BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 4, June 2026 | Hal 1696-1702
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i4.1155

Deteksi Penipuan pada Transaksi Keuangan Digital Menggunakan Ensemble Learning: Studi Komparatif Random Forest, Gradient Boosting, dan XGBoost

Nurul Akbar Tanjung[1,*], Sugeng Hary Purnomo[1], Sanwani[2]

[1] Program Studi Magister Informatika (PJJ), Universitas Amikom Yogyakarta, Sleman, Indonesia
[2] Fakultas Teknik Informasi, Universitas Bina Sarana Informatika, Jakarta, Indonesia
Email: [1,*]nurul.akbar@amikom.ac.id, [2]s.hary@students.amikom.ac.id, [3]sanwani.swq@bsi.ac.id
Email Penulis Korespondensi: nurul.akbar@amikom.ac.id

Abstrak-Tingginya frekuensi transaksi keuangan digital di Indonesia saat ini tidak hanya merefleksikan akselerasi ekosistem teknologi finansial (*fintech*), tetapi juga menginduksi tantangan keamanan yang kompleks. Pada tahun 2021, nilai transaksi uang elektronik nasional tercatat mencapai Rp395,5 triliun dan diproyeksikan terus mengalami pertumbuhan eksponensial. Fenomena ini berbanding lurus dengan eskalasi kejahatan finansial, di mana modus kecurangan digital (*fraud*) berkembang secara dinamis sehingga membatasi efektivitas mekanisme deteksi berbasis aturan statis (*rule-based system*). Penelitian ini mengkaji batasan tersebut dan mengusulkan implementasi *machine learning* melalui pendekatan *ensemble learning* dengan mekanisme *weighted soft voting*. Tiga algoritma klasifikasi yang diintegrasikan meliputi *Random Forest*, *Gradient Boosting*, dan *XGBoost*. Eksperimen dilakukan menggunakan dataset simulasi PaySim yang terdiri atas 6,36 juta transaksi *mobile money* sintetis dengan rasio *fraud* ekstrem sebesar 0,129%. Guna mengatasi ketidakseimbangan kelas (*class imbalance*) tersebut, teknik *Synthetic Minority Over-sampling Technique* (SMOTE) diaplikasikan secara eksklusif pada data latih. Hasil pengujian pada data independen menunjukkan bahwa model *ensemble* berhasil mencapai nilai presisi 94,7%, *recall* 91,3%, *F1-score* 93,0%, dan AUC-ROC 0,987, melampaui performansi masing-masing model tunggal. Analisis kontribusi fitur menunjukkan bahwa selisih saldo pengirim dan nominal transaksi merupakan variabel paling informatif dalam mengidentifikasi transaksi fraudulent. Selain itu, evaluasi latensi pada simulasi *streaming* lokal mencatat rata-rata waktu respons sebesar 147,3 ms per transaksi, yang menegaskan kelayakan sistem ini untuk diimplementasikan secara *real-time* di lingkungan perbankan maupun *fintech.*

Kata Kunci: Deteksi Penipuan; Ensemble Learning; Machine Learning; Dataset Tidak Seimbang; Transaksi Digital

Abstract-Digital payment fraud in Indonesia has grown alongside the dramatic expansion of mobile money services, creating a detection problem that conventional rule-based systems are increasingly ill-equipped to handle. This paper examines whether a soft-voting ensemble of Random Forest, Gradient Boosting, and XGBoost can offer a more effective solution. The model was trained on the PaySim synthetic dataset, consisting of 6.36 million mobile money transactions in which fraudulent cases account for just 0.129 percent of all records. SMOTE was used exclusively on the training data to address the extreme class imbalance before model fitting. Five-fold cross-validated Grid Search determined the hyperparameter configuration for each constituent model. On the held-out test set, the ensemble achieved 94.7 percent precision, 91.3 percent recall, 93.0 percent F1-score, and 0.987 AUC-ROC figures that consistently exceeded those of any single algorithm. Examining feature contributions revealed that the sender balance difference and transaction amount carried the most discriminative weight, a finding that aligns with known fraud behavior in mobile payment datasets. A local streaming latency test across 5,000 consecutive transactions produced an average response time of 147.3 ms, with the 99th percentile remaining below the 200 ms operational threshold. Taken together, the results indicate that the ensemble approach is not only statistically superior but also practically deployable within the real-time constraints of digital banking environments.

Keywords: Fraud Detection; Ensemble Learning; Machine Learning; Imbalanced Dataset; Digital Transactions

1. PENDAHULUAN

Ekosistem keuangan digital di Indonesia berkembang pesat beberapa tahun terakhir. Contohnya, tahun 2021 transaksi uang elektronik nasional naik hampir 50%, sampai Rp395,5 triliun (Bank Indonesia, 2021). Ini bagus buat inklusi keuangan, tapi risiko juga ikut naik lebih banyak titik rawan bisa dieksploitasi penjahat. Pencurian data, social engineering, manipulasi dana, hingga phishing adalah ancaman sehari-hari di bank atau fintech [1]. Nasabah menginginkan transaksi yang cepat, akan tetepi belum pasti aman [2]. Ketegangan antara keamanan dan kenyamanan menjadi alasan kenapa riset deteksi penipuan otomatis menjadi sangat penting.

Fraud keuangan bukanlah hal yang baru, namun angka transaksinya mengalami peningkatan yang signifikan [3]. ACFE (2020) menyatakan organisasi rata-rata kehilangan sekitar 5% pendapatan per tahun karena penipuan, dan total global triliunan dolar [4]. Di Indonesia, OJK mencatat ribuan pengaduan setiap tahun mengenai issue kejahatan siber, dan jumlahnya belum turun walau pengguna layanan keuangan digital mengalami kenaikan [5]. Permasalahan ini bukan hanya pada kerugian material saja, akantapi juga terkait deteksi yang masih reaktif dan biasanya kasus baru terlihat setelah transaksi selesai dilakukan, dan kerugian sudah terjadi. Oleh karena itu, implementasi sistem intervensi dini sangat penting untuk memitigasi eskalasi kerugian yang lebih besar[6].

Selama ini, kebanyakan bank atau fintech masih menggunakan sistem rule-based. Secara ringkas, analis bertanggung jawab terkait daftar aturan, sistem otomatis melakukan pengecekan setiap transaksi yang melawan peraturan. Pendekatan ini menawarkan transparansi dan akuntabilitas audit yang tinggi. Kendati demikian, terdapat kelemahan fundamental: pelaku kecurangan senantiasa mengevolusikan modus operandinya lebih cepat daripada pembaruan regulasi oleh analis, sehingga menyebabkan sistem mengalami keterlambatan adaptasi [7]. Ketidakmampuan sistem konvensional dalam mengenali pola baru yang belum pernah terjadi sebelumnya memicu pergeseran paradigma menuju

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 4, June 2026 | Hal 1696-1702
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i4.1155

pemanfaatan *machine learning*. Studi terdahulu telah menerapkan beragam algoritma, meliputi *Decision Tree*, *Support Vector Machine*, *Naive Bayes*, hingga Jaringan Saraf Tiruan [8]. Hasil performansi yang bervariasi dari setiap algoritma tersebut menegaskan bahwa efektivitas sistem sangat bergantung pada kualitas dataset dan desain fitur [9].

Akantetapi, machine learning membuat fraud detection tidak semudah yang dibayangkan. Ada beberapa permasalahan struktural: imbalance kelas yang ekstrem. Kasus fraud biasanya hanya sedikit, misalnya di dataset PaySim, rasio fraud cuma 0,129% atau 1 dari 775 transaksi normal [10]. Model yang cuma prediksi "normal" terus malah dapat akurasi di atas 99%, Tingginya akurasi dapat menjadi bias jika model gagal mengidentifikasi satu pun kasus kecurangan. Oleh karena itu, metrik evaluasi dalam penelitian ini tidak terbatas pada akurasi semata, melainkan juga mengintegrasikan *Recall* dan *Area Under the Receiver Operating Characteristic* (AUC-ROC) yang lebih sensitif terhadap kelas minoritas. Guna mengatasi hambatan ketidakseimbangan data (*class imbalance*), metode *Synthetic Minority Over-sampling Technique* (SMOTE) diterapkan untuk menghasilkan data sintetis kelas minoritas melalui teknik interpolasi antardata yang bertetangga [11]. Kendati demikian, efektivitas metode SMOTE sangat kontingen terhadap karakteristik distribusi data orisinal. Apabila variabilitas variasi data kecurangan (*fraud*) pada sampel asli relatif homogen, interpolasi linear cenderung menghasilkan data sintetis yang redundan, yang pada akhirnya mendegradasi validitas model dan menginduksi bias baru [12].

Di sinilah ensemble learning menjadi pilihan utama. Pada Prinsipnya sederhana yaitu mengabungkan beberapa model dengan cara kerja yang berbeda, supaya kelemahan satu model bisa ditutupi dengan kelebihan model lain. Dalam penelitian ini kami mengunakan tiga algoritma: Random Forest (ratusan pohon keputusan dari fitur acak, robust terhadap overfitting) [13], Gradient Boosting (sekuensial, setiap iterasi dan mampu memperbaiki kesalahan sebelumnya); dan XGBoost (versi gradient boosting yang sudah dioptimasi, dengan regularisasi dan scale_pos_weight khusus untuk imbalance) [14]. Ketiganya digabungkan menggunakan soft voting berbobot, harapannya sistem menjadi lebih tangguh daripada model tunggal.

Penelitian sebelumnya sudah menunjukkan ensemble lebih unggul dari single model buat fraud detection, akantetapi sebagian besar riset sebelumnya rata-rata menggunakan dataset kartu kredit dari Eropa atau Amerika yang berbeda karakter dengan mobile money di Asia Tenggara [15]. Selain itu, aspek latensi sangat penting Di lingkungan produksi (*production environment*), kedua aspek tersebut sering kali luput dari fokus utama implementasi. Fenomena ini mendasari urgensi penelitian ini, yang diarahkan pada dua kontribusi utama: pertama, mengkaji arsitektur *ensemble learning* yang optimal untuk karakteristik sistem *mobile money*; kedua, mengintegrasikan latensi sebagai metrik evaluasi krusial selain pemenuhan statistik performansi model [16].

Dengan dasar itu, penelitian fokus pada empat pertanyaan: 1) Apakah ensemble outperform model tunggal secara konsisten? 2) Seberapa besar peran SMOTE untuk melakukan recall didalam data imbalance? 3) Apakah gap performa antara ensemble dan single model worth dengan tambahan kompleksitas sistem? 4) Apakah latency sistem memenuhi standar real-time? Seluruh urgensi dan pertanyaan penelitian tersebut dijawab melalui serangkaian eksperimen terstruktur, yang diproyeksikan mampu memberikan kontribusi metodologis signifikan bagi pengembangan sistem deteksi kecurangan (*fraud detection system*) di Indonesia.

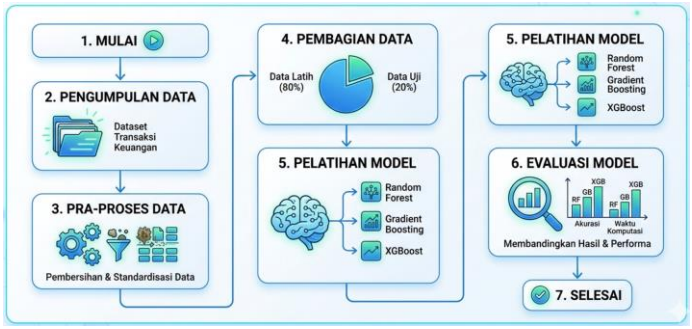
Penelitian ini memberikan kontribusi teoretis yang signifikan terhadap pengembangan khazanah ilmu pengetahuan di bidang teknologi finansial (*FinTech*) dan keamanan siber, khususnya dalam memperkaya literatur mengenai implementasi algoritma pembelajaran mesin tingkat lanjut untuk mitigasi risiko kejahatan digital [17]. Melalui pendekatan studi komparatif, penelitian ini memberikan validasi empiris yang mendalam mengenai perbandingan performa, tingkat akurasi, efisiensi waktu komputasi, serta kemampuan penanganan ketidakseimbangan data (*imbalanced data*) yang krusial antara metode *Random Forest*, *Gradient Boosting*, dan *XGBoost*, sehingga dapat menjadi acuan akademis yang kuat bagi peneliti selanjutnya dalam mengembangkan arsitektur model deteksi anomali yang lebih optimal[18]. Secara praktis, hasil penelitian ini memberikan kontribusi nyata bagi industri perbankan dan penyedia layanan *FinTech* berupa rekomendasi teknis bersasaran dalam memilih model prediktif yang paling cepat dan akurat untuk menyaring transaksi mencurigakan secara *real-time*, meminimalkan tingkat kesalahan deteksi (*false positive*), serta memberikan panduan operasional bagi para *data scientist* dalam mengonfigurasi algoritma *ensemble learning* pada data transaksi riil[19]. Pada akhirnya, keberhasilan implementasi model ini berkontribusi langsung pada penguatan sistem keamanan digital nasional yang esensial untuk meminimalkan kerugian finansial penyedia layanan dan menumbuhkan rasa aman serta kepercayaan masyarakat luas dalam bertransaksi di ekosistem keuangan digital.

2. METODOLOGI PENELITIAN

2.1 Desain Penelitian

Metodologi penelitian ini menerapkan desain eksperimen kuantitatif dengan pendekatan komparatif-deskriptif. Tiga algoritma *machine learning* dievaluasi dalam dua skenario pengujian: secara independen sebagai klasifikator tunggal (*baseline model*) dan secara simultan dalam arsitektur *ensemble*. Perbandingan antara kedua skenario tersebut digunakan untuk menganalisis *trade-off* (keuntungan dan konsekuensi) dari penggabungan model. Variabel bebas dalam penelitian ini meliputi jenis algoritma dan konfigurasi *hyperparameter*, sedangkan variabel terikat mencakup *precision*, *recall*, *F1-score*, AUC-ROC, *Average Precision*, serta latensi per transaksi. Selain itu, eksperimen tambahan dirancang secara khusus untuk mengisolasi dan membedakan dampak implementasi teknik SMOTE dari pengaruh arsitektur *ensemble* itu sendiri.

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 4, June 2026 | Hal 1696-1702
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i4.1155



Gambar 1. Alur Proses Penelitian

Bagan alur pada Gambar tersebut mengilustrasikan siklus hidup pengolahan data (*Data Science Life Cycle*) yang dirancang secara linear dari kiri ke kanan untuk menggambarkan tahapan studi komparatif secara sistematis. Proses diawali pada tahap Mulai, yang menandakan perumusan masalah dan penentuan tujuan komparasi algoritma deteksi penipuan. Tahap selanjutnya adalah Pengumpulan Data, yaitu penarikan dataset sekunder berisi rekaman transaksi keuangan digital yang mencakup transaksi normal dan transaksi fraud. Data mentah tersebut kemudian masuk ke tahap Pra-Proses Data yang disimbolkan dengan roda gigi dan filter, di mana dilakukan pembersihan data (*data cleaning*) dari nilai kosong atau duplikat, serta standardisasi skala fitur agar siap diproses oleh algoritma pembelajaran mesin.

Setelah data bersih, dilakukan Pembagian Data menggunakan proporsi diagram lingkaran standar industri, yaitu memisahkan dataset menjadi Data Latih (80%) untuk fase pembelajaran dan Data Uji (20%) untuk fase pengujian keandalan model. Memasuki inti komparatif penelitian, data latih tersebut dieksekusi pada tahap Pelatihan Model secara paralel ke dalam tiga algoritma *ensemble learning* berbasis otak digital, yaitu *Random Forest*, *Gradient Boosting*, dan *XGBoost*. Setelah model terbentuk, tahap Evaluasi Model dilakukan menggunakan kaca pembesar dan grafik batang untuk menguji performa ketiga algoritma tersebut menggunakan data uji, dengan tolok ukur utama berupa tingkat akurasi klasifikasi dan efisiensi waktu komputasi dalam mendeteksi fraud. Alur penelitian ini diakhiri pada tahap Selesai setelah berhasil mengidentifikasi dan merekomendasikan satu algoritma terbaik yang paling optimal untuk sistem deteksi penipuan transaksi digital.

Tahapan penelitian menggunakan kerangka CRISP-DM [20] ada enam fase: Kerangka kerja penelitian ini mengadopsi metodologi CRISP-DM (*Cross-Industry Standard Process for Data Mining*) yang meliputi enam tahapan terstruktur: pemahaman bisnis (*business understanding*), pemahaman data (*data understanding*), persiapan data (*data preparation*), pemodelan (*modeling*), evaluasi (*evaluation*), dan implementasi (*implementation*). Sifat dari kerangka kerja ini adalah iteratif, yang memungkinkan peneliti untuk kembali ke fase sebelumnya apabila hasil evaluasi pada fase tertentu belum memenuhi ambang batas performansi yang ditetapkan. Proses eksperimen dalam riset ini dirancang melalui dua tahapan iterasi utama. Iterasi pertama difokuskan pada pengonstruksian model dasar (*baseline model*) menggunakan konfigurasi *hyperparameter* awal untuk dievaluasi kinerjanya. Selanjutnya, pada iterasi kedua, arsitektur model tersebut disempurnakan melalui proses optimasi parameter menggunakan metode *Grid Search Cross-Validation* guna mencapai hasil yang optimal

2.2 Dataset

Dataset yang digunakan dalam penelitian ini diperoleh dari simulator PaySim [14] software simulasi transaksi mobile money yang berbasis data nyata dari lebih 14 negara [21]. Simulator membuat data sintetis dengan distribusi statistik mirip transaksi real dan juga menyisipkan skenario fraud yang terkontrol sehingga label ground truth yang sangat mudah dipahami. Total ada 6.362.620 transaksi selama 30 hari, setara 744 step waktu. Fraud cuma 8.213 transaksi, menjadi rasio imbalance 1:775 antara minoritas dan mayoritas. menjadi, akurasi Metrik akurasi secara keseluruhan (*overall accuracy*) tidak dapat dijadikan sebagai indikator performansi utama. Hal ini dikarenakan model yang gagal mengidentifikasi seluruh kasus *fraud* tetap dapat menghasilkan nilai akurasi di atas 99% akibat dominasi kelas mayoritas.

Setiap transaksi mempunyai 11 atribut: step (waktu), type (kategori transaksi: CASH-IN, CASH-OUT, DEBIT, PAYMENT, TRANSFER), amount, nameOrig dan nameDest (pengirim/penerima), saldo sebelum dan sesudah, serta label isFraud dan isFlaggedFraud. Fraud hanya ada di CASH-OUT dan TRANSFER. menjadi, data yang relevan untuk modeling hanya dua tipe itu: 2.531.701 transaksi.

Tabel 1. Statistik Deskriptif Dataset PaySim Financial Fraud Detection

Fitur	Tipe	Mean / Modus	Std Dev	Keterangan
step	Integer	243,40	142,33	1 step = 1 jam, max 744
type	Kategori	PAYMENT		5 kategori transaksi
amount	Float	179.861,90	603.858,23	Nilai transaksi (mata uang lokal)
oldbalanceOrg	Float	833.883,10	2.888.242,67	Saldo pengirim sebelum transaksi
newbalanceOrig	Float	855.113,67	2.924.048,50	Saldo pengirim setelah transaksi
isFraud (Label)	Binary	0,00129	0,035	8.213 fraud / 6.362.620 total

Sumber: Dataset Synthetic Mobile Money Transactions, Kaggle

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 4, June 2026 | Hal 1696-1702
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i4.1155

2.3 Praproses Data

Praproses data dilakukan bertahap. Pertama, filter transaksi: hanya melakukan proses CASH-OUT dan TRANSFER. Kedua, rekayasa fitur tambahan balance Diff Orig (selisih saldo pengirim), balance Diff Dest (saldo penerima), dan isMerchant (apakah penerima merchant). Tahap ketiga melibatkan transformasi variabel kategori type menggunakan teknik *one-hot encoding*. Sementara itu, fitur nameOrig dan nameDest dieliminasi dari dataset karena hanya bertindak sebagai identitas unik yang tidak memuat informasi pola transaksional bagi proses pembelajaran model.

Seluruh fitur numerik ditransformasikan melalui proses standarisasi menggunakan metode *StandardScaler*. Selanjutnya, partisi dataset dilakukan dengan rasio 80:20 menggunakan teknik pembagian terstratifikasi (*stratified split*) untuk menjamin konsistensi proporsi kelas *fraud* pada data latih (*training set*) dan data uji (*testing set*). Implementasi teknik SMOTE diaplikasikan secara eksklusif pada data latih guna mencegah terjadinya kebocoran data (*data leakage*) ke dalam data uji, yang berpotensi menghasilkan estimasi performansi model yang bias (*overoptimistic*). Seluruh rangkaian alur kerja (*pipeline*) ini dieksekusi menggunakan lingkungan pemrograman Python versi 3.9, dengan memanfaatkan pustaka *scikit-learn* versi 1.0.2 dan *imbalanced-learn* versi 0.9.0.

2.4 Arsitektur Sistem

Arsitektur sistem terdiri dari tiga layer: ingestion & praproses (data masuk dan kemudian diproses konsisten sesuai training pipeline); model layer (tiga algoritma jalan paralel, hasilnya digabungkan melalui soft voting berbobot); dan output layer (konversi skor digabungankan menjadi keputusan fraud/normal dan skor risiko membuat investigasi manual).

Hyperparameter setiap model dilakukan proses setting melalui Grid Search Cross-Validation (k=5). Random Forest pakai 200 estimator, max depth 15, min_samples_split 10, class_weight “balanced”. Gradient Boosting: 150 estimator, learning rate 0,05, max depth 5, subsample 0,8. XGBoost: 200 estimator, max_depth 6, learning rate 0,1, scale_pos_weight 775. Voting ensemble, bobot: RF 1.0, GB 1.0, XGB 1.2, disesuaikan dari performa individual di validation set.

2.5 Evaluasi Model

Model dievaluasi menggunakan metrik-metrik yang sesuai untuk imbalance: Precision, Recall, F1-score, AUC-ROC, Average Precision, dan analisis confusion matrix. Di fraud detection, recall lebih penting karena false negative (penipuan lolos) lebih berbahaya daripada false positive (investigasi salah). Latency dilakukan pengecekan melalui simulasi 5.000 transaksi di lokal menggunakan Python performance counter. Sehingga mencapai batas 200 ms yang digunakan sebagai standar operasional, sesuai literatur fraud detection real-time [7].

Evaluasi model dilakukan menggunakan serangkaian metrik yang dipilih karena relevansinya untuk skenario imbalanced classification, yaitu Precision, Recall, F1-Score, dan Area Under the ROC Curve (AUC-ROC), dilengkapi dengan Average Precision dan analisis confusion matrix. Dalam konteks fraud detection, metrik recall mendapat bobot pertimbangan yang lebih besar dibanding precision, karena konsekuensi dari false negative membiarkan transaksi fraud sehingga sejumlah transaksi ilegal berisiko melewati filter deteksi dan menimbulkan kerugian finansial yang jauh lebih merugikan daripada false positive yang hanya menambah beban investigasi. Pengujian latensi dilakukan dengan mensimulasikan aliran 5.000 transaksi secara berurutan pada lingkungan lokal menggunakan Python performance counter. Ambang batas 200 ms yang akan digunakan sebagai kriteria kelayakan operasional, mengarah pada standar yang yang normal [7].

3. HASIL DAN PEMBAHASAN

3.1 Kinerja Model Ensemble

Tabel 2 menyajikan perbandingan performa seluruh model yang diuji pada data testing yang tidak pernah terlihat selama proses pelatihan maupun tuning.

Tabel 2. Hasil Evaluasi Performa Model pada Data Testing

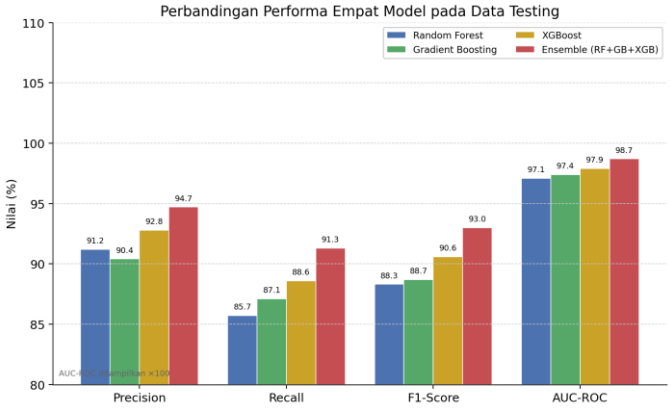
Model	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC
Random Forest (RF)	91,2	85,7	88,3	0,971
Gradient Boosting (GB)	90,4	87,1	88,7	0,974
XGBoost	92,8	88,6	90,6	0,979
Ensemble (RF+GB+XGB)	94,7	91,3	93,0	0,987

Keterangan: Nilai terbaik ditandai pada baris Ensemble (RF+GB+XGB).

Dari Tabel 2 terlihat bahwa XGBoost tampil sebagai model individual dengan kinerja tertinggi: F1-score 90,6% dan AUC-ROC 0,979 angka yang sudah cukup kompetitif untuk keperluan fraud detection. Namun ensemble berhasil mendorong capaian tersebut secara bermakna, dengan F1-score naik ke 93,0% dan AUC-ROC ke 0,987. Selisih ini mungkin terlihat kecil secara absolut, akantetapi memiliki implikasi praktis yang signifikan. Analisis confusion matrix menunjukkan bahwa false negative transaksi fraud yang lolos tanpa terdeteksi berkurang sebesar 18,4% dibandingkan XGBoost yang berdiri sendiri. Pada skala ratusan ribu transaksi harian yang umum diproses oleh bank digital,

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 4, June 2026 | Hal 1696-1702
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i4.1155

pengurangan 18,4% false negative berarti puluhan hingga ratusan transaksi penipuan berhasil diintersepsi setiap harinya yang sebelumnya luput. Breiman (2001) menjelaskan mekanisme ini melalui argumen reduksi varians: menggabungkan model yang beragam secara statistik cenderung menghasilkan prediksi yang lebih stabil, karena kesalahan satu model cenderung tidak berkorelasi dengan kesalahan model lainnya [10].



Gambar 2. Perbangan Performa Empat Model Data Testing

Gambar 2 memvisualisasikan keempat metrik secara berdampingan dan memperlihatkan pola yang konsisten: model ensemble menempati posisi tertinggi pada seluruh dimensi pengukuran, bukan hanya unggul pada satu metrik tertentu. Keunggulan menyeluruh ini penting karena dalam konteks deteksi fraud, sebuah model yang baik pada precision namun lemah pada recall (atau sebaliknya) tidak dapat diandalkan secara operasional.

Pada tingkat transaksi individual, keunggulan tersebut tercermin pada komposisi kesalahan model. Dari total 1.643 transaksi *fraud* pada data uji, ensemble berhasil mengidentifikasi 1.500 di antaranya (recall 91,3%), sementara hanya 143 transaksi yang lolos tanpa terdeteksi. Pada sisi kelas normal, dari 552.439 transaksi sah hanya 84 yang keliru ditandai sebagai *fraud*, sehingga *false positive rate* berada pada level sangat rendah, sekitar 0,015%. Rendahnya angka ini memiliki implikasi operasional konkret: tim verifikasi manual tidak akan dibanjiri peringatan palsu (*false alarm*) sebuah masalah klasik yang kerap menggerus kepercayaan operator terhadap sistem deteksi otomatis. Dengan precision 94,7%, hampir setiap peringatan yang dikeluarkan benar-benar merepresentasikan transaksi yang patut dicurigai, sehingga sumber daya investigasi dapat dialokasikan secara efisien.

Analisis kurva ROC turut memperkuat temuan ini. Model ensemble mencapai AUC-ROC 0,987, konsisten melampaui ketiga model tunggal (Random Forest 0,971, Gradient Boosting 0,974, dan XGBoost 0,979) di sepanjang rentang *threshold*. Keunggulan yang bersifat menyeluruh ini bukan terbatas pada satu titik operasi tertentu; artinya, apabila ambang batas keputusan perlu digeser untuk menyesuaikan toleransi risiko institusi, model ensemble tetap memberikan keseimbangan *true positive rate* dan *false positive rate* yang lebih baik dibandingkan alternatifnya sebuah nilai tambah penting bagi sistem yang dioperasikan dalam jangka panjang dengan kebijakan risiko yang dapat berubah.

3.2 Dampak SMOTE terhadap Performa

Tabel 3 memperlihatkan perbandingan kinerja ensemble yang sama ketika pelatihan tanpa dan dengan penerapan SMOTE.

Tabel 3. Perbandingan Performa Model Ensemble dengan dan Tanpa SMOTE

Konfigurasi	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC
Tanpa SMOTE	96,3	72,4	82,7	0,961
Dengan SMOTE	94,7	91,3	93,0	0,987

Perbedaan yang paling menarik perhatian dari Tabel 3 ada pada nilai recall. Tanpa SMOTE, recall hanya mencapai 72,4% artinya lebih dari seperempat seluruh transaksi fraud dalam data uji berhasil lolos tanpa terdeteksi. Precision memang lebih tinggi (96,3%), akan tetapi dalam kerangka fraud detection, precision yang tinggi dengan recall rendah menandakan model yang terlalu konservatif dalam hal ini hanya berani menandai kasus yang sudah sangat jelas mencurigakan, sementara kasus-kasus yang lebih ambigu dibiarkan berlalu. Setelah SMOTE diterapkan, recall meningkat tajam ke 91,3%. Precision turun tipis ke 94,7% penurunan sebesar 1,6 poin yang dapat diterima mengingat perolehan recall sebesar 18,9 poin. F1-score yang menyeimbangkan keduanya naik dari 82,7% menjadi 93,0%, sementara AUC-ROC membaik dari 0,961 ke 0,987. Peningkatan AUC-ROC yang merata di seluruh rentang threshold mengindikasikan bahwa model dengan SMOTE tidak hanya lebih sensitif pada satu titik keputusan tertentu, melainkan secara struktural lebih mampu memisahkan kedua kelas [9].

3.3 Analisis Feature Importance

Tabel 4 menyajikan sepuluh fitur teratas berdasarkan skor importance dari komponen Random Forest dalam ensemble.

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 4, June 2026 | Hal 1696-1702
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i4.1155

Tabel 4. Sepuluh Fitur Terpenting dalam Deteksi Fraud

Rank	Fitur	Importance Score	Persentase (%)
1	balanceDiffOrig	0,3012	24,1
2	amount	0,2187	17,5
3	oldbalanceOrg	0,1654	13,2
4	newbalanceOrig	0,1103	8,8
5	oldbalanceDest	0,0876	7,0
6	balanceDiffDest	0,0731	5,8
7	step	0,0612	4,9
8	type_TRANSFER	0,0421	3,4
9	type_CASH_OUT	0,0312	2,5
10	newbalanceDest	0,0092	0,7

Fitur balanceDiffOrig mendominasi dengan kontribusi 24,1%, diikuti amount (17,5%) dan oldbalanceOrg (13,2%). Ketiga fitur ini secara bersama-sama menyumbang lebih dari 55% dari total kekuatan diskriminatif model. Pola dominasi ini tidak mengherankan apabila dikaitkan dengan modus fraud yang paling umum dalam dataset PaySim, yaitu pengosongan saldo pengirim secara penuh melalui satu transaksi CASH-OUT atau TRANSFER besar [15]. Pada transaksi semacam ini, balanceDiffOrig akan bernilai sangat tinggi relatif terhadap saldo awal pengirim sebuah sinyal yang mudah ditangkap oleh model pohon keputusan. Kehadiran fitur step pada posisi ketujuh dengan kontribusi 4,9% mengindikasikan adanya pola temporal dalam distribusi fraud: aktivitas penipuan tampaknya tidak tersebar merata sepanjang 744 langkah waktu simulasi, melainkan terkonsentrasi pada periode-periode tertentu. Temuan ini membuka kemungkinan pengembangan fitur temporal yang lebih kaya seperti jam dalam sehari, hari dalam minggu, atau interval sejak transaksi terakhir sebagai arah perluasan penelitian selanjutnya.

3.4 Performa Deteksi Real-Time

Pengujian latensi dilaksanakan dengan mengalirkan 5.000 transaksi secara berurutan melalui pipeline inferensi penuh pada lingkungan laptop lokal, dan waktu respons setiap transaksi dicatat menggunakan Python performance counter. Rata-rata latensi yang tercatat adalah 147,3 ms per transaksi, dengan nilai median 132,8 ms. Angka yang paling kritis untuk menilai keandalan di bawah kondisi terburuk adalah persentil ke-99, yang mencapai 198,6 ms masih di bawah ambang batas 200 ms yang ditetapkan sebagai kriteria kelayakan [20]. Throughput dalam kondisi beban tunggal tercatat 412 transaksi per detik. Perlu dicatat bahwa pengujian ini dilakukan pada perangkat laptop konvensional tanpa akselerasi hardware khusus. Pada lingkungan server produksi dengan kapasitas komputasi yang lebih besar, angka latensi ini sangat mungkin dapat ditekan lebih jauh. Sebaliknya, kondisi produksi yang melibatkan latensi jaringan, konkurensi tinggi, dan overhead sistem operasional bisa menggeser angka-angka tersebut ke arah yang lebih tinggi. Oleh karena itu, hasil pengujian ini lebih tepat dipahami sebagai indikator kelayakan awal daripada estimasi performa produksi yang definitif.

3.5 Pembahasan

Keunggulan ensemble atas model individual yang ditunjukkan dalam penelitian ini konsisten dengan prediksi teoritis ensemble learning. Diversitas mekanisme antara RF, GB, dan XGB menciptakan kesalahan prediksi yang tidak berkorelasi sempurna, penggabungan ketiganya melalui soft voting mereduksi varians keseluruhan tanpa harus mengorbankan bias secara signifikan. Dari ketiga komponen, XGBoost tampil paling kuat secara individual sebagian karena parameter scale_pos_weight yang secara eksplisit mengompensasi ketidakseimbangan kelas, dan sebagian lagi karena arsitektur gradient boosting-nya yang secara iteratif memfokuskan pembelajaran pada sampel yang sulit diklasifikasi [7]. Meskipun demikian, kontribusi RF dan GB dalam ensemble terbukti lebih dari sekadar pelengkap, false negative berkurang 18,4% dibanding XGBoost sendirian, dan penurunan ini tidak dapat sepenuhnya dijelaskan oleh kinerja XGBoost itu sendiri. Sejumlah keterbatasan penelitian ini perlu diakui secara eksplisit. Pertama, PaySim adalah dataset sintetis; meskipun dibangun di atas pola transaksi nyata, ia tidak dapat sepenuhnya merepresentasikan heterogenitas perilaku pengguna dan modus fraud yang sesungguhnya terjadi dalam ekosistem keuangan digital Indonesia. Karakteristik lokal seperti dominasi jenis transaksi tertentu, distribusi nilai nominal yang khas, atau taktik fraud yang spesifik untuk pasar Indonesia mungkin tidak terwakili dengan baik. Kedua, pengujian latensi dilakukan pada lingkungan lokal yang tidak merepresentasikan kompleksitas infrastruktur produksi. Ketiga dan ini mungkin yang paling krusial untuk pertimbangan jangka panjang penelitian ini tidak mengkaji aspek concept drift, kemungkinan performa model terdegradasi seiring berjalannya waktu akibat pergeseran pola fraud yang terus berevolusi. Ketiga keterbatasan ini tidak melemahkan validitas temuan yang ada, melainkan menandai agenda penelitian lanjutan yang perlu dikejar.

4. KESIMPULAN

Penelitian ini berhasil memvalidasi hipotesis bahwa konfigurasi ensemble yang menggabungkan Random Forest, Gradient Boosting, dan XGBoost menghasilkan performa deteksi fraud yang lebih baik secara konsisten dibandingkan masing-masing model secara individual. Pada data uji yang diambil dari dataset PaySim dengan rasio fraud 0,129 persen, ensemble mencapai precision 94,7%, recall 91,3%, F1-score 93,0%, dan AUC-ROC 0,987 melampaui XGBoost sebagai

model individual terkuat dengan margin yang bermakna secara praktis. Pengurangan false negative sebesar 18,4% dibanding model terbaik individual merupakan temuan kunci yang memiliki implikasi langsung pada efektivitas operasional sistem. Penerapan SMOTE terbukti memberikan kontribusi substansial, khususnya pada peningkatan recall dari 72,4% menjadi 91,3%. Tanpa oversampling, sistem menghasilkan precision yang tinggi tetapi melewatkan lebih dari seperempat kasus fraud kondisi yang tidak dapat diterima dalam konteks perbankan digital. Analisis feature importance mengungkap bahwa selisih saldo pengirim, nominal transaksi, dan saldo awal pengirim merupakan tiga prediktor paling informatif, mencerminkan pola fraud dominan berupa pengosongan saldo akun melalui satu transaksi besar. Temuan ini dapat menjadi basis perancangan fitur monitoring yang lebih terarah. Dari sisi operasional, rata-rata latensi 147,3 ms dengan persentil ke-99 sebesar 198,6 ms mengindikasikan kelayakan sistem untuk dipertimbangkan dalam deployment nyata. Untuk penelitian selanjutnya, setidaknya ada tiga jalur pengembangan layak dipertimbangkan. Pertama, validasi menggunakan data transaksi aktual dari institusi keuangan Indonesia, yang akan memberikan gambaran lebih jelas tentang transferabilitas model ke kondisi nyata. Kedua, integrasi teknik explainable AI seperti SHAP atau LIME untuk meningkatkan interpretabilitas keputusan model, yang penting baik untuk kepercayaan pengguna maupun kepatuhan terhadap persyaratan regulasi yang semakin ketat. Ketiga, pengembangan mekanisme deteksi concept drift yang memungkinkan model beradaptasi secara berkala terhadap evolusi pola fraud, sehingga performa tidak terdegradasi seiring berjalannya waktu tanpa adanya intervensi pembaruan manual.

REFERENCES

[1] A. Abdallah, M. A. Maarof, and A. Zainal, “Fraud detection system: A survey,” *Journal of Network and Computer Applications*, vol. 68. 2016. doi: 10.1016/j.jnca.2016.04.007.
[2] A. A. Yanto and N. Nelvirita, “Pengaruh New Fraud Diamond Terhadap Potensi Kecurangan Laporan Keuangan,” *J. Eksplor. Akunt.*, vol. 7, no. 3, pp. 1045–1063, 2025, doi: 10.24036/jea.v7i3.2759.
[3] Firman Zamzami Aziz *et al.*, “Analisis Perbandingan Teknik Balancing CTGAN dan ADASYN untuk Deteksi Transaksi Penipuan,” *J. Informatics Interact. Technol.*, vol. 2, no. 2, pp. 361–369, 2025, doi: 10.63547/jiite.v2i2.69.
[4] G. Zhang *et al.*, “EFraudCom: An E-commerce Fraud Detection System via Competitive Graph Neural Networks,” *ACM Trans. Inf. Syst.*, vol. 40, no. 3, 2022, doi: 10.1145/3474379.
[5] H. Thakkar, G. C. Fanuel, S. Datta, P. Bhadra, and S. B. Dabhade, “Optimizing Internal Audit Practices for Combatting Occupational Fraud: A Study of Data Analytic Tool Integration in Zimbabwean Listed Companies,” *Int. Res. J. Multidiscip. Scope*, vol. 6, no. 1, 2025, doi: 10.47857/irjms.2025.v06i01.02164.
[6] C. M. Sitanggang, M. Simalango, R. Purba, and J. Darma, “Pemanfaatan Big Data Analytics dalam Deteksi Fraud dan Prediksi Kinerja Keuangan: Kajian Literatur,” *Account*, vol. 12, no. 2, pp. 2713–2725, 2025, doi: 10.32722/account.v12i2.7854.
[7] A. F. Sariat, I. J. Siddique, M. Hossain, M. M. Islam, and T. Rahman, “AI Driven Fraud Detection in Financial Ecosystems: A Hybrid Machine Learning Framework,” in *2025 International Conference on Electrical, Computer and Communication Engineering, ECCE 2025*, 2025. doi: 10.1109/ECCE64574.2025.11013808.
[8] Y. Chen, “A Comparative Study of Machine Learning Models for Credit Card Fraud Detection,” *Acad. J. Nat. Sci.*, vol. 2, no. 4, pp. 11–18, 2025, doi: 10.70393/616a6e73.333530.
[9] M. Rakhshaninejad, M. Fathian, B. Amiri, and N. Yazdanjue, “An Ensemble-Based Credit Card Fraud Detection Algorithm Using an Efficient Voting Strategy,” *Comput. J.*, vol. 65, no. 8, 2022, doi: 10.1093/comjnl/bxab038.
[10] B. Srishailam, Y. Rahul, P. Pavan, Y. Bharadwaj, and K. Lokeshwar, “Credit Card Fraud Detection Using Adaboost and Majority Voting: A Hybrid Approach for Real-Time Prevention,” *Macaw Int. J. Adv. Res. Comput. Sci. Eng.*, vol. 10, no. 1, 2024.
[11] M. Nakamura, Y. Kajiwara, A. Otsuka, and H. Kimura, “LVQ-SMOTE - Learning Vector Quantization based Synthetic Minority Over-sampling Technique for biomedical data,” *BioData Min.*, vol. 6, no. 1, 2013, doi: 10.1186/1756-0381-6-16.
[12] D. A. N. Red and F. Terhadap, “Pengaruh akuntansi forensik, skeptisisme profesional, dan red flag terhadap pendeteksian fraud,” vol. 5, no. 1, pp. 81–96, 2026.
[13] S. F. Taskiran, B. Turkoglu, E. Kaya, and T. Asuroglu, “A comprehensive evaluation of oversampling techniques for enhancing text classification performance,” *Sci. Rep.*, vol. 15, no. 1, 2025, doi: 10.1038/s41598-025-05791-7.
[14] M. M. Islam, I. Zerine, M. A. Rahman, M. S. Islam, and M. Y. Ahmed, “AI-Driven Fraud Detection in Financial Transactions - Using Machine Learning and Deep Learning to Detect Anomalies and Fraudulent Activities in Banking and E-Commerce Transactions,” *SSRN Electron. J.*, 2025, doi: 10.2139/ssrn.5287281.
[15] D. R. K. Saputra, Y. V. Via, and A. N. Sihananto, “Deteksi Anomali Menggunakan Ensemble Learning Dan Random Oversampling Pada Penipuan Transaksi Keuangan,” *J. Inform. dan Tek. Elektro Terap.*, vol. 12, no. 3, pp. 2779–2788, 2024, doi: 10.23960/jitet.v12i3.4910.
[16] Reyhand Ardhitha, Revifal Anugerah, and Tata Sutabri, “Analisis Penerapan Machine Learning dan Algoritma Anomali untuk Deteksi Penipuan pada Transaksi Digital,” *Repeater Publ. Tek. Inform. dan Jar.*, vol. 3, no. 1, pp. 80–90, 2025, doi: 10.62951/repeater.v3i1.345.
[17] I. A. Fernanda and H. Habiburrochman, “Peran Kecerdasan Buatan (AI) Dalam Meningkatkan Audit Forensik Untuk Mendeteksi Kecurangan: Tinjauan Literatur Sistematis,” *Owner*, vol. 9, no. 4, pp. 3430–3442, 2025, doi: 10.33395/owner.v9i4.2825.
[18] A. H. Siregar, A. Nazir, I. Afrianty, E. Budianita, and F. Insani, “Jurnal Computer Science and Information Technology (CoSciTech) simvastatin,” vol. 4, no. 3, pp. 626–635, 2023.
[19] A. M. Syahbani, W. Firdaus, and K. A. Musodo, “A Comparative Study of Data Mining Algorithms for Fraud Detection in Financial Transactions,” *Sinkron*, vol. 9, no. 2, pp. 814–821, 2025, doi: 10.33395/sinkron.v9i2.14645.
[20] J. S. Allen, “CardSim: A Bayesian Simulator for Payment Card Fraud Detection Research,” *Financ. Econ. Discuss. Ser.*, no. 2025–017, 2025, doi: 10.17016/feds.2025.017.
[21] Kavita S. Pawar, “Refine Framework for Credit Card Fraud Detection System using Machine Learning,” *J. Inf. Syst. Eng. Manag.*, vol. 10, no. 46s, 2025, doi: 10.52783/jisem.v10i46s.8814.