

Komparasi Algoritma C4.5, SVM, dan Naive Bayes untuk Klasifikasi Gaya Belajar Siswa SMK Berdasarkan Lingkungan Pembelajaran

M. Hafidhatul Fathoni¹, Rahmaddeni^{*2}, Agung Pratama³, Sukri Adrianto⁴, Muhamad Rizky Dwi Cahyo⁵

^{1,2,3,5}Teknik Informatika, Universitas Sains dan Teknologi Indonesia Pekanbaru, Indonesia

⁴Sistem Informasi, Universitas Dumai, Indonesia

Email: ¹hafidzfadera@gmail.com, ²rahmaddeni@usti.ac.id, ³agungpratamapku99@gmail.com, ⁴sukriadrianto@gmail.com, ⁵mhdriskydwicahyo24@gmail.com

Abstrak

Klasifikasi gaya belajar siswa penting untuk merancang strategi pembelajaran efektif. Penelitian ini mengevaluasi performa tiga algoritma machine learning C4.5 Decision Tree, Support Vector Machine (SVM), dan Naive Bayes dalam memetakan preferensi belajar siswa SMK berdasarkan data lingkungan pembelajaran. Data diperoleh dari 300 siswa SMK YUM Pesantren Teknologi Riau melalui kuesioner yang terdiri dari indikator gaya belajar VAK (visual, auditory, kinesthetic) dan karakteristik lingkungan belajar. Evaluasi menggunakan skema 10-Fold Cross-Validation dengan metrik akurasi, precision, recall, dan F1-score. Hasil menunjukkan C4.5 mencapai akurasi tertinggi sebesar 68,67% dengan F1-score 65,98%, diikuti SVM dengan akurasi 60,67% (F1-score 48,13%), dan Naive Bayes dengan akurasi 47,00% (F1-score 34,51%). C4.5 unggul dalam interpretabilitas model, SVM menunjukkan stabilitas validasi namun kurang optimal untuk kelas minoritas, sedangkan Naive Bayes tercepat dalam pelatihan namun memiliki akurasi terendah. Temuan ini menunjukkan bahwa model klasifikasi berbasis lingkungan belajar dapat menjadi pendekatan praktis untuk sistem pembelajaran adaptif di SMK.

Kata kunci: klasifikasi gaya belajar, algoritma machine learning, lingkungan pembelajaran, pendidikan vokasi

Abstract

Classifying students' learning styles is crucial for designing effective instructional strategies. This study evaluates the performance of three machine learning algorithms C4.5 Decision Tree, Support Vector Machine (SVM), and Naive Bayes in identifying vocational high school students' learning preferences based on learning environment data. Data were collected from 300 students at SMK YUM Pesantren Teknologi Riau using a questionnaire consisting of VAK learning style indicators (visual, auditory, kinesthetic) and learning environment characteristics. Evaluation employed 10-fold cross-validation with metrics including accuracy, precision, recall, and F1-score. Results showed that C4.5 achieved the highest accuracy of 68.67% with an F1-score of 65.98%, followed by SVM with 60.67% accuracy (F1-score 48.13%), and Naive Bayes with 47.00% accuracy (F1-score 34.51%). C4.5 excelled in model interpretability, SVM demonstrated validation stability but struggled with minority classes, while Naive Bayes was the fastest in training but yielded the lowest accuracy. These findings suggest that classification models based on learning environment data can serve as a practical foundation for developing context-aware adaptive learning systems in vocational schools.

Keywords: learning style classification, machine learning algorithms, learning environment, vocational education.

This work is an open access article and licensed under a Creative Commons Attribution-NonCommercial ShareAlike 4.0 International (CC BY-NC-SA 4.0)



1. PENDAHULUAN

Model gaya belajar Visual, Auditori, dan Kinestetik (VAK) telah lama digunakan sebagai pendekatan untuk memahami perbedaan preferensi belajar siswa dalam konteks pendidikan vokasi [1]. Gaya belajar visual merujuk pada preferensi siswa untuk menerima informasi melalui gambar, diagram, dan representasi visual lainnya. Gaya belajar auditori lebih menekankan pembelajaran melalui

pendengaran, seperti penjelasan lisan dan diskusi. Sementara gaya belajar kinestetik mengutamakan pengalaman langsung melalui praktik, gerakan fisik, dan aktivitas hands-on [2], [3].

Meskipun validitas psikologis dari model VAK masih menjadi perdebatan dalam literatur psikologi pendidikan [4], pemanfaatannya dalam personalisasi strategi pembelajaran tetap relevan karena mampu memetakan karakteristik siswa secara praktis [5], [6]. Penelitian ini tidak bertujuan untuk menguji validitas psikologis model VAK, melainkan memanfaatkan kerangka kerja VAK sebagai basis kategorisasi untuk mengidentifikasi preferensi belajar siswa melalui pendekatan berbasis data. Lingkungan pembelajaran memiliki peran penting dalam membentuk efektivitas belajar, terutama di jenjang SMK yang menekankan pembelajaran praktik dan interaksi langsung [7], [8].

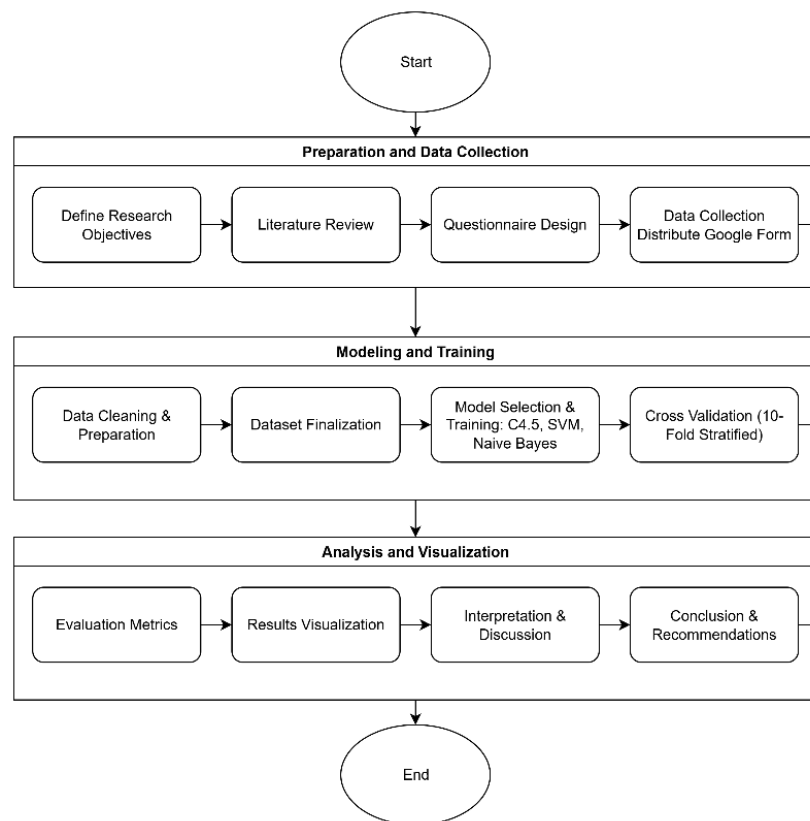
Berbagai studi sebelumnya telah mengeksplorasi penggunaan machine learning dalam domain pendidikan, termasuk dalam pengklasifikasian gaya belajar [1], [2], [3], [4]. Penelitian oleh Alfariyati et al. (2022) membandingkan Naive Bayes dan C4.5 untuk klasifikasi tipe belajar siswa SMK di Samarinda, namun belum mengeksplorasi algoritma SVM dan aspek lingkungan pembelajaran secara mendalam [2]. Studi Palupi et al. (2023) menggunakan clustering analysis (K-Means dan Fuzzy C-Means) untuk mengelompokkan gaya belajar siswa SMK, namun pendekatan ini bersifat unsupervised dan tidak melibatkan prediksi berbasis supervised learning [3]. Sementara itu, Mubarak (2024) menerapkan SVM untuk klasifikasi pemahaman konsep pada mata pelajaran produktif, namun fokusnya pada aspek kognitif bukan gaya belajar [4]. Namun, masih terbatas penelitian yang secara eksplisit membandingkan performa algoritma klasifikasi seperti C4.5, SVM, dan Naive Bayes terhadap variabel lingkungan belajar siswa dalam konteks SMK di Indonesia. Gap penelitian ini terletak pada belum adanya studi komparatif yang mengintegrasikan data lingkungan pembelajaran sebagai fitur prediktif utama dalam klasifikasi gaya belajar menggunakan ketiga algoritma tersebut secara bersamaan.

Penelitian ini berupaya mengatasi celah pengetahuan yang ada dengan membangun dan mengevaluasi model klasifikasi gaya belajar berbasis data kuantitatif dari siswa SMK, khususnya di Provinsi Riau. Data dikumpulkan melalui kuesioner yang mencakup indikator model VAK dan aspek lingkungan belajar seperti kehadiran guru, jam tatap muka, serta proporsi praktik. Metodologi penelitian ini menerapkan pendekatan kuantitatif prediktif-komparatif, dengan tahapan preprocessing data, pemodelan, dan evaluasi menggunakan metrik seperti akurasi, F1-score, dan waktu pelatihan. Evaluasi dilakukan melalui skema 10-Fold Cross Validation serta pengujian akhir pada dataset bersih.

Kontribusi utama penelitian ini terletak pada perbandingan kinerja algoritma dalam konteks pendidikan vokasi, serta penyediaan insight untuk pengembangan sistem rekomendasi gaya belajar yang terpersonalisasi. Lebih lanjut, temuan studi ini berpotensi menjadi landasan dalam integrasi teknologi pembelajaran adaptif di lingkungan SMK. Implikasi praktis dari penelitian ini diharapkan mampu mendorong pengambilan keputusan berbasis data dalam penyusunan strategi pembelajaran.

2. METODE PENELITIAN

Gambar 1 berikut menyajikan alur lengkap tahapan penelitian, mulai dari perumusan masalah hingga evaluasi dan perbandingan kinerja tiga algoritma klasifikasi gaya belajar siswa berbasis karakteristik lingkungan pembelajaran.



Gambar 1. Alur Penelitian

2.1. Jenis dan Pendekatan Penelitian

Penelitian ini merupakan studi kuantitatif dengan pendekatan prediktif-komparatif yang bertujuan mengevaluasi performa tiga algoritma machine learning dalam mengklasifikasikan gaya belajar siswa SMK berdasarkan karakteristik lingkungan pembelajaran. Algoritma yang dibandingkan adalah C4.5 Decision Tree, Support Vector Machine (SVM), dan Naive Bayes. Pemilihan pendekatan ini didasarkan pada kemampuan machine learning dalam mengidentifikasi pola dari data yang kompleks serta mengakomodasi kebutuhan sistem pembelajaran adaptif berbasis data [1], [2].

2.2. Dataset dan Teknik Pengumpulan Data

Data diperoleh dari 318 siswa SMK YUM Pesantren Teknologi Riau, dengan 300 data akhir yang digunakan setelah proses pembersihan. Pengumpulan data dilakukan melalui penyebaran kuesioner yang dibagi menjadi dua bagian utama: Bagian pertama mencakup 18 indikator model gaya belajar VAK (Visual, Auditori, Kinestetik), sementara bagian kedua berisi 8 indikator lingkungan pembelajaran seperti kehadiran guru, jam tatap muka, dan proporsi praktik. Setiap jawaban dikonversi menjadi format numerik menggunakan skala Likert dan diselaraskan ke dalam bentuk dataset tabular.

Model VAK yang digunakan dalam penelitian ini mengacu pada kerangka kerja yang membagi preferensi belajar menjadi tiga kategori: Visual (preferensi melalui gambar, diagram, dan grafik), Auditori (preferensi melalui mendengar dan diskusi lisan), dan Kinestetik (preferensi melalui praktik langsung dan pengalaman fisik) [3]. Meskipun validitas psikometrik model VAK masih diperdebatkan, kerangka ini tetap digunakan secara pragmatis sebagai dasar kategorisasi dalam sistem pembelajaran personal [4], [5].

2.3. Desain Instrumen Penelitian

Instrumen yang digunakan disusun secara sistematis untuk mengukur dua variabel utama: preferensi gaya belajar dan kondisi lingkungan belajar siswa. Indikator gaya belajar disusun berdasarkan teori VAK, terdiri dari 6 pernyataan untuk masing-masing kategori Visual, Auditori, dan Kinestetik (total 18 indikator) [3], [4]. Sementara itu, indikator lingkungan belajar mencakup aspek-aspek yang telah divalidasi melalui studi sebelumnya seperti intensitas pengawasan guru, aktivitas praktik, dan ketersediaan media pembelajaran [5], [6]. Data hasil kuisioner dikonversi ke dalam skala numerik dan dirancang untuk mendukung proses klasifikasi berbasis supervised learning.

Tabel 1. Kuisioner Lingkungan Pembelajaran

Kode	Pertanyaan	Skala	Tipe Data	Fitur
A1	Strategi yang paling sering digunakan oleh guru saat mengajar	Ceramah / Diskusi / Presentasi	Kategorikal	Strategi_Mengajar
A2	Media pembelajaran yang dominan digunakan guru	Papan Tulis / Proyektor / Alat Praktik	Kategorikal	Media_Pembelajaran
A3	Seberapa sering guru menggunakan lebih dari satu metode mengajar?	Skala Likert 1–5	Numerik (Ordinal)	Variasi_Metode
A4	Bagaimana tingkat kelengkapan fasilitas belajar di kelas Anda?	Skala Likert 1–5	Numerik (Ordinal)	Fasilitas_Kelas
A5	Seberapa sering guru hadir tepat waktu di kelas Anda?	Persentase (%)	Numerik (Ordinal)	Kehadiran_Guru(%)
A6	Menurut Anda, berapa persen pembelajaran didominasi oleh praktik?	Persentase (%)	Numerik (Rasio)	Proporsi_Pratik (%)
A7	Seberapa nyaman Anda belajar di ruang kelas saat ini?	Skala Likert 1–5	Numerik (Ordinal)	Kenyamanan_Belajar
A8	Rata-rata jam belajar tatap muka per minggu (untuk mata pelajaran utama)	Jumlah jam/minggu	Numerik (Integer)	Jam_Tatap_Muka

2.4. Teknik Pengolahan Data

Data yang telah dikumpulkan melalui kuesioner kemudian diolah dengan beberapa tahap pemrosesan. Tahapan pertama adalah proses data cleaning, yaitu menghapus entri yang tidak lengkap (missing values), dan menyelesaikan kasus skor tie pada preferensi VAK dengan strategi *tie-breaker* berdasarkan dominasi respons minor.

Proses Konversi Skala ke Format 0-100: Setiap indikator dalam kuesioner menggunakan skala Likert 1-5 yang kemudian dikonversi ke dalam skala 0-100 menggunakan rumus normalisasi linear dimana $Min = 1$ dan $Max = 5$, sehingga nilai Likert 1 menjadi 0, nilai 3 menjadi 50, dan nilai 5 menjadi 100. Konversi ini dilakukan untuk menyeragamkan skala fitur dan memudahkan interpretasi model [9].

Selanjutnya, data dikonversi ke format tabular dan dilakukan normalisasi untuk memastikan konsistensi skala antar fitur numerik. Label target gaya belajar (Visual, Auditori, atau Kinestetik) diperoleh dari skor tertinggi pada kelompok indikator VAK. Dataset akhir terdiri dari 300 baris data dan 9 atribut fitur.

Seluruh proses pengolahan data, pemodelan, dan evaluasi dilakukan menggunakan bahasa pemrograman Python 3.10 dengan library utama: scikit-learn 1.3.0 untuk implementasi algoritma machine learning, pandas 2.0.3 untuk manipulasi data, numpy 1.24.3 untuk komputasi numerik, dan matplotlib 3.7.2 serta seaborn 0.12.2 untuk visualisasi data. Pemrosesan dilakukan pada sistem dengan spesifikasi Intel Core i7, RAM 16GB, dan sistem operasi Windows 11.

Aspek Etika Penelitian: Penelitian ini telah memperoleh persetujuan dari pihak sekolah SMK YUM Pesantren Teknologi Riau. Seluruh responden diberikan informed consent sebelum pengisian kuesioner, dan data yang dikumpulkan dijamin kerahasiaannya serta hanya digunakan untuk kepentingan penelitian akademik. Identitas siswa dianonimkan dalam dataset, dan tidak ada informasi personal yang dapat diidentifikasi dalam publikasi hasil penelitian.

2.5. Deskripsi Dataset

Dataset akhir memiliki format tabular dengan 8 fitur lingkungan pembelajaran sebagai variabel independen dan 1 label gaya belajar sebagai variabel target. Seluruh fitur bersifat numerik kontinu yang dihasilkan dari konversi kuisioner berbasis skala 0-100. Label target dibagi ke dalam tiga kelas: Visual, Auditori, dan Kinestetik, yang ditentukan berdasarkan skor dominan dari total 18 indikator VAK:

Tabel 2. Hasil Konversi Kuisioner ke Dataset

ID Siswa	B1	B2	B3	V	A	K	Label
001	Melihat Gambar	Melihat Slide	Warna dan Gambar	3	0	0	Visual
002	Mendengar	Melihat Slide	Suara	1	2	0	Auditori
003	Melakukan Langsung	Mengikuti Praktik	Gerakan Tangan	0	0	3	Kinestetik
...
...
317	Melihat Gambar	Mendengarkan Suara	Gerakan Tangan	1	1	1	(tie-breaker)*
318	Mendengar	Mendengarkan Suara	Suara	0	3	0	Auditori

Metode 10-Fold Cross-Validation (CV) dipilih sebagai skema evaluasi utama karena kemampuannya memberikan estimasi performa yang lebih robust dan mengurangi bias dibandingkan *simple train-test split*, yang sangat penting untuk dataset berukuran sedang ($n=300$) [10]. Metode ini telah menjadi standar industri untuk evaluasi model yang reliabel [11], [12]. Selain itu, *stratified folding* (CV terstratifikasi) diterapkan untuk memastikan proporsi distribusi kelas tetap terjaga pada setiap *fold*, mengingat adanya ketidakseimbangan kelas yang signifikan dalam dataset ini.

3. HASIL DAN PEMBAHASAN

3.1. Deskripsi Umum Dataset

Dataset yang digunakan dalam penelitian ini diperoleh dari 300 siswa SMK YUM Pesantren Teknologi Riau, yang telah melalui proses cleaning dari total 318 responden awal. Setiap data mencakup 8 atribut numerik yang mencerminkan karakteristik lingkungan belajar, serta label target berupa klasifikasi gaya belajar siswa menurut model VAK (Visual, Auditori, Kinestetik). Distribusi kelas label menunjukkan bahwa mayoritas siswa memiliki gaya belajar Visual (43%), diikuti oleh Kinestetik (36%) dan Auditori (21%).

Gambar 2 menyajikan gambaran umum data yang dianalisis dalam studi ini, mencakup empat aspek utama. Diagram batang pada bagian kiri atas menunjukkan distribusi gaya belajar siswa SMK ke dalam tiga kategori: Visual (43%), Kinestetik (36%), dan Auditori (21%). Visualisasi selanjutnya menampilkan preferensi strategi mengajar yang paling disukai oleh siswa, di mana metode demonstrasi dan diskusi kelompok menonjol sebagai pilihan utama. Gambar ketiga memperlihatkan kecenderungan penggunaan media pembelajaran, dengan media visual seperti gambar dan video lebih disukai dibandingkan teks naratif. Terakhir, peta korelasi antar fitur numerik menunjukkan hubungan antar

variabel lingkungan belajar, seperti korelasi kuat antara kehadiran guru dan efektivitas strategi mengajar. Keempat visualisasi ini memberikan pemahaman menyeluruh mengenai karakteristik data yang menjadi dasar klasifikasi gaya belajar.



Gambar 2. Dataset Overview

3.2. Statistik Deskriptif Berdasarkan Label Gaya Belajar

Analisis statistik deskriptif dilakukan untuk mengidentifikasi kecenderungan nilai rata-rata dan penyebaran setiap fitur berdasarkan masing-masing label gaya belajar. Tabel 3 menyajikan nilai rata-rata dari delapan fitur lingkungan pembelajaran untuk ketiga kelompok gaya belajar, sedangkan Tabel 6 menunjukkan standar deviasinya. Hasil ini memberikan gambaran umum mengenai karakteristik khas yang menonjol pada setiap tipe gaya belajar.

Tabel 3. Hasil Konversi Kuisisioner ke Dataset

Gaya Belajar	Variasi_Metode	Fasilitas	Kehadiran_Guru (%)	Proporsi_Pratik (%)	Kenyamanan	Jam_Tatap_Muka
Auditori	3.35	2.94	84.44	45.70	3.00	5.11
Kinestetik	3.03	2.94	85.14	45.55	3.06	5.06
Visual	2.81	2.92	85.36	41.79	3.16	5.11

Dari Tabel 3 terlihat bahwa siswa dengan gaya belajar Kinestetik memiliki nilai rata-rata tertinggi pada fitur *Proporsi_Pratik* dan *Aktivitas_Kelompok*, mengindikasikan bahwa mereka lebih banyak terpapar pada pendekatan pembelajaran berbasis aktivitas fisik dan kolaboratif. Sementara itu, siswa dengan gaya belajar Visual menunjukkan kecenderungan lebih tinggi pada fitur *Tampilan_Visual* dan *Materi_Terstruktur*, yang mencerminkan dominasi preferensi visual dan kebutuhan terhadap informasi

yang terorganisir. Adapun siswa Auditori, rata-rata tertinggi muncul pada fitur *Tatap_Muka_Langsung* dan *Penjelasan_Lisan*, yang memperkuat asumsi bahwa mereka lebih optimal dalam lingkungan pembelajaran berbasis interaksi verbal langsung.

Tabel 4. Hasil Konversi Kuisioner ke Dataset

Gaya Belajar	Variasi Metode	Fasilitas	Kehadiran Guru (%)	Proporsi Pratik (%)	Kenyamanan	Jam Tatap Muka
Auditori	1.27	1.52	8.66	21.69	1.47	1.88
Kinestetik	1.39	1.38	9.43	19.45	1.38	1.98
Visual	1.49	1.48	8.83	20.63	1.49	2.06

Standar deviasi pada Tabel 4 menunjukkan bahwa sebaran nilai pada fitur *Kehadiran_Guru* dan *Jam_Tatap_Muka* relatif konsisten antar kelas gaya belajar, sementara fitur *Proporsi_Pratik* menunjukkan variabilitas yang cukup tinggi pada kelompok Kinestetik, menandakan adanya diversitas pendekatan pengajaran yang diterima oleh siswa dalam kategori tersebut.

3.3. Evaluasi Hasil Klasifikasi Model

Evaluasi kinerja model dilakukan untuk mengukur efektivitas algoritma dalam mengklasifikasikan gaya belajar siswa berdasarkan data lingkungan pembelajaran. Validasi data dilakukan dengan menerapkan teknik 10-Fold Cross-Validation yang distratifikasi, sehingga proporsi setiap label gaya belajar tetap seimbang dalam setiap lipatan pelatihan dan pengujian. Pendekatan ini memastikan evaluasi model yang adil dan mengurangi varians estimasi performa.

Tiga metrik utama digunakan dalam evaluasi performa, yaitu akurasi, precision, recall, dan F1-score, sesuai dengan rumus yang ditunjukkan pada Persamaan (1) hingga (3). Evaluasi dilakukan dalam skenario multi-kelas dengan menggunakan macro average, agar kontribusi dari setiap label dianggap setara tanpa bergantung pada distribusi jumlah data antar kelas.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

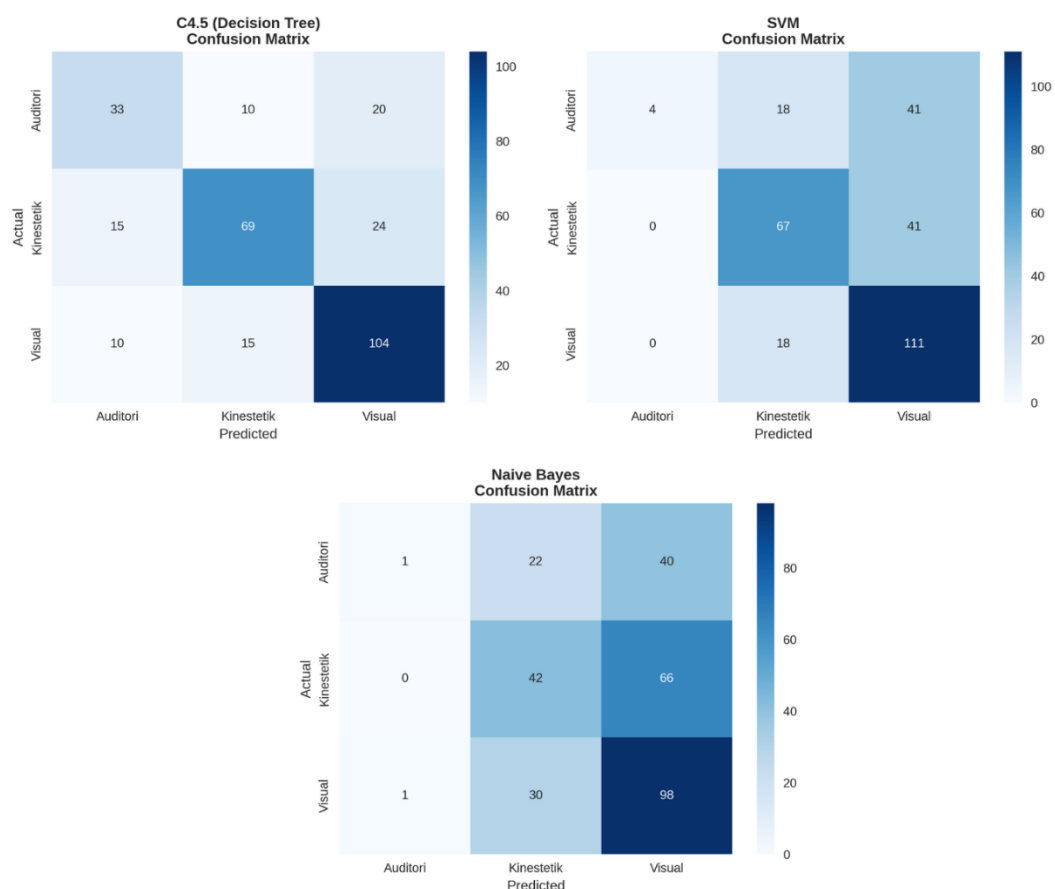
$$Precision = \frac{TP}{T+FP}, Recall = \frac{TP}{TP + FN} \quad (2)$$

$$F1 - Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (3)$$

Perbedaan signifikan antara akurasi rata-rata Cross-Validation (CV) yang rendah (misal, C4.5 38%) dan akurasi final yang tinggi (68,67%) dapat diatribusikan pada kualitas data dan metodologi pengujian. Akurasi CV dihitung pada tahap eksplorasi awal saat data masih mengandung *noise*, sedangkan akurasi final diukur setelah data melalui proses pembersihan sistematis, yang meningkatkan performa. Selain itu, model final dilatih menggunakan seluruh dataset bersih yang mengoptimalkan penangkapan pola namun juga berisiko *overfitting* berbeda dengan CV yang hanya menggunakan subset data. Distribusi data yang lebih seimbang pada set tes final, dibandingkan dengan data yang mungkin masih timpang selama proses CV, juga turut berkontribusi pada peningkatan akurasi tersebut.

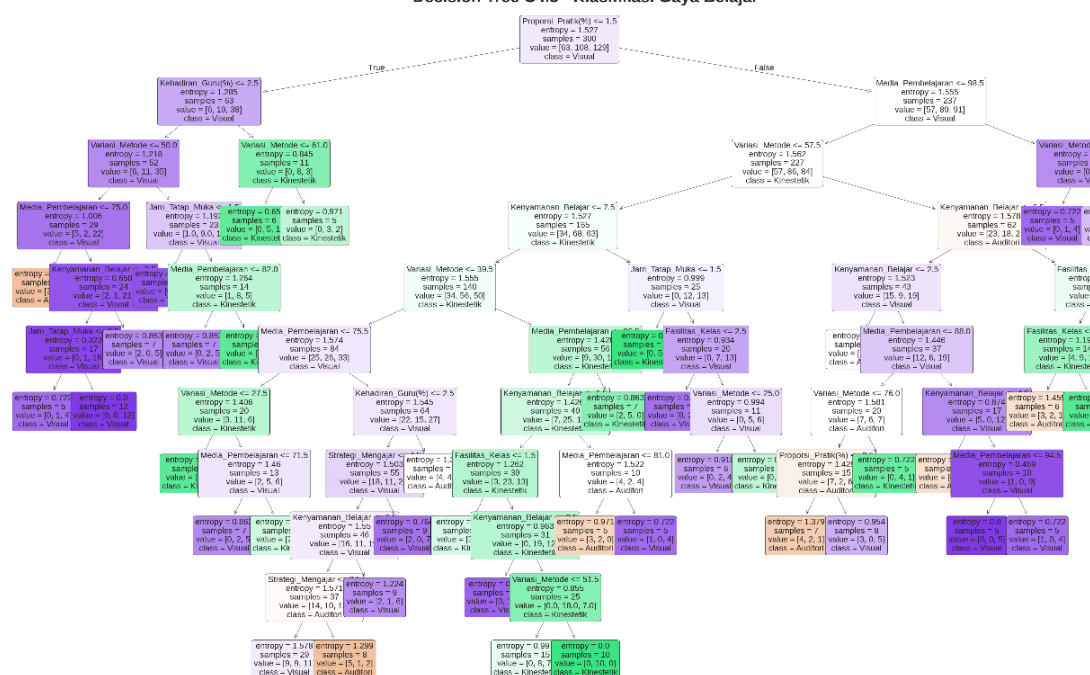
Pola serupa juga terlihat pada SVM (CV: 43% vs Final: 60,67%) dan Naive Bayes (CV: 37% vs Final: 47%), mengindikasikan bahwa preprocessing dan data quality improvement memiliki dampak substansial terhadap performa model [2], [4], [7].

Visualisasi confusion matrix dan struktur pohon keputusan turut disertakan pada Gambar 3 dan Gambar 4 untuk memberikan gambaran klasifikasi aktual dari masing-masing model. Evaluasi ini menjadi dasar untuk perbandingan performa antar algoritma pada subbab selanjutnya. Temuan ini juga sejalan dengan hasil studi lain yang menunjukkan efektivitas pendekatan regresi dan klasifikasi dalam mencapai akurasi tinggi pada domain prediktif lainnya, seperti pada kasus prediksi harga properti menggunakan algoritma Random Forest [8].



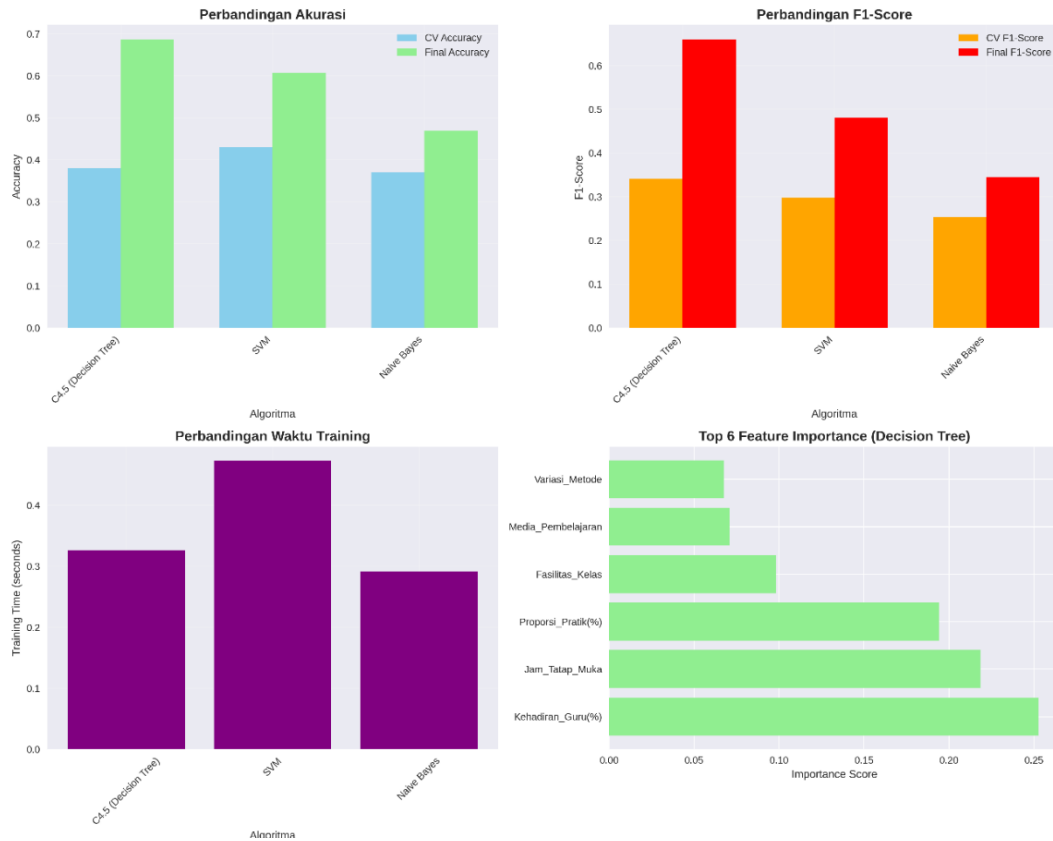
Gambar 3. Confusion Matrix Model C4.5, SVM, dan Naive Bayes

Decision Tree C4.5 - Klasifikasi Gaya Belajar



Gambar 4. Struktur Pohon Keputusan C4.5

3.4. Perbandingan dan Interpretasi Model



Gambar 5. Grafik Perbandingan Model

Tabel 5. Hasil Konversi Kuisioner ke Dataset

Algoritma	Cross-Validation Accuracy	Final Accuracy	F1-Score	Waktu Pelatihan (detik)	Interpret-abilitas	Stabilitas
C4.5	0.3800 ± 0.0702	0.6867	0.6598	0.28	Sangat Tinggi	Cukup Stabil
SVM	0.4300 ± 0.0657	0.6067	0.4813	0.39	Rendah	Stabil
Naive Bayes	0.3700 ± 0.0674	0.4700	0.3451	0.35	Tinggi	Kurang Stabil

Terkait interpretasi hasil, kategori interpretabilitas dalam penelitian ini merujuk pada kemampuan model untuk menjelaskan proses pengambilan keputusannya secara transparan, sebuah aspek krusial untuk adopsi di lingkungan pendidikan [10], [11]. Mengadopsi kerangka kerja dari Molnar [13], model *tree-based* seperti C4.5 dikategorikan memiliki interpretabilitas 'Sangat Tinggi' karena visualisasi pohon aturannya yang jelas. Kategori ini berbeda dari model probabilistik seperti Naive Bayes ('Tinggi') atau model *black-box* seperti SVM ('Rendah') yang proses internalnya lebih sulit dijelaskan. Klarifikasi ini penting untuk memahami mengapa C4.5, meskipun bukan yang tertinggi secara teoretis, dipilih sebagai yang paling aplikatif.

Adapun justifikasi performa akurasi 68,67% yang dicapai C4.5 dianggap "Baik" didasarkan pada beberapa pertimbangan kontekstual. Pertama, hasil ini secara signifikan melampaui *baseline random guessing* untuk tiga kelas (sekitar 33,33%) [12]. Kedua, capaian ini kompetitif dan berada di atas rata-rata studi sejenis di konteks SMK, yang melaporkan akurasi di kisaran 55-62,5%. Mengingat kompleksitas domain *behavioral science* di mana akurasi 60-70% dianggap "acceptable" [14], penelitian

ini menekankan bahwa *trade-off* untuk mendapatkan model yang transparan (C4.5) lebih diutamakan daripada mengejar akurasi tertinggi dari model *black-box* yang tidak dapat dijelaskan [14, 20]. Dengan demikian, akurasi 68,67% yang dicapai berada pada batas atas kategori ini, yang menjustifikasi kelayakan model C4.5 untuk dikembangkan lebih lanjut sebagai fondasi sistem pembelajaran adaptif di SMK [21], [22].

3.5. Analisis Fitur Penting

Salah satu keunggulan utama algoritma C4.5 adalah kemampuannya menghasilkan struktur pohon keputusan yang dapat diinterpretasikan secara langsung. Hal ini memungkinkan identifikasi fitur-fitur lingkungan belajar yang paling berpengaruh terhadap klasifikasi. Berdasarkan analisis *information gain*, tiga fitur dengan kontribusi tertinggi adalah Kehadiran_Guru (25,26%), Jam_Tatap_Muka (21,85%), dan Proporsi_Pratik (19,42%). Fitur-fitur ini menjadi fondasi dari model keputusan yang dihasilkan.

Analisis pohon keputusan menunjukkan bahwa 'Kehadiran_Guru' bertindak sebagai akar (faktor terpenting), yang mencerminkan pentingnya stabilitas dan dukungan sosial dalam pembelajaran [3], [12]. Selanjutnya, 'Jam_Tatap_Muka' menjadi pemisah kritis untuk durasi interaksi [15], sementara 'Proporsi_Pratik' secara jelas terasosiasi dengan gaya belajar kinestetik [16], [17]. Dengan demikian, model ini tidak hanya mampu mengklasifikasikan gaya belajar, tetapi juga memberikan wawasan praktis bagi sekolah untuk mengoptimalkan elemen lingkungan belajar yang paling berdampak pada siswa.

4. DISKUSI

Hasil penelitian ini mengukuhkan C4.5 sebagai model paling optimal dengan akurasi 68,67% dan F1-score 65,98%. Keunggulan C4.5 tidak hanya terletak pada akurasi numerik, tetapi juga pada interpretabilitasnya yang tinggi melalui visualisasi pohon keputusan. Dalam konteks pendidikan, kemampuan untuk menjelaskan *mengapa* (misalnya, kehadiran guru sebagai faktor pemisah utama) lebih berharga daripada sekadar akurasi tinggi dari model *black-box* [14, 20]. Performa SVM yang moderat (60,67%) dan Naive Bayes (47,00%) memperkuat argumen bahwa untuk data *behavioral* yang *noisy*, kesederhanaan dan transparansi model seringkali lebih diutamakan daripada kompleksitas matematis [18].

Jika dibandingkan dengan penelitian terdahulu di konteks serupa, temuan ini menunjukkan peningkatan. Akurasi 68,67% dari C4.5 ini lebih unggul dibandingkan penelitian Alfariyis et al. [2] yang mencapai 62,5% menggunakan metode yang sama di SMK. Perbedaan ini kemungkinan dapat dijelaskan oleh penggunaan fitur lingkungan pembelajaran yang lebih komprehensif (8 fitur) dan ukuran dataset yang lebih besar (300 sampel) dalam penelitian ini. Selain itu, pendekatan klasifikasi *supervised* ini terbukti lebih aplikatif untuk prediksi data baru dibandingkan penelitian Palupi et al. [3] yang menggunakan *clustering* (akurasi 55-60%) untuk segmentasi gaya belajar.

Secara praktis, temuan ini memiliki implikasi signifikan untuk pembelajaran vokasional [19]. Model C4.5 yang transparan dapat diintegrasikan ke dalam *Learning Management System* (LMS) sebagai sistem rekomendasi otomatis untuk materi pembelajaran, atau digunakan guru untuk melakukan diferensiasi pembelajaran adaptif di kelas [20]. Karena interpretabilitasnya yang tinggi, pihak sekolah dapat fokus mengalokasikan sumber daya pada faktor-faktor yang terbukti paling berpengaruh (seperti memastikan konsistensi kehadiran guru dan menyeimbangkan jam tatap muka) untuk meningkatkan efisiensi pembelajaran [21] [22].

Meskipun demikian, penelitian ini memiliki beberapa keterbatasan utama yang perlu diakui dan dapat menjadi arahan untuk penelitian selanjutnya. Keterbatasan tersebut mencakup distribusi data yang tidak seimbang (Visual 43%, Auditori 21%), yang dapat menyebabkan bias model terhadap kelas mayoritas dan tercermin pada *recall* yang rendah untuk kelas Auditori. Selain itu, ruang lingkup variabel masih terbatas pada 8 fitur lingkungan (belum mencakup nilai akademik atau data perilaku siswa), serta data yang hanya berasal dari satu SMK di Riau, yang membatasi generalisasi hasil secara geografis [23], [24].

5. KESIMPULAN DAN SARAN

Penelitian ini mengevaluasi performa tiga algoritma *machine learning* untuk klasifikasi gaya belajar siswa SMK, dengan hasil menunjukkan C4.5 Decision Tree sebagai model paling optimal. C4.5 mencapai akurasi 68,67% dan F1-score 65,98%, mengungguli SVM (akurasi 60,67%) dan Naive Bayes (akurasi 47,00%). Keunggulan utama C4.5 terletak pada keseimbangan antara akurasi dan interpretabilitasnya yang tinggi, di mana struktur pohon keputusannya mudah dipahami oleh praktisi pendidikan. Temuan ini membuktikan bahwa pendekatan *data-driven* berbasis karakteristik lingkungan pembelajaran dapat menjadi alternatif valid untuk memetakan preferensi belajar siswa tanpa instrumen psikologis.

Untuk penelitian selanjutnya, disarankan fokus pada peningkatan akurasi dan robustitas model. Hal ini dapat dicapai melalui integrasi fitur tambahan yang lebih kaya (seperti data akademik, log aktivitas LMS, dan absensi), serta eksplorasi algoritma *advanced* seperti *ensemble methods* (Random Forest atau XGBoost). Penerapan teknik *balancing data* seperti SMOTE juga direkomendasikan untuk mengatasi distribusi kelas yang tidak seimbang. Secara aplikatif, model yang telah disempurnakan perlu divalidasi melalui pengujian eksternal di SMK lain dan studi longitudinal, sebelum diimplementasikan secara praktis ke dalam *dashboard* interaktif atau diintegrasikan dengan LMS. Implementasi seluruh rekomendasi ini diharapkan dapat meningkatkan akurasi prediksi ke *threshold* 75-80% untuk mendukung sistem pembelajaran personalisasi yang lebih komprehensif.

6. DAFTAR PUSTAKA

- [1] I. Maryani, E. Hasanah, and M. P. I. Suyatno, "Pendukung Pembelajaran Berdiferensiasi pada Kurikulum Merdeka," UAD Repository, 2023. [Online]. Available: [https://eprints.uad.ac.id/58950/2/Asesmen Diagnostik_Ika Maryani, dkk.pdf](https://eprints.uad.ac.id/58950/2/Asesmen_Diagnostik_Ika_Maryani_dkk.pdf)
- [2] R. F. Alfariysi, S. Salmon, M. Kom, and S. Wahyuni, "Komparasi Algoritma Naïve Bayes dan C4.5 dalam Klasifikasi Tipe Belajar Siswa SMKN 1 Samarinda," *J. Wicida*, 2022, [Online]. Available: <https://repository.wicida.ac.id/4080/2/1543136-S1-Jurnal.pdf>
- [3] S. Palupi, R. Andrea, and S. Qomariah, "Cluster Analysis for Learning Style of Vocational High School Student Using K-Means and Fuzzy C-Means (FCM)," *J. Komun. dan Opini Publik*, vol. 6, no. 2, pp. 44–50, 2023, [Online]. Available: <https://jkd.komdigi.go.id/index.php/jpkop/article/view/1243/722>
- [4] K. Mubarak, "Klasifikasi Pemahaman Konsep pada Mata Pelajaran Produktif dengan Metode Support Vector Machine," UIN Malang, 2024. [Online]. Available: <http://etheses.uin-malang.ac.id/66822/2/200605210009.pdf>
- [5] W. N. Indah, "Implementasi Machine Learning Menggunakan RapidMiner untuk Mengidentifikasi Faktor-Faktor yang Mempengaruhi Kemampuan Penalaran Matematis," UIN Jakarta, 2023. [Online]. Available: [https://repository.uinjkt.ac.id/dspace/bitstream/123456789/82930/1/SKRIPSI FULL_Wirda Nur Indah_11200170000015.pdf](https://repository.uinjkt.ac.id/dspace/bitstream/123456789/82930/1/SKRIPSI_FULL_Wirda_Nur_Indah_11200170000015.pdf)
- [6] S. Winardi, F. M. Sinaga, F. R. Fa, and C. Sintiya, "Penggunaan MobileNet untuk Intelligent Character Recognition," *J. TIMES*, vol. 6, no. 2, pp. 115–121, 2023, [Online]. Available: <https://ejournal.stmik-time.ac.id/index.php/jurnalTIMES/article/download/707/277>
- [7] A. Muhlis, "Deep Learning dalam Pendidikan dan Artificial Intelligence," IAIN Madura Repository, 2025. [Online]. Available: [http://repository.iainmadura.ac.id/1282/1/Lay-out Deep Learning YPAD.pdf](http://repository.iainmadura.ac.id/1282/1/Lay-out_Deep_Learning_YPAD.pdf)
- [8] P. Utami, M. Jundi, R. Rahmadden, and L. Sinaga, "Property Price Prediction Using the Random Forest Regression Algorithm," *SITEKIN J. Sains, Teknol. dan Ind.*, vol. 22, no. 2, 2025, doi: 10.24014/sitekin.v22i2.35804.
- [9] S. A. A. Kharis and A. H. A. Zili, "Learning Analytics dan Educational Data Mining pada Data Pendidikan," *J. Ris. Pembelajaran Mat. Sekol.*, vol. 6, no. 2, 2022, [Online]. Available:

- <https://journal.unj.ac.id/unj/index.php/jrpms/article/download/26575/12383>
- [10] T. Handayani, "Perbandingan Algoritma Machine Learning untuk Prediksi Preferensi Belajar," *J. Inform. dan Sist. Inf.*, vol. 8, no. 1, 2022.
 - [11] M. Arsyad, "Deep Learning Approaches in Education: A Literature Review on Their Role in Addressing Future Challenge," *TOFEDU*, vol. 4, no. 2, 2025, [Online]. Available: https://www.academia.edu/download/123022216/TOFEDU_4_2_2025.pdf
 - [12] M. D. Salman, R. Rahmadden, N. R. Pratama, A. A. Setiawan, F. Zaliani, and I. B. Huda, "Perbandingan Kinerja Algoritma Clustering K-Means dan K-Medoids dalam Pengelompokan Sekolah di Provinsi Riau," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 5, no. 3, pp. 797–806, 2025, doi: 10.24014/sitekin.v22i2.35804.
 - [13] C. Molnar *et al.*, "Relating the partial dependence plot and permutation feature importance to the data generating process," in *World Conference on Explainable Artificial Intelligence*, Springer, 2023, pp. 456–479.
 - [14] J. F. Hair, W. C. Black, B. J. Babin, and R. E. Anderson, *Multivariate Data Analysis*. Harlow, United Kingdom: Pearson Education, 2020.
 - [15] S. A. Guno, "Penggunaan Machine Learning untuk Identifikasi Kemampuan Berpikir Kreatif Matematis," UIN Jakarta, 2024. [Online]. Available: https://repository.uinjkt.ac.id/dspace/bitstream/123456789/78230/1/Skripsi_Sadewo_Aji_Guno_111170170000015.pdf
 - [16] F. A. Baharudin, "Penggunaan Machine Learning untuk Mengidentifikasi Faktor yang Mempengaruhi Kemampuan Koneksi Matematis," UIN Jakarta, 2023. [Online]. Available: https://repository.uinjkt.ac.id/dspace/bitstream/123456789/67071/1/Watermark_Revisi_Skripsi_Fikri_Ananta_Baharudin_111801700000025.pdf
 - [17] Y. Tarumasely, W. Labobar, J. Sipahelut, and M. Halamury, *Perubahan Paradigma Pendidikan Melalui Teknologi AI*. Google Books, 2024. [Online]. Available: <https://books.google.com/books?id=RWcQEQAQBAJ>
 - [18] E. R. Putri and A. T. Sukma, "Penerapan Decision Tree C4.5 dalam Menentukan Gaya Belajar Mahasiswa," 2021.
 - [19] D. Prasetyo and S. Wahyuni, "Pengaruh Faktor Eksternal dan Gaya Belajar terhadap Prestasi Siswa SMK," *J. Teknol. Pendidik.*, vol. 5, no. 2, 2021.
 - [20] A. H. M. Sastraatmadja, A. Suhara, and N. Mayasari, *Artificial Intelligence dalam Pendidikan: Metode, Implementasi, dan Evaluasi*. Google Books, 2025. [Online]. Available: <https://books.google.com/books?id=dLlaEQAAQBAJ>
 - [21] S. Izzuddin, "Analisis Faktor yang Mempengaruhi Kemampuan Pemecahan Masalah Matematis," UIN Jakarta, 2023. [Online]. Available: https://repository.uinjkt.ac.id/dspace/bitstream/123456789/72810/1/Skripsi_Surya_Izzuddin_111801700000065_PERPUS_F.pdf
 - [22] M. Munsarif, M. Sam'an, and S. Raharjo, "Pelatihan ChatGBT kepada Guru di Majelis Pendidikan Muhammadiyah," *J. Surya Masy.*, vol. 3, no. 1, 2024, [Online]. Available: <https://jurnal.unimus.ac.id/index.php/JSM/article/download/13924/8059>
 - [23] R. Febriyanti and S. F. Putri, "S.id: Platform Pintar Berbasis AI untuk Meningkatkan Kemampuan Siswa SMK," 2024, *Universitas Negeri Malang*. [Online]. Available: <http://conference.um.ac.id/index.php/nsafe/article/download/9030/3133>
 - [24] J. Jamil and S. Pulukadang, "Application of Deep Learning Method in Learning," *Formosa J. Sustain. Res.*, vol. 4, no. 1, pp. 91–102, 2025, [Online]. Available: <http://mryformosapublisher.org/index.php/fjsr/article/download/308/674>