Vol. 8, No. 2, Agustus 2026, pp 333-345
https://doi.org/10.36590/jika.v8i2.2042
http://salnesia.id/index.php/jika
jika@salnesia. id, p-ISSN: 2337-9847, e-ISSN: 2686-2883
Penerbit: Sarana Ilmu Indonesia (salnesia)

ARTIKEL PENELITIAN

Analisis Keakuratan Hasil Penelitian START Triase oleh *AI ChatGPT* di IGD RSUD Gunung Jati

Accuracy and Agreement of ChatGPT Based START Triage in the Emergency Department of Gunung Jati Hospital

Efa Dwina[1*], Agil Putra Tri Kartika[1], Rizaluddin Akbar[1], Widuri Azmi Atmaji[1]
[1]Program Studi Ilmu Keperawatan, Universitas Muhammadiyah Cirebon, Cirebon, Indonesia

Abstract

Triage is an essential process in emergency care services aimed at determining patient treatment priorities based on the severity of clinical conditions. This study aimed to analyze the accuracy and level of agreement of Artificial Intelligence (AI), particularly ChatGPT, in triage classification at the Emergency Department of RSUD Gunung Jati compared with medical personnel as the reference standard. This study employed a quantitative observational approach with a retrospective cross-sectional design. Data were obtained from 420 adult patient medical records collected between March and May 2025. Triage classification was conducted using the Simple Triage and Rapid Treatment (START) method by both medical personnel and AI ChatGPT. Data analysis was performed using a confusion matrix and diagnostic tests, including sensitivity, specificity, accuracy, positive predictive value (PPV), negative predictive value (NPV), as well as agreement analysis using Cohen's kappa and weighted Cohen's kappa. The results showed that AI ChatGPT achieved an overall accuracy of 72.1% (95% CI: 67.8–76.4). The Cohen's kappa value was 0.313, while the weighted Cohen's kappa values were 0.338 using linear weighting and 0.384 using quadratic weighting, indicating a fair level of agreement. These findings suggested that ChatGPT had the potential as a supportive system in triage classification; however, it could not yet fully replace the role of medical personnel and still required further development, particularly in cases involving higher levels of emergency severity.

Keywords: *artificial intelligence, triage, patient treatment*

Article history:

Submitted 19 April 2026
Accepted 16 Mei 2026
Published 04 Agustus 2026

PUBLISHED BY:
Sarana Ilmu Indonesia (salnesia)
Address:
Jl. Dr. Ratulangi No. 75A, Baju Bodoa, Maros Baru, Kab. Maros, Provinsi Sulawesi Selatan, Indonesia

Email:
info@salnesia.id, jika@salnesia.id
Phone:
+62 85255155883

Abstrak

Triase merupakan proses penting dalam pelayanan kegawatdaruratan yang bertujuan menentukan prioritas penanganan pasien berdasarkan tingkat keparahan kondisi klinis. Penelitian ini bertujuan menganalisis tingkat keakuratan dan kesesuaian Artificial Intelligence (AI), khususnya ChatGPT, dalam klasifikasi triase di Instalasi Gawat Darurat RSUD Gunung Jati dibandingkan dengan tenaga medis sebagai baku acuan. Penelitian menggunakan pendekatan kuantitatif observasional dengan desain retrospektif *cross-sectional*. Data penelitian diperoleh dari 420 rekam medis pasien dewasa periode Maret–Mei 2025. Proses klasifikasi triase dilakukan menggunakan metode *Simple Triage and Rapid Treatment* (START) oleh tenaga medis dan AI ChatGPT. Analisis data dilakukan menggunakan confusion matrix serta uji diagnostik yang meliputi sensitivitas, spesifisitas, akurasi, *positive predictive value* (PPV), *negative predictive value* (NPV), dan uji kesesuaian menggunakan *Cohen's kappa* serta *weighted Cohen's kappa*. Hasil penelitian menunjukkan bahwa AI ChatGPT memiliki tingkat akurasi sebesar 72,1% (95% CI: 67,8–76,4). Nilai *Cohen's kappa* sebesar 0,313, sedangkan *weighted Cohen's kappa* sebesar 0,338 pada pembobotan linear dan 0,384 pada pembobotan kuadratik, yang menunjukkan tingkat kesesuaian kategori *fair*. Temuan tersebut menunjukkan bahwa ChatGPT memiliki potensi sebagai sistem pendukung dalam klasifikasi triase, namun belum dapat menggantikan peran tenaga medis secara penuh dan masih memerlukan pengembangan lebih lanjut, terutama pada kasus dengan tingkat kegawatan yang lebih tinggi.

Kata Kunci: kecerdasan buatan, triase, penanganan pasien

*Penulis Korespondensi:
Efa Dwina, email: efadwinaa@gmail.com

Highlight:

- AI ChatGPT menunjukkan akurasi yang memadai dan berpotensi dimanfaatkan sebagai alat bantu untuk mendukung klasifikasi triase *Simple Triage and Rapid Treatment* (START) di Instalasi Gawat Darurat (IGD) secara konsisten dan sistematis.
- Performa dan tingkat kesesuaian AI ChatGPT terhadap penilaian tenaga medis masih berada dalam kategori sedang (*fair agreement*), serta memperlihatkan keterbatasan dalam mengenali kasus dengan tingkat kegawatan tinggi.
- AI ChatGPT belum dapat menggantikan peran tenaga medis sebagai pengambil keputusan utama di IGD, sehingga pemanfaatannya disarankan sebatas sistem pendukung keputusan (*decision support tool*) dengan tetap memerlukan verifikasi dan pengawasan medis.

PENDAHULUAN

Instalasi Gawat Darurat (IGD) merupakan unit pelayanan rumah sakit yang berfokus pada penanganan pasien dengan kondisi kegawatdaruratan dan memerlukan tindakan medis secara cepat dan tepat. Pelayanan di IGD menuntut ketepatan pengambilan keputusan klinis dalam menentukan prioritas penanganan sesuai tingkat kegawatan pasien (Widodo dan Riani, 2026). Salah satu sistem yang digunakan dalam menentukan prioritas tersebut adalah triase, yaitu proses pengelompokan pasien berdasarkan tingkat kegawatan untuk menentukan prioritas pelayanan (Pranoto dan

Wibowo, 2020). Metode *Simple Triage and Rapid Treatment* (START) menjadi salah satu metode triase yang banyak digunakan karena memiliki prosedur yang sederhana, cepat, dan mudah diterapkan pada situasi kegawatdaruratan.

Namun, pelaksanaan triase masih berpotensi mengalami ketidaktepatan akibat berbagai faktor, seperti tingginya jumlah pasien, keterbatasan sumber daya, serta perbedaan pengalaman tenaga medis. Kondisi tersebut dapat menyebabkan terjadinya *undertriage* maupun *overtriage* yang berdampak terhadap kualitas pelayanan kegawatdaruratan (Harrigan et al., 2023). *Undertriage* berisiko menimbulkan keterlambatan penanganan pada pasien dengan kondisi berat, sedangkan *overtriage* dapat meningkatkan penggunaan sumber daya secara tidak efisien dan mengurangi ketersediaan pelayanan bagi pasien yang lebih membutuhkan (Huabbangyang et al., 2023).

Penelitian sebelumnya menunjukkan bahwa sistem AI dalam klasifikasi triase memiliki spesifisitas sebesar 88,2% dalam mengidentifikasi pasien *non-emergency*. Akan tetapi, masih ditemukan *false positive* sebesar 11,8% yang menunjukkan terjadinya *overtriage*, yaitu kondisi ketika pasien *non-emergency* diklasifikasikan sebagai pasien *emergency*. Kondisi tersebut berpotensi meningkatkan penggunaan sumber daya di unit gawat darurat secara kurang efisien, meskipun risiko klinisnya dinilai lebih rendah dibandingkan *undertriage* karena keselamatan pasien tetap lebih terjaga. Temuan tersebut menunjukkan bahwa pemanfaatan AI sebagai sistem pendukung keputusan klinis masih memerlukan evaluasi dan penerapan secara hati-hati (Kartika et al., 2025).

Perkembangan *Artificial Intelligence* (AI), khususnya model berbasis bahasa seperti ChatGPT, membuka peluang pemanfaatan teknologi dalam mendukung pengambilan keputusan klinis (Ayoub et al., 2023). ChatGPT memiliki kemampuan dalam memahami dan mengolah informasi berbasis teks secara cepat dan sistematis sehingga berpotensi dimanfaatkan sebagai alat bantu dalam klasifikasi triase (Zuhairini et al., 2025). Meskipun demikian, pemanfaatan AI dalam konteks kegawatdaruratan masih memerlukan evaluasi lebih lanjut, terutama terkait tingkat keakuratan dan kesesuaiannya dibandingkan dengan penilaian tenaga medis sebagai baku acuan.

Beberapa penelitian menunjukkan bahwa penggunaan AI dalam pelayanan kesehatan berpotensi meningkatkan efisiensi dan ketepatan pengambilan keputusan klinis. Namun, penelitian mengenai penggunaan ChatGPT dalam klasifikasi triase START masih terbatas, terutama yang membandingkan secara langsung hasil klasifikasi AI dengan tenaga medis sebagai baku acuan di IGD (Lee et al., 2025). Hingga saat ini, penelitian yang secara khusus mengevaluasi tingkat keakuratan dan kesesuaian AI ChatGPT dalam klasifikasi triase START dibandingkan tenaga medis sebagai baku acuan di IGD masih terbatas, khususnya pada setting pelayanan kesehatan di Indonesia. Selain itu, penelitian serupa pada fasilitas pelayanan kesehatan di Indonesia juga masih belum banyak dilakukan. Berdasarkan kondisi tersebut, penelitian ini memiliki kebaruan dalam mengevaluasi tingkat keakuratan dan kesesuaian AI ChatGPT pada klasifikasi triase START dibandingkan dengan tenaga medis di IGD RSUD Gunung Jati Kota Cirebon (Sandmann et al., 2024).

Penelitian ini dilakukan untuk mengevaluasi tingkat keakuratan dan kesesuaian hasil klasifikasi triase START oleh AI ChatGPT dibandingkan dengan penilaian tenaga medis di IGD RSUD Gunung Jati Kota Cirebon. Analisis dilakukan menggunakan *confusion matrix*, uji diagnostik yang meliputi sensitivitas, spesifisitas, akurasi, *positive predictive value* (PPV), dan *negative predictive value* (NPV), serta uji kesesuaian menggunakan Cohen's kappa. Penelitian ini diharapkan dapat memberikan gambaran

objektif mengenai potensi AI sebagai sistem pendukung dalam klasifikasi triase sekaligus menilai keterbatasannya dalam mendukung pengambilan keputusan klinis di IGD.

METODE

Penelitian ini menggunakan pendekatan kuantitatif dengan rancangan observasional analitik dan desain retrospektif. Data penelitian diperoleh dari rekam medis pasien yang berkunjung ke Instalasi Gawat Darurat (IGD) RSUD Gunung Jati Kota Cirebon selama periode Maret–Mei 2025. Penelitian ini bertujuan mengevaluasi tingkat keakuratan dan kesesuaian klasifikasi triase *Simple Triage and Rapid Treatment* (START) yang dihasilkan oleh AI ChatGPT dibandingkan dengan penilaian tenaga medis sebagai baku acuan.

Sampel penelitian dipilih menggunakan teknik *simple random sampling* dengan bantuan *Microsoft Excel*. Sebanyak 420 rekam medis pasien dewasa berusia ≥18 tahun yang memenuhi kriteria inklusi dianalisis dalam penelitian ini. Kriteria inklusi meliputi kelengkapan data klinis berupa usia, jenis kelamin, keluhan utama, riwayat penyakit, tingkat kesadaran, serta tanda vital seperti tekanan darah, frekuensi nadi, frekuensi pernapasan, suhu tubuh, dan saturasi oksigen. Adapun kriteria eksklusi dalam penelitian ini adalah rekam medis dengan data klinis yang tidak lengkap, tidak terbaca dengan jelas, atau memiliki informasi yang tidak sesuai dengan kebutuhan penelitian.

Data rekam medis yang memenuhi kriteria dimasukkan ke dalam AI ChatGPT versi GPT-4 untuk menghasilkan klasifikasi triase berdasarkan metode *Simple Triage and Rapid Treatment* (START), kemudian dibandingkan dengan hasil klasifikasi tenaga medis di IGD. Label baku acuan (*ground truth*) dalam penelitian ini ditentukan berdasarkan hasil klasifikasi triase yang dilakukan oleh perawat triase IGD dan tercatat pada rekam medis pasien. Data klinis dimasukkan ke dalam AI menggunakan format yang terstruktur dan seragam guna meminimalkan bias interpretasi selama proses klasifikasi. Prompt yang digunakan memuat informasi klinis pasien berupa usia, jenis kelamin, keluhan utama, tingkat kesadaran, tanda vital, serta riwayat penyakit penyerta, kemudian AI diminta menentukan klasifikasi triase berdasarkan metode START. Setiap data pasien dimasukkan menggunakan format prompt yang sama untuk menjaga konsistensi proses klasifikasi.

Analisis data dilakukan menggunakan IBM SPSS Statistics serta perhitungan tambahan melalui *confusion matrix*. Analisis univariat digunakan untuk mendeskripsikan karakteristik subjek dalam bentuk distribusi frekuensi dan persentase. Perbandingan hasil klasifikasi antara tenaga medis dan AI ChatGPT disajikan melalui tabel silang (*cross-tabulation/confusion matrix*). Kinerja AI dievaluasi menggunakan uji diagnostik yang meliputi sensitivitas, spesifisitas, akurasi, *positive predictive value* (PPV), dan *negative predictive value* (NPV). Tingkat kesesuaian antara kedua metode klasifikasi dianalisis menggunakan Cohen's kappa.

Uji Cohen's kappa digunakan untuk menilai tingkat kesepakatan antara hasil klasifikasi triase START oleh AI ChatGPT dan penilaian tenaga medis sebagai baku acuan. Penggunaan uji ini sesuai dengan jenis data kategorik karena mampu mengukur derajat kesesuaian dengan mempertimbangkan kemungkinan kesepakatan yang terjadi secara kebetulan (Hanegraaf et al., 2024).

Penelitian ini telah memperoleh persetujuan etik dari Komite Etik RSUD Gunung Jati Kota Cirebon dengan nomor No.044/LAIKETIK/KEPPKRSGJ/VIII/2025. Seluruh data pasien yang digunakan dalam penelitian dijaga kerahasiaannya dengan

menyamarkan identitas subjek dan hanya digunakan untuk kepentingan penelitian. Akses terhadap data penelitian dibatasi hanya kepada peneliti guna menjaga keamanan dan kerahasiaan informasi pasien. Selain itu, penggunaan desain retrospektif dalam penelitian ini memungkinkan adanya keterbatasan terkait kelengkapan dan kualitas data rekam medis yang tersedia.

HASIL DAN PEMBAHASAN

Data demografi pasien yang masuk IGD

Berdasarkan Tabel 1, mayoritas subjek berada pada kelompok usia dewasa 30–59 tahun sebanyak 258 subjek (61,4%) dengan rerata usia 45 ± 15,1 tahun dan rentang usia 18–94 tahun. Subjek didominasi oleh laki-laki sebanyak 222 orang (52,9%).

Tabel 1. Data demografi pasien yang masuk IGD RSUD Gunung Jati Kota Cirebon periode Maret–Mei 2025 (n=420)

Parameter	Kategori	F (%)	Mean (±SD)	Min-Max
Usia	Remaja (13- 19 tahun)	7 (1,7)	45 ±15,1	18–94
	Dewasa muda (20-29 tahun)	90 (21,4)		
	Dewasa (30-59 tahun)	258 (61,4)		
	Lansia (60 tahun)	65 (15,5)		
Tanda–tanda Vital	Sistol (mmhg)		123 ±28,5	73-255
	Diastol (mmhg)		79 ±17,6	41-162
	Nadi (kali/menit)		101 ±23,9	39-193
	Pernapasan (kali/menit)		24 ±10,2	12-46
	Suhu		36,8 ±1,2	35,2-41
	SpO_2 (%)		97 ±6	82-100
	Skor GCS		15 ±1,2	3-15
Jenis Kelamin	Laki–laki	222 (52,9)		
	Perempuan	198 (47,1)		
Komorbid	Hipertensi	43 (10,2)		
	Diabetes melitus	46 (11)		
	Penyakit jantung koroner	88 (21)		
	Gagal ginjal kronik	15 (3,6)		
	Infeksi pernapasan	57 (13,6)		
	Stroke	3 (0,7)		
	Dyspepsia	168 (40)		
Keluhan Utama	Nyeri perut	50 (11,9)		
	Nyeri dada	125 (29,8)		
	Sesak napas	69 (16,4)		
	Demam	77 (18,3)		
	Mual atau muntah	38 (9,0)		
	Perdarahan	45 (10,7)		
	Badan lemas	16 (3,8)		

Keterangan: Data sekunder yang diperoleh dari rekam medis pasien Instalasi Gawat Darurat RSUD Gunung Jati Kota Cirebon periode Maret–Mei 2025

Tabel 1 pada pemeriksaan tanda vital juga menunjukkan rerata tekanan darah sistolik sebesar 123 ± 28,5 mmHg, tekanan darah diastolik 79 ± 17,6 mmHg, frekuensi nadi 101 ± 23,9 kali/menit, frekuensi pernapasan 24 ± 10,2 kali/menit, suhu tubuh 36,8

± 1,2°C, saturasi oksigen 97 ± 6,0%, serta skor GCS 15 ± 1,2. Komorbid yang paling banyak ditemukan adalah dyspepsia sebanyak 168 subjek (40,0%), sedangkan keluhan utama terbanyak yaitu nyeri dada sebanyak 125 subjek (29,8%).

Karakteristik subjek dalam penelitian ini menunjukkan bahwa pasien yang datang ke IGD memiliki kondisi klinis yang beragam, baik dari sisi usia, tanda vital, komorbiditas, maupun keluhan utama. Variasi kondisi tersebut menggambarkan bahwa proses triase memerlukan penilaian yang cepat, tepat, dan konsisten untuk menentukan prioritas pelayanan sesuai tingkat kegawatan pasien. Penilaian triase tidak hanya dipengaruhi oleh satu indikator klinis, tetapi juga memerlukan interpretasi menyeluruh terhadap kombinasi tanda vital, tingkat kesadaran, serta kondisi penyerta pasien (NCHS, 2022).

Keluhan utama yang paling sering ditemukan berupa nyeri dada, demam, dan sesak napas, yang umumnya berkaitan dengan kondisi kardiovaskular maupun gangguan respirasi dan memerlukan penanganan segera di IGD. Selain itu, keberadaan komorbid seperti penyakit jantung koroner, diabetes mellitus, dan infeksi pernapasan berpotensi meningkatkan kompleksitas penilaian klinis karena dapat memperburuk kondisi pasien secara cepat. Kondisi tersebut menunjukkan bahwa ketepatan klasifikasi triase menjadi aspek penting dalam pelayanan kegawatdaruratan guna mengurangi risiko keterlambatan penanganan maupun kesalahan prioritas pelayanan. Oleh karena itu, pemanfaatan *Artificial Intelligence* (AI) sebagai sistem pendukung keputusan klinis berpotensi membantu meningkatkan konsistensi proses triase, meskipun penggunaannya tetap memerlukan evaluasi dan pengawasan tenaga medis (Sax et al., 2025). Keberagaman karakteristik pasien dalam penelitian ini juga menunjukkan bahwa sistem triase memerlukan konsistensi penilaian untuk mendukung pengambilan keputusan klinis secara cepat dan tepat di IGD.

Perbandingan kategori triase START antara Rumah Sakit dan AI ChatGPT

Tabel 2 menunjukkan distribusi kategori triase START berdasarkan penilaian Rumah Sakit dan *AI ChatGPT*. Pada kedua penilaian, kategori yang paling dominan adalah kuning. Namun, *AI ChatGPT* cenderung mengklasifikasikan lebih banyak pasien ke dalam kategori merah dan hijau dibandingkan dengan Rumah Sakit. Temuan ini menunjukkan adanya perbedaan pola distribusi kategori triase antara kedua metode penilaian.

Tabel 2. Distribusi kategori triage START menurut Rumah Sakit dan *AI ChatGPT*

Kategori	Rumah Sakit		*AI ChatGPT*	
	n	%	n	%
Hijau	7	1,7	33	7,9
Kuning	353	84,0	278	66,2
Merah	60	14,3	109	26,0
Hitam	0	0	0	0
Total	**420**	**100,0**	**420**	**100,0**

Keterangan: n = jumlah subjek; % = persentase dari total 420 subjek; Distribusi kategori triase disajikan berdasarkan penilaian Rumah Sakit dan AI ChatGPT

Berdasarkan Tabel 2, kategori triase START yang paling dominan pada penilaian Rumah Sakit adalah kategori kuning sebanyak 353 subjek (84,0%), diikuti kategori merah sebanyak 60 subjek (14,3%) dan hijau sebanyak 7 subjek (1,7%). Pola yang sama juga tampak pada penilaian *AI ChatGPT,* di mana kategori kuning tetap menjadi

kategori terbanyak, yaitu 278 subjek (66,2%), diikuti kategori merah sebanyak 109 subjek (26,0%) dan hijau sebanyak 33 subjek (7,9%). Pada kedua metode penilaian tidak ditemukan pasien dalam kategori hitam.

Dominasi kategori kuning pada penilaian Rumah Sakit dan *AI ChatGPT* menunjukkan bahwa sebagian besar pasien dalam penelitian ini berada pada tingkat kegawatan yang memerlukan penanganan, namun belum termasuk dalam kondisi paling kritis. Dalam pelayanan IGD, triase berfungsi untuk mengidentifikasi tingkat keparahan pasien, menentukan prioritas pelayanan, serta mendukung pemanfaatan sumber daya secara lebih efektif (Uly et al., 2025). Oleh karena itu, distribusi kategori triase pada Tabel 2 dapat digunakan sebagai gambaran awal mengenai pola kegawatan pasien sebelum dilakukan analisis lanjutan terhadap akurasi dan kesesuaian klasifikasi.

Meskipun kategori kuning menjadi kategori terbanyak pada kedua metode penilaian, *AI ChatGPT* menunjukkan proporsi kategori merah dan hijau yang lebih tinggi dibandingkan Rumah Sakit. Perbedaan ini mengindikasikan adanya kemungkinan pergeseran klasifikasi antara hasil *AI* dan penilaian tenaga medis. Dalam konteks triase, pergeseran klasifikasi perlu diperhatikan karena dapat mengarah pada *undertriage* maupun *overtriage*. *Undertriage* berisiko menyebabkan pasien dengan kondisi serius tidak segera memperoleh prioritas penanganan, sedangkan *overtriage* dapat menyebabkan penggunaan sumber daya pelayanan menjadi kurang efisien. Oleh sebab itu, hasil distribusi ini perlu dilanjutkan dengan analisis tabel silang, uji diagnostik, dan uji kesesuaian untuk menilai ketepatan serta konsistensi klasifikasi *AI ChatGPT* terhadap baku acuan Rumah Sakit.

Penjelasan ini didukung oleh Margaret (2023) yang mengaitkan *mistriage* dengan definisi *undertriage* dan *overtriage* berbasis intervensi serta kebutuhan sumber daya, Pada penelitian Sari dan Fajarini (2022) menunjukkan bahwa mistriage dapat berupa undertriage maupun overtriage, serta Sax et al. (2025) yang membahas faktor terkait *undertriage* dan *overtriage* pada pasien trauma.

Tabel 3. Tabel silang kategori triase START antara Rumah Sakit dan *AI ChatGPT*

Triase RS	***AI ChatGPT***			
	Hijau	**Kuning**	**Merah**	**Total**
Hijau	5	2	0	7
Kuning	28	257	68	353
Merah	0	19	41	60
Total	**420**	**100,0**	**420**	**100,0**

Keterangan: Tabel disajikan dalam bentuk *confusion matrix*

Berdasarkan Tabel 3, tingkat kesesuaian klasifikasi triase START antara Rumah Sakit dan *AI ChatGPT* paling banyak ditemukan pada kategori kuning, yaitu sebanyak 257 kasus, kemudian diikuti kategori merah sebanyak 41 kasus dan kategori hijau sebanyak 5 kasus. Ketidaksesuaian klasifikasi terutama tampak pada pasien yang menurut Rumah Sakit termasuk kategori kuning, namun oleh *AI ChatGPT* diklasifikasikan sebagai hijau sebanyak 28 kasus dan sebagai merah sebanyak 68 kasus. Selain itu, terdapat 19 kasus yang menurut Rumah Sakit dikategorikan sebagai merah, tetapi oleh *AI ChatGPT* dinilai sebagai kuning. Pada kedua metode penilaian tidak ditemukan kasus dalam kategori hitam.

Tabel 3 menunjukkan bahwa kesesuaian paling banyak terjadi pada kategori kuning. Artinya, sebagian besar pasien yang dinilai kuning oleh Rumah Sakit juga dinilai kuning oleh *AI ChatGPT*. Namun, masih terdapat beberapa perbedaan

klasifikasi, terutama pada pasien yang menurut Rumah Sakit termasuk kategori kuning tetapi oleh *AI* dinilai sebagai hijau atau merah. Selain itu, ada juga pasien yang menurut Rumah Sakit termasuk kategori merah tetapi oleh *AI* dinilai sebagai kuning. Perbedaan ini menunjukkan bahwa *AI ChatGPT* belum sepenuhnya konsisten dalam menentukan kategori triase pada beberapa kasus.

Tabel silang atau *confusion matrix* pada penelitian ini digunakan untuk menggambarkan pola kesesuaian dan ketidaksesuaian klasifikasi triase START antara penilaian Rumah Sakit dan *AI ChatGPT* (Rainio, 2024). Melalui tabel ini, dapat diketahui kategori yang paling sering diklasifikasikan secara sama oleh kedua metode, sekaligus kategori yang mengalami pergeseran.

Dalam evaluasi model klasifikasi, *confusion matrix* berperan penting karena mampu menampilkan distribusi hasil prediksi pada setiap kelas dan menjadi dasar untuk menilai performa klasifikasi secara lebih rinci (Hicks et al., 2022). Oleh karena itu, Tabel 3 tidak hanya menunjukkan jumlah kasus yang sesuai dan tidak sesuai, tetapi juga menjadi dasar awal sebelum dilakukan perhitungan uji diagnostik dan uji kesesuaian. Hal ini sejalan dengan Rainio (2024) yang menjelaskan bahwa *confusion matrix* digunakan dalam evaluasi klasifikasi biner maupun multikelas untuk menilai performa model secara terstruktur.

Dalam pelayanan gawat darurat, perbedaan klasifikasi seperti ini penting diperhatikan karena dapat memengaruhi prioritas penanganan pasien. Jika pasien yang seharusnya masuk kategori lebih gawat dinilai lebih ringan, maka dapat terjadi undertriage dan berisiko menyebabkan keterlambatan penanganan (Sax et al., 2025). Sebaliknya, jika pasien yang sebenarnya tidak terlalu gawat dinilai lebih berat, maka dapat terjadi overtriage yang dapat membuat penggunaan sumber daya IGD menjadi kurang efisien (Sax et al., 2025). Oleh karena itu, hasil tabel silang ini perlu dilanjutkan dengan uji diagnostik dan uji kappa untuk menilai secara lebih jelas tingkat akurasi dan kesesuaian *AI ChatGPT* dibandingkan dengan penilaian Rumah Sakit.

Tabel 4. Parameter uji diagnostik *AI ChatGPT* terhadap penilaian triase Rumah Sakit

Kategori	Sensitivitas	Spesifisitas	Akurasi	PPV	NPV	False Negative	False Positive
Hijau	71,4	93,2	92,9	15,2	99,5	28,6	6,8
Kuing	72,8	68,7	72,1	92,4	32,4	27,2	31,3
Merah	68,3	81,1	79,3	37,6	93,9	31,7	18,9

Keterangan: Nilai sensitivitas, spesifisitas, akurasi, positive predictive value (PPV), dan negative predictive value (NPV) diperoleh dengan menggunakan hasil triase Rumah Sakit sebagai baku acuan

Berdasarkan Tabel 4, hasil uji diagnostik menunjukkan bahwa performa AI ChatGPT dalam mengklasifikasikan triase START berbeda pada setiap kategori. Pada kategori hijau, *AI ChatGPT* menunjukkan sensitivitas 71,4%, spesifisitas 93,2%, akurasi 92,9%, PPV 15,2%, NPV 99,5%, false negative 28,6%, dan false positive 6,8%. Pada kategori kuning, diperoleh sensitivitas 72,8%, spesifisitas 68,7%, akurasi 72,1%, PPV 92,4%, NPV 32,4%, false negative 27,2%, dan false positive 31,3%. Sementara itu, pada kategori merah, sensitivitas sebesar 68,3%, spesifisitas 81,1%, akurasi 79,3%, PPV 37,6%, NPV 93,9%, false negative 31,7%, dan false positive 18,9%. Secara umum, hasil ini menunjukkan bahwa *AI ChatGPT* memiliki performa yang cukup baik pada beberapa parameter, terutama pada kategori kuning dan hijau, tetapi masih memiliki

keterbatasan dalam mendeteksi kategori merah. Kategori hitam tidak dapat dievaluasi karena tidak ditemukan kasus pada kedua metode penilaian.

Pada parameter sensitivitas, menunjukkan bahwa kemampuan *AI* dalam mengenali kategori merah masih lebih rendah dibandingkan kategori lainnya. Kondisi ini perlu menjadi perhatian karena kategori merah merepresentasikan pasien dengan tingkat kegawatan lebih tinggi, sehingga kegagalan mendeteksi kategori ini dapat berimplikasi pada keterlambatan prioritas penanganan (Sax et al., 2025).

Penilaian ini diarahkan untuk melihat kemampuan *AI ChatGPT* dalam mengenali pasien yang benar-benar termasuk ke dalam kategori triase tertentu menurut penilaian Rumah Sakit sebagai baku acuan (Hoyer dan Zapf, 2021). Dalam penelitian ini, sensitivitas menjadi penting karena berkaitan dengan kemampuan *AI* dalam mendeteksi pasien yang seharusnya memperoleh prioritas sesuai tingkat kegawatannya. Sensitivitas merupakan salah satu ukuran utama dalam studi akurasi diagnostik karena menggambarkan kemampuan suatu metode dalam mengenali kasus positif berdasarkan standar acuan.

Pada parameter spesifisitas, pembahasan berfokus pada kemampuan *AI ChatGPT* dalam mengidentifikasi pasien yang memang tidak termasuk dalam kategori triase tertentu. Spesifisitas penting karena menunjukkan sejauh mana *AI* mampu menghindari pemberian label kategori yang tidak sesuai (Hoyer dan Zapf, 2021).

Pada penelitian ini, spesifisitas yang lebih tinggi pada kategori hijau menunjukkan bahwa *AI* relatif baik dalam membedakan pasien yang bukan kategori hijau. Namun, spesifisitas yang lebih rendah pada kategori kuning memperlihatkan bahwa *AI* masih cenderung mengklasifikasikan sebagian pasien ke kategori kuning meskipun menurut baku acuan pasien tersebut tidak termasuk dalam kategori tersebut. Oleh karena itu, spesifisitas perlu dibaca bersama sensitivitas agar performa *AI* tidak hanya dilihat dari kemampuan mengenali kategori tertentu, tetapi juga dari kemampuannya menyingkirkan kategori yang tidak tepat (Lim, 2021).

Pada parameter akurasi, pembahasan diarahkan untuk menilai ketepatan klasifikasi *AI ChatGPT* secara keseluruhan dibandingkan dengan penilaian Rumah Sakit sebagai baku acuan. Akurasi memiliki peran penting karena memberikan gambaran umum mengenai proporsi hasil klasifikasi *AI* yang sesuai pada seluruh data penelitian (Hicks et al., 2022). Berbeda dengan sensitivitas dan spesifisitas yang hanya menilai kemampuan *AI* dalam mengenali atau menyingkirkan kategori tertentu, akurasi menunjukkan performa secara menyeluruh (Rainio, 2024). Namun, pada penelitian ini, akurasi perlu ditafsirkan secara hati-hati karena distribusi kategori triase tidak seimbang dan didominasi oleh kategori kuning. Dengan demikian, akurasi tidak dapat dijadikan satu-satunya dasar penilaian, melainkan perlu dibaca bersama parameter lain seperti sensitivitas, spesifisitas, PPV, NPV, false negative, dan false positive agar *performa AI ChatGPT* dapat dipahami secara lebih komprehensif.

Pada parameter *positive predictive value* (PPV) dan *negative predictive value* (NPV), penilaian difokuskan pada tingkat kepercayaan terhadap hasil klasifikasi *AI ChatGPT*. PPV menunjukkan kemungkinan bahwa pasien yang diklasifikasikan *AI* ke dalam suatu kategori benar-benar sesuai dengan penilaian Rumah Sakit, sedangkan NPV menunjukkan kemungkinan bahwa pasien yang dinyatakan tidak termasuk suatu kategori memang benar tidak termasuk kategori tersebut (Hoyer dan Zapf, 2021).

Pada penelitian ini, PPV yang lebih tinggi pada kategori kuning menunjukkan bahwa klasifikasi *AI* pada kategori tersebut relatif lebih dapat dipercaya. Sebaliknya, PPV yang lebih rendah pada kategori hijau dan merah menunjukkan bahwa hasil positif pada kedua kategori tersebut masih perlu ditafsirkan secara hati-hati (Hicks et al.,

2022). NPV yang lebih baik pada kategori hijau dan merah menunjukkan bahwa keputusan *AI* untuk menolak kategori tersebut relatif lebih konsisten, sedangkan rendahnya NPV kategori kuning dapat dipengaruhi oleh dominannya kategori tersebut dalam data. Oleh karena itu, PPV dan NPV perlu dibaca bersama parameter diagnostik lain agar performa *AI ChatGPT* dapat dipahami secara lebih menyeluruh (Hoyer dan Zapf, 2021).

Pada parameter *false negative* dan *false positive*, pembahasan berfokus pada bentuk kesalahan klasifikasi yang masih ditemukan pada hasil triase *AI ChatGPT* (Chen et al., 2024). *False negative* menggambarkan kondisi ketika pasien yang seharusnya masuk dalam kategori tertentu menurut Rumah Sakit tidak dikenali oleh *AI*, sedangkan *false positive* menunjukkan kondisi ketika *AI* memasukkan pasien ke kategori tertentu padahal menurut baku acuan pasien tersebut tidak termasuk dalam kategori tersebut (Hicks et al., 2022).

Dalam penelitian ini, *false negative* pada kategori merah menjadi perhatian karena dapat mengarah pada *undertriage*, yaitu kondisi ketika pasien dengan tingkat kegawatan lebih tinggi dinilai lebih ringan dari seharusnya. Sebaliknya, *false positive* pada kategori kuning dapat menunjukkan kecenderungan *overtriage,* yaitu pasien diklasifikasikan ke tingkat prioritas yang tidak sesuai dengan kondisi sebenarnya (Chen et al., 2024). Dalam konteks IGD, *undertriage* berisiko menyebabkan keterlambatan penanganan, sedangkan *overtriage* dapat berdampak pada penggunaan sumber daya yang kurang efisien. Oleh karena itu, kedua parameter ini penting untuk menilai keamanan dan ketepatan *AI ChatGPT* sebagai alat bantu triase, terutama pada kategori dengan tingkat kegawatan tinggi (Huabbangyang et al., 2023).

Tabel 5. Hasil uji kesesuaian antara triase Rumah Sakit dan AI ChatGPT

Uji Kesesuaian	Nilai	*p-value*	Interpretasi
Cohen's Kappa	0,313	<0,001	Fair

Keterangan: Nilai *Cohen's kappa* diperoleh dari analisis SPSS

Berdasarkan Tabel 5, hasil uji kesesuaian menggunakan Cohen's kappa menunjukkan nilai 0,313 dengan *p-value* <0,001. Hasil tersebut menunjukkan adanya kesesuaian yang bermakna secara statistik antara klasifikasi triase oleh tenaga medis dan AI ChatGPT. Namun, nilai kappa yang berada pada kategori *fair agreement* menunjukkan bahwa tingkat kesesuaian AI terhadap penilaian tenaga medis masih belum optimal.

Uji Cohen's kappa digunakan karena penelitian ini membandingkan dua hasil klasifikasi yang bersifat kategorik, yaitu hasil triase tenaga medis dan AI ChatGPT. Metode ini dinilai lebih tepat dibandingkan persentase kesesuaian biasa karena mempertimbangkan kemungkinan kesamaan hasil yang dapat terjadi secara kebetulan (Rainio, 2024). Hasil penelitian ini sejalan dengan temuan pada *confusion matrix* dan uji diagnostik yang masih menunjukkan adanya pergeseran klasifikasi pada beberapa kategori triase, terutama antara kategori kuning dan merah.

Temuan tersebut menunjukkan bahwa AI ChatGPT berpotensi digunakan sebagai alat bantu dalam klasifikasi triase, tetapi penggunaannya tetap memerlukan verifikasi tenaga medis untuk mengurangi risiko ketidaktepatan klasifikasi (Dettori & Norvell, 2020). Hasil ini juga sesuai dengan literatur sebelumnya yang menyatakan bahwa nilai Cohen's kappa pada rentang 0,21–0,40 termasuk dalam kategori *fair agreement* (Li et al., 2023).

KESIMPULAN

Hasil penelitian ini menunjukkan bahwa AI ChatGPT memiliki potensi sebagai sistem pendukung dalam klasifikasi triase *Simple Triage and Rapid Treatment* (START) di Instalasi Gawat Darurat, karena menunjukkan tingkat keakuratan dan kesesuaian tertentu terhadap penilaian tenaga medis. Namun, konsistensi hasil klasifikasi yang dihasilkan AI masih belum sepenuhnya optimal, sehingga penggunaannya belum dapat menggantikan peran tenaga medis sebagai pengambil keputusan utama dalam menentukan tingkat kegawatan pasien. Oleh karena itu, pemanfaatan AI ChatGPT lebih tepat digunakan sebagai alat bantu untuk mendukung proses triase secara lebih konsisten dan sistematis dalam pelayanan kegawatdaruratan. Temuan penelitian ini menunjukkan bahwa pengembangan dan evaluasi lebih lanjut masih diperlukan, khususnya dalam meningkatkan ketepatan klasifikasi pada kasus dengan tingkat kegawatan tinggi. Selain memiliki implikasi praktis dalam mendukung pengambilan keputusan klinis di IGD, hasil penelitian ini juga dapat menjadi bahan pertimbangan bagi fasilitas pelayanan kesehatan dalam pengembangan kebijakan terkait pemanfaatan teknologi *Artificial Intelligence* (AI) sebagai sistem pendukung pelayanan kegawatdaruratan. Penelitian selanjutnya disarankan melibatkan jumlah sampel yang lebih besar, variasi kasus klinis yang lebih beragam, serta penerapan pada berbagai setting pelayanan kesehatan guna memperoleh gambaran yang lebih komprehensif mengenai keandalan AI dalam mendukung proses triase di IGD.

UCAPAN TERIMA KASIH

Penulis menyampaikan terima kasih kepada Universitas Muhammadiyah Cirebon dan RSUD Gunung Jati Kota Cirebon atas dukungan, kerja sama, serta kontribusi yang diberikan selama proses penelitian hingga penyusunan naskah publikasi ini.

DAFTAR PUSTAKA

Ayoub, M., Ballout, A.A., Zayek, R.A., Ayoub, N.F., 2023. Mind + Machine: ChatGPT as a Basic Clinical Decisions Support Tool. Cureus Journal of Medical Science 15(8), 8–11. Https://Doi.Org/10.7759/Cureus.43690

Chen, Q., Qin, Y., Jin, Z., Zhao, X., He, J., Wu, C., 2024. Enhancing Performance of The National Field Triage Guidelines using Machine Learning: Development of A Prehospital Triage Model to Predict Severe Trauma. Journal of Medical Internet Research 26, 1–15. Https://Doi.Org/10.2196/58740

Dettori, J.R., Norvell, D.C., 2020. Kappa and Beyond: Is There Agreement? Global Spine Journal 10(4), 499–501. Https://Doi.Org/10.1177/2192568220911648

Hanegraaf, P., Wondimu, A., Mosselman, J.J., Jong, R.D., Abogunrin, S., Queiros, L., Lane, M., Postma, M.J., 2024. Reviewer Reliability of Human Literature Reviewing and Implications Assisted for The Introduction of Machine Systematic Reviews: A Mixed Methods Review. BMJ Open 14(3), 1–10. Https://Doi.Org/10.1136/Bmjopen-2023-076912

Harrigan, M.E., Boremski, P.A., Collier, B.R., Tegge, A.N., Gillen, J.R., 2023. Impact of Nonphysician, Technology-Guided Alert Level Selection on Rates of Appropriate Trauma Triage in The United States: A Before and After Study. Journal of Trauma and Injury 36(3), 231–241. https://doi.org/10.20408/jti.2023.0020

Hicks, S.A., Strümke, I., Thambawita, V., Hammou, M., Riegler, M.A., Halvorsen, P., Parasa, S., 2022. On Evaluation Metrics for Medical Applications of Artificial Intelligence. Scientific Reports 12(5979), 1–9. Https://Doi.Org/10.1038/S41598-022-09954-8

Hoyer, A., Zapf, A., 2021. Studies for The Evaluation of Diagnostic Tests. Deutsches Arzteblatt 118, 555-560. Https://Doi.Org/10.3238/Arztebl.M2021.0224

Huabbangyang, T., Rojsaengroeng, R., Tiyawat, G., Silakoon, A., Vanichkulbodee, A., On, J.S., Buathong, S., 2023. Associated Factors of Under and Over-Triage Based on The Emergency Severity Index; A Retrospective Cross- Sectional Study. Archives of Academic Emergency Medicine 11(1), 1–11. https://doi.org/10.22037/aaem.v11i1.2076

Kartika, A.P.T., Akbar, R., Atmaji, W.A., Dwina, E., Gennie, A., Krisna, A.A., 2025. Triase Instalasi Gawat Daruat Berbasis Artificial Intelligence Berdasarkan Emergency Severity Index. Jurnal Ners 10(1), 2522–2529. https://journal.universitaspahlawan.ac.id/index.php/ners/article/view/54523

Lee, S., Jung, S., Park, J.H., Cho, H., Moon, S., Ahn, S., 2025. Performance of ChatGPT, Gemini and Deepseek for Non-Critical Triage Support Using Real-World Conversations in Emergency Department. BMC Emergency Medicine 25(1), 1-20. https://doi.org/10.1186/s12873-025-01337-2

Li, M., Gao, Q., Yu, T., 2023. Kappa Statistic Considerations in Evaluating Inter-Rater Reliability Between Two Raters: Which, When, and Context Matters. BMC Cancer 23(1), 1–5. https://doi.org/10.1186/s12885-023-11325-z

Lim, C., 2021. Methods for Evaluating The Accuracy of Diagnostic Tests. Cardiovascular Prevention and Pharmacotherapy 3(1), 15–20. https://e-jcpp.org/journal/view.php?doi=10.36011/cpp.2021.3.e2

Margaret, E., 2023. Evaluation of The Version 4 of the Emergency Severity Index in US Emergency Departments for The Rate of Mistriage. JAMA Netw Open 6(3), 1-19. Https://Doi.Org/10.1001/Jamanetworkopen.2023.3404

[NCHS] National Center for Health Statistics., 2022. National Hospital Ambulatory Medical Care Survey: 2022 Emergency Department Summary Tables. National Center for Health Statistics, Washington.

Pranoto, Y.A., Wibowo, S.A., 2020. Aplikasi Desktop Sistem Triase untuk Pendukung Prioritas Tingkat Kegawatan. Jurnal Ilmu Komputer dan Teknok Informatika 3(1), 1–6. https://ejournal.itn.ac.id/index.php/mnemonic/article/view/2319

Rainio, O., Teuho, J., Klen, R., 2024. Evaluation Metrics and Statistical Tests for Machine Learning. Scientific Reports 14(6086), 1–14. Https://Doi.Org/10.1038/S41598-024-56706-X

Sandmann, S., Riepenhausen, S., Plagwitz, L., Varghese, J., 2024. Systematic Analysis of ChatGPT, Google Search and Llama 2 for Clinical Decision Support Tasks. Nature Communications 15, 1–8. Https://Doi.Org/10.1038/S41467-024-46411-8

Sari, S.R., Fajarini, M., 2022. The Emergency Severity Indeks (ESI) Usage: Triage Accuracy and Causes Of Mistriage. Jurnal Ilmu Kesehatan 7(S1), 243–248. Https://Doi.Org/10.30604/Jika.V7is1.1190

Sax, D.R., Warton, E.M., Mark, D.G., Reed, M.E., 2025. Emergency Department Triage Accuracy and Delays in Care for High-Risk Conditions. JAMA Network Open 8(5), 1–12. Https://Doi.Org/10.1001/Jamanetworkopen.2025.8498

Uly, R.G.Z., Sriyono, S., Qona'ah, A., 2025. Triage in Emergency Management in the Emergency Departement: A Systematic Review. Indonesian Journal of Global Health Research 7(2), 527–534.

https://jurnal.globalhealthsciencegroup.com/index.php/IJGHR/article/view/5607

Widodo, S,P., Riani, D.A., 2026. Literatur Desain Instalasi Gawat Darurat (IGD) pada Rumah Sakit di Bekasi. Jurnal Riset Ilmiah 3(2), 513–522. https://manggalajournal.org/index.php/SINERGI/article/view/2328

Zuhairini, R., Natasyah, W., Nurfadillah, W., Putri, I.F., Sapikah, N., 2025. Pemanfaatan Artificial Intelligence dalam Pengambilan Keputusan Klinis dan Manajemen Pasien Gawat Darurat di Bidang Anestesiologi: Sebuah Scoping Review. Jurnal Siti Rufaidah 3(4), 310-326. https://journal.ppniunimman.org/index.php/JASIRA/article/view/271