGPT.

Each model was independently provided with the same case vignette, including vital signs, age, chief complaint, and ECG image. Models were instructed to classify each case dichotomously as "STEMI" or "not STEMI." No intermediate or probabilistic outputs were recorded.

Outcome Measures

The primary outcome was diagnostic accuracy of each model, defined as the proportion of correctly classified cases compared to the gold-standard diagnosis (STEMI vs. not STEMI) confirmed in the EMR. Secondary outcomes included:

- Sensitivity: proportion of true STEMI cases correctly identified;
- Specificity: proportion of true non-STEMI cases correctly excluded.

Statistical Analysis

Results were tabulated separately for the STEMI (n = 28) and non-STEMI (n = 28) cohorts. Sensitivity, specificity, and accuracy were calculated with 95% confidence intervals (Wilson method). Accuracy was computed as the proportion of correctly classified cases across all 56 patients. All analyses were descriptive, given the exploratory nature of the study.

RESULTS

Study Population

A total of 56 patient cases were included in the analysis, comprising 28 confirmed ST-elevation myocardial infarction (STEMI) cases and 28 confirmed not STEMI cases. Among the STEMI group, 22 (78.6%) patients were male and 6 (21.4%) were female, with ages ranging from 38 to 89 years. The most common presenting complaint in STEMI patients was chest pain, reported in 21 of 28 cases (75%), followed by abdominal pain, syncope, weakness, or dyspnea. The non-STEMI cohort included 19 (67.9%) males and 9 (32.1%) females, with ages ranging from 34 to 85 years. Presenting complaints included chest pain, dizziness, seizures, abdominal pain, and flulike symptoms. Baseline vital signs demonstrated broad variability in both cohorts, reflecting the heterogeneity of clinical presentations in the ED setting. No image quality issues arose from ECG submission, and all model responses were binary due to the submitted prompt. No ambiguous responses were output from any of the three LLMs.

Model Performance on STEMI Cases

The ability of each AI model to correctly classify STEMI cases varied considerably (Table 1):

- ChatGPT-4.0 correctly identified only 6 of 28 STEMI cases (21.4% sensitivity;
- Gemini 2.5 Pro correctly identified 19 of 28 STEMI cases (67.9% sensitivity);
- ECG-GPT + ChatGPT hybrid model correctly identified 20 of 28 STEMI cases (71.4% sensitivity).

While ChatGPT-4.0 frequently failed to recognize STEMI despite suggestive clinical and ECG findings, Gemini 2.5 Pro and the hybrid model demonstrated substantially greater sensitivity.

Model	Sensitivity (STEMI, 95% CI)	Specificity (not STEMI 95% CI)	Accuracy (95% CI)
ChatGPT-4.0	21.4% (10.2–39.5)	92.9% (77.4–98.0)	57.1% (44.1–69.2)
Gemini 2.5 Pro	67.9% (49.3–82.1)	28.6% (15.3–47.1)	48.2% (35.7–61.0)
Hybrid ECG-GPT + ChatGPT	71.4% (52.9–84.7)	96.4% (82.3–99.4)	83.9% (72.2–91.3)

Table 1. Diagnostic performance of AI models in identifying STEMI.

Model Performance on Non-STEMI Cases

Performance on the 28 non-STEMI cases showed a contrasting pattern:

- ChatGPT-4.0 correctly classified 26 of 28 non-STEMI cases (92.9% specificity);
- Gemini 2.5 Pro correctly classified only 8 of 28 non-STEMI cases (28.6% specificity), with frequent false-positive STEMI classifications;
- ECG-GPT + ChatGPT hybrid model correctly classified 27 of 28 non-STEMI cases (96.4% specificity).

Thus, while Gemini 2.5 Pro increased sensitivity, it did so at the cost of markedly reduced specificity. In contrast, the hybrid model preserved both high specificity and improved sensitivity.

Overall Accuracy

When combining both cohorts (n = 56), overall diagnostic accuracy was:

- ChatGPT-4.0: 57.1% (32/56 correct);
- Gemini 2.5 Pro: 48.2% (27/56 correct);
- ECG-GPT + ChatGPT hybrid: 83.9% (47/56 correct).

Confidence intervals are presented in Table 1. The hybrid model's performance intervals did not overlap substantially with those of ChatGPT-4.0 or Gemini 2.5 Pro, supporting its superior diagnostic balance. Performance metrics are illustrated in Figure 2, and the case selection process is summarized in Figure 1.

Key Observations

- ChatGPT-4.0 demonstrated strong specificity but extremely poor sensitivity, under recognizing STEMI cases;
- Gemini 2.5 Pro demonstrated improved sensitivity but produced frequent false positives, lowering specificity;
- Hybrid ECG-GPT + ChatGPT achieved the best balance, with both high sensitivity and specificity and the highest overall accuracy.

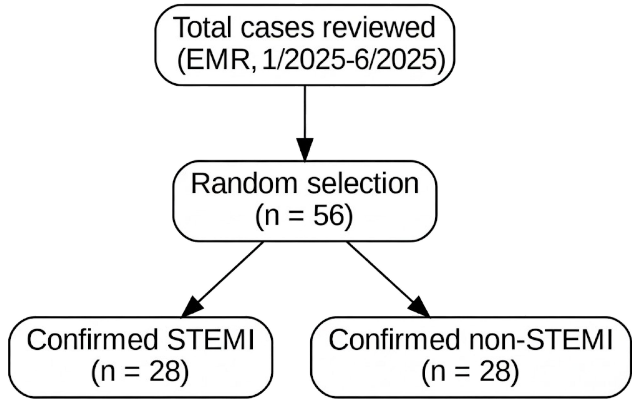

Figure 1. Case selection flow diagram.

Note. Flow diagram illustrating case inclusion. From electronic medical records (1/2025-6/2025), 56 cases were randomly selected, comprising 28 confirmed STEMI and 28 confirmed not STEMI patients.

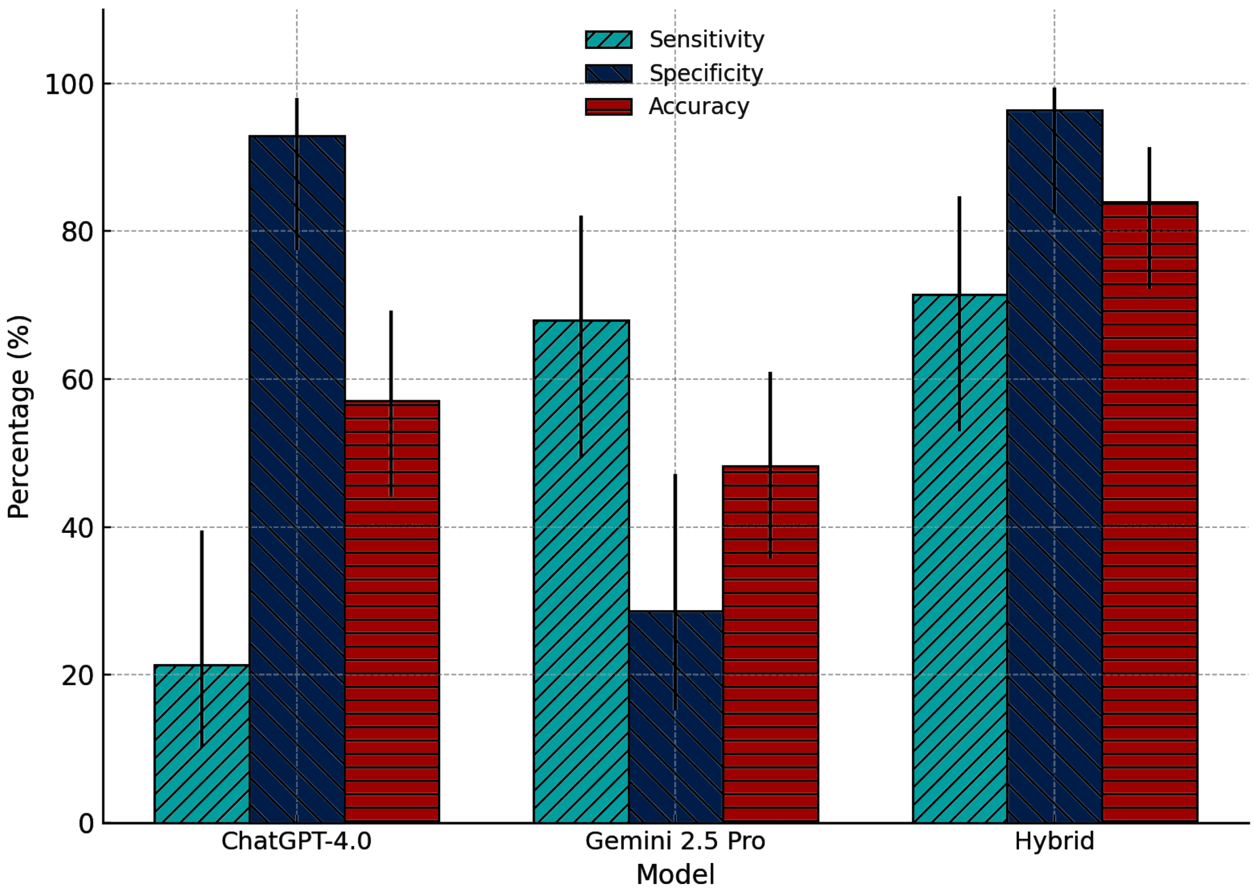

Figure 2. Performance of AI models in STEMI classification.

Note. Bar chart showing sensitivity, specificity, and overall accuracy (with 95% confidence intervals) for ChatGPT-4.0, Gemini 2.5 Pro, and the hybrid ECG-GPT + ChatGPT model. Distinct hatching patterns and colors indicate the three performance metrics.

DISCUSSION

Principal Findings

In this study, we evaluated the diagnostic performance of three artificial intelligence (AI) models, ChatGPT-4.0, Gemini 2.5 Pro, and a hybrid ECG-GPT + ChatGPT system, in classifying ST-elevation myocardial infarction (STEMI) versus non-STEMI presentations using deidentified emergency department cases. The hybrid model achieved the highest overall diagnostic accuracy (83.9%), outperforming both general-purpose language models. While ChatGPT-4.0 exhibited high specificity (92.9%) in ruling out not STEMI, it showed poor sensitivity (21.4%), highlighting a limitation in recognizing true STEMI cases. In contrast, Gemini 2.5 Pro improved sensitivity (67.9%) but at the cost of a steep decline in specificity (28.6%), producing frequent false positives. Confidence intervals further highlight this advantage: the hybrid model achieved 83.9% accuracy (72.2–91.3), compared with 57.1% (44.1– 69.2) for ChatGPT-4.0 and 48.2% (35.7–61.0) for Gemini 2.5 Pro, demonstrating a clear performance margin in this exploratory sample.

Comparison With Prior Studies

Our findings align with prior work demonstrating the promise of AI in prehospital and ED cardiac triage. Earlier studies reported that an AI-driven mini-12-lead ECG device deployed in ambulances in Taiwan achieved 94.1% sensitivity, 99.4% specificity, and a 99.2% overall accuracy, while also reducing time to physician notification from 113 seconds to 37 seconds (Chen et al., 2022). Similarly, large-scale validation studies of deep learning ECG classifiers have reported very accurate results, performing on par with human cardiologists in controlled settings (Hannun et al., 2019). However, most existing models have been domain-specific (CNN or LSTM-based ECG interpreters) and do not completely incorporate clinical context such as vital signs or patient-reported symptoms.

Our study is distinct in comparing general-purpose LLMs, which are readily available to EMTs and medical personnel, with a specialized hybrid approach that integrates ECG-focused AI. The poor sensitivity of ChatGPT-4.0 illustrates the limitations of LLMs when applied directly to clinical classification tasks without domain-specific training. Conversely, Gemini 2.5 Pro's higher sensitivity but poor specificity suggests that LLMs may overfit to the presence of suggestive symptoms (e.g., chest pain), generating false alarms. The hybrid approach, by contrast, suggests a promising pathway: combining domain-trained models with general reasoning models to capture both structured ECG patterns and unstructured clinical context.

Clinical Implications

The ability to rapidly and accurately identify STEMI in prehospital and ED settings is important, as early reperfusion therapy directly correlates with reduced morbidity and mortality (Rao et al., 2025; Nallamothu et al., 2015). False negatives (missed STEMIs) can delay reperfusion and worsen outcomes, while false positives can overwhelm catheterization laboratories with unnecessary activations (Garvey et al., 2012; Larson et al., 2007; Squire et al., 2014). Our results suggest that an integrated hybrid AI system may enhance frontline responders' diagnostic capabilities while minimizing unnecessary over triage. Such systems could be embedded into ambulance monitors or emergency department workflows to support rapid triage and expedite "door-to-balloon" times (Squire et al., 2014).

Strengths and Limitations

A strength of this study is the inclusion of both ECG images and clinical variables, reflecting real-world ED presentations rather than isolated waveform data. Cases were also balanced by age distribution, minimizing demographic skew. However, several limitations must be acknowledged. First, the sample size was modest (n=56), limiting generalizability. Second, the dataset was retrospective and restricted to a single health system, which may not completely capture variability in ECG presentations or patient populations. Third, AI models were tested in a static classification framework, without iterative refinement or probabilistic confidence outputs, which may underestimate their potential performance in real deployment. Finally, we did not benchmark against human paramedic or physician interpretations, which would provide useful context for relative performance.

Future Directions

Future research should evaluate hybrid AI models in larger, prospective, and multicenter datasets, ideally incorporating real-time ambulance and ED workflows. Integration of additional multimodal inputs—such as laboratory biomarkers, prior ECGs, and risk scores—may further enhance accuracy. Moreover, comparative studies against human experts are essential to establish clinical utility and safety. Importantly, research should also examine usability and workflow integration, ensuring that AI systems support rather than overwhelm frontline responders.

CONCLUSION

This study highlights both the potential and limitations of current AI models for STEMI detection. While general-purpose LLMs alone were insufficient, a hybrid ECG-GPT + ChatGPT model achieved superior diagnostic accuracy, balancing sensitivity and specificity. This suggests that multimodal AI integration may offer a powerful tool for prehospital and ED triage. These results are hypothesis-generating and calling for prospective, multicenter, adequately powered validation—with human comparators and validated reporting standards such as TRIPOD-AI (Collins et al., 2024).

REFERENCES

Chen, K. W., Wang, Y. C., Liu, M. H., Tsai, B. Y., Wu, M. Y., Hsieh, P. H., Wei, J. T., Shih, E. S. C., Shiao, Y. T., Hwang, M. J., Wu, Y. L., Hsu, K. C., & Chang, K. C. (2022). Artificial intelligence-assisted remote detection of ST-elevation myocardial infarction using a mini-12-lead electrocardiogram device in prehospital ambulance care. *Frontiers in Cardiovascular Medicine, 9*, Article 1001982. https://doi.org/10.3389/fcvm.2022.1001982

Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., et al. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, 385, Article e078378. https://doi.org/10.1136/bmj-2023-078378

Garvey, J. L., Monk, L., Granger, C. B., Studnek, J. R., Roettig, M. L., & Jollis, J. G. (2012). Rates of cardiac catheterization cancelation for ST-segment elevation myocardial infarction after activation by emergency medical services or emergency physicians: Results from the North Carolina Catheterization Laboratory Activation Registry. *Circulation, 125*(2), 308–313. https://doi.org/10.1161/CIRCULATIONAHA.110.007039

Gregory, P., Lodge, S., Kilner, T., & Paget, S. (2019). Accuracy of ECG chest electrode placements by paramedics: An observational study. *British Paramedic Journal, 4*(3), 51–52. https://doi.org/10.29045/14784726.2019.12.4.3.51

Hannun, A. Y., Rajpurkar, P., Haghpanahi, M., Tison, G. H., Bourn, C., Turakhia, M. P., & Ng, A. Y. (2019). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine, 25*(1), 65-69. https://doi.org/10.1038/s41591-018-0268-3

Khunte, A., Sangha, V., Oikonomou, E. K., Dhingra, L. S., et al. (2026). Artificial intelligence based automated interpretation of images of electrocardiograms: Development and multinational validation of ECG-GPT. *European Heart Journal—Digital Health, 7*(3). https://doi.org/10.1093/ehjdh/ztag031

Lai, J. H. Y., Lui, C. T., Chan, T. W. T., Wong, B. C. P., Tsui, M. S. H., Wan, B. K. A., & Mok, K. L. (2024). Diagnostic accuracy of a prehospital electrocardiogram rule-based algorithm for ST elevation myocardial infarction: Results from a population-wide project. *Hong Kong Medical Journal, 30*(4), 271–280. https://doi.org/10.12809/hkmj2310827

Larson, D. M., Menssen, K. M., Sharkey, S. W., Duval, S., Schwartz, R. S., Harris, J., Meland, J. T., Unger, B. T., & Henry, T. D. (2007). "False-positive" cardiac catheterization laboratory activation among patients with suspected ST-segment elevation myocardial infarction. *JAMA, 298*(23), 2754–2760. https://doi.org/10.1001/jama.298.23.2754

Mencl, F., Wilber, S., Frey, J., & Bhalla, M. C. (2013). Paramedic ability to recognize ST-segment elevation myocardial infarction on prehospital electrocardiograms. *Prehospital Emergency Care, 17*(2), 203–210. https://doi.org/10.3109/10903127.2012.755585

Nallamothu, B. K., Normand, S. L., Wang, Y., Hofer, T. P., Brush, J. E., Jr., Messenger, J. C., Bradley, E. H., Rumsfeld, J. S., & Krumholz, H. M. (2015). Relation between door-to-balloon times and mortality after primary percutaneous coronary intervention over time: A retrospective study. *Lancet, 385*(9973), 1114–1122. https://doi.org/10.1016/S0140-6736(14)61932-2

Pilbery, R., Teare, M. D., Goodacre, S., & Morris, F. (2016). The recognition of STEMI by paramedics and the effect of computer interpretation: A randomised crossover feasibility study. *Emergency Medicine Journal, 33*(7), 471–476. https://doi.org/10.1136/emermed-2015204988

Rao, S. V., O'Donoghue, M. L., Ruel, M., Rab, T., Tamis-Holland, J. E., Alexander, J. H., Baber, U., Baker, H., Cohen, M. G., Cruz-Ruiz, M., Davis, L. L., de Lemos, J. A., DeWald, T. A., Elgendy, I. Y., Feldman, D. N., Goyal, A., Isiadinso, I., … Williams, M. S. (2025). 2025 ACC/AHA/ACEP/NAEMSP/SCAI guideline for the management of patients with acute coronary syndromes: A report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. *Circulation, 151*(13), e771–e862. https://doi.org/10.1161/CIR.0000000000001309

Squire, B. T., Tamayo-Sarver, J. H., Rashi, P., Koenig, W., & Niemann, J. T. (2014). Effect of prehospital cardiac catheterization lab activation on door-to-balloon time, mortality, and false positive activation. *Prehospital Emergency Care, 18*(1), 1–8. https://doi.org/10.3109/10903127.2013.836263

APPENDIX

Verbatim LLM Prompt

You are reviewing a de-identified emergency department presentation for research purposes only. You will be provided with patient age, sex, vital signs, chief complaint, and a 12-lead ECG image. Classify the case as either STEMI or not STEMI based on the ECG and clinical information.

Do not provide treatment recommendations.

Do not request additional information.

Choose only one final classification: STEMI or not STEMI.

Case:

- Age: [age]
- Sex: [sex]
- Chief complaint: [chief complaint]
- Blood pressure: [BP]
- Heart rate: [HR]
- Respiratory rate: [RR]
- Oxygen saturation: [SpO2]
- ECG image: [uploaded ECG image]