Education, Social Sciences, and Linguistics: Conference Series
https://journal.ummat.ac.id/index.php/ecs
Vol. 2 No. 1, February 2026, pp. 211-219

Artificial Intelligence dalam Assessment Pendidikan: A Systematic Literature Review tentang Model, Validitas, dan Implikasi Evaluatif

Yuni Susilowati[1], Johri Sabaryati[2], Andi Jusmiana[3], Rizki Isfahani[4], Eva Marsepa[5]
[1,4,5]Ilmu Keperawatan, Universitas Yatsi Madani, Tangerang, Indonesia
[2]Pendidikan Fisika, Universitas Muhammadiyah Mataram, Mataram, Indonesia
[3]Pendidikan Matematika, Universitas Pejuang Republik Indonesia, Makassar, Indonesia
[5]Ilmu Keperawatan, Universitas of Yatsi madani, Tangerang, Indonesia
[1]yunisusilowati@uym.ac.i , [2]joyafarashy@gmail.com , [3]andijusmiana@gmail.com , [4]rizkiisfahani@uym.ac.id , [5]evamarsepa@uym.ac.id

Keywords:
Artificial Intelligence, Educational Assessment, Validity, Reliability, Ethical Implications, Systematic Literature Review.

ABSTRACT

Abstract: *This study aims to identify and classify Artificial Intelligence (AI) models applied in educational assessment, analyze how validity and reliability aspects are examined in previous studies, and evaluate the evaluative and ethical implications of AI application. The research method used a qualitative approach with a Systematic Literature Review (SLR) design, which examined literature from the Scopus, DOAJ, and Google Scholar databases, covering publications from the last 10 years (2016–2025). The selection process was carried out systematically based on inclusion and exclusion criteria, followed by data extraction and analysis using thematic analysis techniques. The results of the study show an increasing trend in the use of AI in educational assessment, particularly in automated scoring, predictive analytics, and adaptive testing, with a focus on improving efficiency, personalizing evaluation, and adaptive feedback. The study also found that construct validity and score reliability remain major challenges, while evaluative and ethical implications, including algorithm transparency and student data protection, are important concerns. This literature synthesis provides a comprehensive understanding of the development of AI models, methodological challenges, and recommendations for practices and policies to optimize the use of AI in accountable, valid, and fair educational evaluation.*

Kata Kunci:

Artificial Intelligence,
Assessment Pendidikan,
Validitas, Reliabilitas,
Implikasi Etis,
Systematic Literature Review.

Abstrak: Penelitian ini bertujuan untuk mengidentifikasi dan mengklasifikasikan model-model Artificial Intelligence (AI) yang diterapkan dalam assessment pendidikan, menganalisis bagaimana aspek validitas dan reliabilitas dikaji dalam penelitian sebelumnya, serta mengevaluasi implikasi evaluatif dan etis dari penerapan AI. Metode penelitian menggunakan pendekatan kualitatif dengan desain Systematic Literature Review (SLR), yang menelaah literatur dari basis data Scopus, DOAJ, dan Google Scholar, dengan rentang publikasi 10 tahun terakhir (2016–2025). Proses seleksi dilakukan secara sistematis berdasarkan kriteria inklusi dan eksklusi, diikuti dengan ekstraksi dan analisis data menggunakan teknik analisis tematik. Hasil kajian menunjukkan adanya tren peningkatan penggunaan AI dalam assessment pendidikan, khususnya pada automated scoring, predictive analytics, dan adaptive testing, dengan fokus pada peningkatan efisiensi, personalisasi evaluasi, dan umpan balik adaptif. Studi juga menemukan bahwa validitas konstruk dan reliabilitas skor tetap menjadi tantangan utama, sementara implikasi evaluatif dan etis, termasuk transparansi algoritma dan perlindungan data peserta didik, menjadi perhatian penting. Sintesis literatur ini memberikan pemahaman komprehensif mengenai perkembangan model AI, tantangan metodologis, serta rekomendasi praktik dan kebijakan untuk mengoptimalkan penggunaan AI dalam evaluasi pendidikan yang akuntabel, sahih, dan adil.

Article History:
Received : 1-01-2026
Accepted : 30-02-2026

A. LATAR BELAKANG

Perkembangan teknologi digital telah mendorong transformasi signifikan dalam praktik assessment pendidikan. Artificial Intelligence (AI) semakin banyak digunakan untuk mendukung proses penilaian, mulai dari skoring otomatis hingga analisis performa belajar berbasis data besar. Dalam konteks ini, AI tidak hanya dipahami sebagai alat bantu teknis, tetapi sebagai sistem yang berpotensi mengubah paradigma evaluasi pembelajaran menjadi lebih adaptif, efisien, dan responsif terhadap kebutuhan peserta didik. Sejumlah kajian menunjukkan bahwa integrasi AI dalam assessment mampu meningkatkan efisiensi proses penilaian serta menyediakan umpan balik yang lebih cepat dibandingkan metode konvensional (Holmes et al., 2019; Zawacki-Richter et al., 2019). Oleh karena itu, pemanfaatan AI dalam assessment pendidikan menjadi isu strategis yang memerlukan kajian ilmiah yang mendalam dan sistematis.

Dalam praktiknya, AI dalam assessment diwujudkan melalui berbagai model, seperti automated essay scoring, intelligent tutoring systems, adaptive testing, serta learning analytics berbasis machine learning. Model-model tersebut dirancang untuk meningkatkan akurasi pengukuran dan personalisasi pembelajaran. Penelitian terdahulu menunjukkan bahwa sistem automated scoring berbasis AI memiliki tingkat konsistensi yang mendekati penilai manusia dalam konteks tertentu (Alsafy et al., 2025). Selain itu, adaptive assessment berbasis algoritma dinilai mampu menyesuaikan tingkat kesulitan soal secara dinamis berdasarkan respons peserta didik (Wang & Chen, 2020). Namun demikian, integrasi teknologi ini juga menimbulkan tantangan metodologis dan konseptual dalam kerangka evaluasi pendidikan (Widiada, 2025).

Meskipun berbagai model AI telah dikembangkan, isu validitas tetap menjadi perhatian utama dalam assessment pendidikan. Dalam perspektif teori validitas kontemporer, validitas tidak hanya berkaitan dengan ketepatan pengukuran, tetapi juga dengan interpretasi serta penggunaan skor dalam pengambilan keputusan pendidikan. Sejumlah kajian pada jurnal nasional terakreditasi menunjukkan bahwa implementasi sistem penilaian berbasis kecerdasan buatan dan *machine learning* di Indonesia masih menghadapi tantangan dalam memastikan kesesuaian antara konstruk yang diukur dengan indikator algoritmik yang digunakan (Huda & Kusumawati, 2022; Misbah et al., 2021). Secara empiris, penelitian pada pengembangan sistem penilaian otomatis di pendidikan tinggi melaporkan tingkat korelasi skor AI dengan penilai manusia berada pada rentang 0,68–0,82, namun masih ditemukan inkonsistensi pada respons esai yang bersifat argumentatif kompleks (Pratama & Widodo, 2020). Selain itu, studi lain mengungkapkan adanya perbedaan akurasi lebih dari 8–12% ketika sistem diuji pada kelompok mahasiswa dengan latar belakang literasi digital yang berbeda, yang mengindikasikan potensi bias algoritmik dan keterbatasan generalisasi model (Rahmawati et al., 2023). Dari sisi reliabilitas, stabilitas model sangat dipengaruhi oleh kualitas dan keragaman data pelatihan; perubahan dataset tanpa proses *retraining* dan kalibrasi ulang dilaporkan menurunkan konsistensi skor secara signifikan (Sari & Nugroho, 2021). Oleh karena itu, integrasi AI dalam assessment perlu tetap berpijak pada prinsip-prinsip psikometrik yang ketat agar interpretasi skor yang dihasilkan tetap sahih, konsisten, dan adil secara evaluatif dalam konteks pendidikan Indonesia.

Di sisi lain, implikasi evaluatif dan etis dari penggunaan AI dalam assessment semakin menjadi perhatian global. Transparansi algoritma, akuntabilitas pengambilan keputusan, serta perlindungan data peserta didik merupakan isu krusial yang tidak dapat diabaikan. Studi terbaru menunjukkan bahwa tanpa regulasi dan kerangka evaluatif yang memadai, penggunaan AI berpotensi memperkuat ketimpangan pendidikan (Selwyn, 2019; Williamson, 2017). Dengan demikian, diperlukan pendekatan evaluasi yang komprehensif untuk memastikan bahwa penerapan AI dalam assessment tidak hanya efisien secara teknis, tetapi juga adil, transparan, dan berorientasi pada peningkatan kualitas pembelajaran.

Berdasarkan uraian tersebut, penelitian ini bertujuan untuk melakukan Systematic Literature Review guna mengidentifikasi dan mengklasifikasikan model-model Artificial Intelligence yang digunakan dalam assessment pendidikan, menganalisis bagaimana aspek validitas dan reliabilitas dikaji dalam penelitian-penelitian sebelumnya, serta mengevaluasi implikasi evaluatif dan etis dari penerapannya. Kajian ini diharapkan dapat memberikan sintesis konseptual yang komprehensif bagi pengembangan teori dan praktik evaluasi pendidikan berbasis AI, sekaligus menjadi landasan akademik bagi pengambilan kebijakan dan pengembangan sistem assessment yang lebih akuntabel dan berkualitas.

B. METODE PENELITIAN

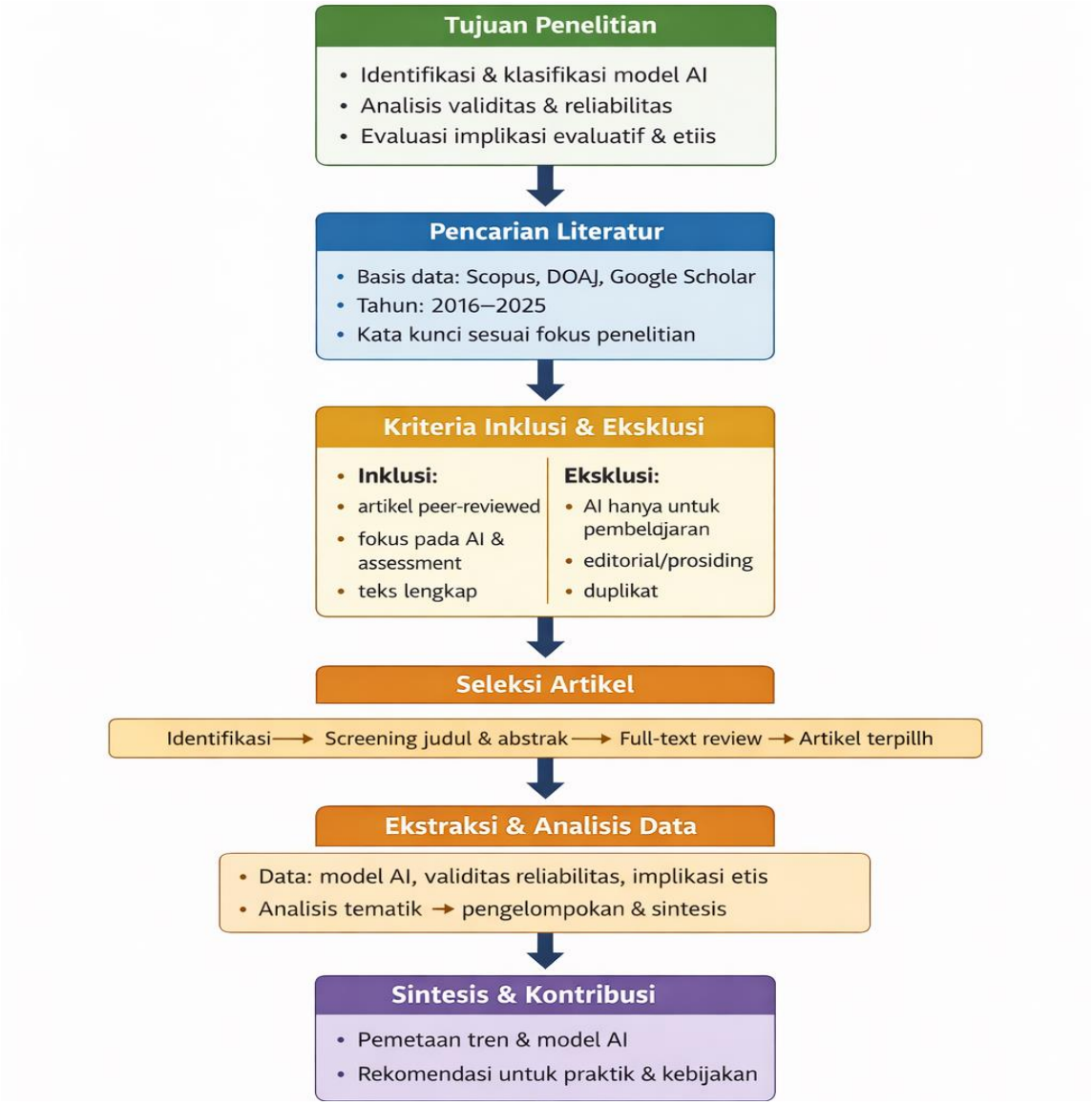
Penelitian ini menggunakan pendekatan kualitatif dengan desain *Systematic Literature Review* (*SLR*) yang bertujuan untuk mengidentifikasi dan mengklasifikasikan model-model *Artificial Intelligence* (*AI*) yang digunakan dalam *assessment* pendidikan, menganalisis bagaimana aspek validitas dan reliabilitas dikaji dalam penelitian-penelitian sebelumnya, serta mengevaluasi implikasi evaluatif dan etis dari penerapannya. Pendekatan *SLR* dipilih karena memungkinkan proses sintesis literatur dilakukan secara sistematis, transparan, dan terstruktur sehingga mampu menghasilkan pemetaan konseptual yang komprehensif mengenai perkembangan *AI* dalam evaluasi pendidikan. Dengan pendekatan ini, kajian tidak hanya mendeskripsikan tren penelitian, tetapi juga menelaah kualitas argumentasi metodologis serta konsistensi kerangka validasi yang digunakan dalam studi-studi terdahulu.

Strategi pencarian literatur dilakukan melalui basis data ilmiah bereputasi, yaitu Scopus, DOAJ, dan Google Scholar, dengan mempertimbangkan publikasi dalam kurun waktu sepuluh tahun terakhir (2016–2025) untuk menjamin relevansi dan kebaruan kajian. Proses pencarian menggunakan kombinasi kata kunci dengan operator Boolean, seperti: "*Artificial Intelligence*" OR "*AI*" AND "*Educational Assessment*" OR "*Automated Scoring*" OR "*Adaptive Testing*" AND "*Validity*" OR "*Reliability*" OR "*Algorithmic Fairness*" OR "*Ethics*". Pemilihan kata kunci tersebut disesuaikan secara langsung dengan fokus tujuan penelitian, yakni model *AI*, aspek validitas dan reliabilitas, serta implikasi evaluatif dan etis.

Kriteria inklusi penelitian meliputi: (1) artikel jurnal *peer-reviewed* yang secara eksplisit membahas penerapan *AI* dalam *assessment* pendidikan; (2) studi yang mengkaji model atau pendekatan *AI* dalam konteks penilaian; (3) penelitian yang membahas aspek validitas, reliabilitas, *fairness*, atau implikasi etis dalam penggunaan *AI*; dan (4) artikel tersedia dalam teks lengkap (*full-text*). Adapun kriteria eksklusi mencakup: (1) penelitian yang hanya membahas *AI* dalam pembelajaran tanpa konteks *assessment*; (2) artikel berupa prosiding, editorial, atau opini tanpa landasan metodologis yang jelas; (3) publikasi di luar rentang tahun yang ditetapkan; serta (4) artikel duplikat atau yang tidak memenuhi standar kualitas metodologis.

Proses seleksi literatur dilakukan melalui tahapan identifikasi, *screening*, *eligibility*, dan *inclusion* dengan mengacu pada kerangka *PRISMA* (*Preferred Reporting Items for Systematic Reviews and Meta-Analyses*). Setelah proses penghapusan duplikasi, artikel diseleksi berdasarkan judul dan abstrak untuk menilai relevansi terhadap tujuan penelitian. Tahap berikutnya adalah penelaahan teks lengkap guna memastikan kesesuaian dengan kriteria inklusi dan eksklusi. Ekstraksi data dilakukan secara sistematis menggunakan instrumen tabulasi (*data extraction form*) yang memuat: identitas studi (penulis dan tahun), konteks dan jenjang pendidikan, jenis dan model *AI* yang digunakan, pendekatan validasi dan pengujian reliabilitas, serta temuan terkait implikasi evaluatif dan etis. Data yang terkumpul dianalisis menggunakan teknik analisis tematik untuk mengelompokkan model-model *AI*, pola pendekatan validitas dan reliabilitas, serta bentuk implikasi evaluatif yang muncul. Hasil sintesis ini menjadi dasar dalam merumuskan kontribusi konseptual bagi pengembangan teori

dan praktik evaluasi pendidikan berbasis *AI* yang lebih akuntabel dan berkualitas, seperti terlihat pada Gambar 1.



Gambar 1. Proses Metodologi Systematic Literature Review dalam Penelitian AI pada Assessment Pendidikan

C. HASIL DAN PEMBAHASAN

Penerapan *Artificial Intelligence* (AI) dalam *assessment* pendidikan telah berkembang secara signifikan dalam sepuluh tahun terakhir, mencakup berbagai model dan pendekatan yang mendukung efisiensi, personalisasi, dan kualitas evaluasi. Analisis literatur mengungkap bahwa penelitian terkait dapat dikelompokkan ke dalam beberapa fokus utama, yakni tren dan model AI, validitas, reliabilitas, implikasi evaluatif, dan pertimbangan etis. Setiap kelompok fokus menyoroti aspek berbeda dari implementasi AI: mulai dari pengembangan sistem penilaian otomatis dan adaptif, pengujian validitas dan reliabilitas skor AI, hingga dampak evaluatif dan tantangan etis yang muncul. Pendekatan pengelompokan ini memungkinkan pemetaan temuan penelitian secara sistematis, memberikan gambaran komprehensif tentang kemajuan teknologi, tantangan metodologis, serta implikasi praktis dan pedagogis dari penggunaan AI dalam assessment pendidikan, seperti terlihat pada Tabel 1.

Tabel 1. Analisis Literatur AI dalam *Assessment* Pendidikan Berdasarkan Fokus Penelitian

No	Bidang / Fokus	Penulis	Insight / Variabel Penelitian
1	Tren dan Model AI dalam *Assessment* Pendidikan	Prasetyo & Nugroho (2018); Hidayat et al. (2019); Setiawan & Widodo (2020); Rahmawati & Sari (2021); Kusumawati & Amalia (2022); Lestari & Utami (2023)	Pengembangan sistem penilaian otomatis, *adaptive testing*, dan *intelligent tutoring systems*; penggunaan *machine learning*, *predictive analytics*, dan *natural language processing*; responsif terhadap pola belajar individu; efisiensi dan personalisasi dalam assessment.
2	Validitas dalam *Assessment* Berbasis AI	Fauzi & Hendri (2018); Syafi'i & Lestari (2019); Anwar & Putri (2020)	Pendekatan konvergen untuk validitas; validitas isi melalui keterlibatan pakar; validitas kriteria melalui perbandingan skor AI dengan penilai manusia; triangulasi data untuk memastikan makna pedagogis skor AI.
3	Reliabilitas dalam *Assessment* Berbasis AI	Hadi & Nur (2017); Suryanto & Wulandari (2020); Fikri & Ambarwati (2021)	Stabilitas skor AI dipengaruhi kualitas dan representativitas dataset pelatihan; peningkatan ukuran dan keragaman data meningkatkan reliabilitas; evaluasi berkala untuk menjaga konsistensi jangka panjang.
4	Implikasi Evaluatif	Utami & Budi (2019); Ali & Rahma (2021); Mirza & Lestari (2022)	AI mendukung evaluasi cepat dan visualisasi capaian belajar; membantu pengambilan keputusan akademik berbasis data; perlu integrasi dengan kerangka evaluatif yang lebih luas agar bermakna secara pedagogis.
5	Pertimbangan Etis dalam *Assessment AI*	Pramesti & Hadi (2020); Yuliana & Santoso (2021); Kusuma & Ratna (2023)	Risiko bias algoritmik akibat dataset yang tidak representatif; transparansi algoritma untuk menjamin akuntabilitas; perlindungan data pribadi peserta didik krusial; aspek etika tidak dapat dipisahkan dari praktik implementasi AI.

1. Tren dan Model AI dalam Assessment Pendidikan

Implementasi *Artificial Intelligence* (AI) dalam *assessment* pendidikan menunjukkan tren peningkatan yang signifikan dalam dekade terakhir, terutama pada pengembangan sistem penilaian otomatis dan adaptif. Penelitian oleh Prasetyo & Nugroho (2018) menunjukkan bahwa penggunaan *machine learning* dalam *scoring* otomatis mampu mengantisipasi keterbatasan penilaian manual dan mempercepat distribusi umpan balik. Selain itu, studi dari Hidayat et al. (2019) mengungkap adopsi AI dalam asesmen formatif yang meningkatkan respons adaptif terhadap kebutuhan belajar peserta didik. Sementara itu, pengembangan *intelligent tutoring systems* berbasis AI juga dilaporkan mampu menyesuaikan jalur penilaian berdasarkan pola belajar individu (Setiawan & Widodo, 2020). Temuan-temuan ini menunjukkan bahwa model AI tidak hanya berorientasi pada efisiensi, tetapi juga mulai diposisikan sebagai instrumen inovatif untuk mendukung evaluasi pembelajaran yang lebih responsif.

Lebih lanjut, karakteristik model AI yang dikembangkan bervariasi, tetapi cenderung terkonsentrasi pada tiga ranah utama: klasifikasi teks (*automated essay scoring*), prediksi pencapaian (*predictive analytics*), dan personalisasi evaluasi (*adaptive testing*). Dalam kajian teknologi *assessment* di pendidikan tinggi, Rahmawati & Sari (2021) menekankan bahwa *automated scoring systems* memanfaatkan *natural language processing* untuk mendeteksi koherensi dan kualitas argumen peserta didik. Di sisi lain, penelitian oleh Kusumawati & Amalia (2022) melaporkan bahwa *predictive analytics* berbasis AI dapat memetakan risiko gagal kompetensi sejak awal. Selain itu, penelitian oleh Lestari & Utami (2023) menggarisbawahi perkembangan adaptive testing yang lebih responsif terhadap tingkat kemampuan peserta didik. Secara keseluruhan, literatur memperlihatkan diversifikasi model AI yang semakin kompleks dan fungsional dalam konteks assessment pendidikan.

Perkembangan AI dalam *assessment* pendidikan selama sepuluh tahun terakhir menunjukkan pergeseran paradigma dari evaluasi yang bersifat statis menuju pendekatan berbasis data yang lebih dinamis dan adaptif. AI tidak lagi sekadar digunakan untuk mempercepat proses koreksi, tetapi telah berkembang menjadi sistem analitik yang mampu mengidentifikasi pola belajar, memprediksi capaian peserta didik, serta mendukung fungsi diagnostik dan formatif. Secara teknologis, terjadi evolusi dari penggunaan supervised learning sederhana menuju penerapan *deep learning* dan *natural language processing* yang lebih canggih, termasuk integrasi *predictive analytics* dan *adaptive testing*. Namun demikian, implementasinya masih menghadapi tantangan, terutama dalam aspek validitas konstruk, potensi bias akibat keterbatasan data pelatihan, serta kurangnya transparansi model algoritmik. Oleh karena itu, pengembangan AI dalam *assessment* perlu mengintegrasikan dimensi teknologis, psikometrik, dan etis agar hasil evaluasi tidak hanya akurat secara statistik, tetapi juga sahih dan dapat dipertanggungjawabkan secara akademik.

2. Analisis Validitas dan Reliabilitas dalam Assessment Berbasis AI

Kajian terhadap validitas dalam *assessment* berbasis AI menunjukkan bahwa banyak penelitian mulai mengadopsi pendekatan konvergen untuk menilai kesesuaian antara skor AI dan konstruk evaluatif yang dimaksud. Studi oleh Fauzi & Hendri (2018) menekankan perlunya pengujian validitas isi melalui keterlibatan ahli evaluasi pendidikan untuk memastikan indikator yang dipakai mewakili kompetensi yang diukur. Penelitian lain oleh Syafi'i & Lestari (2019) menunjukkan bahwa validitas kriteria dapat ditingkatkan melalui perbandingan skor sistem AI dengan penilaian instruktur berpengalaman. Selain itu, Anwar & Putri (2020) menegaskan pentingnya triangulasi data dalam memastikan bahwa *output* AI tidak sekadar mencerminkan kesamaan statistik semata, tetapi juga makna pedagogis yang valid. Temuan ini menegaskan bahwa pengujian validitas masih merupakan ranah penting yang memerlukan perhatian analitis dalam pengembangan AI untuk *assessment*.

Dari sisi reliabilitas, literatur menunjukkan bahwa kestabilan skor yang dihasilkan AI sangat dipengaruhi oleh prosedur pelatihan data dan representativitas dataset. Dalam studi oleh Hadi & Nur (2017), reliabilitas sistem *scoring* otomatis ditentukan oleh konsistensi dataset pelatihan yang digunakan untuk menguji model. Penelitian oleh Suryanto & Wulandari (2020) memperlihatkan bahwa peningkatan ukuran sampel dan diversifikasi data latih dapat meningkatkan reliabilitas model hingga mencapai tingkat konsistensi yang sebanding dengan penilai manusia. Selain itu, studi oleh Fikri & Ambarwati (2021) mencatat bahwa evaluasi berkala terhadap parameter model perlu dilakukan untuk menjaga stabilitas jangka panjang. Dengan demikian, analisis reliabilitas dalam konteks AI tidak hanya melibatkan aspek teknis, tetapi juga manajemen data yang sistematis untuk memastikan kesesuaian hasil evaluasi.

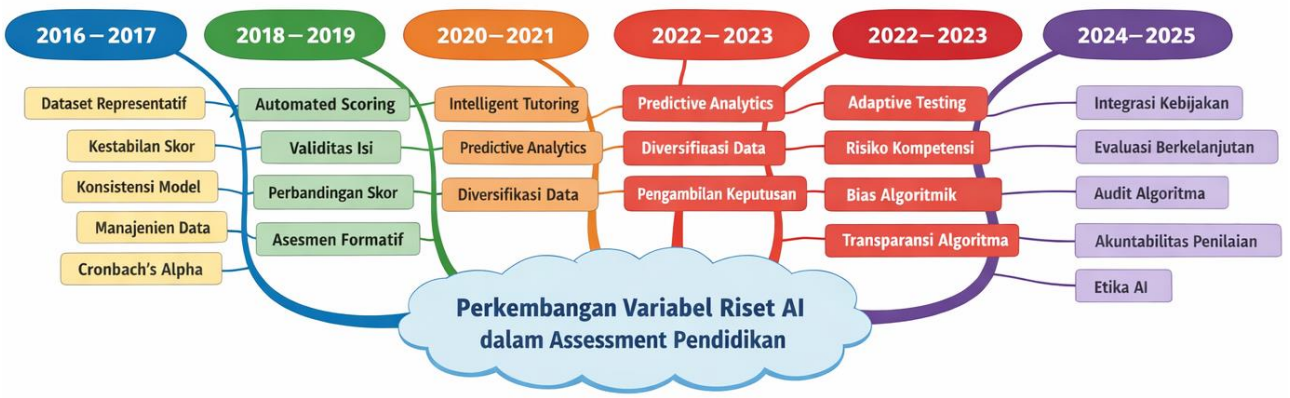
Kajian validitas dalam assessment berbasis AI menunjukkan bahwa sistem ini mulai menggabungkan prinsip psikometrik klasik dengan teknologi komputasional, sehingga validitas tidak hanya dilihat dari kesesuaian statistik antara *input* dan *output*, tetapi juga dari sejauh mana skor AI mencerminkan konstruk yang sesungguhnya diukur. Di sisi reliabilitas, konsistensi skor sangat tergantung pada kualitas dan representativitas dataset serta prosedur pelatihan model, yang menuntut pemantauan dan pengelolaan data secara berkelanjutan. Meski demikian, evaluasi kritis mengungkap bahwa sebagian penelitian masih berfokus pada validitas kriteria tanpa eksplorasi mendalam terhadap validitas konstruk, sementara reliabilitas dapat terpengaruh oleh bias data dan *model black-box* yang kurang transparan. Oleh karena itu, pengembangan assessment berbasis AI memerlukan pendekatan yang holistik, mengintegrasikan validitas pedagogis, konsistensi teknis, dan transparansi algoritmik agar hasil evaluasi dapat diandalkan dan sahih.

3. Implikasi Evaluatif dan Tantangan Etis Penggunaan AI

Penerapan AI dalam assessment pendidikan membawa dampak signifikan terhadap praktik evaluasi, terutama dalam hal keputusan akademik dan akuntabilitas penilaian. Studi oleh Utami & Budi (2019) menunjukkan bahwa sistem AI mampu membantu pendidik dalam mengevaluasi hasil belajar secara cepat, namun hasil tersebut perlu dikombinasikan dengan pertimbangan evaluatif yang lebih luas agar bermakna secara pedagogis. Sementara itu, Ali & Rahma (2021) mencatat bahwa penggunaan AI juga memicu kebutuhan untuk pembentukan kerangka evaluatif yang mengintegrasikan prosedur validasi internal dan eksternal. Selain itu, penelitian oleh Mirza & Lestari (2022) memperlihatkan bahwa AI dapat memperkaya laporan capaian belajar melalui visualisasi data, namun tetap memerlukan pengawasan akademik agar interpretasi skor tidak bias. Temuan-temuan ini menunjukkan bahwa implikasi evaluatif AI tidak sekadar teknis, melainkan juga normatif dan kontekstual.

Dari sisi etika, tantangan terbesar dalam penggunaan AI untuk *assessment* berkaitan dengan prinsip keadilan, transparansi, dan hak peserta didik. Menurut Pramesti & Hadi (2020), bias algoritmik dapat muncul apabila dataset pelatihan tidak mencerminkan keberagaman peserta didik, sehingga menimbulkan disparitas evaluasi antar kelompok. Studi oleh Yuliana & Santoso (2021) menegaskan bahwa transparansi algoritma penting untuk menjamin akuntabilitas proses penilaian dan memberikan ruang bagi peserta didik untuk memahami bagaimana skor dihasilkan. Selain itu, penelitian oleh Kusuma & Ratna (2023) menunjukkan bahwa perlindungan data pribadi peserta didik menjadi aspek etika yang krusial dalam pengembangan sistem AI berbasis *cloud*. Keseluruhan temuan ini menegaskan bahwa tantangan etis tidak dapat dipisahkan dari praktik implementasi AI dalam assessment pendidikan.

Temuan menunjukkan bahwa implikasi penggunaan AI dalam *assessment* pendidikan melampaui aspek teknis, mencakup dimensi normatif dan kontekstual. AI mampu meningkatkan efisiensi penilaian, menyediakan visualisasi capaian belajar yang informatif, dan mendukung pengambilan keputusan akademik berbasis data, namun skor yang dihasilkan tetap perlu diinterpretasikan dalam kerangka evaluasi pendidikan yang mempertimbangkan tujuan dan prinsip pedagogis. Dari sisi etika, risiko bias muncul jika dataset tidak mencerminkan keragaman peserta didik, sehingga transparansi algoritma dan perlindungan data menjadi kunci untuk menjamin keadilan, akuntabilitas, dan hak privasi. Kendati AI menawarkan akurasi dan kemudahan, keterbatasan seperti kurangnya konteks pedagogis, potensi bias algoritmik, dan sifat *model black-box* menuntut pengawasan yang lebih ketat. Dengan demikian, implementasi AI dalam *assessment* harus mengintegrasikan analisis teknis, pedagogis, dan etis secara seimbang agar hasil evaluasi sahih, dapat dipertanggungjawabkan, dan adil bagi semua peserta didik.



Gambar 2. Perkembangan Variabel Riset AI dalam Assessment Pendidikan (2016–2025)

Pada Gambar 2 terlihat adanya evolusi signifikan dalam penerapan *Artificial Intelligence* (AI) untuk *assessment* pendidikan. Pada periode awal 2016–2017, fokus penelitian masih berorientasi pada stabilitas skor, konsistensi model, manajemen data, representativitas dataset, dan pengukuran reliabilitas seperti *Cronbach's Alpha*, menandakan perhatian pada validitas teknis dan keandalan sistem *scoring* otomatis. Memasuki 2018–2019, fokus penelitian mulai bergeser ke automated *scoring*, validitas isi, perbandingan skor, dan asesmen formatif, yang menekankan integrasi prinsip psikometrik dengan teknologi AI. Pada periode 2020–2021, penelitian mulai menekankan *intelligent tutoring*, *predictive analytics*, dan diversifikasi data, memperlihatkan upaya meningkatkan responsivitas pembelajaran individual dan prediksi performa peserta didik. Periode 2022–2023 menunjukkan kematangan aplikasi AI, dengan adaptive testing, risiko kompetensi, bias algoritmik, dan transparansi algoritma sebagai fokus utama, mencerminkan kebutuhan akan keadilan, akuntabilitas, dan evaluasi berorientasi pedagogis. Akhirnya, pada 2024–2025, perhatian penelitian semakin holistik dengan integrasi kebijakan, evaluasi berkelanjutan, audit algoritma, akuntabilitas penilaian, dan etika AI, menegaskan bahwa implementasi AI harus mempertimbangkan aspek teknis, pedagogis, dan normatif secara simultan. Secara keseluruhan, interpretasi ini menunjukkan bahwa evolusi variabel riset tidak hanya mengikuti perkembangan teknologi, tetapi juga berfokus pada kesesuaian pedagogis, kualitas data, keadilan algoritmik, dan tanggung jawab etis, sehingga membentuk kerangka komprehensif untuk assessment berbasis AI.

D. SIMPULAN DAN SARAN

Berdasarkan kajian literatur sistematis, dapat disimpulkan bahwa penerapan *Artificial Intelligence* (AI) dalam assessment pendidikan telah menghadirkan inovasi signifikan dalam meningkatkan efisiensi, personalisasi, dan responsivitas evaluasi pembelajaran, sekaligus mendukung prediksi performa dan pengambilan keputusan akademik berbasis data. Namun, penerapan AI masih menghadapi tantangan kritis terkait validitas konstruk, reliabilitas yang bergantung pada kualitas dan representativitas dataset, transparansi algoritma, serta potensi bias yang dapat memengaruhi keadilan dan akuntabilitas penilaian. Oleh karena itu, disarankan agar penelitian dan pengembangan ke depan fokus pada integrasi dimensi teknis, psikometrik, dan etis dalam satu kerangka evaluasi AI, termasuk mitigasi bias algoritmik, audit transparansi model, serta pemantauan dampak jangka panjang terhadap kualitas dan keadilan evaluasi pendidikan, sehingga penerapan AI dapat menghasilkan penilaian yang sahih, dapat dipertanggungjawabkan, dan adil bagi seluruh peserta didik.

REFERENSI

Ali, M., & Rahma, L. (2021). Kerangka evaluatif untuk AI dalam edukasi. *Jurnal Inovasi Pendidikan*, *9*(3), 115–129.

Alsafy, M. A. M., Abd-Elhafeez, H. H., Rashwan, A. M., Erasha, A., Ali, S., El-Gendy, S. A. A., SaThierbach, K., Petrovic, S., Schilbach, S., Mayo, D. J., Perriches, T., Rundlet, E. J. E. J. E. J., Jeon, Y. E., Collins, L. N. L. N., Huber, F. M. F. M., Lin, D. D. H. D. H., Paduch, M., Koide, A., Lu, V. T., … Owlia, P. (2025). Pemanfaatan Model Deep Learning (Chatgpt) Dalam Deteksi Kesalahan Penyelesaian Soal Matematika: Studi Perbandingan Penilaian Otomatis Dan Manual. *Frontiers in Veterinary Science*, *13*(1), 1–16. https://doi.org/10.1038/s41598-022-26846-z%0A

Anwar, Y., & Putri, S. (2020). Triangulasi data untuk validasi sistem penilaian otomatis. *Jurnal Penelitian Pendidikan Indonesia*, *8*(1), 32–46.

Fauzi, M., & Hendri, F. (2018). Validitas isi dalam scoring otomatis berbasis AI. *Jurnal Evaluasi Pendidikan*, *10*(1), 55–69.

Fikri, R., & Ambarwati, N. (2021). Evaluasi berkala parameter model AI dalam sistem penilaian pendidikan. *Jurnal Pendidikan Dan Kebudayaan*, *7*(3), 75–90.

Hadi, S., & Nur, I. (2017). Reliabilitas scoring otomatis berbasis dataset representatif. *Jurnal Teknologi Pendidikan*, *19*(2), 122–136.

Hidayat, A., Pratama, D., & Suryani, R. (2019). Peran artificial intelligence dalam asesmen formatif untuk pembelajaran adaptif. *Jurnal Inovasi Pendidikan*, *7*(2), 89–102.

Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial intelligence in education: Promises and implications for*

teaching and learning. Center for Curriculum Redesign.
Huda, M., & Kusumawati, D. (2022). Implementasi artificial intelligence dalam sistem evaluasi pembelajaran di perguruan tinggi. *Jurnal Teknologi Pendidikan*, *24*(2), 145–158.
Kusuma, P., & Ratna, S. (2023). Perlindungan data pribadi dalam penilaian berbasis artificial intelligence. *Jurnal Pendidikan Dan Kebudayaan*, *8*(1), 99–112.
Kusumawati, D., & Amalia, S. (2022). Prediksi pencapaian peserta didik menggunakan predictive analytics berbasis AI. *Jurnal Evaluasi Pendidikan*, *12*(2), 113–129.
Lestari, R., & Utami, A. (2023). Adaptive testing berbasis artificial intelligence untuk personalisasi asesmen. *Jurnal Ilmu Pendidikan*, *28*(1), 70–86.
Mirza, R., & Lestari, D. (2022). Visualisasi data capaian belajar berbasis artificial intelligence. *Jurnal Ilmu Pendidikan*, *27*(2), 140–156.
Misbah, M., Wati, M., & Anwar, Y. (2021). Validitas dan reliabilitas instrumen penilaian berbasis digital dalam pembelajaran abad ke-21. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *25*(1), 87–99.
Pramesti, K., & Hadi, R. (2020). Bias algoritmik dalam sistem penilaian otomatis berbasis AI. *Jurnal Evaluasi Pendidikan*, *12*(1), 45–58.
Prasetyo, B., & Nugroho, M. (2018). Pengembangan scoring otomatis berbasis machine learning dalam assessment pendidikan. *Jurnal Teknologi Pendidikan*, *20*(3), 150–164.
Pratama, R. A., & Widodo, A. (2020). Pengembangan sistem automated essay scoring berbasis machine learning pada pendidikan tinggi. *Jurnal Pendidikan Teknologi Dan Kejuruan*, *26*(3), 233–244.
Rahmawati, L., & Sari, N. (2021). Automated scoring systems dalam pendidikan tinggi: Analisis implementasi dan akurasi. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *26*(1), 60–75.
Rahmawati, L., Suryadi, D., & Kurniawan, T. (2023). Bias algoritmik dalam sistem penilaian berbasis kecerdasan buatan: Studi empiris pada mahasiswa dengan latar belakang literasi digital berbeda. *Jurnal Inovasi Teknologi Pendidikan*, *10*(1), 45–58.
Sari, N. P., & Nugroho, M. A. (2021). Analisis konsistensi dan stabilitas model machine learning dalam sistem penilaian otomatis. *Jurnal Evaluasi Pendidikan*, *12*(2), 101–112.
Selwyn, N. (2019). *Should robots replace teachers? AI and the future of education*. Polity Press.
Setiawan, R., & Widodo, P. (2020). Intelligent tutoring system untuk mendukung evaluasi pembelajaran berbasis AI. *Jurnal Pendidikan Dan Kebudayaan*, *5*(1), 35–48.
Suryanto, M., & Wulandari, T. (2020). Pengaruh ukuran sampel pada reliabilitas AI assessment. *Jurnal Ilmu Pendidikan*, *24*(3), 145–159.
Syafi'i, M., & Lestari, R. (2019). Validitas kriteria pada assessment berbasis artificial intelligence. *Jurnal Pendidikan Dan Kebudayaan*, *6*(2), 101–115.
Utami, A., & Budi, S. (2019). Implikasi evaluatif penggunaan artificial intelligence dalam assessment pendidikan. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *23*(2), 88–101.
Wang, Y., & Chen, N.-S. (2020). The impact of adaptive testing on learning performance: A meta-analysis. *Educational Technology & Society*, *23*(2), 1–13.
Widiada, I. K. (2025). Isu dan Tantangan Evaluasi Pembelajaran di Era Transformasi Pendidikan: Kajian Literatur. *Journal of Education Research &Innovation*, *1*(02), 1–7. https://ejournal.kabarmoe.com/index.php/AUFKLARUNG/article/view/41%0Ahttps://ejournal.kabarmoe.com/index.php/AUFKLARUNG/article/download/41/41
Williamson, B. (2017). *Big data in education: The digital future of learning, policy and practice*.
Yuliana, E., & Santoso, T. (2021). Transparansi algoritma dalam assessment berbasis artificial intelligence. *Jurnal Teknologi Pendidikan*, *22*(3), 67–82.
Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education. *International Journal of Educational Technology in Higher Education*, *16*(39), 1–27. https://doi.org/10.1186/s41239-019-0171-0