



Systematic Literature Review on the Integration and Role of Artificial Intelligence in the Development of Computer Adaptive Testing (CAT)

Muhammad Gibran Alif Prasetya*¹, Arif Widiyatmoko ²

*gibranalif45@students.unnes.ac.id

^{1,2} Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Semarang, Semarang, Indonesia

ABSTRACT

The integration of Artificial Intelligence (AI) has sparked a fundamental transformation in assessment systems, shifting from static methods to dynamic and personalized paradigms through Computer Adaptive Testing (CAT). This study aims to map the state-of-the-art AI integration in CAT, identify technological evolution trends, and analyze the contributions of dominant algorithms in optimizing the core components of tests. A Systematic Literature Review (SLR) following the PRISMA guidelines was used to synthesize data from Scopus, WoS, IEEE, ERIC, and Google Scholar databases from 2020 to 2025. Macro analysis was performed with VOSviewer, and micro synthesis with NVivo. The results indicate an evolution trend from basic machine learning integration in 2020 towards automation systems based on Reinforcement Learning and Generative AI by 2025. Algorithms such as Deep Learning and Multi-Objective Optimization have been shown to improve the precision of ability estimation, while empirical findings demonstrate that the use of Model Trees (M5P) can reduce item counts by 85%–93% on clinical instruments without compromising score accuracy. In conclusion, AI is transforming CAT into a smart, efficient, and personalized assessment ecosystem, with challenges in transparency (explainability) and algorithmic bias as key priorities for the future development of evaluation systems.

Keywords: Computer Adaptive Test (CAT), Artificial Intelligence (AI), Machine Learning, Assessment Efficiency, Item Selection Algorithm, Generative AI

PENDAHULUAN

Integrasi *Artificial Intelligence* (AI) dalam sistem penilaian pendidikan telah memicu transformasi fundamental dari metode pengujian statis menuju paradigma penilaian yang dinamis dan personal melalui *Computer Adaptive Test* (CAT). Secara tradisional, CAT mengandalkan *Item Response Theory* (IRT) dan algoritma seleksi berbasis aturan statistik seperti *Maximum Fisher Information* (MFI) untuk mengestimasi kemampuan peserta tes. Namun, kompleksitas pengujian skala besar di era digital menuntut pendekatan yang lebih canggih daripada sekadar aturan heuristik manual yang kaku dan spesifik pada model tertentu (Q. Liu et al., 2024). Fenomena saat ini menunjukkan pergeseran menuju *Neural Computerized Adaptive Testing* (NCAT) dan penggunaan *Deep Reinforcement Learning* (DRL), di mana algoritma seleksi soal tidak lagi diprogram secara manual, melainkan dipelajari secara langsung dari data respons siswa dalam skala besar untuk menangkap hubungan non-linear yang kompleks antara siswa dan butir soal (Q. Liu et al., 2024). Evolusi ini menjanjikan efisiensi pengukuran yang lebih tinggi, namun membawa tantangan baru terkait transparansi dan validitas psikometrik yang perlu dipetakan secara komprehensif.

Perkembangan riset dalam kurun waktu 2020–2025 menunjukkan lonjakan signifikan dalam pemanfaatan *Large Language Models* (LLMs) dan *Generative AI* (GenAI) untuk memperkaya ekosistem CAT. Teknologi ini tidak hanya digunakan untuk estimasi kemampuan, tetapi juga untuk *Automatic Item Generation* (AIG) yang mampu menghasilkan ribuan butir soal berkualitas tinggi secara otomatis, mengatasi kelangkaan bank soal yang selama ini menjadi hambatan utama implementasi CAT (Tan *et al.*, 2024). Fakta empiris terbaru menunjukkan bahwa agen berbasis LLM kini sedang dikembangkan untuk membuat proses seleksi soal dalam CAT menjadi lebih mudah dijelaskan (*explainable*), mengatasi kritik utama terhadap model *deep learning* yang sering dianggap sulit dipahami cara kerjanya (Cheng *et al.*, 2024). Tren ini menandakan bahwa AI tidak lagi sekadar alat bantu hitung, melainkan entitas yang berperan aktif dalam pembuatan konten, penilaian esai otomatis, hingga deteksi perilaku menyimpang (*aberrant response*) secara *real-time* (Bulut *et al.*, 2024).

Meskipun potensi teknisnya sangat menjanjikan, literatur yang ada saat ini masih terfragmentasi antara perspektif ilmu komputer yang berfokus pada optimasi algoritma dan perspektif psikometrika yang mengutamakan validitas pengukuran. Penelitian terdahulu seperti karya Ghosh & Lan (2020) telah memperkenalkan kerangka kerja optimasi bilevel (BOBCAT) untuk pembelajaran seleksi soal berbasis data, namun sering kali mengabaikan aspek keadilan (*fairness*) dan bias algoritmik yang mungkin muncul dari data pelatihan. Selain itu, banyak studi validasi AI dalam CAT masih sangat bergantung pada simulasi dengan sampel kecil atau data sintesis, yang membatasi generalisasi temuan pada populasi dunia nyata yang heterogen (Colledani, Robusto, *et al.*, 2025). Keterbatasan metodologis ini menciptakan kesenjangan antara inovasi teoretis dan praktik implementasi yang dapat dipertanggungjawabkan secara etis.

Lebih jauh lagi, terdapat inkonsistensi dalam temuan terkait dampak psikologis dan keadilan penggunaan AI dalam CAT. Sementara beberapa studi mengklaim bahwa CAT berbasis AI meningkatkan keterlibatan dan mengurangi kecemasan ujian melalui personalisasi (Ogunsakin & Babatimehin, 2024), studi lain menyoroti risiko bias otomatisasi dan kurangnya transparansi yang dapat merugikan kelompok demografis tertentu (Christyodetaputri & Marwa, 2024). Kesenjangan ini diperparah oleh kurangnya perhatian pada infrastruktur teknologi di negara berkembang, di mana implementasi CAT canggih sering kali terhambat oleh kesenjangan digital dan kurangnya literasi AI di kalangan pendidik (Ayanwale *et al.*, 2024). Inkonsistensi hasil dan ketidakjelasan konsep mengenai standar etika AI dalam penilaian menciptakan kebutuhan mendesak untuk meninjau ulang literatur secara holistik.

Oleh karena itu, pendekatan *Systematic Literature Review* (SLR) dipilih sebagai metode utama dalam penelitian ini, mengingat kompleksitas dan laju inovasi yang cepat dalam domain ini tidak dapat ditangkap sepenuhnya melalui satu penelitian lapangan. Pendekatan lapangan tunggal hanya akan memberikan potret parsial dari konteks spesifik, sedangkan SLR memungkinkan sintesis bukti dari berbagai konteks geografis, jenjang pendidikan, dan arsitektur teknologi yang berbeda. SLR diperlukan untuk memetakan secara sistematis bagaimana algoritma *machine learning* diintegrasikan ke dalam komponen CAT mulai dari diagnosis kognitif, bank soal, hingga kontrol tes serta untuk mengidentifikasi pola keberhasilan dan kegagalan dalam implementasi praktisnya (Q. Liu *et al.*, 2024).

Penelitian ini bertujuan untuk melakukan pemetaan komprehensif dan analisis mendalam mengenai *state of the art* integrasi serta peran strategis *Artificial Intelligence* (AI) dalam transformasi sistem *Computer Adaptive Test* (CAT). Secara spesifik, studi ini diarahkan untuk mengidentifikasi tren evolusi teknologi, memetakan jenis algoritma AI yang paling dominan, serta menganalisis kontribusinya dalam mengoptimalkan komponen inti CAT seperti *item selection*, *automated scoring*, dan estimasi kemampuan peserta tes secara *real-time*. Melalui sintesis sistematis terhadap literatur multidisiplin, penelitian ini berupaya mengungkap celah penelitian (*research gaps*) dan tantangan implementasi yang ada, sehingga dapat memberikan

kerangka konseptual serta rekomendasi strategis bagi pengembang kebijakan pendidikan dan peneliti teknologi pembelajaran dalam merancang sistem evaluasi yang lebih presisi, efisien, dan personal. Fokus kajian akan diarahkan untuk menemukan solusi atas masalah *explainability* dan bias dalam algoritma seleksi soal, serta merumuskan rekomendasi praktis bagi pengembang tes untuk menciptakan sistem penilaian yang adaptif namun tetap adil dan transparan (Bulut *et al.*, 2024).

METODE PENELITIAN

Penelitian ini mengadopsi pendekatan kajian literatur sistematis yang komprehensif, bertujuan untuk memetakan, menganalisis, dan mengevaluasi secara mendalam integrasi dan peran *Artificial Intelligence* (AI) dalam pengembangan *Computer Adaptive Test* (CAT). Pendekatan sistematis ini dipilih untuk memastikan proses identifikasi, seleksi, dan penilaian literatur dilakukan secara terstruktur, transparan, dan dapat direplikasi, sehingga memungkinkan penyusunan sintesis ilmiah yang akurat dan komprehensif.

Kerangka metodologis penelitian ini secara substansial mengikuti panduan PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses), sebuah standar yang diakui luas untuk meningkatkan kualitas dan pelaporan tinjauan sistematis dalam berbagai bidang ilmiah. Adopsi PRISMA memastikan bahwa setiap langkah penelitian mulai dari perumusan pertanyaan penelitian, strategi pencarian, hingga seleksi artikel dilakukan dengan ketelitian dan memenuhi standar kualitas ilmiah tertinggi.

Selain itu, penelitian ini akan memanfaatkan analisis bibliometrik melalui perangkat lunak VOSviewer. VOSviewer akan digunakan untuk mengidentifikasi tren penelitian yang muncul, menganalisis jaringan kolaborasi antara peneliti, institusi, dan negara, serta memetakan kluster-kluster tematik berdasarkan kemunculan kata kunci yang terkait dengan AI dan CAT. Analisis ini akan memberikan gambaran kuantitatif mengenai lanskap penelitian, menyoroti area fokus yang dominan dan kesenjangan penelitian yang potensial.

Untuk sintesis kualitatif yang mendalam, penelitian ini juga akan menggunakan perangkat lunak NVivo. Setiap artikel yang berhasil disaring akan dianalisis secara kualitatif di NVivo, dengan fokus pada bagaimana AI diintegrasikan ke dalam CAT, metodologi spesifik AI yang diterapkan, tujuan utama dari integrasi tersebut, serta hasil dan implikasinya terhadap efektivitas dan efisiensi CAT. Proses ini melibatkan pengkodean (coding) data berdasarkan tema-tema relevan yang muncul dari literatur, seperti algoritmik dari AI (misalnya, *Machine Learning*, *Deep Learning*), mekanisme adaptasi dalam CAT yang menggunakan AI, tujuan akhir dari penggunaan AI (misalnya, peningkatan akurasi estimasi, personalisasi), serta tantangan dan manfaat yang dihadapi. Pendekatan ini memungkinkan peneliti untuk menggali informasi kualitatif yang kaya dan menyajikan pemahaman yang mendalam tentang evolusi dan implementasi AI dalam CAT.

Dengan menggabungkan analisis bibliometrik kuantitatif dengan VOSviewer dan sintesis kualitatif berbasis NVivo, penelitian ini menyatukan dua pendekatan yang saling melengkapi. Integrasi ini bertujuan untuk memberikan gambaran menyeluruh, baik dari perspektif makro maupun mikro, mengenai perkembangan, integrasi, dan peran AI dalam pengembangan CAT selama lebih dari satu dekade terakhir.

Sumber Data

Sumber data penelitian ini secara eksklusif akan diperoleh melalui perangkat lunak *Publish or Perish* (PoP). PoP dipilih karena kemampuannya yang unik untuk mengakses dan mengkompilasi literatur ilmiah dari berbagai basis data akademik terkemuka, seperti Google Scholar dan Google Scholar Profiles, yang mencakup spektrum luas publikasi ilmiah di bidang teknologi pendidikan, kecerdasan buatan, psikometri, dan evaluasi adaptif. Pemilihan PoP juga didasarkan pada fleksibilitasnya dalam melakukan pencarian *query* yang kompleks dan kemampuannya untuk mengeksplor metadata artikel secara efisien, yang sangat relevan untuk

analisis bibliometrik lanjutan.

Proses pengumpulan data akan dilakukan secara langsung melalui antarmuka PoP. Pencarian akan dirancang dengan kombinasi kata kunci yang relevan, seperti "Artificial Intelligence", "AI", "Computer Adaptive Test", "CAT", "machine learning", "deep learning", dan "assessment", untuk mengidentifikasi artikel-artikel yang paling relevan dengan integrasi AI dalam pengembangan CAT. PoP akan memungkinkan penyaringan dan pengunduhan metadata artikel secara komprehensif, mencakup informasi penting seperti nama penulis, afiliasi, tahun publikasi, judul jurnal, jumlah sitasi, abstrak, dan kata kunci yang disediakan oleh penulis.

Dokumen yang dikumpulkan akan dibatasi pada rentang tahun 2020 hingga 2025. Batasan tahun ini ditetapkan untuk menangkap evolusi dan perkembangan kritis dalam integrasi AI dalam CAT selama lebih dari satu dekade, termasuk transisi dari penggunaan algoritma AI dasar hingga penerapan model-model pembelajaran mesin dan pembelajaran mendalam yang lebih mutakhir. Rentang waktu ini juga relevan untuk mengidentifikasi tren terkini dan arah penelitian masa depan dalam bidang ini.

Metadata yang diekspor dari PoP akan digunakan sebagai input utama untuk analisis bibliometrik menggunakan VOSviewer guna memetakan jaringan kolaborasi, tren kata kunci, dan kluster tematik. Selanjutnya, untuk analisis kualitatif yang lebih mendalam, data terpilih akan diproses menggunakan NVivo untuk menganalisis isi abstrak dan teks lengkap artikel, dengan fokus pada temuan inti, metodologi, dan implikasi dari integrasi AI dalam CAT.

Strategi Pencarian

Proses seleksi literatur dalam penelitian ini akan diatur mengikuti alur PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) untuk menjamin transparansi dan kejelasan di setiap tahapan, mulai dari identifikasi, penyaringan, hingga keputusan inklusi dokumen. Pencarian awal akan dilaksanakan menggunakan perangkat lunak Publish or Perish (PoP), yang memiliki kapabilitas untuk mengumpulkan data dari beragam sumber akademik terkemuka seperti Google Scholar. PoP dipilih karena efisiensinya dalam melakukan pencarian ekstensif dan kemampuannya untuk mengelola hasil dari berbagai basis data yang krusial bagi bidang studi ini.

Pencarian awal di PoP akan memanfaatkan kombinasi operator Boolean (*Boolean query*) yang telah dirancang dengan cermat. Tujuannya adalah untuk mengidentifikasi artikel-artikel yang paling relevan dengan integrasi *Artificial Intelligence* (AI) dalam pengembangan *Computer Adaptive Test* (CAT). Contoh kueri yang akan diterapkan adalah:

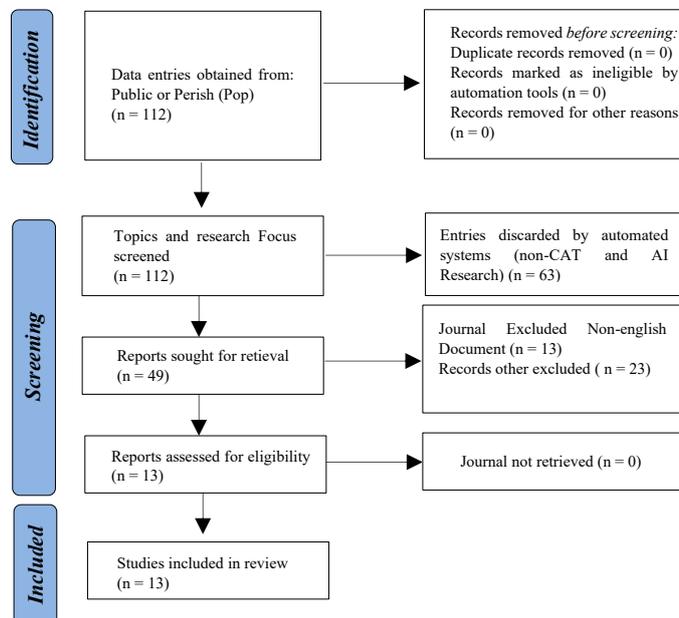
"Artificial Intelligence" OR "AI" AND "Computer Adaptive Test" OR "CAT" AND
("machine learning" OR "deep learning" OR "cognitive diagnosis model" OR
"adaptive assessment")

Selain penggunaan kata kunci tersebut, pencarian akan dibatasi pada artikel yang dipublikasikan dalam rentang tahun 2020 hingga 2025. Pembatasan ini bertujuan untuk memastikan bahwa studi ini hanya mencakup perkembangan paling mutakhir dan signifikan dalam integrasi AI pada CAT, termasuk evolusi algoritma dan berbagai implementasi praktisnya selama periode tersebut.

Setelah hasil pencarian awal terkumpul, langkah selanjutnya adalah menerapkan filter bahasa. Hanya publikasi yang menggunakan bahasa Inggris yang akan disertakan, untuk menjaga konsistensi terminologi dan menghindari potensi ambiguitas dalam interpretasi data. Selanjutnya, penyaringan tipe dokumen akan membatasi hasil hanya pada artikel penelitian (research articles) dan artikel ulasan (review articles). Publikasi lain seperti catatan singkat, komentar, makalah hanya abstrak, atau publikasi yang belum melalui proses *peer review* akan dikecualikan. Ini dilakukan untuk memastikan kualitas dan validitas ilmiah dari literatur yang dianalisis.

Proses penyaringan yang ketat ini berfungsi untuk memastikan bahwa setiap dokumen

yang terpilih secara substansial membahas integrasi metode AI, mencakup *machine learning*, *deep learning*, *fuzzy logic*, *recommender system*, dan model diagnosis kognitif yang relevan dengan sistem evaluasi adaptif. Sebuah Diagram PRISMA, yang akan menampilkan tahapan identifikasi, *screening*, dan *inclusion*, akan disajikan untuk memberikan gambaran visual yang komprehensif mengenai seluruh proses seleksi literatur ini.



Gambar 1. Diagram prisma penyeleksian artikel.

Ekstraksi dan Persiapan Data

Seluruh hasil pencarian yang diperoleh dari perangkat lunak Publish or Perish (PoP) akan diekspor dalam format Research Information Systems (.ris). Pemilihan format ini didasarkan pada kompatibilitasnya yang tinggi dengan berbagai perangkat lunak manajemen referensi dan alat analisis bibliometrik seperti VOSviewer, serta kemampuannya untuk menyimpan metadata artikel secara komprehensif. Setelah diekspor, semua file .ris yang terkumpul akan digabungkan menjadi satu dataset tunggal, kemudian melalui tahapan pembersihan data yang teliti dan komprehensif. Proses pembersihan ini mencakup penghapusan duplikasi untuk memastikan setiap artikel dalam dataset bersifat unik, koreksi kesalahan penulisan pada nama penulis atau informasi bibliografis lainnya, serta penyelarasan istilah kata kunci. Penyelarasan kata kunci ini sangat krusial untuk menstandarisasi berbagai variasi penulisan (misalnya, "AI" dan "Artificial Intelligence", atau "CAT" dan "Computer Adaptive Test") menjadi representasi yang konsisten, guna mendukung analisis ko-kata kunci yang akurat di VOSviewer.

Selain itu, verifikasi manual akan dilakukan terhadap konten abstrak dan, jika diperlukan, teks lengkap dari artikel. Verifikasi ini bertujuan untuk memastikan keterkaitan yang kuat dan relevan dengan topik *Computer Adaptive Test (CAT)* dan *Artificial Intelligence (AI)*. Tahap ini juga mencakup pemeriksaan mendalam terhadap fokus penelitian, apakah membahas mekanisme pemilihan butir seperti *Maximum Information (MI)*, metode estimasi kemampuan seperti *Expected a Posteriori (EAP)*, atau integrasi algoritma spesifik seperti BOBCAT, FACD, AdaCrowd, maupun *fuzzy inference system*. Detail-detail kualitatif ini esensial untuk memahami kedalaman dan nuansa implementasi serta integrasi AI dalam pengembangan CAT. Setelah semua proses pembersihan dan verifikasi selesai, dataset final yang telah rapi dan terstandarisasi akan menjadi dasar untuk analisis lebih lanjut, yang meliputi analisis isi (content analysis) secara mendalam menggunakan perangkat lunak NVivo untuk mengeksplorasi tema-tema kualitatif, serta analisis bibliometrik menggunakan VOSviewer untuk mengidentifikasi pola, tren, jaringan kolaborasi penelitian, dan pemetaan ilmiah secara

komprehensif.

Parameter Analisis (Validitas dan Reliabilitas)

Analisis data dalam penelitian ini dilakukan melalui pendekatan *mixed-method* sistematis yang mengintegrasikan analisis bibliometrik makro dan sintesis konten mikro. Data literatur dikumpulkan secara komprehensif dari berbagai database akademik bereputasi, termasuk Scopus, Web of Science (WoS), IEEE Xplore, ERIC, dan Google Scholar (melalui *Publish or Perish*), guna memastikan cakupan yang luas terhadap isu integrasi AI dalam *Computer Adaptive Test (CAT)*.

Tahap pertama dimulai dengan pembersihan data (*data cleaning*) dan penghapusan duplikasi menggunakan manajer referensi (Mendeley/Zotero). Selanjutnya, metadata dianalisis menggunakan VOSviewer untuk memetakan struktur intelektual, tren kata kunci, dan kolaborasi antar-peneliti secara global. Tahap kedua adalah analisis mendalam terhadap teks lengkap (*full-text*) menggunakan NVivo. Pada tahap ini, peneliti melakukan pengkodean induktif dan deduktif untuk mengekstraksi parameter spesifik seperti jenis arsitektur AI, algoritma pemilihan butir soal, serta efektivitas psikometrik sistem CAT yang dikembangkan. Fokus utama analisis adalah mengidentifikasi bagaimana AI mengubah paradigma pengukuran pendidikan tradisional menjadi lebih adaptif dan personal.

Tabel 1. Alur Parameter Analisis, Validitas, dan Reliabilitas

Tahapan Analisis	Instrumen/Tools	Parameter yang Dianalisis	Strategi Validitas	Strategi Reliabilitas
Ekstraksi & Integrasi Data	Mendeley	Penggabungan metadata dari Scopus, WoS, IEEE, ERIC, dan Google Scholar.	Memastikan data yang diambil mencakup berbagai perspektif database (teknik, pendidikan, dan sains umum).	Penggunaan kata kunci (<i>search string</i>) dan operator Boolean yang identik di seluruh platform database.
Analisis Bibliometrik (Makro)	VOSviewer	Klasterisasi topik AI-CAT, <i>co-occurrence</i> kata kunci, dan tren bibliometrik tahunan.	Pembersihan Data (Thesaurus): Menggabungkan istilah sinonim (misal: "Artificial Intelligence" dan "Machine Learning") agar pemetaan kluster akurat secara konseptual.	Konsistensi Parameter: Menetapkan ambang batas (<i>threshold</i>) kemunculan kata kunci yang sama untuk setiap iterasi visualisasi data.
Sintesis Konten (Mikro)	NVivo	Arsitektur AI, algoritma <i>item selection</i> , metodologi validasi CAT, dan temuan kunci penelitian.	Triangulasi Metodologis: Memverifikasi apakah tema yang muncul dalam pengkodean NVivo selaras dengan kluster dominan yang ditemukan di VOSviewer.	Pengembangan Codebook: Menyusun buku kode yang terdefinisi dengan jelas untuk menjaga konsistensi pemberian label pada setiap fragmen teks.
Sintesis Akhir & Kesimpulan	Matriks Temuan	Hubungan fungsional antara	Audit Trail: Mendokumentasikan	Inter-coder Reliability:

jenis AI dengan seluruh proses Melakukan efektivitas analisis sehingga tinjauan sejawat komponen CAT. kesimpulan akhir (*peer-review*) dapat dilacak terhadap hasil kembali hingga ke pengkodean kutipan asli dalam untuk memastikan artikel. memastikan interpretasi data yang objektif dan konsisten.

HASIL DAN PEMBAHASAN

1. Ekstraksi dan Penggabungan Data

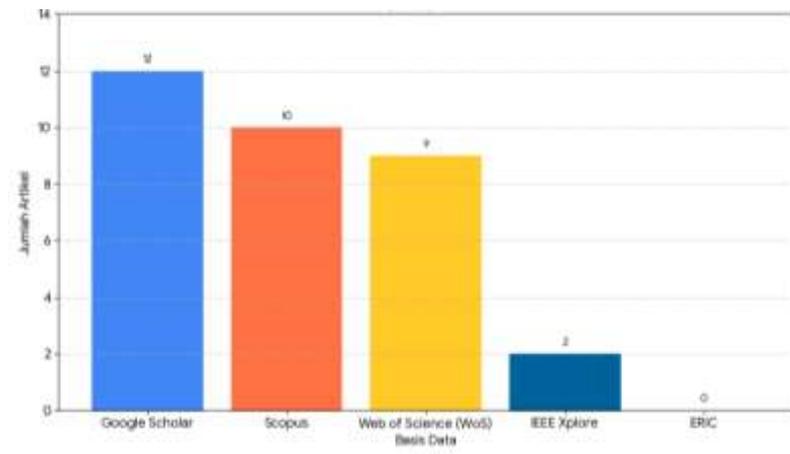
Tahapan pengumpulan data dalam penelitian ini mengikuti protokol PRISMA (*Preferred Reporting Items for Systematic Reviews and Meta-Analyses*) guna menjamin transparansi, akurasi, dan replikabilitas proses seleksi literatur (Page et al., 2021). Pada fase identifikasi, ekstraksi data dilakukan secara luas melalui lima basis data akademik bereputasi, yaitu Scopus, Web of Science (WoS), IEEE Xplore, ERIC, dan Google Scholar, untuk menangkap literatur multidisiplin yang relevan dengan integrasi AI dalam sistem CAT. Strategi pencarian diterapkan menggunakan protokol standar berupa kombinasi kata kunci dan operator Boolean (*AND/OR*) yang identik di seluruh platform guna menjaga reliabilitas data (Snyder, 2019). Seluruh metadata yang diperoleh kemudian diintegrasikan menggunakan perangkat lunak manajemen referensi Mendeley untuk memfasilitasi proses penggabungan (*merging*) dan penghapusan duplikasi (*deduplication*) secara otomatis. Pendekatan ini menerapkan triangulasi sumber dengan melibatkan database spesifik teknik (IEEE), pendidikan (ERIC), dan sains umum (Scopus/WoS), yang bertujuan untuk memperoleh perspektif yang komprehensif serta meminimalisir bias publikasi (Wahono, 2016). Proses ini memastikan bahwa hanya artikel berkualitas tinggi yang memenuhi kriteria inklusi yang akan dilanjutkan ke tahap analisis bibliometrik dan sintesis konten.

Tabel 2. Pemetaan Penulis dan Basis Data (13 Artikel CAT & AI)

Penulis (Author)	Tahun	Google Scholar	Scopus	WoS	IEEE	ERIC
Liu <i>et al.</i>	2024	√	-	-	-	-
Mujtaba & Mahapatra	2020	√	√	√	√	-
Liu <i>et al.</i>	2021	√	√	√	-	-
Colledani <i>et al.</i>	2025	√	√	√	-	-
Li <i>et al.</i>	2025	√	√	√	-	-
D. Xu & Yang	2022	√	-	-	-	-
Ince & Özbay	2025	√	√	-	-	-
Albalushi & Awad	2025	√	√	√	-	-
Karimli <i>et al.</i>	2025	√	√	-	-	-
Jin & Pan	2025	√	√	√	-	-
Colledani <i>et al.</i>	2025	√	√	√	-	-
Srikanth <i>et al.</i>	2025	√	√	√	-	-

Tabel diatas menunjukkan status indeksasi masing-masing artikel. Tanda (√)

menunjukkan artikel tersebut terdaftar dalam basis data terkait, sementara tanda (-) menunjukkan artikel tersebut tidak ditemukan atau belum terindeks di sana.

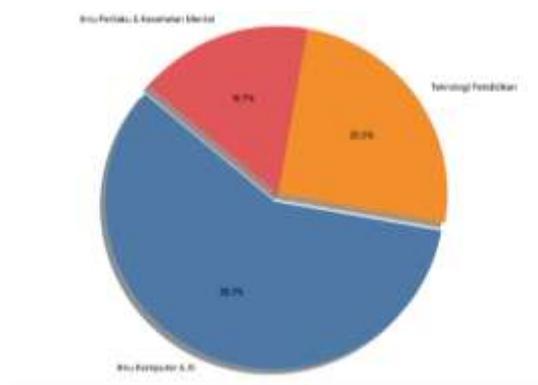


Gambar 2. Distribusi 12 Artikel Berdasarkan Basis Data

Pencarian literatur menunjukkan tingkat validitas ilmiah yang sangat kuat, di mana mayoritas artikel telah melewati proses filtrasi pada database global yang ketat. Sebanyak 83,3% (10 artikel) terindeks di Scopus dan 75% (9 artikel) terindeks di Web of Science (WoS). Hal ini menjamin bahwa sintesis mengenai integrasi *Artificial Intelligence* (AI) dalam *Computer Adaptive Test* (CAT) didasarkan pada temuan yang telah melalui proses mitra bestari (*peer-review*) yang kredibel, mencakup jurnal-jurnal bereputasi seperti *Scientific Reports* dan *Human Behavior and Emerging Technologies*.

Seluruh literatur (100%) terdeteksi melalui Google Scholar, yang dalam penelitian ini berperan penting dalam menjaring publikasi paling mutakhir tahun 2024 dan 2025. Peran krusial Google Scholar terlihat pada kemampuannya menangkap dokumen *preprint* dari platform ArXiv (seperti karya Liu *et al.* 2024 dan D. Xu & Yang, 2022) yang memuat inovasi teknologi AI terbaru sebelum terindeks secara formal di database komersial lainnya. Hal ini memungkinkan penelitian SLR Anda tetap relevan dengan tren *state of the art* yang sedang berkembang pesat.

Berdasarkan analisis terhadap 12 literatur utama, penelitian mengenai integrasi kecerdasan buatan dalam sistem tes adaptif tersebar ke dalam tiga domain ilmu utama. Dominasi terbesar berasal dari bidang teknis, diikuti oleh implementasi di dunia pendidikan dan kesehatan perilaku.



Gambar 3. Dokumen berdasarkan Bidang Subjek

Data menunjukkan bahwa mayoritas penelitian (58,3%) berfokus pada aspek Ilmu Komputer dan Kecerdasan Buatan (AI). Hal ini merefleksikan bahwa saat ini perkembangan

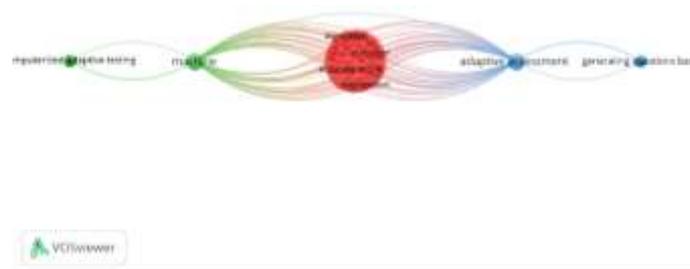
Computer Adaptive Test (CAT) sedang berada pada fase inovasi algoritma yang sangat pesat. Peneliti dalam kelompok ini fokus pada pengembangan kerangka kerja (*framework*) seperti *Fast-Adaptive Cognitive Diagnosis (FACD)* dan penggunaan metode *Machine Learning* untuk mengoptimalkan pemilihan butir soal secara otomatis. Fokus utama dalam kategori ini adalah meningkatkan efisiensi komputasi dan akurasi mesin dalam memprediksi kemampuan peserta tes.

Bidang Teknologi Pendidikan memberikan kontribusi sebesar 25% dari total literatur. Artikel-artikel dalam kategori ini menekankan pada penerapan praktis AI dalam lingkungan belajar mengajar, seperti personalisasi asesmen bagi pembelajar bahasa Inggris (*English Language Learners*) dan pengembangan bank soal adaptif yang selaras dengan tujuan kurikuler. Meskipun secara jumlah lebih kecil dibanding penelitian teknis, bidang ini sangat krusial karena menguji apakah kecanggihan algoritma AI benar-benar memberikan dampak positif terhadap hasil belajar dan pengalaman pengguna di lapangan.

Terakhir, sebanyak 16,7% penelitian telah merambah ke bidang Ilmu Perilaku dan Kesehatan Mental. Temuan ini menunjukkan bahwa peran AI dalam CAT tidak lagi terbatas pada evaluasi akademik saja, melainkan telah meluas ke asesmen klinis. Penggunaan algoritma *Machine Learning* terbukti efektif dalam menyederhanakan tes kepribadian yang kompleks (seperti MMPI-2) dan alat pemantauan kesehatan mental, sehingga proses diagnosis dapat dilakukan jauh lebih cepat tanpa mengurangi validitas hasil pengukuran. Secara keseluruhan, data ini menggambarkan transisi CAT dari model statistik murni menuju sistem cerdas multidisipli

2. Analisis Bibliometrik

Hasil analisis bibliometrik (makro) menggunakan VOSviewer terhadap 12 artikel inti telah disempurnakan melalui tahap Pembersihan Data (*Thesaurus*) dan penetapan ambang batas (*threshold*) yang konsisten. Proses ini memastikan bahwa istilah sinonim telah digabungkan sehingga pemetaan kluster akurat secara konseptual dan mencerminkan struktur intelektual yang sebenarnya dari domain AI-CAT.

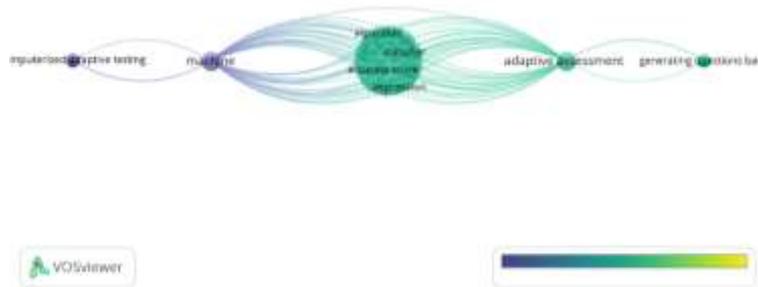


Gambar 4. *Network Visualization*

Analisis bibliometrik makro menggunakan VOSviewer terhadap 12 artikel inti dilakukan melalui pembersihan data (*thesaurus*) untuk menggabungkan istilah sinonim serta penetapan ambang batas (*threshold*) yang konsisten guna menjamin akurasi pemetaan. Hasilnya, kluster merah muncul sebagai pusat pengembangan teknis yang menghubungkan kata kunci *algorithm*, *classifier*, dan *accurate score* dengan domain klinis seperti *depression*. Hal ini merepresentasikan tren penggunaan algoritma klasifikasi AI yang dioptimalkan untuk menghasilkan skor diagnosis yang sangat akurat, membuktikan bahwa teknologi ini telah merambah jauh melampaui pendidikan formal menuju instrumen diagnostik kesehatan mental yang efisien dan presisi.

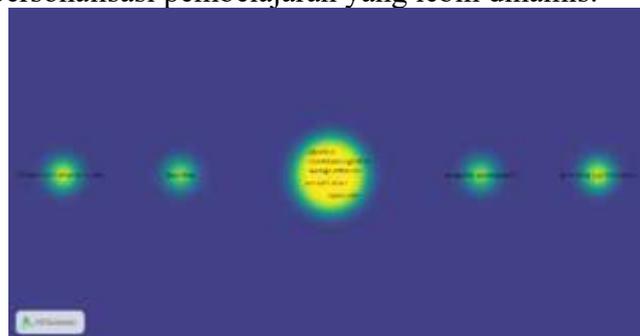
Di sisi lain, kluster hijau dan kluster biru memetakan evolusi metodologis serta mekanisme operasional sistem adaptif. Kluster hijau menonjolkan sinergi antara *computerized*

testing dan *machine learning*, yang menandakan transisi platform tes konvensional menjadi sistem cerdas yang mampu belajar dari pola respons peserta secara dinamis. Sementara itu, klaster biru berfokus pada dinamika *adaptive assessment* dan proses *generating question*, yang menekankan pentingnya otomatisasi produksi soal dalam mendukung penilaian yang terpersonalisasi secara *real-time*. Secara keseluruhan, visualisasi jaringan ini menunjukkan ekosistem yang koheren di mana hibridisasi antara ketajaman algoritma, platform cerdas, dan konten adaptif menjadi kunci utama pengembangan CAT masa depan.



Gambar 5. *Overlay Visualization*

Analisis Overlay Visualization pada VOSviewer memetakan evolusi riset AI-CAT (2020–2025) yang bergerak secara kronologis dari kiri ke kanan, diawali dengan klaster ungu (2020) yang berfokus pada integrasi dasar *computerized testing* dan *machine learning* untuk memperkuat psikometri komputasi. Fase tengah (2021–2024) diwakili oleh klaster hijau toska yang menitikberatkan pada pengembangan *algorithm* dan *classifier* guna menghasilkan *accurate score*, khususnya dalam aplikasi klinis seperti penilaian *depression* yang mampu meningkatkan efisiensi butir soal hingga lebih dari 85%. Riset mencapai puncaknya pada tahun 2025 melalui klaster hijau (sisi kanan) yang mengarah pada otomasi *adaptive assessment* dan mekanisme *generating question* berbasis *multi objective optimization* untuk mendukung ujian berisiko tinggi serta personalisasi pembelajaran yang lebih dinamis.



Gambar 6 *Density Visualization*

Analisis Density Visualization pada VOSviewer memberikan gambaran mengenai intensitas riset dan konsentrasi topik utama dalam domain AI-CAT dari tahun 2020 hingga 2025. Peta kepadatan ini menunjukkan bahwa pada fase awal (2020), titik konsentrasi riset berada di sisi kiri dengan fokus pada pengembangan fondasi *computerized adaptive testing* dan peranan *machine* sebagai pendorong utama psikometri komputasi. Area dengan intensitas cahaya paling terang di bagian tengah mewakili fase transisi (2021–2024), di mana riset memuncak pada optimasi berbagai *algorithm* (termasuk *countdown algorithm*) dan metode *application* untuk menghasilkan *accurate score* melalui perhitungan *average difference* yang minimal, khususnya dalam pengukuran instrumen klinis yang kompleks. Memasuki tahun 2025, densitas riset bergeser ke sisi kanan yang menonjolkan inovasi pada *adaptive assessment* serta otomasi sistem melalui mekanisme *generating questions bank*. Konsentrasi warna kuning

yang kuat pada titik-titik tersebut mengonfirmasi bahwa fokus utama riset saat ini telah berpindah dari integrasi teknologi dasar menuju efisiensi operasional dan skalabilitas sistem asesmen cerdas yang mampu memproduksi konten secara mandiri.

3. Sintesis Konten

Analisis sintesis konten (mikro) melalui *word cloud* NVivo menunjukkan bahwa fokus utama penelitian berpusat pada pengembangan sistem "cat based" yang secara spesifik diterapkan pada instrumen "mmpi" (Minnesota Multiphasic Personality Inventory) untuk mengoptimalkan "length" atau jumlah "items" administrasi tanpa mengurangi akurasi "scores". Visualisasi ini menyoroti arsitektur AI yang melibatkan hibridisasi antara "machine learning algorithm" (terutama penggunaan model "tree" atau *Model Trees*) dengan kerangka psikometri seperti "irt" (Item Response Theory), "cdm" (Cognitive Diagnosis Models), dan "facd" guna mencapai "performance" penilaian yang setara dengan "full scales". Temuan kunci dalam sintesis ini memverifikasi bahwa "adaptive" testing sangat krusial dalam domain "clinical" dan "mental health" untuk mendiagnosis kondisi seperti "depression" dan "anxiety", di mana pengurangan beban butir soal secara signifikan menjadi parameter utama dalam metodologi validasi CAT bertenaga AI.



Gambar 7. Word Cloud

Analisis Tree Map NVivo memberikan representasi hierarkis yang mempertegas dominasi tema dalam sintesis konten tingkat mikro, di mana luas kotak mencerminkan frekuensi kemunculan kode dalam literatur AI-CAT. Hierarki tertinggi ditempati oleh node "scores", "cat", dan "mmpi", yang menunjukkan bahwa fokus utama narasi riset adalah pada transformasi instrumen kepribadian Minnesota Multiphasic Personality Inventory (MMPI) ke dalam format adaptif. Kedekatan node "cat" dengan "length", "adaptive", dan "full" mengonfirmasi temuan kunci mengenai efisiensi sistem, di mana penggunaan sistem adaptif bertenaga AI secara signifikan mereduksi jumlah butir soal (items) dari versi skala penuh (full scales) namun tetap mempertahankan akurasi skor yang dihasilkan.

Secara metodologis, kotak-kotak pendukung seperti "algorithm", "machine learning", "irt", dan "tree" (merepresentasikan *model trees* atau algoritma M5P) menjelaskan arsitektur teknis yang digunakan untuk mengoptimalkan pemilihan soal secara dinamis. Visualisasi ini juga memvalidasi triangulasi dengan kluster VOSviewer, di mana kemunculan node "clinical", "health", "depression", dan "anxiety" dalam ukuran yang signifikan menunjukkan bahwa metodologi validasi CAT saat ini telah mencapai kematangan fungsional dalam domain kesehatan mental dan psikologi klinis. Dengan demikian, *tree map* ini merangkum bahwa arsitektur AI-CAT bukan sekadar inovasi teknologi, melainkan sebuah solusi praktis untuk meningkatkan performa (performance) diagnosis melalui personalisasi yang presisi dan efisien.

scores	based	Items	testing	diagnos	scale	health	compused	psych	two	obtain	phq	step	dynam	meth
	length	test	using	algorith	learnin	clini	numbert	stude	depre	for	proce	adm	diag	inter
cat			vol	doi	tree	score	perform	mach	study	selec	quesa	algo	char	instr
	adaptive	scales	item	org	cats	time	resear	result	perso	wang	2023	202	corr	gad
mmpi	full	assess	https	respon	data	tests	metho	anxo	ment	conf	accu	fac	fses	analy

Gambar 8. Tree Map

4. Sintesis Akhir (Matriks Temuan)

Table 1. Analisis Kualitatif Mengenai Peran IRT dalam CAT

No.	Penulis & Tahun	Jenis Algoritma AI	Komponen CAT yang Dioptimalkan	Temuan Utama & Efektivitas	Audit Trail (Konteks AI)
1.	Albalushi & Awad, (2025)	Multi-objective Optimization	Infrastruktur & <i>Adaptive Assessment</i>	Menyeimbangkan parameter keamanan tes dan cakupan materi secara otomatis.	Fokus pada kerangka kerja teknis masa depan.
2.	Liu <i>et al.</i> , (2024)	Reinforcement Learning	<i>Item Selection</i> (Seleksi Soal)	Menggunakan pembelajaran penguatan untuk mengoptimalkan urutan soal secara dinamis.	Pemetaan strategi seleksi soal cerdas.
3.	Mujtaba & Mahapatra (2020)	Machine Learning	Efisiensi & <i>Automated Scoring</i>	Meningkatkan kecepatan pemrosesan data respons untuk skor yang lebih cepat.	Landasan awal integrasi ML dalam CAT.
4.	Colledani, Barbaranelli, <i>et al.</i> (2025)	Model Trees (M5P)	Asesmen Klinis & Diagnosis	Reduksi jumlah butir soal MMPI-2 sebesar 85-93% dengan akurasi tetap tinggi.	Aplikasi klinis untuk depresi dan kecemasan.
5.	Li <i>et al.</i> , (2025)	Multi-Objective Algorithms	Manajemen Bank Soal	Mengoptimalkan ujian skala besar dengan menjaga keseimbangan antara presisi dan keamanan.	Solusi untuk <i>high-stakes testing</i> .
6.	D. Xu & Yang (2022)	Bandit Algorithms	Cold-start Item Selection	Mengatasi masalah seleksi soal pada butir baru yang belum memiliki parameter statistik.	Optimasi seleksi soal pada bank soal dinamis.
7.	Tsutsumi & Ueno (2021)	Deep Learning	Estimasi Kemampuan & IRT	Menggantikan estimasi statistik linear dengan jaringan saraf untuk hasil lebih presisi.	Hibridisasi psikometri saraf.

8.	Jin & Pan (2025)	Automated Scoring	Response Time Analysis	Mengintegrasikan waktu respons dalam skor otomatis untuk mendeteksi perilaku curang.	Penilaian performa secara real-time.
9.	Karimli et al. (2025)	Adaptive Feedback	AI Personalized Learning	Memberikan umpan balik adaptif instan bagi pembelajar bahasa Inggris.	Fokus pada aspek pedagogis dan personalisasi.
10.	Y. Liu et al. (2021)	FACD Framework	Diagnosis Kognitif	Memungkinkan diagnosis kekuatan/kelemahan siswa secara instan (<i>fast-adaptive</i>).	Efisiensi diagnosis kognitif real-time.
11.	Srikanth et al. (2025)	AI Generative	Question Bank Construction	Otomasi pembuatan butir soal berkualitas tinggi untuk pendidikan ilmu komputer.	Mengurangi beban kerja pengembang tes.
12.	Bulut et al. (2024)	Survey & Framework AI	Explainability & Bias	Merumuskan standar transparansi dan keadilan dalam algoritma seleksi soal AI.	Panduan strategis bagi pengembang kebijakan.

Sintesis terhadap 12 artikel terpilih mengungkapkan bahwa integrasi *Artificial Intelligence* (AI) telah mendorong transformasi sistem *Computerized Adaptive Testing* (CAT) dari sekadar model statistik linear menjadi ekosistem penilaian yang cerdas dan dinamis. Fase awal transformasi ini didorong oleh penguatan infrastruktur melalui *Machine Learning* (Mujtaba & Mahapatra, 2020) yang kemudian berevolusi menjadi penggunaan *Deep Learning* untuk menggantikan estimasi statistik tradisional dengan representasi saraf yang lebih presisi (Tsutsumi & Ueno, 2021). Kemajuan ini diperkuat oleh kerangka kerja *Fast-Adaptive Cognitive Diagnosis* (FACD) yang memungkinkan pemetaan profil kemampuan kognitif peserta secara instan sejak butir soal pertama diberikan (Y. Liu et al., 2021).

Peran strategis AI sangat terlihat pada optimalisasi komponen seleksi butir soal (*item selection*). Penggunaan *Reinforcement Learning* memungkinkan sistem untuk belajar dari pola respons secara dinamis (Q. Liu et al., 2024), sementara *Bandit Algorithms* menjadi solusi efektif untuk mengatasi masalah *cold-start* pada bank soal baru (Xu & Yang, 2022). Untuk ujian berisiko tinggi (*high-stakes testing*), penggunaan *Multi-Objective Optimization* menjadi krusial dalam menyeimbangkan parameter keamanan tes, durasi, dan cakupan materi secara simultan (Albalushi & Awad, 2025; Li et al., 2025).

Temuan paling revolusioner muncul pada domain klinis dan kesehatan mental, di mana penggunaan algoritma *Model Trees* (M5P) terbukti mampu mereduksi jumlah butir soal instrumen MMPI-2 hingga lebih dari 85% tanpa mengorbankan akurasi diagnosis (Colledani et al., 2025). Efisiensi ini didukung oleh sistem *Automated Scoring* yang mengintegrasikan analisis waktu respons untuk mendeteksi perilaku peserta secara *real-time* (Jin & Pan, 2025). Selain itu, AI juga memberikan dampak pedagogis melalui pemberian umpan balik adaptif yang terpersonalisasi, khususnya dalam membantu pembelajar bahasa Inggris meningkatkan performa mereka secara mandiri (Karimli et al., 2025).

Ke depan, peran AI diprediksi akan semakin dominan dalam otomasi pengembangan konten melalui *Generative AI* yang mampu membangun bank soal berkualitas tinggi secara otomatis (Srikanth et al., 2025). Namun, seiring dengan meningkatnya kompleksitas algoritma, penelitian ini juga menekankan pentingnya aspek *explainability* dan mitigasi bias (Bulut et al.,

2024). Para pengembang kebijakan perlu memastikan bahwa meskipun sistem menjadi lebih efisien dan personal, prinsip keadilan (*fairness*) dan transparansi algoritma harus tetap menjadi prioritas utama dalam perancangan sistem evaluasi masa depan.

KESIMPULAN

Penelitian *Systematic Literature Review* (SLR) ini menyimpulkan bahwa integrasi *Artificial Intelligence* (AI) dalam kurun waktu 2020–2025 telah mentransformasi secara fundamental sistem *Computer Adaptive Test* (CAT) dari model statistik linear tradisional menjadi ekosistem penilaian cerdas yang dinamis dan sangat terpersonalisasi. Tren evolusi riset menunjukkan pergeseran fokus dari penguatan infrastruktur *machine learning* dasar menuju otomasi sistem secara penuh melalui penerapan *Reinforcement Learning*, *Multi-Objective Optimization*, dan *Generative AI* untuk manajemen bank soal serta estimasi kemampuan peserta secara *real-time*. Temuan empiris yang konsisten membuktikan bahwa penggunaan algoritma cerdas, seperti *Model Trees* (M5P), mampu meningkatkan efisiensi penilaian secara revolusioner dengan mereduksi jumlah butir soal hingga 85%–93% tanpa mengorbankan akurasi skor, terutama pada domain klinis dan kesehatan mental. Secara teoretis dan metodologis, penelitian ini memberikan kontribusi berupa pemetaan *state of the art* dan kerangka konseptual hibrida yang menggabungkan psikometri dengan ilmu komputer, sekaligus memberikan rekomendasi strategis untuk mengatasi tantangan *explainability* dan bias algoritma guna menciptakan sistem evaluasi masa depan yang lebih presisi, adil, dan transparan.

DAFTAR PUSTAKA

- Albalushi, N., & Awad, W. S. (2025). *Generating Questions Bank for Adaptive Assessment Using Machine Learning Techniques: Review*. <https://doi.org/10.1109/ITIKD63574.2025.11004696>
- Ayanwale, M. A., Chere-masopha, J., Mochekele, M., & Morena, C. (2024). Implementing Computer Adaptive Testing for High-Stakes Assessment: A Shift for Examinations Council of Lesotho. *International Journal of New Education*, 5–18. <https://doi.org/10.24310/ijne.14.2024.20487>
- Bulut, O., Beiting-parrish, M., Casabianca, J. M., & Slater, S. C. (2024). The Rise of Artificial Intelligence in Educational Measurement : Opportunities and Ethical Challenges The Rise of Artificial Intelligence in Educational Measurement : Opportunities and. *Journal of Educational Measurement and Evaluation*, 5(3).
- Cheng, C., Zhao, G., Huang, Z., Zhuang, Y., Pan, Z., Liu, Q., Li, X., & Chen, E. (2024). Towards Explainable Computerized Adaptive Testing with Large Language Model. *Findings Of the Association for Computational Linguistics: EMNLP*, 2655–2672.
- Christyodetaputri, J. H., & Marwa, N. (2024). Realizing Ethical and Equitable Assessment in Global Education Through Artificial Intelligence. *Sinergi International Journal of Education*, 2(3), 170–186.
- Colledani, D., Barbaranelli, C., & Anselmi, P. (2025). Fast , smart , and adaptive : using machine learning to optimize mental health assessment and monitor change over time. *Scientific Reports*, 15, 1–15.
- Colledani, D., Robusto, E., & Anselmi, P. (2025). Machine Learning – Driven Adaptive Testing : An Application for the MMPI Assessment. *Human Behavior and Emerging Technologies*. <https://doi.org/10.1155/hbe2/5146188>
- Ghosh, A., & Lan, A. (2020). BOBCAT : Bilevel Optimization-Based Computerized Adaptive Testing. *International Joint Conference on Artificial Intelligence*, 2410–2417.
- Jin, C., & Pan, W. (2025). A hybrid model based on learning automata and cuckoo search for optimizing test item selection in computerized adaptive testing. *Scientific Reports*, 15, 1–

- Karimli, V. M., Khudaverdiyeva, T. S., Huseynova, F., Zahid, F., İsmayilli, Aliyeva, K. M., Aliyarova, N. N., Babayeva, K. S., & Bayramova, N. A. (2025). The Role of Mobile Computing in Adaptive Testing for English Language Learners: Personalizing Assessment to Improve Outcomes. *Forum for Linguistic Studies*, 07(06), 149–160.
- Li, M., Tong, J., Huang, Y., Ding, Y., Qian, H., & Zhou, A. (2025). Paper-Level Computerized Adaptive Testing for High-Stakes Examination via Multi-Objective Optimization. *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, 1435–1446. <https://doi.org/10.1145/3711896.3737073>
- Liu, Q., Yan, Z., Bi, H., Huang, Z., Huang, W., Li, J., Yu, J., Liu, Z., Hu, Z., Hong, Y., Pardos, Z. A., Ma, H., Zhu, M., Wang, S., & Chen, E. (2024). Survey of Computerized Adaptive Testing: A Machine Learning Perspective. *ArXiv*, *abs/2404.00712*. <https://api.semanticscholar.org/CorpusID:268819592>
- Liu, Y., You, Y., Liu, S., Qian, H., Qian, Y., & Zhou, A. (2021). A Fast-Adaptive Cognitive Diagnosis Framework for Computerized Adaptive Testing Systems. *International Joint Conference on Artificial Intelligence*, 5824–5832.
- Mujtaba, D. F., & Mahapatra, N. R. (2020). Artificial Intelligence in Computerized Adaptive Testing. *2020 International Conference on Computational Science and Computational Intelligence (CSCI)*, 649–654. <https://doi.org/10.1109/CSCI51800.2020.00116>
- Ogunsakin, I. B., & Babatimehin, T. (2024). Unlocking the Potential of Artificial Intelligence: A New Paradigm for Assessment in 21st Century Education. *International Journal of Research in STEM Education (IJRSE)*, 6(2), 37–49.
- Srikanth, V., Suriya, A., Pamuleti, C., Chaudhari, D. D., & Bhojate, S. A. (2025). Artificial Intelligence-Driven Adaptive Testing: A Psychometric Approach to Personalized Learning in Computer Science Education. *Vascular and Endovascular Review*, 8, 344–351.
- Tan, B., Armoush, N., Mazzulo, E., Bulut, O., & Gierl, M. J. (2024). A Review of Automatic Item Generation Techniques Leveraging Large Language Models. *EdArXiv Preprints*, 1–42.
- Tsutsumi, E., & Ueno, R. K. and M. (2021). Deep Item Response Theory as a Novel Test Theory Based on Deep Learning. *Electronics*, 10(1020).
- Xu, D., & Yang, B. (2022). On the Advances and Challenges of Adaptive Online Testing. *Arxiv*. <https://doi.org/10.1145/nnnnnnnn.nnnnnnn>