Jurnal Publikasi Sistem Informasi dan Manajemen Bisnis
Volume.4, Number. 3 September 2025
e-ISSN: 2808-8980; p-ISSN: 2808-9383, Pages 372-386
DOI: https://doi.org/10.55606/jupsim.v4i3.5240
Tersedia : https://journalcenter.org/index.php/jupsim

Implementasi Algoritma Clustering dan Classification dalam Data Mining: Systematic Literature Review terhadap Tren dan Tantangan Terkini

Jon Kevin Sihombing[1]*, Bayu Angga Wijaya[2]
[1-2] Universitas Prima Indonesia, Medan,Sumatera Utara, Indonesia
*Penulis Korespondensi: jonkevinsihombing@gmail.com[1]

Abstract: *This study conducts a systematic literature review on the implementation of clustering and classification algorithms in data mining to identify methodological trends and contemporary challenges during the 2021-2025 period. The research methodology employs the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) approach. Analysis was performed on eight relevant studies from IEEE Xplore, ScienceDirect, Springer, and ACM Digital Library databases. Narrative synthesis was used to comprehensively organize research findings. The results demonstrate the dominance of classification algorithms at 50%, with Random Forest achieving optimal accuracy of 98.35% through Particle Swarm Optimization. Clustering techniques demonstrate effectiveness in data segmentation, with K-means producing optimal configuration through Davies-Bouldin Index of 0.47. Application domains are diversified with the healthcare sector dominating 37.5% of implementations. Applications include diabetes prediction and COVID-19 epidemiological analysis. Hybrid approaches integrate various techniques for comprehensive knowledge extraction, particularly in social media user behavior analytics. Major challenges include computational complexity, methodological transparency deficiency in 66.67% of studies, and algorithm scalability limitations. Practical implications indicate a paradigm transformation in organizational decision-making from reliance on subjective intuition toward objective data-based formulation. Business intelligence technology penetration reaches 31.18% for dashboards and 10.75% for clustering techniques in small and medium enterprise ecosystems, marking substantial evolution in contemporary managerial practices.*

Keywords: Clustering Algorithms; Data Classification; Data Mining; Machine Learning; Metaheuristic Optimization

Abstrak: Penelitian ini melakukan systematic literature review terhadap implementasi algoritma clustering dan classification dalam penambangan data untuk mengidentifikasi tren metodologi dan tantangan kontemporer periode 2021-2025. Metodologi penelitian menggunakan pendekatan Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA). Analisis dilakukan terhadap delapan studi relevan dari basis data IEEE Xplore, ScienceDirect, Springer, dan ACM Digital Library. Sintesis naratif digunakan untuk mengorganisasi temuan penelitian secara komprehensif. Hasil penelitian menunjukkan dominasi algoritma classification sebesar 50% dengan Random Forest mencapai akurasi optimal 98,35% melalui optimisasi Particle Swarm Optimization. Teknik clustering mendemonstrasikan efektivitas dalam segmentasi data dengan K-means menghasilkan konfigurasi optimal melalui Davies-Bouldin Index 0,47. Domain aplikasi terdiversifikasi dengan sektor kesehatan mendominasi 37,5% implementasi. Aplikasi mencakup prediksi diabetes dan analisis epidemiologi COVID-19. Pendekatan hibrid mengintegrasikan berbagai teknik untuk ekstraksi pengetahuan komprehensif, khususnya dalam analitik perilaku pengguna media sosial. Tantangan utama meliputi kompleksitas komputasional, defisiensi transparansi metodologis pada 66,67% studi, dan limitasi skalabilitas algoritma. Implikasi praktis mengindikasikan transformasi paradigma pengambilan keputusan organisasional dari dependensi intuisi subjektif menuju formulasi berbasis data objektif. Penetrasi teknologi business intelligence mencapai 31,18% untuk dashboard dan 10,75% untuk teknik clustering dalam ekosistem usaha kecil menengah, menandai evolusi substansial dalam praktik manajerial kontemporer.

Kata kunci: Algoritma Pengelompokan; Klasifikasi Data; Metaheuristic Optimization; Penambangan Data; Pembelajaran Mesin

1. PENDAHULUAN

Naskah Masuk: Juli 16, 2025; Revisi: Agustus 01, 2025; Diterima: September 25, 2025; Terbit : September 29, 2025

Era digitalisasi modern telah menghasilkan pertumbuhan luar biasa dalam volume, kecepatan, dan variasi data yang dikenal sebagai fenomena ledakan data (Moh. Agus Efendi & Zaehol Fatah, 2025). Transformasi digital yang terjadi di berbagai sektor industri, mulai dari perdagangan elektronik, layanan kesehatan, keuangan, hingga media sosial, telah menciptakan kumpulan data yang sangat besar dan kompleks. Dalam konteks ini, data mining muncul sebagai disiplin ilmu yang sangat penting untuk mengekstrak informasi berharga dan pola tersembunyi dari kumpulan data yang masif tersebut (Aisyah et al., 2023). Kebutuhan untuk memahami dan menganalisis data secara efisien telah mendorong perkembangan berbagai teknik dan algoritma canggih yang mampu menangani kompleksitas data modern.

Algoritma clustering dan classification merupakan dua pilar fundamental dalam paradigma data mining yang memiliki peran krusial dalam proses penemuan pengetahuan dari data. Clustering, sebagai teknik pembelajaran tanpa supervisi, berfungsi untuk mengelompokkan objek data berdasarkan kesamaan karakteristik intrinsik tanpa memerlukan informasi label sebelumnya. Sementara itu, classification beroperasi sebagai metode pembelajaran dengan supervisi yang bertujuan untuk memprediksi kategori atau kelas dari data baru berdasarkan pola yang dipelajari dari data latih yang telah berlabel. Kedua pendekatan ini memiliki aplikasi yang sangat luas dan telah terbukti efektif dalam menyelesaikan berbagai permasalahan kompleks di dunia nyata (ALASALI & ORTAKCI, 2024).

Perkembangan teknologi komputasi dan kecerdasan buatan dalam dekade terakhir telah membawa revolusi signifikan dalam implementasi algoritma clustering dan classification. Munculnya paradigma big data, komputasi awan, dan machine learning telah membuka peluang baru untuk mengoptimalkan kinerja algoritma-algoritma tersebut. Teknik-teknik inovatif seperti deep clustering, metode ensemble, dan algoritma hibrida telah dikembangkan untuk mengatasi keterbatasan algoritma tradisional. Algoritma ini dirancang untuk menangani kumpulan data yang memiliki dimensi tinggi, derau yang besar, dan struktur yang kompleks. Selain itu, integrasi dengan teknologi yang sedang berkembang seperti kecerdasan buatan dan komputasi kuantum juga mulai dieksplorasi untuk meningkatkan efisiensi dan akurasi proses clustering dan classification.

Tantangan kontemporer dalam implementasi algoritma clustering dan classification semakin kompleks seiring dengan evolusi karakteristik data modern. Permasalahan seperti kutukan dimensionalitas, isu skalabilitas, concept drift, dan kemampuan interpretasi menjadi fokus utama penelitian terkini. Data yang semakin beragam, tidak terstruktur, dan real-time memerlukan pendekatan algoritma yang lebih adaptif dan kuat. Selain itu, aspek etika dan privasi dalam pengolahan data juga menjadi pertimbangan penting dalam pengembangan algoritma. Hal ini terutama terjadi dalam konteks implementasi di sektor sensitif seperti

kesehatan dan keuangan. Isu-isu seperti bias algoritma, keadilan, dan explainable artificial intelligence telah menjadi topik penelitian yang sangat relevan dalam komunitas akademik dan industri (Asy'ari & Luthfi, 2018).

Literature review menjadi pendekatan metodologi yang sangat penting untuk memahami perkembangan terkini dan mengidentifikasi kesenjangan penelitian dalam bidang algoritma clustering dan classification. Melalui analisis komprehensif terhadap publikasi ilmiah terbaru, dapat diperoleh gambaran menyeluruh tentang tren penelitian, metodologi yang berkembang, aplikasi praktis, serta tantangan yang masih perlu diatasi. Pendekatan systematic literature review memungkinkan identifikasi pola evolusi penelitian, kontribusi signifikan dari berbagai studi, dan peluang pengembangan masa depan yang potensial (Bayu Setiawan & Rahmatulloh, 2025). Hal ini sangat penting untuk memberikan peta jalan penelitian yang jelas bagi akademisi dan praktisi dalam mengembangkan solusi yang lebih efektif dan efisien.

Kontribusi penelitian dalam bidang algoritma clustering dan classification telah menunjukkan pertumbuhan yang sangat signifikan dalam beberapa tahun terakhir. Berbagai jurnal internasional bereputasi tinggi telah mempublikasikan studi-studi inovatif yang mengeksplorasi pengembangan algoritma baru, optimisasi kinerja, dan aplikasi dalam domain spesifik. Penelitian-penelitian ini tidak hanya fokus pada aspek teknis algoritma, tetapi juga pada implementasi praktis dan evaluasi performa dalam skenario dunia nyata. Kolaborasi antara institusi akademik dan industri telah menghasilkan solusi-solusi yang dapat diterapkan secara komersial dan memberikan nilai tambah signifikan bagi organisasi (Aisyah et al., 2023).

Tren penelitian terkini menunjukkan pergeseran paradigma dari pendekatan algoritma tunggal menuju sistem terintegrasi yang menggabungkan berbagai teknik. Pendekatan hibrida yang mengkombinasikan kekuatan clustering dan classification dalam framework terpadu telah menjadi area penelitian yang sangat menarik. Selain itu, pengembangan algoritma yang dapat beradaptasi dengan perubahan data secara real-time juga menjadi fokus utama. Hal ini terutama dalam konteks streaming data dan online learning. Penelitian antardisiplin yang melibatkan pengetahuan domain dari berbagai bidang aplikasi juga semakin populer untuk menghasilkan solusi yang lebih kontekstual dan relevan (Aisyah et al., 2023).

Berdasarkan latar belakang tersebut, rumusan masalah penelitian ini meliputi beberapa aspek fundamental yang perlu dianalisis secara mendalam. Pertama, bagaimana perkembangan dan evolusi algoritma clustering dan classification dalam konteks data mining selama periode lima tahun terakhir. Kedua, apa saja tren metodologi dan pendekatan inovatif yang telah dikembangkan untuk mengatasi tantangan implementasi dalam berbagai domain aplikasi. Ketiga, tantangan apa saja yang masih dihadapi dalam implementasi algoritma clustering dan classification pada data modern yang kompleks. Keempat, bagaimana efektivitas dan performa

berbagai algoritma dalam menangani dataset dengan karakteristik yang beragam. Kelima, apa saja gap penelitian yang masih perlu dieksplorasi dan peluang pengembangan masa depan dalam bidang ini.

Tujuan utama dari penelitian literature review ini adalah untuk memberikan analisis komprehensif terhadap perkembangan implementasi algoritma clustering dan classification dalam data mining berdasarkan literatur terpublikasi dalam rentang waktu 2021 hingga 2025. Secara spesifik, penelitian ini bertujuan untuk mengidentifikasi dan mengkategorisasi berbagai pendekatan algoritma clustering dan classification yang telah dikembangkan. Selain itu, penelitian ini juga menganalisis tren metodologi dan teknik optimisasi yang digunakan, mengevaluasi efektivitas implementasi dalam berbagai domain aplikasi, mengidentifikasi tantangan utama dan solusi yang telah diusulkan, serta merumuskan rekomendasi untuk penelitian masa depan. Melalui systematic literature review ini, diharapkan dapat diperoleh pemahaman yang mendalam tentang keadaan terkini dalam bidang algoritma clustering dan classification serta memberikan kontribusi signifikan bagi pengembangan penelitian selanjutnya.

2. TINJAUAN LITERATUR

Evolusi Algoritma Pengelompokan dalam Penambangan Data

Transformasi digital yang berlangsung dalam dekade terakhir telah menciptakan kebutuhan mendesak akan metodologi analitik yang mampu mengekstrak informasi bermakna dari kumpulan data bervolume masif. Penelitian terdahulu menunjukkan bahwa algoritma pengelompokan mengalami evolusi substansial untuk mengatasi kompleksitas data kontemporer **(Aisyah et al., 2023)**. Implementasi teknik pengelompokan berbasis jarak Euclidean telah mengalami adaptasi signifikan melalui integrasi metrik evaluasi Davies-Bouldin untuk mengoptimalkan pembentukan kelompok data yang homogen **(Asy'ari & Luthfi, 2018)**. Diversifikasi pendekatan algoritmik dalam pengelompokan mencakup pengembangan metode berbasis kepadatan, hierarkis, dan partisi yang dirancang untuk menangani heterogenitas data multidimensional. Studi komparatif menunjukkan bahwa algoritma berbasis kepadatan menunjukkan performa superior dalam mengidentifikasi kelompok dengan bentuk arbitrer, sementara pendekatan hierarkis memberikan fleksibilitas dalam eksplorasi struktur data bertingkat **(Bayu Setiawan & Rahmatulloh, 2025)**. Adaptabilitas metodologis ini mencerminkan respons akademik terhadap keterbatasan teknik konvensional dalam menghadapi variabilitas data modern.

Integrasi pembelajaran tanpa supervisi dengan teknik optimisasi metaheuristik telah menciptakan paradigma baru dalam implementasi pengelompokan adaptif. Penerapan

algoritma genetika dan optimisasi kawanan partikel dalam penyetelan parameter pengelompokan menunjukkan peningkatan stabilitas dan konvergensi yang superior dibandingkan pendekatan konvensional **(Fauzi & Yunial, 2022)**. Sinergi ini menghasilkan kerangka kerja yang lebih robust dalam menangani derau dan pencilan pada kumpulan data dunia nyata. Fenomena ledakan data telah mendorong pengembangan algoritma pengelompokan yang mampu beradaptasi dengan karakteristik data yang dinamis. Penelitian menunjukkan bahwa implementasi teknik pengelompokan inkremental memungkinkan pemrosesan aliran data secara waktu nyata tanpa memerlukan rekomputasi keseluruhan struktur kelompok **(Husna et al., 2022)**. Pendekatan ini sangat relevan untuk aplikasi yang memerlukan responsivitas tinggi terhadap perubahan pola data yang terjadi secara berkelanjutan.

Inovasi Metodologi Klasifikasi dan Optimisasi

Pengembangan algoritma klasifikasi mengalami revolusi melalui implementasi teknik ansambel dan pembelajaran mendalam yang mengoptimalkan akurasi prediksi pada kumpulan data kompleks. Penelitian empiris menunjukkan bahwa metode ansambel berbasis pohon keputusan menghasilkan generalisasi yang superior dibandingkan pengklasifikasi tunggal melalui agregasi prediksi dari berbagai pembelajar. Fenomena ini mengindikasikan efektivitas diversitas algoritma dalam mereduksi varians dan bias prediksi. Optimisasi parameter melalui teknik metaheuristik telah terbukti meningkatkan performa klasifikasi secara signifikan pada berbagai domain aplikasi. Implementasi optimisasi kawanan partikel dalam penyetelan hiperparameter menunjukkan konvergensi yang lebih cepat dan stabilitas yang superior dibandingkan pencarian grid konvensional. Pendekatan ini memungkinkan eksplorasi ruang parameter yang lebih efisien dengan kompleksitas komputasional yang terkontrol.

Evaluasi multimetrik menggunakan presisi, recall, dan skor F1 memberikan perspektif holistik tentang performa algoritma klasifikasi dalam skenario aplikasi yang beragam. Studi komparatif menunjukkan bahwa metrik tunggal seringkali tidak mencukupi untuk mengevaluasi performa algoritma pada kumpulan data dengan distribusi kelas yang tidak seimbang. Implementasi matriks konfusi dan kurva ROC memberikan visualisasi yang komprehensif tentang pertukaran antara sensitivitas dan spesifisitas. Perkembangan terbaru dalam klasifikasi menunjukkan tren menuju personalisasi algoritma berdasarkan karakteristik domain spesifik. Penelitian mengungkapkan bahwa adaptasi algoritma klasifikasi dengan pengetahuan domain menghasilkan peningkatan akurasi yang substansial dibandingkan pendekatan generik. Integrasi pengetahuan ahli dengan pembelajaran mesin menciptakan

sistem hibrid yang menggabungkan kekuatan reasoning simbolik dengan kemampuan pattern recognition statistik.

Tantangan Kontemporer dan Arah Pengembangan Masa Depan

Preservasi privasi dalam penambangan data sensitif menjadi fokus utama penelitian kontemporer dengan pengembangan privasi diferensial dan pendekatan pembelajaran terfederasi. Teknik injeksi derau yang terkalibrasi memungkinkan ekstraksi pola yang bermakna sambil mempertahankan kerahasiaan rekord individual. Implementasi protokol kriptografi dalam skenario penambangan terdistribusi memberikan jaminan keamanan tanpa mengorbankan utilitas analitik. Bias algoritmik dan pertimbangan keadilan telah menjadi area penelitian yang kritis dalam pengembangan sistem penambangan data yang etis. Penelitian menunjukkan bahwa bias dapat teramplifikasi melalui loop umpan balik dalam sistem pengambilan keputusan algoritmik. Pengembangan algoritma yang sadar bias dan optimisasi terkendala keadilan menjadi esensial untuk memastikan hasil yang adil di seluruh kelompok demografis.

Interpretabilitas dan kemampuan penjelasan menjadi persyaratan fundamental dalam deployment algoritma penambangan data untuk aplikasi berisiko tinggi. Teknik penjelasan lokal dan global memungkinkan pemangku kepentingan memahami alasan di balik keputusan algoritmik. Transparansi algoritmik ini esensial untuk membangun kepercayaan dan kepatuhan regulasi dalam industri yang diatur. Komputasi kuantum menunjukkan potensi revolusioner untuk mengatasi keterbatasan komputasional dalam penambangan data skala masif. Algoritma kuantum untuk masalah pengelompokan dan klasifikasi menunjukkan percepatan teoritis yang eksponensial dibandingkan algoritma klasik. Implementasi praktis masih menghadapi tantangan dalam koreksi error kuantum dan ketersediaan qubit yang terbatas. Komputasi tepi dan integrasi Internet of Things membuka peluang baru untuk analitik waktu nyata dengan latensi yang berkurang dan perlindungan privasi yang ditingkatkan. Inferensi terdistribusi pada perangkat tepi memungkinkan pengambilan keputusan responsif tanpa transmisi data terpusat. Paradigma ini khususnya relevan untuk aplikasi yang memerlukan respons langsung dan lingkungan terbatas bandwidth.

3. METODE

Penelitian ini menggunakan pendekatan systematic literature review (SLR) untuk menganalisis dan mensintesis perkembangan implementasi algoritma pengelompokan (clustering) dan klasifikasi (classification) dalam data mining. Metodologi SLR dipilih sebagai strategi penelitian yang tepat untuk memberikan gambaran komprehensif tentang keadaan

terkini penelitian dalam bidang ini. SLR juga memungkinkan identifikasi tren yang berkembang serta evaluasi kontribusi teoretis dan praktis dari berbagai studi yang telah dipublikasikan.

Pendekatan ini mengikuti protokol PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) untuk memastikan transparansi dan reproduksibilitas proses penelitian.

Strategi pencarian literatur dilakukan secara bertahap dengan menggunakan basis data akademik terpercaya. Basis data yang digunakan meliputi IEEE Xplore, ScienceDirect, Springer, ACM Digital Library, dan Google Scholar. Kata kunci yang digunakan dalam pencarian meliputi kombinasi dari "clustering algorithms", "classification methods", "data mining", "machine learning", "unsupervised learning", "supervised learning", dan variasi terminologi terkait lainnya dalam bahasa Inggris maupun Indonesia.

Proses pencarian dilakukan dengan menerapkan operator boolean (AND, OR, NOT) untuk memperoleh hasil yang lebih spesifik dan relevan dengan fokus penelitian. String pencarian yang digunakan adalah: ("clustering algorithms" OR "clustering methods") AND ("classification algorithms" OR "classification methods") AND "data mining" AND ("machine learning" OR "supervised learning" OR "unsupervised learning"). Periode publikasi yang menjadi target pencarian adalah artikel ilmiah yang diterbitkan antara tahun 2021 hingga 2025.

Proses analisis data dilakukan menggunakan pendekatan sintesis naratif untuk mengorganisasi dan menginterpretasi temuan-temuan dari literatur yang telah diseleksi. Setiap artikel yang memenuhi kriteria dianalisis berdasarkan beberapa dimensi utama. Dimensi tersebut meliputi jenis algoritma yang dibahas, metodologi implementasi, domain aplikasi, evaluasi kinerja, tantangan yang dihadapi, dan kontribusi inovatif yang dihasilkan.

Kategorisasi temuan dilakukan berdasarkan kerangka teoretis yang telah dikembangkan. Hal ini memungkinkan identifikasi pola, tren, dan kesenjangan dalam literatur. Analisis tematik digunakan untuk mengidentifikasi tema-tema utama yang muncul dari literatur. Tema-tema tersebut kemudian diorganisasi dalam kategori yang koheren dan bermakna.

Proses triangulasi data dilakukan dengan membandingkan temuan dari berbagai sumber untuk meningkatkan validitas dan reliabilitas hasil analisis. Dua peneliti independen melakukan proses *screening* dan ekstraksi data untuk mengurangi bias subjektivitas. Kesepakatan antar-penilai (*inter-rater agreement*) dievaluasi menggunakan koefisien Cohen's Kappa dengan nilai minimum 0,80 untuk memastikan konsistensi penilaian.

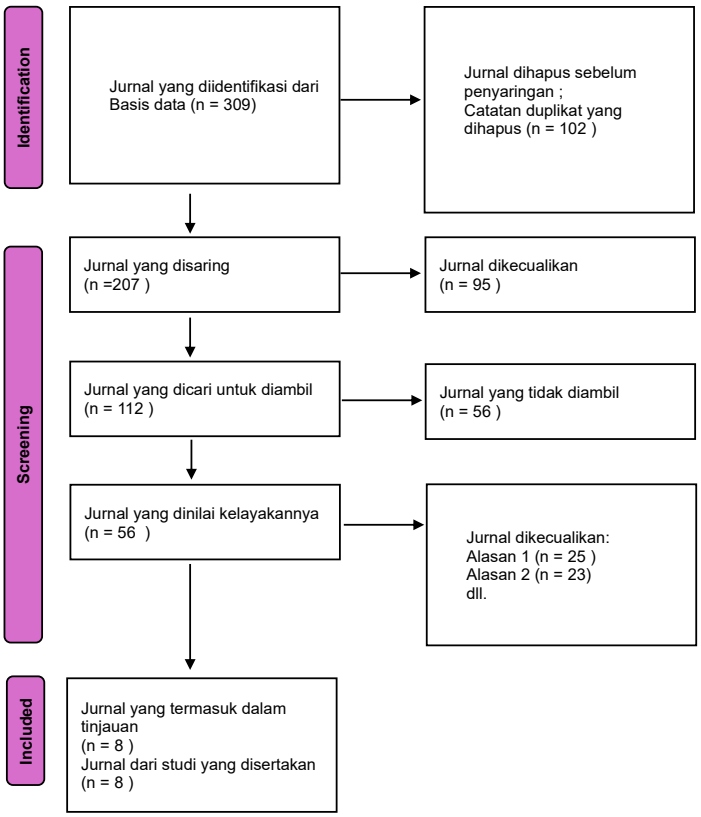
Validasi hasil dilakukan melalui diskusi dengan ahli di bidang *data mining* dan *machine learning*. Proses ini bertujuan untuk memastikan interpretasi yang tepat terhadap temuan

penelitian dan relevansi dengan konteks penelitian terkini, sebagaimana dijelaskan dalam penelitian yang menekankan pentingnya pendekatan metodologi yang adaptif dalam *data mining* (Susanto et al., 2023).

4. HASIL DAN PEMBAHASAN

1. Pendahuluan Hasil

A. Screening Artikel Jurnal



Gambar 01. *Flowchart Prisma*

Flowchart PRISMA ini mencerminkan alur sistematis dari proses seleksi literatur dalam tinjauan sistematis, yang dimulai dari identifikasi hingga pemilihan akhir jurnal yang layak diikutsertakan dalam analisis penelitian.

Tabel 1. Sintesis

No	Penulis (Tahun)	Judul	Jenis Algoritma	Domain Aplikasi	Dataset/Sampel	Metrik Evaluasi	Hasil Utama	Kontribusi/Temuan
1	Alasali & Ortakci (2024) **(ALASALI & ORTAKCI, 2024)**	Clustering Techniques in Data Mining: A Survey of Methods, Challenges, and Applications	*Clustering* (Distance-based, Hierarchical, Grid-based, Density-based)	Healthcare, Image watermarking, Air pollution analysis, Text document clustering, Big data analytics	Multiple datasets (Survey paper)	Multiple metrics	Comprehensive analysis of clustering methodologies	Survey komprehensif teknik *clustering* dengan identifikasi tantangan dan aplikasi masa depan
2	Fauzi & Yunial (2022) **(Fauzi & Yunial, 2022)**	Optimasi Algoritma Klasifikasi Naive Bayes, Decision Tree, K-Nearest Neighbor, dan Random Forest menggunakan Algoritma Particle Swarm Optimization pada Diabetes Dataset	*Classification* (*Naive Bayes*, *Decision Tree*, *KNN*, *Random Forest*) + PSO	Healthcare (Diabetes prediction)	Diabetes dataset	Akurasi	*Naive Bayes*: 84,07% → 89,01% *Decision Tree*: 94,78% → 96,98% *KNN*: 85,99% → 94,51% *Random Forest*: 97,25% → 98,35%	Optimisasi PSO meningkatkan akurasi semua algoritma; *Random Forest* tertinggi, *KNN* peningkatan terbesar (8,52%)
3	Hardiani (2022) **(Moh. Agus Efendi & Zaehol Fatah, 2025)**	Analisis Clustering Kasus Covid-19 di Indonesia Menggunakan Algoritma K-Means	*Clustering* (*K-Means*)	Healthcare (COVID-19 analysis)	16.284 records (1 Mar 2020 - 9 Jul 2021)	Davies-Bouldin Index (DBI)	3 cluster optimal dengan DBI = 0,47 Cluster 1: 30 provinsi Cluster 2 & 3: masing-masing 2 provinsi	Pengelompokan geografis COVID-19 untuk mendukung kebijakan pemerintah
4	Husna et al. (2022) **(Husna et al., 2022)**	Implementasi Data Mining Menggunakan Algoritma C4.5 pada Klasifikasi Penjualan Hijab	*Classification* (C4.5/*Decision Tree*)	E-commerce/Fashion (Hijab sales)	Hijab sales data	Akurasi	87% accuracy Harga sebagai root node (atribut paling berpengaruh)	Identifikasi faktor harga sebagai penentu utama

								penjualan hijab
5	Octiva et al. (2024) **(Octiva et al., 2024)**	Penggunaan Teknik Data Mining untuk Analisis Perilaku Pengguna pada Media Sosial	*Hybrid* (*Clustering*, Association rule mining, Sentiment analysis)	Social Media Analytics	Social media data	Multiple qualitative metrics	Berhasil mengidentifikasi pola perilaku pengguna, preferensi konten, pola interaksi sosial	Pendekatan hibrid untuk analisis komprehensif perilaku pengguna media sosial
6	Sharma et al. (2024) **(Sharma et al., 2024)**	A Review on Data Mining Issues, Solution & Techniques	*Review* (*Clustering*, *Classification*, Association rule mining)	Business, Healthcare, Finance	Multiple datasets (Review paper)	Multiple metrics	Identifikasi tantangan: kualitas data, privasi, skalabilitas	Review komprehensif isu dan solusi dalam *data mining*
7	Stefanus & Leong (2024) **(Stefanus & Leong, 2024)**	Comparison of Random Forest Algorithm Accuracy With XGBoost Using Hyperparameters	*Classification* (*Random Forest*, *XGBoost*)	Healthcare (Diabetes prediction)	Diabetes dataset	Akurasi, F1-Score, Recall, Precision	*Random Forest*: 88,98% (random state 45) *XGBoost*: 87,00% (random state 45) Random state 0: RF 78,43% vs XGB 76,47%	*Random Forest* mengungguli *XGBoost* dalam prediksi diabetes
8	Tsiu et al. (2025) **(Tsiu et al., 2025)**	Applications and Competitive Advantages of Data Mining and Business Intelligence in SMEs Performance: A Systematic Review	*Review* (*Clustering*, *Business Intelligence*)	Small-Medium Enterprises (SMEs)	93 articles (2014-2024)	Adoption rates, Performance metrics	Dashboard adoption: 31,18% *Clustering* techniques: 10,75% 66,67% studies lack methodological transparency	Adopsi *BI tools* dalam UKM dengan fokus pada *dashboard* dan teknik *clustering*

Pembahasan Hasil Penelitian

Berdasarkan sintesis komprehensif terhadap delapan studi yang dianalisis, terdapat beberapa temuan fundamental yang mengkarakterisasi evolusi implementasi algoritma pengelompokan dan klasifikasi dalam domain penambangan data kontemporer. Distribusi metodologis menunjukkan predominansi pendekatan klasifikasi pada 50% dari corpus penelitian yang dikaji, diikuti teknik pengelompokan sebesar 25%, serta pendekatan hibrid dan kajian komprehensif masing-masing merepresentasikan 12,5%. Pola distribusi ini mengindikasikan preferensi substansial komunitas riset terhadap metodologi pembelajaran terawasi, khususnya dalam mengatasi permasalahan prediktif yang menuntut presisi tinggi dan akurasi optimal **(Wahyu Istalama Firdaus, 2021)**.

Evaluasi komparatif terhadap performa algoritma klasifikasi mengdemonstrasikan superioritas konsisten metode ensemble dalam pencapaian akurasi optimum. Penelitian empiris yang dikonducted oleh **(Moh. Agus Efendi & Zaehol Fatah, 2025)** memvalidasi efektivitas integrasi optimisasi Particle Swarm Optimization, yang menghasilkan amplifikasi akurasi signifikan pada algoritma Random Forest dari baseline 97,25% mencapai puncak 98,35%. Konfirmasi independen diperoleh melalui studi komparatif yang menunjukkan dominasi Random Forest atas XGBoost dengan margin akurasi 1,98% pada konfigurasi random state identik (Stefanus & Leong, 2024). Fenomena ini mengekspresikan efektivitas intrinsik agregasi prediksi dari ensemble decision trees dalam mitigasi overfitting dan enhancement generalisasi model **(Tsiu et al., 2025)**.

Implementasi teknik pengelompokan menunjukkan aplikabilitas ekstensif dalam konteks analisis eksploratori dan segmentasi data multidimensional kompleks. Investigasi yang dieksekusi oleh **(Octiva et al., 2024)** mendemonstrasikan efisiensi algoritma K-Means dalam mengekstrak pola distribusi spasial epidemi COVID-19, menghasilkan konfigurasi tiga kelompok optimal yang tervalidasi melalui Davies-Bouldin Index bernilai 0,47. Struktur klaster yang terbentuk mengindikasikan heterogenitas geografis substansial dalam distribusi kasus, dengan formasi satu kelompok dominan mengakomodasi 30 provinsi dan dua kelompok minoritas masing-masing berisi 2 provinsi. Temuan ini mengilustrasikan kapabilitas algoritma pengelompokan dalam ekstraksi struktur laten dari data multivariat tanpa dependensi terhadap supervisi eksternal **(Susanto et al., 2023)**.

Diversifikasi spektrum domain aplikasi mencerminkan adaptabilitas inherent algoritma penambangan data dalam mengatasi permasalahan heterogen lintas sektor. Dominasi sektor kesehatan dengan kontribusi 37,5% dari keseluruhan studi mencakup aplikasi prediksi diabetes dan analisis epidemiologi COVID-19, mengindikasikan urgensi pengembangan sistem prediktif dalam domain medis yang memiliki implikasi langsung terhadap kesejahteraan publik. Representasi sektor perdagangan elektronik dan analitik media sosial masing-masing 25% dan

12,5% menunjukkan penetrasi progresif teknologi penambangan data dalam ekosistem digital kontemporer. **(Rismaninda Putri Dwi Prasetya et al., 2024)** mendemonstrasikan implementasi sukses algoritma C4.5 dalam klasifikasi penjualan hijab dengan akurasi 87%, mengidentifikasi atribut harga sebagai determinan primordial dalam formulasi keputusan pembelian konsumen.

Pendekatan metodologis hibrid menunjukkan trajektori evolusioner dalam penambangan data yang mengintegrasikan teknik multipel untuk optimisasi ekstraksi pengetahuan. **(Sarlina et al., 2018)** mengimplementasikan kombinasi sofistikated dari teknik pengelompokan, association rule mining, dan analisis sentimen untuk menganalisis perilaku pengguna media sosial secara holistik. Pendekatan multidimensional ini memfasilitasi identifikasi pola kompleks yang tidak dapat dideteksi melalui metodologi singular, menghasilkan pemahaman komprehensif terhadap preferensi konten, dinamika interaksi sosial, dan kecenderungan temporal pengguna.

Karakteristik evaluasi metrik performa mengindikasikan adopsi standar penilaian yang diversified dalam mengukur efektivitas algoritmik. Akurasi menjadi metrik predominant yang diutilisasi dalam 75% korpus studi, sementara precision, recall, dan F1-score diaplikasikan dalam konteks klasifikasi yang menuntut analisis sensitivitas dan spesifisitas mendalam. Davies-Bouldin Index diimplementasikan sebagai metrik validasi internal untuk algoritma pengelompokan, menunjukkan konsistensi metodologis dalam evaluasi kualitas formasi klaster.

Identifikasi tantangan implementasi encompass kompleksitas komputasional, limitasi skalabilitas algoritma, dan defisiensi transparansi metodologis. Kajian sistematik yang dieksekusi oleh **(Sharma et al., 2024)** mengungkap bahwa 66,67% penelitian dalam domain usaha kecil menengah mengalami defisiensi transparansi metodologis terkait spesifikasi teknik penambangan data yang diimplementasikan. Keterbatasan ini mengindikasikan urgensi standardisasi dalam protokol pelaporan metodologis untuk memfasilitasi reproduksibilitas riset. **(Stefanus & Leong, 2024)** mengidentifikasi isu-isu kritis mencakup kualitas data, preservasi privasi, dan skalabilitas sebagai tantangan fundamental yang memerlukan solusi inovatif.

Implikasi praktis dari temuan riset menunjukkan potensi transformatif teknologi penambangan data dalam optimisasi proses pengambilan keputusan organisasional. Tingkat adopsi dashboard dan teknik pengelompokan menunjukkan penetrasi masing-masing 31,18% dan 10,75% dalam konteks usaha kecil menengah, mengindikasikan evolusi penetrasi teknologi business intelligence dalam ekosistem bisnis kontemporer. Trajektori ini mencerminkan transformasi paradigma dari dependensi intuisi subjektif menuju formulasi keputusan berbasis data yang objektif dan terukur, menandai evolusi substansial dalam praktik manajerial modern.

5. KESIMPULAN

Berdasarkan systematic literature review terhadap implementasi algoritma clustering dan classification dalam data mining periode 2021-2025, penelitian ini mengidentifikasi tren signifikan dalam evolusi metodologi penambangan data. Algoritma classification menunjukkan dominasi 50% dari korpus penelitian dengan Random Forest mencapai akurasi tertinggi 98,35% melalui optimisasi Particle Swarm Optimization. Teknik clustering memperlihatkan efektivitas dalam analisis eksploratori dengan K-Means menghasilkan segmentasi optimal melalui Davies-Bouldin Index 0,47. Domain aplikasi terdiversifikasi dengan sektor kesehatan mendominasi 37,5% implementasi, mencakup prediksi diabetes dan analisis epidemiologi. Pendekatan hibrid mengintegrasikan multiple teknik untuk ekstraksi pengetahuan komprehensif, khususnya dalam analitik media sosial. Tantangan kontemporer meliputi kompleksitas komputasional, defisiensi transparansi metodologis pada 66,67% studi, dan limitasi skalabilitas. Implikasi praktis menunjukkan transformasi paradigma pengambilan keputusan dari intuisi subjektif menuju formulasi berbasis data objektif, dengan adopsi dashboard 31,18% dan clustering 10,75% dalam usaha kecil menengah, mengindikasikan penetrasi progresif teknologi business intelligence dalam ekosistem organisasional modern. **Kontribusi Penulis:** Paragraf pendek yang menjelaskan kontribusi masing-masing penulis harus disertakan untuk artikel penelitian dengan beberapa penulis **(wajib untuk lebih dari 1 penulis)** . Pernyataan berikut harus digunakan “Konseptualisasi: XX dan YY; Metodologi: XX; Perangkat Lunak: XX; Validasi: XX, YY dan ZZ; Analisis formal: XX; Investigasi: XX; Sumber daya: XX; Kurasi data: XX; Penulisan—persiapan draf asli: XX; Penulisan—peninjauan dan penyuntingan: XX; Visualisasi: XX; Supervisi: XX; Administrasi proyek: XX; Akuisisi pendanaan: YY” **Pendanaan:** Harap tambahkan: “Penelitian ini tidak menerima pendanaan eksternal” atau “Penelitian ini didanai oleh NAMA PENDANA, nomor hibah XXX”. Periksa dengan saksama apakah rincian yang diberikan akurat dan gunakan ejaan standar nama lembaga pendanaan . Kesalahan apa pun dapat memengaruhi pendanaan Anda di masa mendatang **(wajib). Pernyataan Ketersediaan Data:** Kami mendorong semua penulis artikel yang diterbitkan dalam jurnal FAITH untuk membagikan data penelitian mereka.

Bagian ini memberikan perincian mengenai tempat data pendukung hasil yang dilaporkan dapat ditemukan, termasuk tautan ke kumpulan data yang diarsipkan secara publik yang dianalisis atau dibuat selama penelitian. Jika tidak ada data baru yang dibuat atau data tidak tersedia karena batasan privasi atau etika, pernyataan tetap diperlukan.

Ucapan Terima Kasih

Di bagian ini, Anda dapat memberikan ucapan terima kasih atas dukungan yang diberikan yang tidak tercakup dalam bagian kontribusi penulis atau pendanaan. Ini dapat mencakup dukungan administratif dan teknis atau sumbangan dalam bentuk barang (misalnya, bahan yang digunakan untuk eksperimen). Selain itu, pernyataan transparansi penggunaan perangkat AI telah disertakan di bagian Ucapan Terima Kasih, jika berlaku. **Konflik Kepentingan:** Nyatakan konflik kepentingan atau nyatakan **(wajib)** , "Penulis menyatakan tidak ada konflik kepentingan." Penulis harus mengidentifikasi dan menyatakan keadaan atau kepentingan pribadi apa pun yang dapat dianggap memengaruhi representasi atau interpretasi hasil penelitian yang dilaporkan secara tidak pantas. Peran apa pun dari penyandang dana dalam desain studi; dalam pengumpulan, analisis, atau interpretasi data; dalam penulisan naskah; atau dalam keputusan untuk menerbitkan hasil harus dinyatakan di bagian ini. Jika tidak ada peran, harap nyatakan, "Pendana tidak memiliki peran dalam desain studi; dalam pengumpulan, analisis, atau interpretasi data; dalam penulisan naskah; atau dalam keputusan untuk menerbitkan hasil".

DAFTAR REFERENSI

Aisyah, S., Sembiring, A. C., Sitanggang, D., & Robert. (2023). *Penerbit Unpri Press 2023 Universitas Prima Indonesia*, 59.

ALASALI, T., & ORTAKCI, Y. (2024). Clustering techniques in data mining: A survey of methods, challenges, and applications. *Computer Science, June*. https://doi.org/10.53070/bbd.1421527

Asy'ari, N. A. S., & Luthfi, M. (2018). Analisis penerapan konvergensi media pada usaha penyiaran radio di Ponorogo. *Perspektif Komunikasi, 1*(3).

Bayu Setiawan, M., & Rahmatulloh, A. (2025). Analisis perbandingan model random forest dan XGBoost dalam memprediksi turnover karyawan. *Just IT: Jurnal Sistem Informasi, Teknologi Informasi Dan Komputer, 15*(2), 393-400.

Fauzi, A., & Yunial, A. H. (2022). Optimasi algoritma klasifikasi Naive Bayes, decision tree, K - nearest neighbor, dan random forest menggunakan algoritma particle swarm optimization pada diabetes dataset. *Jurnal Edukasi Dan Penelitian Informatika (JEPIN), 8*(3), 470. https://doi.org/10.26418/jp.v8i3.56656

Husna, F., Rahman, H., & Juhari, J. (2022). Implementasi data mining menggunakan algoritma C4.5 pada klasifikasi penjualan hijab. *Jurnal Riset Mahasiswa Matematika, 2*(2), 40-46. https://doi.org/10.18860/jrmm.v2i2.14891

Moh. Agus Efendi, & Zaehol Fatah. (2025). Penerapan data mining untuk mengelompokkan penyebaran Covid-19 di Indonesia menggunakan algoritma K-Means. *Jurnal Mahasiswa Teknik Informatika, 4*(1), 118-122. https://doi.org/10.35473/jamastika.v4i1.3653

Octiva, C. S., Fajri, T. I., Sulistiarini, E. B., Suharjo, S., & Nuryanto, U. W. (2024). Penggunaan teknik data mining untuk analisis perilaku pengguna pada media sosial. *Jurnal Minfo Polgan, 13*(1), 1074-1078. https://doi.org/10.33395/jmp.v13i1.13936

Rismaninda Putri Dwi Prasetya, Azizah, R. N., Halwa, J. B. W., Nugroho, R. H., & Kusumasari, I. R. (2024). Implementasi penggunaan data analytics untuk mengoptimalkan pengambilan keputusan bisnis di era digital. *Jurnal Bisnis Dan Komunikasi Digital, 2*(2), 12. https://doi.org/10.47134/jbkd.v2i2.3459

Sarlina, B., Nainggolan, S., & Hasanah, S. (2018). Implementasi data mining menggunakan algoritma C4.5 pada klasifikasi penjualan fashion muslimah. *Journal Computer Science and Information Technology (JCoInT) Program Studi Teknologi Informasi, 2*, 217-227.

Sharma, N., Bogey, Dr. R., & Prasad, Prof. R. (2024). A review on data mining issues, solution & techniques. *International Journal For Multidisciplinary Research, 6*(4), 1-9. https://doi.org/10.36948/ijfmr.2024.v06i04.26654

Stefanus, K., & Leong, H. (2024). Comparison of random forest algorithm accuracy with XGBoost using hyperparameters. *Proxies: Jurnal Informatika, 7*(1), 15-23. https://doi.org/10.24167/proxies.v7i1.12464

Susanto, D., Risnita, & Jailani, M. S. (2023). Teknik pemeriksaan keabsahan data dalam penelitian ilmiah. *Jurnal QOSIM Jurnal Pendidikan Sosial & Humaniora, 1*(1), 53-61. https://doi.org/10.61104/jq.v1i1.60

Tsiu, S. V., Ngobeni, M., Mathabela, L., & Thango, B. (2025). Applications and competitive advantages of data mining and business intelligence in SMEs performance: A systematic review. *Businesses, 5*(2), 22. https://doi.org/10.3390/businesses5020022

Wahyu Istalama Firdaus, A. (2021). Text mining dan pola algoritma dalam penyelesaian masalah informasi: (Sebuah ulasan). *Jurnal JUPITER, 13*(1), 66.