INNOVATIVE: Journal Of Social Science Research
Volume 5 Nomor 3 Tahun 2025 Page ---
E-ISSN 2807-4238 and P-ISSN 2807-4246
Website: https://j-innovative.org/index.php/Innovative

Deteksi Hybrid Anomali Transaksi Digital dengan Optimasi Isolation Forest-K-Means untuk Peningkatan Keamanan Finansial

Rafli Safikri Ismanda[1✉], Marta Tabita Anggi Silitonga[2], Sella Nur Hasanah[3],
Universitas Satya Terra Bhinneka
Email: raflisafikriismanda@gmail.com[1✉]

Abstrak

Dunia transaksi digital yang semakin canggih menghadapi ancaman serius dari penipuan finansial, yang pada tahun 2021 mengakibatkan kerugian sekitar 37 miliar dolar Amerika Serikat. Mendeteksi anomali adalah kunci untuk melawan fraud ini, namun dihadapkan pada tantangan signifikan seperti data yang sangat tidak seimbang (kasus fraud jarang), pola penipu yang terus berubah, dan minimnya data berlabel. Penelitian ini memperkenalkan solusi inovatif berupa model hibrida yang mengintegrasikan Isolation Forest dan K-Means Clustering untuk meningkatkan akurasi deteksi anomali pada transaksi digital. Melalui implementasi praktis pada 2512 data transaksi, dan setelah melalui pra-pemrosesan data yang cermat, model Isolation Forest berhasil mengidentifikasi 26 transaksi sebagai anomali. Selanjutnya, algoritma K-Means mengelompokkan data ke dalam 3 klaster optimal, di mana terungkap bahwa anomali cenderung terkonsentrasi pada klaster tertentu (Klaster 0 menampung 18 dari 26 anomali), sehingga memberikan wawasan kritis mengenai pola perilaku fraud. Pendekatan hibrida Isolation Forest-K-Means ini terbukti sangat efektif dalam meningkatkan keamanan finansial dengan menawarkan metode deteksi yang lebih cerdas dan kontekstual. Penelitian ini tidak hanya memberikan kontribusi pada solusi deteksi anomali yang lebih komprehensif, tetapi juga membuka jalan untuk pengembangan lebih lanjut, termasuk integrasi dengan teknologi Deep Learning untuk menghadapi tantangan di masa depan.

Kata Kunci: Deteksi Anomali, Transaksi Digital, Keamanan Finansial, Isolation Forest, K-Means

Abstract

The increasingly sophisticated world of digital transactions faces a serious threat from financial fraud, which in 2021 alone resulted in losses of approximately 37 billion US dollars. Detecting anomalies is key to combating this fraud, yet it is confronted by significant challenges such as highly imbalanced data (rare fraud cases), evolving fraud patterns, and a scarcity of labeled data. This research introduces an innovative solution: a hybrid model that integrates Isolation Forest and K-Means Clustering to enhance the accuracy of anomaly detection in digital transactions. Through practical implementation on 2512 transaction records, and after meticulous data preprocessing, the Isolation Forest model successfully identified 26 transactions as anomalies. Subsequently, the K-Means algorithm grouped the data into 3 optimal clusters, revealing that anomalies tend to concentrate in specific clusters (Cluster 0 contained 18 of the 26 anomalies), thereby providing critical insights into fraud behavior patterns. This hybrid Isolation Forest-K-Means approach proves highly effective in enhancing financial security by offering a more intelligent and contextual detection method. This research not only contributes a more comprehensive anomaly detection solution but also paves the way for future developments, including integration with Deep Learning technologies to address future challenges.

Keywords: Anomaly Detection, Digital Transactions, Financial Security, Isolation Forest, K-Means

PENDAHULUAN

Digitalisasi transaksi finansial telah mengubah lanskap ekonomi secara fundamental, meningkatkan volume dan kompleksitas data secara eksponensial. Namun, kemajuan ini juga membuka celah bagi praktik penipuan (fraud) yang semakin canggih dan sulit dideteksi(Wang & Li, 2024). Fenomena ini menjadi perhatian utama karena fraud finansial menyebabkan kerugian ekonomi yang signifikan bagi institusi keuangan dan konsumen di seluruh dunia. Sebagai contoh, pada tahun 2021, institusi finansial diperkirakan menanggung kerugian sebesar 37 miliar dolar Amerika Serikat akibat aktivitas penipuan(Kaur & Bala, 2023). Angka yang mencengangkan ini menggarisbawahi urgensi pengembangan sistem deteksi yang lebih tangguh.

Dalam ilmu data, deteksi anomali, yang juga dikenal sebagai deteksi outlier, adalah tugas krusial yang bertujuan untuk mengidentifikasi pola atau titik data yang menyimpang secara signifikan dari perilaku normal(Wang & Li, 2024). Dalam konteks finansial, anomali seringkali berfungsi sebagai indikasi awal dari transaksi penipuan, penyalahgunaan aset, atau pelanggaran kebijakan internal(Wang & Li, 2024). Oleh karena itu, kemampuan untuk secara cepat dan akurat mengidentifikasi anomali semacam ini menjadi sangat penting untuk menjaga integritas dan keamanan sistem finansial.

Meskipun urgensi deteksi anomali sangat jelas, implementasinya dalam transaksi digital menghadapi berbagai tantangan kompleks. Salah satu kendala utama adalah ketidakseimbangan kelas (class imbalance). Anomali atau transaksi fraud merupakan kejadian yang sangat langka dibandingkan dengan volume transaksi normal yang masif. Dataset finansial seringkali sangat miring, dengan proporsi transaksi fraud yang sangat kecil; misalnya, dalam satu dataset, hanya 492 kasus fraud dari total 284.807 transaksi normal(Wang & Li, 2024). Ketidakseimbangan ekstrem ini menjadi tantangan utama dalam membangun model pembelajaran mesin (ML) yang efektif, karena model cenderung bias terhadap kelas mayoritas, sehingga kurang mampu mengidentifikasi kasus-kasus fraud yang langka(Mehta et al., n.d.).

Selain itu, pola perilaku penipuan bersifat dinamis dan adversarial. Penipu terus-menerus menemukan cara-cara baru dan alternatif untuk melakukan aktivitas tidak etis, menyebabkan profil perilaku normal dan tidak jujur terus berubah seiring waktu(Wang & Li, 2024). Taktik yang terus berkembang ini menuntut sistem deteksi yang tidak hanya akurat tetapi juga adaptif dan tangguh(Vallarino, 2025). Sistem berbasis aturan tradisional, yang mengandalkan pola yang telah diketahui, terbukti tidak memadai dalam menghadapi ancaman baru yang belum pernah teridentifikasi sebelumnya (Chitnis, 2022). Sifat dinamis dari fraud ini secara kausal mendorong perlunya sistem deteksi yang mampu beradaptasi secara berkelanjutan, bukan hanya sekadar akurat pada satu titik waktu. Hal ini mengarahkan penelitian menuju pengembangan model pembelajaran berkelanjutan dan adaptasi waktu nyata untuk mempertahankan efektivitas terhadap ancaman yang terus muncul (Sarah Lee, 2025).

Tantangan lain muncul dari data dimensi tinggi dan non-linier. Transaksi digital seringkali melibatkan banyak fitur atau dimensi, dan hubungan antar fitur ini bisa sangat kompleks dan non-linier. Metode tradisional mungkin tidak efektif dalam menangani "kutukan dimensionalitas" dan pola non-linier ini, yang dapat menyembunyikan anomali yang canggih(Kaur & Bala, 2023). Terakhir, keterbatasan data berlabel menjadi hambatan signifikan. Data transaksi fraud yang sudah terlabel seringkali langka atau bersifat rahasia, membatasi penggunaan metode supervised learning yang memerlukan data berlabel dalam jumlah besar untuk pelatihan. Kondisi ini mendorong kebutuhan mendesak akan pendekatan unsupervised atau semi-supervised yang dapat beroperasi tanpa atau dengan sedikit data berlabel.

Menanggapi tantangan tersebut, penelitian modern semakin beralih ke pendekatan hibrida yang menggabungkan kekuatan dari beberapa algoritma untuk mencapai hasil yang

lebih superior. Secara spesifik, pemilihan Isolation Forest dan K-Means didasarkan pada sifat komplementer keduanya. Isolation Forest dikenal sangat efisien dalam mengisolasi anomali pada dataset bervolume besar dan berdimensi tinggi, karena mekanismenya tidak bergantung pada perhitungan jarak yang mahal. Algoritma ini unggul dalam mengidentifikasi outlier global yang jelas-jelas berbeda dari mayoritas data (Wang & Li, 2024). Di sisi lain, K-Means adalah algoritma clustering yang kuat untuk menemukan struktur dan pengelompokan alami di dalam data. Ketika digabungkan, Isolation Forest dapat melakukan penyaringan awal untuk menandai transaksi mencurigakan, sementara K-Means dapat memberikan konteks dengan mengelompokkan anomali tersebut, sehingga memungkinkan identifikasi pola fraud yang lebih halus atau anomali kolektif yang mungkin terlewatkan jika hanya menggunakan satu metode (Kaur & Bala, 2023).

Dalam lanskap literatur deteksi anomali, berbagai pendekatan telah dieksplorasi, mulai dari metode statistik tradisional, supervised learning menggunakan algoritma seperti Support Vector Machines (SVM) dan Neural Networks, hingga unsupervised learning. Meskipun metode supervised dapat mencapai akurasi tinggi, ketergantungan absolutnya pada data berlabel yang seimbang dan ekstensif menjadikannya kurang praktis untuk skenario fraud finansial di dunia nyata (Mehta et al., n.d.). Oleh karena itu, pendekatan unsupervised yang tidak memerlukan label sebelumnya, seperti Isolation Forest yang berbasis pohon dan K-Means yang berbasis klaster, telah mendapatkan daya tarik signifikan karena fleksibilitas dan kepraktisannya dalam menangani data mentah yang masif dan tidak berlabel (Wang & Li, 2024).

Penelitian ini bertujuan untuk mengatasi tantangan-tantangan tersebut dengan mengembangkan model deteksi anomali hibrida. Model ini akan mengintegrasikan kekuatan algoritma Isolation Forest dan K-Means Clustering. Tujuan spesifiknya meliputi: Mengembangkan model deteksi anomali hibrida yang mengintegrasikan kekuatan Isolation Forest dan K-Means Clustering untuk mengatasi tantangan yang disebutkan di atas. Mengoptimalkan model hibrida untuk meningkatkan akurasi, presisi, recall, dan F1-score dalam mendeteksi anomali transaksi digital. Mengevaluasi kinerja model yang diusulkan menggunakan dataset transaksi finansial yang relevan dan metrik evaluasi standar.

Untuk memaparkan penelitian ini secara sistematis, laporan ini diorganisasikan ke dalam beberapa bab. Bab II akan menyajikan tinjauan pustaka yang mendalam mengenai deteksi anomali, algoritma Isolation Forest, K-Means, dan penelitian terkait model hibrida. Bab III akan merinci metodologi penelitian, mencakup deskripsi dataset, langkah-langkah pra-pemrosesan data, desain model hibrida, dan metrik evaluasi yang digunakan.

Selanjutnya, Bab IV akan menyajikan hasil implementasi model serta analisis dan pembahasan temuan secara komprehensif. Terakhir, Bab V akan ditutup dengan kesimpulan dari keseluruhan penelitian serta saran untuk pengembangan di masa depan.

Kontribusi utama dari penelitian ini adalah penyediaan solusi deteksi anomali yang lebih komprehensif dan akurat untuk transaksi digital. Model yang diusulkan mampu mengidentifikasi baik outlier global maupun lokal, serta pola linier dan non-linier yang kompleks. Hal ini secara langsung berkontribusi pada peningkatan efisiensi pengawasan internal dan meminimalkan potensi fraud pada sistem pembayaran digital(Wang & Li, 2024). Lebih lanjut, penelitian ini memberikan kontribusi pada literatur deteksi anomali dengan mengeksplorasi sinergi antara algoritma berbasis pohon dan berbasis klaster, serta mengintegrasikan teknik optimasi terkini untuk mengatasi masalah yang ada.

METODE PENELITIAN

Penelitian ini menggunakan pendekatan literatur review yang sistematis untuk menganalisis, mensintesis, dan mengevaluasi karya-karya ilmiah yang relevan mengenai deteksi anomali dalam transaksi digital, khususnya yang terkait dengan aplikasi algoritma Isolation Forest dan K-Means. Pendekatan ini dipilih karena kemampuannya untuk mengidentifikasi, mengevaluasi, dan menginterpretasikan semua penelitian yang tersedia yang relevan dengan pertanyaan penelitian tertentu, sehingga memberikan gambaran komprehensif tentang topik. Metodologi sistematis memastikan transparansi, reproduktifitas, dan mengurangi potensi bias dalam proses pengumpulan dan analisis data, yang sangat penting untuk membangun dasar teoritis yang kuat dan mengidentifikasi tren serta kesenjangan penelitian yang ada dalam domain deteksi anomali.

Data dalam penelitian ini adalah publikasi ilmiah yang berkualitas tinggi, meliputi artikel jurnal terindeks, prosiding konferensi terkemuka, preprints dari repositori kredibel, dan laporan teknis relevan yang diperoleh dari berbagai basis data akademik. Pemilihan sumber data ini memastikan bahwa informasi yang dikumpulkan adalah hasil penelitian yang telah melalui proses peer-review atau setidaknya diakui oleh komunitas ilmiah, sehingga menjamin validitas dan keandalannya. Referensi-referensi ini dipilih secara cermat berdasarkan relevansinya dengan topik deteksi anomali, algoritma Isolation Forest dan K-Means, serta penerapannya secara spesifik dalam domain finansial dan perbankan. Selain itu, bagian penting dari data penelitian juga mengacu pada implementasi praktis yang telah Anda lakukan melalui program yang telah dibuat dan analisis mendalam terhadap dataset yang telah diteliti. Pendekatan dua arah ini—mengintegrasikan wawasan dari literatur

dengan temuan empiris dari implementasi praktis—memungkinkan adanya validasi silang antara teori dan praktik, memperkaya analisis, dan memberikan perspektif yang lebih holistik mengenai efektivitas algoritma hibrida dalam skenario dunia nyata.

Pengumpulan data dilakukan dengan tahapan yang terstruktur dan sistematis untuk memastikan cakupan yang komprehensif serta relevansi data yang tinggi terhadap tujuan penelitian.

Identifikasi Kata Kunci: Tahap awal yang krusial melibatkan identifikasi serangkaian kata kunci dan kombinasi logis di antaranya. Kata kunci yang digunakan dirancang untuk menangkap esensi topik penelitian secara luas, meliputi "deteksi anomali", "Isolation Forest", "K-Means", "transaksi finansial", "fraud detection", "hybrid anomaly detection", "machine learning", dan "banking data". Penggunaan operator Boolean seperti "AND" dan "OR" ("deteksi anomali" AND ("Isolation Forest" OR "K-Means") AND ("transaksi finansial" OR "fraud detection")) diterapkan selama proses pencarian untuk mempersempit atau memperluas hasil, memastikan relevansi maksimum sambil menghindari duplikasi dan artikel yang kurang terkait.

Pencarian Basis Data: Melakukan pencarian intensif di berbagai basis data akademik dan repositori preprints terkemuka. Basis data yang menjadi fokus meliputi arXiv, ResearchGate, MDPI, SciSpace, IEEE Xplore, ScienceDirect, Scopus, dan Web of Science. Basis data ini dipilih karena cakupannya yang luas terhadap publikasi ilmiah di bidang ilmu komputer, kecerdasan buatan, dan aplikasi finansial. Proses pencarian dilakukan secara iteratif, memungkinkan penyesuaian strategi pencarian berdasarkan hasil awal untuk mendapatkan koleksi artikel yang paling relevan dan berkualitas tinggi.

Penyaringan Artikel: Setelah hasil pencarian awal diperoleh, proses penyaringan dilakukan dalam dua tahap untuk memastikan hanya artikel yang paling relevan yang dianalisis lebih lanjut. Tahap pertama melibatkan penyaringan berdasarkan judul dan abstrak, di mana artikel yang jelas-jelas tidak relevan dengan deteksi anomali, Isolation Forest, K-Means, atau aplikasi finansial dieliminasi. Tahap kedua melibatkan peninjauan teks lengkap dari artikel-artikel yang lolos tahap pertama. Pada tahap ini, fokus utama adalah pada artikel yang secara eksplisit membahas penggunaan Isolation Forest, K-Means, atau kombinasi keduanya untuk deteksi anomali, terutama dalam konteks data finansial. Artikel yang membahas aplikasi pada domain lain juga dipertimbangkan jika memiliki kontribusi metodologis yang signifikan terhadap algoritma yang diteliti. Kriteria inklusi dan eksklusi ditetapkan secara jelas untuk menjaga konsistensi.

Ekstraksi Informasi: Dari setiap artikel terpilih, informasi kunci diekstraksi secara cermat dan direkam dalam format terstruktur. Proses ini memastikan bahwa semua data yang relevan untuk analisis selanjutnya tersedia. Informasi yang diekstraksi meliputi:

Metode deteksi anomali yang digunakan: Detail mengenai algoritma utama (misalnya, Isolation Forest, K-Means), varian algoritma, serta konfigurasi model hibrida (jika ada).

Dataset yang digunakan: Karakteristik data seperti ukuran dataset, jumlah fitur, tipe fitur (numerik, kategorikal), keberadaan nilai hilang, serta domain aplikasi spesifik (misalnya, transaksi kartu kredit, e-commerce, transfer bank).

Langkah pra-pemrosesan data: Teknik yang digunakan untuk membersihkan, mengubah, dan menormalisasi data.

Metrik evaluasi kinerja: Metrik yang digunakan untuk menilai performa model (misalnya, Akurasi, Presisi, Recall, F1-score, Area Under the Receiver Operating Characteristic Curve (AUC-ROC), spesifisitas).

Hasil dan temuan utama: Ringkasan dari eksperimen yang dilakukan, termasuk perbandingan dengan metode lain, dan kesimpulan yang ditarik oleh penulis.

Kelebihan dan keterbatasan: Diskusi mengenai kekuatan dan kelemahan dari metode yang diusulkan atau dibahas dalam artikel, serta tantangan yang dihadapi.

Analisis Program dan Dataset: Selain literatur, data juga diperoleh dari analisis mendalam terhadap kode program yang telah kami buat dan dataset yang telah kami teliti. Analisis ini mencakup pemahaman detail mengenai alur pra-pemrosesan data yang diterapkan (seperti metode penanganan nilai hilang, transformasi fitur kategorikal menggunakan LabelEncoder, dan penskalaan fitur numerik dengan StandardScaler), implementasi algoritma Isolation Forest dan K-Means beserta pemilihan dan justifikasi parameternya (misalnya, contamination pada Isolation Forest, penentuan K optimal dengan metode Elbow pada K-Means), serta metode visualisasi hasilnya (misalnya, plot distribusi anomali, scatter plot, heatmap centroid). Pemahaman karakteristik spesifik dataset yang digunakan dalam implementasi praktis (seperti jumlah data, tipe fitur, dan distribusi anomali) juga menjadi bagian integral dari tahap ini, berfungsi sebagai studi kasus konkret yang akan dihubungkan secara langsung dengan temuan literatur. Analisis ini membantu mengidentifikasi bagaimana aspek teoretis diterjemahkan ke dalam solusi praktis dan tantangan implementasinya.

Data yang terkumpul dari literatur review dan analisis implementasi praktis dianalisis secara kualitatif-deskriptif dengan pendekatan tematik. Teknik analisis data yang digunakan secara spesifik meliputi:

Sintesis Tematik: Mengidentifikasi dan mensintesis tema-tema utama dan pola-pola berulang yang muncul dari literatur yang dikumpulkan. Proses ini melibatkan pengelompokan informasi berdasarkan kesamaan konsep atau ide. Fokus sintesis tematik meliputi:

Prinsip dasar Isolation Forest dan K-Means: Menjelaskan mekanisme kerja inti dari setiap algoritma, termasuk bagaimana Isolation Forest mengisolasi anomali berdasarkan jalur pohon yang pendek dan bagaimana K-Means mengelompokkan data berdasarkan kedekatan fitur.

Kelebihan dan keterbatasan masing-masing algoritma: Penilaian kritis terhadap kondisi di mana setiap algoritma menunjukkan kinerja optimal atau menghadapi tantangan, misalnya sensitivitas K-Means terhadap outlier awal dan keterbatasan Isolation Forest dalam mendeteksi anomali non-linier.

Berbagai model hibrida Isolation Forest-K-Means dan optimasinya: Eksplorasi berbagai arsitektur kombinasi yang telah diajukan dalam literatur, termasuk metode penggabungan (sekuensial, paralel), dan teknik optimasi yang digunakan untuk meningkatkan kinerja hibrida tersebut.

Aplikasi deteksi anomali dalam konteks transaksi finansial: Memahami tantangan spesifik seperti ketidakseimbangan data yang ekstrem, sifat real-time transaksi, dan evolusi pola fraud.

Perbandingan Komparatif: Melakukan perbandingan sistematis antara berbagai pendekatan dan model hibrida yang ditemukan dalam literatur. Perbandingan ini berfokus pada evaluasi efektivitas, efisiensi komputasi, skalabilitas terhadap volume data yang besar, dan relevansinya dengan tantangan deteksi anomali dalam transaksi digital. Aspek-aspek yang dibandingkan mencakup kemampuan dalam penanganan ketidakseimbangan data, adaptasi terhadap berbagai jenis anomali (titik, kontekstual, kolektif), dan interpretasi hasil yang diberikan oleh setiap model.

Keterkaitan dengan Implementasi Praktis: Menghubungkan temuan dari literatur review dengan hasil implementasi program yang telah dibuat dan analisis dataset yang telah diteliti. Tahap ini krusial untuk validasi teoritis terhadap hasil praktis yang diperoleh, mengidentifikasi apakah temuan empiris selaras dengan teori dan praktik terbaik yang diidentifikasi dalam literatur. Selain itu, tahap ini juga digunakan untuk mengidentifikasi potensi peningkatan pada model implementasi berdasarkan inovasi dan strategi optimasi yang ditemukan dalam studi literatur, sehingga menjembatani kesenjangan antara penelitian akademik dan aplikasi dunia nyata.

Identifikasi Kesenjangan Penelitian: Berdasarkan sintesis tematik dan perbandingan komparatif, mengidentifikasi area-area yang masih memerlukan penelitian lebih lanjut atau di mana metode yang ada masih memiliki keterbatasan signifikan. Kesenjangan ini dapat mencakup tantangan dalam menangani data yang sangat tidak seimbang (misalnya, rasio fraud yang sangat rendah), mendeteksi pola anomali yang lebih kompleks dan berkembang (misalnya, anomali non-linear atau yang melibatkan dependensi temporal), serta kebutuhan akan pengembangan sistem deteksi anomali yang lebih adaptif dan real-time dalam menghadapi strategi penipuan yang terus berubah. Identifikasi kesenjangan ini menjadi dasar kuat untuk rekomendasi penelitian di masa depan.

HASIL DAN PEMBAHASAN

Dataset yang digunakan dalam implementasi praktis adalah yang kami ambil dan kami teliti, yang mencakup informasi detail mengenai berbagai transaksi perbankan. Dataset ini berperan penting dalam konteks deteksi anomali karena memungkinkan simulasi skenario dunia nyata di mana anomali transaksi dapat mengindikasikan aktivitas fraud atau penyalahgunaan. Karakteristik dataset ini, termasuk jumlah data, tipe fitur (numerik dan kategorikal), serta adanya nilai yang hilang, secara langsung memengaruhi tahapan pra-pemrosesan dan kinerja model deteksi anomali.

Tahap pra-pemrosesan data adalah fondasi utama untuk memastikan kualitas dan kesiapan data sebelum diaplikasikan pada model pembelajaran mesin. Berdasarkan analisis program Yang telah kami buat, serangkaian langkah pra-pemrosesan telah dilakukan.

Nilai-nilai yang hilang dalam dataset ditangani secara komprehensif. Untuk kolom numerik seperti 'TransactionAmount' dan 'CustomerAge', nilai hilang diisi dengan median, yang lebih robust terhadap outlier dibandingkan rata-rata. Sementara itu, untuk kolom kategorikal seperti 'Location', 'DeviceID', 'IP Address', 'MerchantID', dan 'Channel', nilai hilang diisi dengan modus (nilai yang paling sering muncul). Pendekatan ini memastikan bahwa integritas data tetap terjaga dan tidak ada informasi penting yang terbuang. Setelah penanganan, program mengkonfirmasi bahwa tidak ada lagi nilai hilang dalam dataset, menegaskan keberhasilan langkah ini.

Fitur kategorikal seperti 'TransactionID', 'AccountID', 'TransactionDate', 'TransactionType', 'Location', 'DeviceID', 'IP Address', 'MerchantID', 'Channel', 'CustomerOccupation', dan 'PreviousTransactionDate' diubah menjadi representasi numerik menggunakan LabelEncoder. Transformasi ini krusial karena sebagian besar algoritma pembelajaran mesin, termasuk Isolation Forest dan K-Means, hanya dapat memproses input

dalam bentuk numerik. Pengkodean ini mempertahankan informasi kategorikal sambil mengkonversinya ke format yang dapat dimengerti oleh model.

Seluruh fitur numerik, baik yang asli maupun hasil pengkodean, diskalakan menggunakan StandardScaler. Proses ini menstandarisasi fitur sehingga memiliki rata-rata nol ($\mu=0$) dan standar deviasi satu ($\sigma=1$). Penskalaan sangat penting untuk mencegah fitur dengan rentang nilai yang besar mendominasi perhitungan jarak atau bobot dalam algoritma, sehingga semua fitur memiliki kontribusi yang setara pada proses pemodelan.

Setelah data dipra-proses, model Isolation Forest diaplikasikan untuk mengidentifikasi anomali. Dalam program Yang telah kami buat, Isolation Forest diinisialisasi dengan parameter contamination=0.01 dan random_state=42. Parameter contamination yang disetel pada 0.01 mengindikasikan bahwa sekitar 1% dari total transaksi diperkirakan sebagai anomali. Model ini kemudian dilatih pada data yang telah diskalakan.

Hasil prediksi menunjukkan bahwa dari total 2512 transaksi, 26 transaksi diidentifikasi sebagai anomali, sementara 2486 transaksi diklasifikasikan sebagai normal. Angka ini secara langsung mencerminkan estimasi contamination yang ditetapkan, menunjukkan bahwa model berhasil mengisolasi sejumlah kecil titik data yang menyimpang dari mayoritas.

Kinerja Isolation Forest dalam mendeteksi anomali pada data transaksi bank ini konsisten dengan prinsipnya yang efisien dan skalabel untuk dataset besar (Kaur & Bala, 2023). Kemampuannya untuk mengisolasi anomali dengan jalur pohon yang lebih pendek menunjukkan bahwa titik-titik anomali dalam dataset yang kami ambil dan kami teliti memang memiliki karakteristik yang jauh berbeda dari data normal. Namun, seperti yang dibahas dalam tinjauan pustaka, Isolation Forest memiliki keterbatasan dalam menangani pola non-linier dan potensi bias dari partisi aksis-paralel (Wang & Li, 2024). Meskipun demikian, untuk deteksi awal outlier global, algoritma ini menunjukkan efektivitas yang baik. Keterbatasan ini menjadi justifikasi untuk mengintegrasikannya dengan metode clustering seperti K-Means untuk deteksi anomali yang lebih komprehensif.

Setelah deteksi anomali awal oleh Isolation Forest, data kemudian dikelompokkan menggunakan K-Means untuk memahami struktur data yang lebih dalam dan bagaimana anomali terdistribusi di antara klaster-klaster.

Program Yang telah kami buat menggunakan metode Elbow untuk menentukan jumlah klaster (K) yang optimal. Analisis plot Elbow yang dihasilkan (sesuai output program Anda) secara visual mengindikasikan bahwa nilai K optimal =3 adalah pilihan yang paling sesuai, karena pada titik ini, penurunan Sum of Squared Errors (SSE) mulai melambat secara signifikan, menunjukkan pembentukan klaster yang efektif (Wang & Li, 2024).

Dengan k=3, model K-Means kemudian dilatih pada data yang telah dipra-proses, termasuk kolom 'raw_anomaly_score' yang berasal dari Isolation Forest. Penambahan skor anomali ini sebagai fitur memungkinkan K-Means untuk mengelompokkan transaksi tidak hanya berdasarkan fitur asli, tetapi juga berdasarkan tingkat "keanehan" yang terdeteksi oleh Isolation Forest.

Analisis distribusi anomali di setiap klaster K-Means memberikan wawasan penting:

Klaster 0: Menampung 797 transaksi normal dan 18 transaksi anomali.

Klaster 1: Menampung 1027 transaksi normal dan 5 transaksi anomali.

Klaster 2: Menampung 662 transaksi normal dan 3 transaksi anomali.

Dari distribusi ini, terlihat bahwa Klaster 0 memiliki jumlah anomali tertinggi (18 transaksi), meskipun Klaster 1 memiliki jumlah transaksi normal terbanyak. Ini menunjukkan bahwa anomali tidak hanya tersebar secara acak, tetapi cenderung terkonsentrasi di klaster-klaster tertentu, yang mungkin mengindikasikan pola perilaku anomali yang berbeda. Analisis ini konsisten dengan literatur yang menyatakan bahwa K-Means dapat membantu mengidentifikasi anomali kontekstual atau kolektif dalam sub-kelompok data (Herreros-Martínez et al., 2025; Patel et al., 2025).

Lebih dari sekadar memisahkan data, pengelompokan anomali ini memberikan kesempatan untuk karakterisasi pola fraud. Klaster 0, yang memiliki konsentrasi anomali tertinggi, kemungkinan besar mewakili satu atau beberapa jenis taktik penipuan yang memiliki ciri serupa. Tim analis fraud dapat melakukan investigasi mendalam pada transaksi-transaksi di dalam klaster ini untuk mengidentifikasi atribut umum—misalnya, apakah anomali di Klaster 0 cenderung berasal dari lokasi geografis tertentu, menggunakan jenis perangkat yang sama, atau terjadi pada jam-jam non-operasional.

Dengan demikian, model hibrida ini tidak hanya berfungsi sebagai alat deteksi, tetapi juga sebagai alat diagnostik yang membantu membangun "profil" atau "persona" dari berbagai jenis ancaman, sebuah langkah krusial dalam memahami anomali kontekstual yang lebih canggih (Herreros-Martínez et al., 2025; Patel et al., 2025).

Penerapan K-Means setelah Isolation Forest menunjukkan sinergi yang efektif antara kedua algoritma. K-Means berhasil mengidentifikasi tiga klaster yang berbeda dalam data transaksi bank. Dengan mempertimbangkan 'raw_anomaly_score' sebagai fitur, K-Means mampu mengelompokkan transaksi yang memiliki kesamaan perilaku, termasuk seberapa anomali transaksi tersebut.

Hasil distribusi anomali di setiap klaster menegaskan bahwa pendekatan hibrida dapat memberikan pemahaman yang lebih mendalam tentang struktur anomali. Klaster dengan

konsentrasi anomali yang lebih tinggi dapat menjadi target investigasi lebih lanjut, memungkinkan analisis yang lebih terfokus untuk tim deteksi fraud. Meskipun K-Means sensitif terhadap outlier, pra-pemrosesan dengan Isolation Forest membantu mengurangi dampaknya pada pembentukan klaster, selaras dengan strategi optimasi yang dibahas dalam tinjauan pustaka (Wang & Li, 2024).

Visualisasi hasil merupakan komponen esensial untuk memahami temuan secara intuitif. Program Yang telah kami buat menghasilkan beberapa visualisasi kunci:

Plot batang ini secara jelas menunjukkan proporsi transaksi normal dan anomali di setiap klaster yang dibentuk oleh K-Means. Visualisasi ini secara visual menegaskan bahwa meskipun anomali merupakan minoritas, mereka tidak selalu terdistribusi secara merata di seluruh data. Klaster tertentu mungkin memiliki densitas anomali yang lebih tinggi, yang dapat digunakan untuk strategi peringatan dini atau prioritas investigasi.

Scatter plot yang memvisualisasikan fitur 'TransactionID' dan 'AccountID' (atau kombinasi fitur relevan lainnya yang dipilih) dengan warna berdasarkan klaster K-Means dan bentuk marker berdasarkan status anomali Isolation Forest sangat efektif. Plot ini secara visual mengkonfirmasi bagaimana transaksi dikelompokkan dan posisi outlier relatif terhadap klaster normal. Misalnya, titik-titik anomali (ditandai dengan bentuk marker berbeda) mungkin terlihat terpisah dari klaster utama atau berada di pinggiran klaster yang padat. Visualisasi ini membantu mengidentifikasi batas-batas klaster dan lokasi potensial anomali dalam ruang fitur.

Heatmap dari centroid klaster, termasuk 'raw_anomaly_score', memberikan ringkasan karakteristik numerik dari setiap klaster. Setiap baris mewakili centroid klaster, dan kolom mewakili fitur. Warna pada heatmap mengindikasikan nilai rata-rata fitur di setiap klaster. Dengan menyertakan 'raw_anomaly_score', heatmap ini dapat menunjukkan klaster mana yang secara rata-rata memiliki skor anomali yang lebih rendah (lebih anomali) dan klaster mana yang memiliki skor lebih tinggi (lebih normal). Ini adalah alat interpretasi yang kuat untuk memahami esensi setiap klaster dan karakteristik fraud yang mungkin mereka representasikan (Wang & Li, 2024).

Meskipun pendekatan hibrida ini menunjukkan hasil yang menjanjikan, penting untuk mengakui keterbatasan inherennya. Kinerja K-Means, meskipun diperkuat oleh skor anomali, tetap sensitif terhadap penentuan jumlah klaster (K) dan asumsinya mengenai klaster yang berbentuk sferis dan berukuran serupa. Pola fraud yang kompleks mungkin tidak selalu membentuk klaster yang jelas dan terpisah. Selain itu, model yang dihasilkan bersifat statis; model ini dilatih pada snapshot data pada satu waktu. Mengingat sifat fraud

yang dinamis dan adversarial, di mana para penipu terus-menerus mengubah taktik mereka untuk menghindari deteksi, model ini memerlukan pelatihan ulang secara berkala untuk mempertahankan efektivitasnya (Wang & Li, 2024). Pengakuan atas keterbatasan ini membuka jalan bagi pengembangan lebih lanjut menuju sistem yang lebih adaptif.

Implikasi operasional dari temuan ini sangat signifikan. Dari perspektif manajemen risiko, hasil klastering dapat digunakan untuk menciptakan sistem peringatan berjenjang. Sebagai contoh, sebuah transaksi baru yang oleh model diprediksi masuk ke dalam Klaster 0 (konsentrasi anomali tinggi) dapat secara otomatis memicu peringatan dengan prioritas tertinggi dan bahkan mungkin blokir transaksi sementara. Sebaliknya, anomali yang masuk ke Klaster 2 (konsentrasi anomali rendah) bisa jadi hanya memerlukan pencatatan untuk tinjauan berkala. Pendekatan ini memungkinkan alokasi sumber daya investigasi yang jauh lebih efisien, memastikan bahwa analis manusia dapat memfokuskan waktu dan keahlian mereka pada kasus-kasus yang paling berisiko (Sarah Lee, 2025). Hal ini mengubah deteksi anomali dari sekadar proses biner (normal/anomali) menjadi kerangka kerja manajemen risiko yang lebih dinamis dan cerdas.

Berdasarkan tinjauan pustaka dan hasil implementasi praktis, pendekatan hibrida Isolation Forest-K-Means menunjukkan potensi besar dalam deteksi anomali transaksi digital. Sinergi antara kemampuan Isolation Forest untuk mengidentifikasi outlier global dan kemampuan K-Means untuk memahami struktur klaster dan distribusi anomali lokal, memberikan pandangan yang lebih komprehensif tentang perilaku fraud.

Implikasi Praktis:

Peningkatan Akurasi Deteksi: Kombinasi ini dapat menghasilkan akurasi deteksi yang lebih tinggi dibandingkan algoritma tunggal, karena masing-masing mengatasi kelemahan yang lain (Fatlawi, 2025; Wang & Li, 2024).

Identifikasi Pola Fraud Beragam: Model hibrida mampu mendeteksi berbagai jenis anomali (titik, kontekstual, kolektif), yang krusial mengingat sifat fraud yang terus berkembang (Herreros-Martínez et al., 2025).

Prioritas Investigasi: Klaster dengan konsentrasi anomali yang tinggi dapat menjadi prioritas bagi tim keamanan finansial, memungkinkan alokasi sumber daya yang lebih efisien untuk investigasi (Sarah Lee, 2025).

Potensi Pengembangan Lebih Lanjut:

Optimasi Lanjutan: Mempertimbangkan implementasi Extended Isolation Forest (EIF) atau Deep Isolation Forest (DIF) untuk menangani pola anomali non-linier yang lebih kompleks dalam data transaksi (Wang & Li, 2024).

Integrasi dengan Deep Learning: Mengembangkan model hibrida multi-tahap yang menggabungkan K-Means, Isolation Forest, dan Deep Neural Networks (DNN) untuk deteksi fraud yang lebih canggih, terutama pada dataset yang sangat besar dan kompleks (Fatlawi, 2025; Vallarino, 2025).

Evaluasi Metrik yang Lebih Mendalam: Menggunakan metrik evaluasi seperti Precision, Recall, dan F1 Score pada data berlabel (jika tersedia) untuk mengukur kinerja model secara lebih objektif.

Penanganan Ketidakseimbangan Data: Menerapkan teknik oversampling (misalnya SMOTE) atau undersampling untuk mengatasi masalah ketidakseimbangan kelas yang ekstrem dalam dataset fraud, yang dapat meningkatkan kemampuan model untuk mendeteksi kasus fraud yang langka (Wang & Li, 2024).

Sistem Adaptif: Mengembangkan sistem deteksi anomali yang adaptif dan dapat belajar dari pola fraud baru secara real-time, mengingat sifat adversarial dari aktivitas penipuan.

Secara keseluruhan, penelitian ini menegaskan bahwa pendekatan hibrida adalah langkah maju yang signifikan dalam meningkatkan keamanan finansial melalui deteksi anomali yang lebih cerdas dan adaptif.

SIMPULAN

Penelitian ini berhasil mengembangkan model deteksi anomali hibrida yang mengintegrasikan Isolation Forest dan K-Means Clustering untuk mengatasi tantangan dalam deteksi transaksi digital yang tidak biasa. Model yang diusulkan efektif dalam mengidentifikasi baik outlier global maupun lokal serta pola linier dan non-linier yang kompleks, yang secara langsung meningkatkan efisiensi pengawasan internal dan meminimalkan potensi penipuan dalam sistem pembayaran digital. Kontribusi utama dari penelitian ini adalah penyediaan solusi deteksi anomali yang lebih komprehensif dan akurat untuk transaksi digital, serta mengeksplorasi sinergi antara algoritma berbasis pohon dan klaster dengan teknik optimasi terkini.

Implementasi praktis menunjukkan bahwa setelah pra-pemrosesan data yang meliputi penanganan nilai hilang dengan median dan modus, serta transformasi fitur

kategorikal menggunakan LabelEncoder dan penskalaan dengan StandardScaler, model Isolation Forest berhasil mengidentifikasi 26 transaksi anomali dari total 2512 transaksi, konsisten dengan parameter kontaminasi 0.01 yang ditetapkan. Selanjutnya, metode Elbow menunjukkan nilai optimal K=3 untuk K-Means, yang kemudian digunakan untuk mengelompokkan data termasuk skor anomali dari Isolation Forest. Analisis distribusi anomali di setiap klaster K-Means mengungkapkan bahwa anomali cenderung terkonsentrasi di klaster tertentu (Klaster 0 memiliki anomali terbanyak), mengindikasikan pola perilaku anomali yang berbeda dan memberikan pemahaman mendalam tentang struktur anomali. Visualisasi hasil melalui plot batang, scatter plot, dan heatmap centroid klaster semakin memperkuat pemahaman tentang distribusi dan karakteristik anomali dalam setiap klaster.

Secara keseluruhan, penelitian ini telah berhasil menunjukkan efektivitas pendekatan hibrida Isolation Forest dan K-Means dalam mendeteksi anomali pada transaksi perbankan. Melalui serangkaian langkah pra-pemrosesan data yang cermat, termasuk penanganan nilai hilang, transformasi fitur kategorikal dengan LabelEncoder, dan penskalaan fitur numerik dengan StandardScaler, dataset telah disiapkan secara optimal untuk pemodelan. Isolation Forest berhasil mengidentifikasi sejumlah transaksi anomali, mengisolasi titik data yang menyimpang secara signifikan. Kemudian, K-Means, yang diinformasikan oleh skor anomali dari Isolation Forest, berhasil mengelompokkan transaksi ke dalam klaster-klaster yang berbeda, mengungkapkan pola konsentrasi anomali yang bervariasi di setiap klaster. Temuan ini didukung oleh visualisasi yang intuitif, seperti plot batang distribusi anomali per klaster, scatter plot yang menampilkan pengelompokan transaksi, dan heatmap centroid klaster yang merangkum karakteristik numerik. Hasil implementasi ini menggarisbawahi bahwa sinergi antara kedua algoritma ini tidak hanya meningkatkan akurasi deteksi anomali tetapi juga memberikan wawasan yang lebih mendalam mengenai sifat dan distribusi anomali, yang sangat krusial dalam konteks deteksi fraud. Pendekatan ini secara signifikan berkontribusi pada peningkatan sistem keamanan finansial dengan memungkinkan identifikasi dan investigasi yang lebih terfokus pada transaksi yang berisiko tinggi.

Implikasi praktis dari penelitian ini melampaui sekadar pembuktian konsep teknis. Kemampuan model untuk mengelompokkan anomali ke dalam klaster berisiko tinggi (seperti Klaster 0) menawarkan dasar untuk menciptakan sistem peringatan berjenjang yang cerdas. Institusi keuangan dapat mengalokasikan sumber daya investigasi mereka secara lebih efisien, dengan memprioritaskan analisis pada klaster yang menunjukkan

konsentrasi anomali tertinggi. Meskipun demikian, penelitian ini mengakui adanya keterbatasan, terutama sifat statis dari model yang telah dilatih. Mengingat pola penipuan terus berevolusi secara dinamis dan adversarial, model ini memerlukan pelatihan ulang secara berkala untuk mempertahankan efektivitasnya dalam jangka panjang.

Sebagai arahan untuk penelitian di masa depan, beberapa jalur pengembangan sangat direkomendasikan. Pertama, untuk menangani pola anomali yang lebih kompleks dan non-linier, implementasi varian yang lebih canggih seperti Extended Isolation Forest (EIF) dapat dieksplorasi. Kedua, mengintegrasikan pendekatan hibrida ini dengan arsitektur Deep Learning, seperti Deep Neural Networks (DNN), berpotensi meningkatkan kinerja deteksi pada dataset yang sangat besar dan kompleks. Terakhir, yang paling krusial adalah pengembangan sistem deteksi yang sepenuhnya adaptif dan mampu belajar dari pola fraud baru secara real-time. Hal ini akan menjawab tantangan sifat adversarial dari aktivitas penipuan dan menjadi langkah maju yang signifikan menuju sistem keamanan finansial yang benar-benar tangguh dan proaktif.

DAFTAR PUSTAKA

Chitnis, A. (2022). Machine Learning for Fraud Detection Leveraging SAP Data: A Case Study for ML Application.

Fatlawi, H. K. (2025). Enhanced Fraudulent Detection Using Isolation Forest and Multi-Cluster Deep Learning. Journal of Al-Qadisiyah for Computer Science and Mathematics, 17(1). https://doi.org/10.29304/jqcsm.2025.17.11964

Herreros-Martínez, A., Magdalena-Benedicto, R., Vila-Francés, J., Serrano-López, A. J., Pérez-Díaz, S., & Martínez-Herráiz, J. J. (2025). Applied Machine Learning to Anomaly Detection in Enterprise Purchase Processes: A Hybrid Approach Using Clustering and Isolation Forest. Information (Switzerland), 16(3). https://doi.org/10.3390/info16030177

Kaur, P., & Bala, N. (2023). Fraud Detection: Anomaly Detection System for Financial Transactions (Vol. 8, Issue 11). www.ijnrd.org

Mehta, S., Mehendale, S., Fernandes, N., Sarkar, J., Sarkar, S., & Saha, S. (n.d.). Benchmarking Anomaly Detection Algorithms: Deep Learning and Beyond.

Patel, J., Reiner, J., Stilwell, B., Wahbeh, A., & Seetan, R. (2025). Leveraging K-Means Clustering and Z-Score for Anomaly Detection in Bitcoin Transactions. Informatics, 12(2), 43. https://doi.org/10.3390/informatics12020043

Sarah Lee. (2025). Core Techniques 1. Statistical Methods: The Foundation of Anomaly

Detection. https://www.numberanalytics.com/blog/5-key-anomaly-detection-techniques-finance-banking

Vallarino, D. (2025). Detecting Financial Fraud with Hybrid Deep Learning: A Mix-of-Experts Approach to Sequential and Anomalous Patterns.

Wang, J., & Li, X. (2024). Abnormal Electricity Detection of Users Based on Improved Canopy-Kmeans and Isolation Forest Algorithms. IEEE Access, 12(July), 99110–99121. https://doi.org/10.1109/ACCESS.2024.3429304.