KAMPUS AKADEMIK PUBLISING
Jurnal Sains Student Research (JSSR)
Vol.4, No.5 Oktober 2026
e-ISSN: 3025-9851, p-ISSN : 3025-986X, Hal 1385-1400
DOI : https://doi.org/10.61722/jssr.v4i5.12752

PREDIKSI FRAUD TRANSAKSI DOMPET DIGITAL MENGGUNAKAN ALGORITMA CATBOOST

Mochamad Vikry Firdaus[1], Puspita Nurul Sabrina[2], Herdi Ashaury[3]
[1,2,3] Teknik Informatika, Universitas Jenderal Achmad Yani
Jl. Terusan Jend. Sudirman, Cibeber, Kec. Cimahi Sel., Kota Cimahi, Jawa Barat
*Penulis Korespondensi: vikryfidaus22@if.unjani.ac.id

Abstract. *The growth of digital transactions via digital wallet and mobile money services has heightened the need for fraud detection systems capable of accurately identifying suspicious transactions. A major challenge in fraud detection is class imbalance, where the volume of fraudulent transactions is significantly lower than that of legitimate ones. This study aims to analyze and compare the performance of Random Forest and CatBoost algorithms in detecting fraud within digital transactions using the PaySim dataset. Random Forest served as the benchmark model, while CatBoost was the primary model evaluated through both baseline configurations and hyperparameter tuning. Class imbalance was addressed using class weighting, and the dataset was partitioned using an 80:20 train-test split combined with stratified sampling. Model performance was evaluated using Accuracy, Precision, Recall, F1-Score, ROC-AUC, and PR-AUC, alongside an analysis of False Positives and False Negatives. The results indicate that while Random Forest achieved a Recall of 98.78%, it produced 18,284 False Positives, resulting in a Precision of 8.15%. The CatBoost Baseline model achieved a Precision of 79.19%, Recall of 85.27%, F1-Score of 82.12%, ROC-AUC of 99.92%, and PR-AUC of 90.93%. Following hyperparameter tuning, the CatBoost Tuned model improved its Recall to 92.08%, ROC-AUC to 99.96%, and PR-AUC to 94.46%, while reducing the number of False Negatives from 242 to 130 transactions; however, Precision dropped to 71.46% and the F1-Score to 80.48%. These findings demonstrate that CatBoost Tuned possesses superior fraud detection capabilities—particularly in maximizing Recall and minimizing False Negatives—whereas CatBoost Baseline offers a better balance between Precision and Recall at the classification threshold used in this study.*
Keywords: *fraud detection, digital transactions, mobile money, CatBoost, Random Forest, imbalanced dataset, machine learning*

Abstrak. Pertumbuhan transaksi digital melalui layanan dompet digital dan *mobile money* meningkatkan kebutuhan terhadap sistem deteksi fraud yang mampu mengidentifikasi transaksi mencurigakan secara akurat. Salah satu tantangan utama dalam deteksi fraud adalah ketidakseimbangan kelas (*class imbalance*), di mana jumlah transaksi fraud jauh lebih sedikit dibandingkan dengan transaksi non-fraud. Penelitian ini bertujuan untuk menganalisis dan membandingkan kinerja algoritma Random Forest dan CatBoost dalam mendeteksi fraud pada transaksi digital menggunakan dataset PaySim. Random Forest digunakan sebagai model pembanding, sedangkan CatBoost digunakan sebagai model utama yang dievaluasi melalui konfigurasi *baseline* dan *hyperparameter tuning*. Ketidakseimbangan kelas ditangani menggunakan *class weighting*, sedangkan dataset dibagi menggunakan *train-test split* dengan rasio 80:20 dan metode *stratified sampling*. Kinerja model dievaluasi menggunakan Accuracy, Precision, Recall, F1-Score, ROC-AUC, PR-AUC, serta analisis *False Positive* dan *False Negative*. Hasil penelitian menunjukkan bahwa Random Forest memperoleh Recall sebesar 98,78%, tetapi menghasilkan 18.284 *False Positive* dengan Precision sebesar 8,15%. CatBoost Baseline memperoleh Precision sebesar 79,19%, Recall sebesar 85,27%, F1-Score sebesar 82,12%, ROC-AUC sebesar 99,92%, dan PR-AUC sebesar 90,93%. Setelah dilakukan *hyperparameter tuning*, CatBoost Tuned mengalami peningkatan Recall menjadi 92,08%, ROC-AUC menjadi 99,96%, dan PR-AUC menjadi 94,46%, serta mampu menurunkan jumlah *False Negative* dari 242 menjadi 130 transaksi. Namun, Precision menurun menjadi 71,46% dan F1-Score menjadi 80,48%. Hasil tersebut menunjukkan bahwa CatBoost Tuned memiliki kemampuan deteksi fraud yang lebih kuat, khususnya dalam meningkatkan Recall dan mengurangi *False Negative*, sedangkan CatBoost Baseline memberikan keseimbangan Precision dan Recall yang lebih baik pada *classification threshold* yang digunakan dalam penelitian.
Kata kunci: *deteksi fraud, transaksi digital, mobile money, CatBoost, Random Forest, imbalanced dataset, machine learning*

Naskah Masuk:01 September 2026; Revisi: 02 September 2026; Diterima: 03 September 2026; Terbit: 21 September 2026

1. LATAR BELAKANG

Perkembangan teknologi informasi dan komunikasi telah mendorong perubahan pada sektor keuangan melalui berbagai layanan financial technology (fintech). Salah satu bentuk penerapannya adalah layanan pembayaran digital dan dompet digital yang memungkinkan pengguna melakukan transaksi tanpa uang tunai. Perkembangan tersebut memberikan kemudahan karena transaksi dapat dilakukan secara cepat dan praktis, tetapi pada saat yang sama memperbesar kebutuhan terhadap mekanisme keamanan yang mampu mendeteksi aktivitas transaksi yang mencurigakan. [1], [2].

Fraud pada transaksi digital dapat menimbulkan kerugian finansial dan operasional serta memengaruhi kepercayaan pengguna terhadap penyedia layanan. Tantangan deteksi semakin kompleks karena pola transaksi fraud dapat menyerupai transaksi normal. Sistem berbasis aturan tetap juga memiliki keterbatasan ketika pola fraud berubah atau ketika terdapat variasi perilaku transaksi yang sangat besar. [3], [4].

Machine learning menjadi salah satu pendekatan yang banyak digunakan untuk fraud detection karena model dapat mempelajari hubungan antara karakteristik transaksi dan label fraud dari data historis. Pada data keuangan, permasalahan utama yang harus diperhatikan adalah class imbalance. Pada kondisi ini, jumlah transaksi fraud jauh lebih kecil daripada transaksi normal sehingga Accuracy dapat memberikan gambaran yang menyesatkan apabila digunakan sebagai satu-satunya ukuran performa. [5], [3].

Ensemble learning merupakan pendekatan yang relevan untuk permasalahan tersebut. Random Forest menggunakan sejumlah decision tree melalui pendekatan bagging, sedangkan CatBoost merupakan algoritma gradient boosting. Perbedaan mekanisme pembelajaran tersebut memungkinkan keduanya menghasilkan karakteristik prediksi yang berbeda. Dalam penelitian ini Random Forest digunakan sebagai model pembanding, sedangkan CatBoost menjadi model utama. [6], [7].

Dataset PaySim digunakan karena menyediakan simulasi transaksi mobile money dengan volume data besar dan distribusi kelas yang sangat tidak seimbang. Penelitian ini menggunakan atribut transaksi yang tersedia setelah preprocessing, menghapus nameOrig dan nameDest, melakukan Label Encoding pada type, dan tidak melakukan feature engineering maupun transaction aggregation. [8].

Penelitian ini berfokus pada perbandingan Random Forest dengan CatBoost menggunakan data dan skenario evaluasi yang sama, perbandingan CatBoost dengan CatBoost Tuned, serta evaluasi melalui Accuracy, Precision, Recall, F1-Score, ROC-AUC, PR-AUC, False Positive, dan False Negative. [4], [9], [10], [11].

Kontribusi penelitian adalah memberikan evaluasi yang tidak hanya berorientasi pada nilai Accuracy atau ROC-AUC, tetapi juga mempertimbangkan trade-off Precision–Recall dan karakteristik kesalahan klasifikasi. Analisis tambahan melalui penghapusan fitur dan single-feature validation dilakukan untuk mengkaji robustness performa model. [12], [11].

2. KAJIAN TEORITIS

2.1 Data Mining

Data mining adalah proses menemukan pola, hubungan, atau informasi bermakna dari data besar menggunakan metode komputasi dan analisis. Dalam analisis transaksi, data mining mengolah data dalam jumlah besar untuk mengidentifikasi pola yang sulit diamati langsung dan berperan penting dalam deteksi fraud melalui pemanfaatan pola historis transaksi [3]. Tahapannya meliputi pengumpulan dan pemahaman data, pembersihan data, transformasi/representasi data, pemodelan, serta evaluasi hasil. Kualitas data dan representasi atribut memengaruhi pola yang ditemukan algoritma. Pada

transaksi digital, data mining diarahkan pada klasifikasi dengan atribut transaksi sebagai prediktor dan status fraud sebagai label.

2.2 Dompet Digital

Dompet digital merupakan bentuk teknologi finansial yang memungkinkan aktivitas keuangan elektronik, termasuk pembayaran dan transfer dana. Perkembangan fintech mendorong layanan keuangan yang lebih mudah, cepat, dan aksesibel [1], [2]. Peningkatan pembayaran digital meningkatkan volume transaksi dan kebutuhan mekanisme keamanan yang efektif. Dalam penelitian ini, karakteristik transaksi dompet digital direpresentasikan menggunakan dataset PaySim yang mensimulasikan mobile money dan menyediakan label fraud [8].

2.3 Fraud Transaksi Dompet Digital

Fraud transaksi digital adalah transaksi tidak sah untuk memperoleh keuntungan ilegal atau menimbulkan kerugian pihak lain. Pola fraud dapat menyerupai transaksi normal sehingga deteksi menjadi menantang karena model harus membedakan fraud dan transaksi sah berdasarkan informasi yang tersedia [3], [4]. Karakteristik utama data fraud adalah ketidakseimbangan kelas, karena transaksi fraud jauh lebih sedikit daripada non-fraud, sehingga model cenderung mempelajari kelas mayoritas. Masalah imbalanced data dapat ditangani melalui resampling atau pendekatan pembelajaran yang mempertimbangkan biaya kesalahan kelas [9], [10], [11]. Selain itu, pola fraud dapat berubah sehingga karakteristik historis tidak selalu identik dengan periode berikutnya [4].

2.4 Machine Learning

Machine learning adalah pendekatan komputasional yang memungkinkan sistem mempelajari pola dari data untuk menghasilkan prediksi atau keputusan tanpa aturan eksplisit untuk setiap observasi. Mitchell mendefinisikan machine learning sebagai proses ketika program komputer meningkatkan kinerjanya pada suatu tugas melalui pengalaman [5]. Penelitian ini menggunakan supervised learning karena dataset memiliki label target isFraud. Model mempelajari hubungan atribut transaksi dengan label fraud/non-fraud untuk memprediksi data baru. Pada data transaksi digital, machine learning dapat mengidentifikasi hubungan nonlinier dan pola kompleks, tetapi keberhasilan model juga dipengaruhi kualitas data, representasi fitur, distribusi kelas, konfigurasi model, dan metode evaluasi [14], [29].

2.5 Ensemble Learning

\ Ensemble learning menggabungkan beberapa model dasar untuk menghasilkan prediksi yang lebih kuat, stabil, dan mampu menggeneralisasi lebih baik. Bagging membangun sejumlah model secara relatif independen menggunakan sampel berbeda; Random Forest merupakan implementasinya dengan banyak decision tree dan pemilihan fitur acak [6]. Boosting membangun model secara bertahap, di mana model berikutnya memperbaiki kesalahan model sebelumnya. CatBoost merupakan algoritma boosting berbasis decision tree yang efektif untuk fitur kategorikal [7], [13].

2.6 Algoritma CatBoost

CatBoost (Categorical Boosting) merupakan algoritma gradient boosting berbasis decision tree untuk data tabular, termasuk data dengan fitur kategorikal, yang menggunakan Ordered Target Statistics dalam pemrosesan informasi kategorikal dan Ordered Boosting untuk mengurangi risiko target leakage serta overfitting selama pembelajaran [7], [13]; pada pendekatan boosting, model dibangun secara bertahap, di mana setiap iterasi menambahkan model baru untuk memperbaiki kesalahan model sebelumnya, sehingga secara konseptual model pada iterasi ke-m dapat dituliskan sebagai

$$Fm(x) = Fm - 1(x) + \eta \cdot hm(x)$$

dengan $F_m(x)$ merupakan model pada iterasi ke-m, $F_{m-1}(x)$ merupakan model pada iterasi sebelumnya, $h_m(x)$ merupakan decision tree yang ditambahkan pada iterasi ke-m, dan η merupakan learning rate yang mengontrol kontribusi model baru. CatBoost mengembangkan prinsip boosting tersebut dengan mekanisme yang dirancang untuk menghasilkan estimasi yang lebih terkontrol pada fitur kategorikal dan mengurangi kebocoran informasi target [7].

Dalam penelitian ini, CatBoost digunakan sebagai model utama dalam dua konfigurasi, yaitu CatBoost Baseline dan CatBoost Tuned. CatBoost Baseline digunakan sebagai acuan performa awal, sedangkan CatBoost Tuned menggunakan konfigurasi hyperparameter hasil proses penyesuaian parameter pada eksperimen penelitian.

2.7 Algoritma Random Forest

Random Forest merupakan algoritma ensemble berbasis bagging yang dikembangkan oleh Breiman, yang membangun sejumlah decision tree menggunakan bootstrap sampling dan memilih subset fitur secara acak pada setiap proses pemisahan node, lalu menggabungkan hasil prediksi seluruh pohon untuk menghasilkan prediksi akhir [6]; untuk klasifikasi, keputusan akhir diperoleh melalui majority voting, sehingga jika $h_t(x)$ adalah prediksi pohon ke-t dan T adalah jumlah pohon, prediksi kelas dapat dituliskan sebagai:

$$\hat{y} = \text{mode}(h_1(x), h_2(x), \dots, h_T(x)))$$

di mana penggunaan bootstrap sampling dan random feature selection menghasilkan keragaman antarpohon, mengurangi korelasi antarpohon, serta meningkatkan stabilitas dan mengurangi risiko overfitting dibandingkan satu decision tree [6]; pada penelitian ini, Random Forest digunakan sebagai model pembanding dan, karena dataset memiliki ketidakseimbangan kelas, model dilatih dengan pendekatan class weighting agar kelas fraud memperoleh perhatian lebih besar selama proses pembelajaran, serta dipilih sebagai baseline karena merupakan algoritma ensemble yang banyak digunakan pada klasifikasi data tabular dan fraud detection [6], [17]–[19].

2.8 Confusion Matrix dan Metrik Evaluasi

Confusion Matrix merupakan tabel yang digunakan untuk menggambarkan hubungan antara kelas aktual dan kelas hasil prediksi model klasifikasi. Pada klasifikasi biner, terdapat empat komponen utama, yaitu True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN). TP menunjukkan transaksi fraud yang berhasil diprediksi sebagai fraud, TN menunjukkan transaksi non-fraud yang diprediksi sebagai non-fraud, FP menunjukkan transaksi non-fraud yang diprediksi sebagai fraud, sedangkan FN menunjukkan transaksi fraud yang diprediksi sebagai non-fraud.

Keempat komponen tersebut menjadi dasar perhitungan metrik evaluasi yang digunakan dalam penelitian ini. Accuracy mengukur proporsi seluruh prediksi yang benar terhadap seluruh observasi, dengan rumus:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision mengukur proporsi prediksi positif yang benar-benar merupakan kelas fraud. Nilai precision yang tinggi menunjukkan bahwa lebih sedikit transaksi non-fraud yang salah ditandai sebagai fraud [31].

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall atau True Positive Rate (TPR) mengukur kemampuan model dalam menemukan transaksi fraud yang sebenarnya. Recall yang tinggi menunjukkan semakin sedikit transaksi fraud yang terlewat atau diklasifikasikan sebagai non-fraud.

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1-Score merupakan rata-rata harmonik antara precision dan recall. Metrik ini digunakan untuk melihat keseimbangan antara ketepatan prediksi kelas fraud dan kemampuan model menemukan kelas fraud.

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Selain metrik tersebut, False Positive Rate (FPR) digunakan dalam pembentukan Receiver Operating Characteristic (ROC) Curve. FPR menunjukkan proporsi kelas negatif yang salah diklasifikasikan sebagai positif.

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

ROC Curve menggambarkan hubungan antara True Positive Rate dan False Positive Rate pada berbagai nilai classification threshold. Area di bawah kurva tersebut disebut ROC-AUC. Secara konseptual, ROC-AUC dapat dinyatakan sebagai luas di bawah kurva TPR terhadap FPR [12].

$$ROC - AUC = \int_0^1 T\,PR(FPR)\,d(FPR)$$

Nilai ROC-AUC yang semakin mendekati 1 menunjukkan kemampuan diskriminasi yang semakin baik, sedangkan nilai sekitar 0,5 menunjukkan kemampuan yang mendekati prediksi acak. Pada dataset dengan ketidakseimbangan kelas yang tinggi, ROC-AUC perlu dibaca bersama metrik yang lebih berfokus pada kelas positif.

Precision-Recall Curve menggambarkan hubungan antara precision dan recall pada berbagai nilai threshold. Kurva ini memberikan perhatian lebih langsung terhadap kinerja kelas positif, yaitu transaksi fraud. Area di bawah kurva tersebut disebut PR-AUC dan dapat digunakan sebagai metrik tambahan pada kondisi class imbalance.

$$PR - AUC = \int_0^1 P\,recision(Recall)\,d(Recall)$$

Semakin tinggi nilai PR-AUC, semakin baik kemampuan model mempertahankan precision dan recall pada berbagai threshold. Penggunaan ROC-AUC dan PR-AUC secara bersama-sama memberikan gambaran yang lebih lengkap mengenai kemampuan diskriminasi model serta performanya terhadap kelas fraud [12].

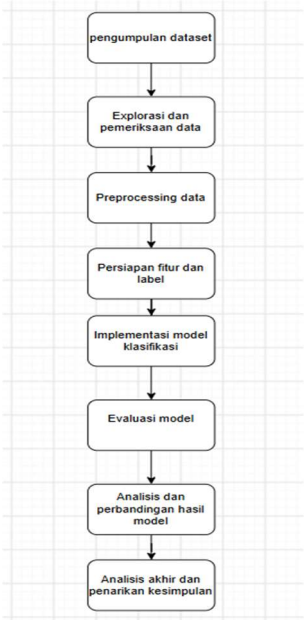
Dengan demikian, evaluasi pada penelitian ini menggunakan Accuracy, Precision, Recall, F1-Score, ROC-AUC, PR-AUC, serta komponen Confusion Matrix berupa TP, TN, FP, dan FN. Kombinasi metrik tersebut memungkinkan evaluasi dilakukan tidak hanya berdasarkan jumlah prediksi benar, tetapi juga berdasarkan kemampuan model dalam mendeteksi kelas fraud dan karakteristik kesalahan klasifikasinya [9],[10],[11],[12], [31]

3. METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dengan desain eksperimen komputasional untuk menganalisis kemampuan algoritma *machine learning* dalam

mendeteksi fraud pada transaksi digital. Penelitian menggunakan dataset PaySim sebagai data penelitian dan membandingkan kinerja algoritma CatBoost dengan Random Forest. CatBoost digunakan sebagai model utama yang terdiri atas model CatBoost dan CatBoost Tuned, sedangkan Random Forest digunakan sebagai model pembanding.

Secara umum, tahapan penelitian terdiri atas pengumpulan dataset, eksplorasi dan pemeriksaan data, *preprocessing* data, persiapan fitur dan label, implementasi model klasifikasi, evaluasi model, analisis dan perbandingan hasil model, serta penarikan kesimpulan.



Gambar 1. Metode Penelitian

3.1 Pengumpulan Data

Data yang digunakan dalam penelitian ini merupakan data sekunder berupa dataset PaySim yang diperoleh dari platform Kaggle. PaySim merupakan dataset simulasi transaksi mobile money yang dikembangkan untuk penelitian deteksi fraud pada sistem pembayaran digital. Dataset tersebut memiliki sejumlah atribut yang menggambarkan karakteristik transaksi, seperti jenis transaksi, nominal transaksi, saldo pengirim sebelum dan sesudah transaksi, serta saldo penerima sebelum dan sesudah transaksi. Dataset juga menyediakan atribut isFraud sebagai label yang menunjukkan apakah suatu transaksi termasuk fraud atau non-fraud [31].

Dataset PaySim terdiri atas 6.362.620 transaksi dengan 11 atribut. Dari keseluruhan transaksi tersebut terdapat 8.213 transaksi fraud atau sekitar 0,13%, sedangkan sebagian besar transaksi lainnya merupakan transaksi non-fraud. Kondisi tersebut menunjukkan bahwa dataset memiliki tingkat ketidakseimbangan kelas yang tinggi sehingga sesuai untuk menguji kemampuan model klasifikasi dalam mendeteksi kelas minoritas.

Dataset kemudian diimpor ke lingkungan Google Colab menggunakan bahasa pemrograman Python untuk dilakukan eksplorasi dan pengolahan lebih lanjut. Tahap pemeriksaan awal dilakukan untuk mengetahui struktur dataset, tipe data, distribusi kelas, keberadaan nilai hilang, serta data duplikat.

3.2 Preprocessing

Tahap preprocessing dilakukan untuk mempersiapkan dataset sebelum digunakan dalam proses pemodelan. Berdasarkan hasil pemeriksaan data, dataset PaySim yang

digunakan tidak memiliki missing value maupun data duplikat sehingga tidak dilakukan proses penghapusan data berdasarkan kedua kondisi tersebut.

Tahap selanjutnya adalah menghapus atribut nameOrig dan nameDest. Kedua atribut tersebut merupakan identifier yang menunjukkan identitas pengirim dan penerima transaksi sehingga tidak digunakan sebagai prediktor dalam proses klasifikasi. Penghapusan kedua atribut tersebut dilakukan untuk menghindari penggunaan identifier sebagai informasi prediktif.

Atribut type yang memiliki tipe data kategorikal kemudian ditransformasikan menjadi bentuk numerik menggunakan metode Label Encoding. Transformasi ini dilakukan agar atribut tersebut dapat digunakan sebagai masukan pada proses pembelajaran model.

Setelah proses preprocessing, dataset dibagi menjadi variabel fitur (X) dan variabel target (y). Variabel X berisi atribut transaksi yang digunakan sebagai masukan model, sedangkan variabel y menggunakan atribut isFraud sebagai label klasifikasi dengan dua kelas, yaitu fraud dan non-fraud.

Karena distribusi kelas pada dataset sangat tidak seimbang, penelitian ini menggunakan pendekatan class weighting pada proses pelatihan Random Forest dan CatBoost. Pendekatan tersebut memberikan bobot yang lebih besar terhadap kelas fraud sehingga model memberikan perhatian lebih terhadap kelas minoritas tanpa melakukan perubahan terhadap distribusi asli dataset [28], [30].

Selanjutnya, data dibagi menjadi data latih dan data uji menggunakan metode train-test split dengan rasio 80:20. Parameter stratify digunakan untuk mempertahankan proporsi kelas fraud dan non-fraud pada data latih maupun data uji. Pembagian data menggunakan random_state = 42 agar eksperimen dapat direproduksi.

3.3 Pemodelan CatBoost

CatBoost digunakan sebagai model utama dalam penelitian ini. CatBoost merupakan algoritma gradient boosting berbasis decision tree yang dirancang untuk menghasilkan performa yang baik pada data tabular dan memiliki mekanisme khusus dalam menangani atribut kategorikal [16], [25]. Dalam penelitian ini, CatBoost digunakan untuk mempelajari pola transaksi yang membedakan kelas fraud dan non-fraud pada dataset PaySim.

Pemodelan CatBoost dilakukan dalam dua konfigurasi, yaitu CatBoost Baseline dan CatBoost Tuned. CatBoost Baseline digunakan untuk memperoleh performa awal model sebelum dilakukan penyesuaian parameter, sedangkan CatBoost Tuned digunakan untuk mengetahui pengaruh penyesuaian hyperparameter terhadap kemampuan model dalam mendeteksi transaksi fraud.

3.3.1 CatBoost Baseline

CatBoost Baseline dibangun menggunakan konfigurasi awal sebagai acuan performa sebelum dilakukan proses *hyperparameter tuning*. Model dilatih menggunakan data latih yang telah melalui tahap *preprocessing* dan pembagian data dengan rasio 80:20 secara *stratified*. Untuk menangani ketidakseimbangan kelas, model menggunakan pendekatan class weighting, sehingga kelas fraud memperoleh bobot yang lebih besar selama proses pembelajaran.

CatBoost Baseline kemudian menghasilkan probabilitas prediksi untuk masing-masing kelas menggunakan fungsi predict_proba. Probabilitas tersebut digunakan dalam proses evaluasi, khususnya untuk menghitung ROC-AUC dan PR-AUC, karena kedua metrik tersebut membutuhkan informasi probabilitas pada berbagai nilai *classification threshold*.

Hasil pengujian CatBoost Baseline digunakan sebagai baseline performance untuk mengetahui apakah proses penyesuaian *hyperparameter* pada tahap berikutnya mampu meningkatkan performa model.

3.3.2 CatBoost Tuned

Setelah CatBoost Baseline diperoleh, dilakukan hyperparameter tuning untuk menyesuaikan model dengan karakteristik dataset PaySim, khususnya dalam mendeteksi kelas fraud minoritas. Parameter yang disesuaikan meliputi *iterations*, *depth*, *learning_rate*, *class_weights*, dan *l2_leaf_reg*. Tuning dilakukan secara manual dengan beberapa kombinasi parameter, dilatih pada data latih yang sama, lalu dievaluasi pada data uji untuk memilih konfigurasi terbaik. CatBoost Tuned kemudian dibandingkan dengan CatBoost Baseline menggunakan Accuracy, Precision, Recall, F1-Score, ROC-AUC, PR-AUC, serta jumlah FP dan FN. Hasilnya, tuning meningkatkan Recall dari 85,27% menjadi 92,08%, ROC-AUC dari 99,92% menjadi 99,96%, dan PR-AUC dari 90,93% menjadi 94,46%, serta menurunkan False Negative dari 242 menjadi 130 transaksi. Namun, Precision menurun dari 79,19% menjadi 71,46% dan F1-Score dari 82,12% menjadi 80,48%. Dengan demikian, CatBoost Tuned lebih sensitif dalam mendeteksi fraud, tetapi konsekuensinya prediksi positif tidak tepat meningkat, dan model ini digunakan sebagai model hasil optimasi untuk dibandingkan dengan CatBoost Baseline serta Random Forest.

3.4 Pemodelan Random Forest

Random Random Forest digunakan sebagai model pembanding (baseline) dalam penelitian ini; algoritma ensemble learning berbasis bagging ini membangun sejumlah decision tree menggunakan sampel data dan subset fitur secara acak, lalu menggabungkan prediksi seluruh pohon untuk menghasilkan prediksi akhir [24]. Model dilatih dengan data latih yang sama seperti CatBoost, dan untuk menangani ketidakseimbangan kelas diterapkan parameter class_weight = 'balanced' agar kelas fraud memperoleh bobot lebih besar. Konfigurasi yang digunakan adalah n_estimators = 50, max_depth = 10, class_weight = 'balanced', dan random_state = 42, di mana n_estimators menentukan jumlah pohon, max_depth membatasi kedalaman pohon untuk mengurangi overfitting, dan random_state menjaga konsistensi eksperimen. Random Forest selanjutnya dibandingkan dengan CatBoost Baseline dan CatBoost Tuned untuk melihat perbedaan performa ensemble berbasis bagging dan boosting pada dataset PaySim.

3.5 Evaluasi

Evaluasi dilakukan setelah pelatihan ketiga model, yaitu Random Forest, CatBoost Baseline, dan CatBoost Tuned, menggunakan data uji yang sama agar perbandingan objektif. Metrik yang digunakan meliputi Accuracy, Precision, Recall, F1-Score, ROC-AUC, PR-AUC, serta Confusion Matrix (TP, TN, FP, FN). Accuracy mengukur proporsi prediksi benar; Precision mengukur ketepatan prediksi fraud; Recall mengukur kemampuan menemukan fraud; F1-Score menyeimbangkan Precision dan Recall; ROC-AUC menilai diskriminasi model pada berbagai threshold melalui predict_proba; dan PR-AUC digunakan karena dataset PaySim sangat tidak seimbang sehingga lebih fokus pada kelas fraud minoritas. Analisis FP dan FN, terutama FN, penting untuk mengetahui kesalahan deteksi fraud. Penentuan model terbaik tidak hanya berdasarkan Accuracy dan ROC-AUC, tetapi juga Precision, Recall, F1-Score, PR-AUC, serta jumlah FP dan FN agar evaluasi kemampuan deteksi fraud lebih menyeluruh.

4. HASIL DAN PEMBAHASAN

4.1 Pengumpulan Data

Dataset yang digunakan dalam penelitian ini adalah PaySim, yaitu dataset simulasi transaksi *mobile money* yang diperoleh dari platform Kaggle. Dataset digunakan sebagai representasi transaksi pada sistem pembayaran digital dan telah dilengkapi dengan label isFraud yang menunjukkan apakah suatu transaksi termasuk fraud atau non-fraud [31].

Dataset awal terdiri atas 6.362.620 transaksi dan 11 atribut. Berdasarkan hasil pemeriksaan distribusi kelas, terdapat 8.213 transaksi fraud atau sekitar 0,13% dari keseluruhan transaksi. Kondisi tersebut menunjukkan bahwa dataset memiliki ketidakseimbangan kelas yang sangat tinggi, dengan transaksi non-fraud sebagai kelas mayoritas.

Karakteristik tersebut menjadi salah satu pertimbangan utama dalam pemodelan karena model tidak cukup hanya memiliki kemampuan menghasilkan prediksi keseluruhan yang tinggi, tetapi juga harus mampu mengenali transaksi fraud sebagai kelas minoritas.

Tabel 1 Deskripsi Dataset

No	Nama Atribut	Deskripsi	Relasi terhadap Label Fraud
1	step	Waktu transaksi	Digunakan untuk melihat pola waktu transaksi
2	type	Jenis transaksi	Memiliki pengaruh besar terhadap fraud karena fraud hanya muncul pada transaksi tertentu
3	amount	Nominal transaksi	Nominal besar berpotensi menunjukkan aktivitas mencurigakan
4	nameOrig	Identitas pengirim	Tidak digunakan karena hanya berupa identifier
5	oldbalanceOrg	Saldo awal pengirim	Digunakan untuk mendeteksi perubahan saldo abnormal
6	newbalanceOrig	Saldo akhir pengirim	Sangat berpengaruh terhadap pola fraud
7	nameDest	Identitas penerima	Tidak digunakan karena hanya berupa identifier
8	oldbalanceDest	Saldo awal penerima	Digunakan untuk melihat perubahan saldo penerima
9	newbalanceDest	Saldo akhir penerima	Digunakan untuk analisis pola transaksi
10	isFraud	Label fraud transaksi	Target klasifikasi penelitian
11	isFlaggedFraud	Indikator transaksi mencurigakan	Memiliki pengaruh kecil terhadap model

4.2 Preprocessing

Tahap *preprocessing* dilakukan untuk menyiapkan dataset sebelum digunakan dalam proses pemodelan. Proses ini mencakup pemeriksaan kualitas data, penghapusan atribut yang tidak digunakan, transformasi data kategorikal, pemisahan fitur dan label, serta pembagian dataset menjadi data latih dan data uji.

Pada tahap ini, atribut nameOrig dan nameDest tidak digunakan dalam proses pemodelan karena keduanya berfungsi sebagai identifier pengirim dan penerima transaksi. Atribut type yang merupakan data kategorikal kemudian ditransformasikan menjadi bentuk numerik menggunakan Label Encoding.

Setelah proses tersebut, atribut isFraud digunakan sebagai variabel target, sedangkan atribut transaksi lainnya yang dipertahankan digunakan sebagai variabel prediktor.

Karena dataset memiliki ketidakseimbangan kelas yang tinggi, proses pelatihan model menggunakan pendekatan class weighting. Pendekatan ini memberikan perhatian lebih besar terhadap kelas fraud tanpa melakukan perubahan terhadap distribusi asli dataset.

4.3 Pembagian Data

Dataset dibagi menjadi data training dan testing menggunakan metode *train-test split* dengan proporsi 80:20. Pembagian dilakukan menggunakan parameter stratify sehingga proporsi kelas fraud dan non-fraud tetap dipertahankan pada data training maupun testing.

Penggunaan *stratified split* penting karena dataset memiliki ketidakseimbangan kelas yang sangat tinggi. Tanpa mempertahankan proporsi kelas, terdapat risiko distribusi kelas pada data training dan testing tidak merepresentasikan distribusi dataset secara keseluruhan.

Parameter random_state = 42 digunakan untuk menjaga konsistensi pembagian data sehingga setiap model memperoleh data training dan testing yang sama.

Hal tersebut menjadi penting dalam penelitian komparatif karena perbedaan performa yang diperoleh dapat lebih diarahkan pada perbedaan algoritma dan konfigurasi model, bukan karena model mendapatkan pembagian dataset yang berbeda.

4.4 Pemodelan CatBoost

Pemodelan CatBoost dilakukan dalam dua konfigurasi, yaitu CatBoost Baseline dan CatBoost Tuned. Kedua model menggunakan data training yang sama dan dievaluasi menggunakan data testing yang sama.

4.4.1 CatBoost Baseline

CatBoost Baseline digunakan sebagai model awal untuk memperoleh gambaran performa CatBoost sebelum dilakukan proses *hyperparameter tuning*.

Model menggunakan *class weighting* untuk memberikan perhatian lebih terhadap kelas fraud. Setelah proses pelatihan, model menghasilkan probabilitas prediksi yang digunakan untuk menghitung ROC-AUC dan PR-AUC.

Berdasarkan hasil eksperimen, CatBoost Baseline menghasilkan:

Tabel 2 Hasil Pemodelan CatBoost

Metrik	CatBoost Baseline
Precision	**79,19%**
Recall	**85,27%**
F1-Score	**82,12%**
ROC-AUC	**99,92%**
PR-AUC	**90,93%**

Hasil tersebut menunjukkan bahwa CatBoost Baseline telah mampu memberikan performa yang baik dalam membedakan transaksi fraud dan non-fraud. Nilai Precision sebesar 79,19% menunjukkan bahwa sebagian besar transaksi yang diprediksi sebagai fraud memang merupakan transaksi fraud.

Sementara itu, Recall sebesar 85,27% menunjukkan bahwa model mampu mendeteksi sebagian besar transaksi fraud yang terdapat dalam data testing.

4.4.2 CatBoost Tuned

CatBoost Tuned merupakan model CatBoost yang telah melalui proses hyperparameter tuning. Tuning dilakukan untuk memperoleh konfigurasi parameter yang lebih sesuai dengan karakteristik dataset PaySim.

Parameter yang disesuaikan meliputi iterations, depth, learning_rate, class_weights, dan l2_leaf_reg. Proses tuning dilakukan melalui beberapa percobaan konfigurasi secara manual.

Hasil pengujian CatBoost Tuned adalah sebagai berikut:

Tabel 3 Hasil Pemodelan CatBoost Tuned

Metrik	CatBoost Tuned
Precision	71,46%
Recall	92,08%
F1-Score	80,48%
ROC-AUC	99,96%
PR-AUC	94,46%

Dibandingkan dengan CatBoost Baseline, CatBoost Tuned mengalami peningkatan Recall dari 85,27% menjadi 92,08%. Peningkatan tersebut menunjukkan bahwa tuning membuat model lebih sensitif dalam mengenali transaksi fraud.

ROC-AUC juga meningkat dari 99,92% menjadi 99,96%, sedangkan PR-AUC meningkat dari 90,93% menjadi 94,46%. Peningkatan PR-AUC menjadi temuan penting karena metrik tersebut lebih berfokus pada hubungan Precision dan Recall pada kelas positif.

Namun, peningkatan Recall disertai dengan penurunan Precision dari 79,19% menjadi 71,46% dan F1-Score dari 82,12% menjadi 80,48%. Kondisi tersebut menunjukkan adanya *trade-off* antara kemampuan menangkap lebih banyak fraud dan ketepatan prediksi fraud.

Selain itu, jumlah False Negative turun dari 242 menjadi 130 transaksi. Dengan demikian, CatBoost Tuned lebih mampu mengurangi transaksi fraud yang lolos dari deteksi.

4.5 Pemodelan Random Forest

Random Forest digunakan sebagai model pembanding terhadap CatBoost. Model menggunakan konfigurasi:

- n_estimators = 50
- max_depth = 10
- class_weight = 'balanced'
- random_state = 42

Penggunaan class_weight='balanced' dilakukan untuk memberikan bobot lebih besar terhadap kelas fraud.

Hasil pengujian Random Forest adalah sebagai berikut:

Tabel 4 Hasil Pemodelan Random Forest

Metrik	Nilai
Accuracy	98,56%
Precision	8,15%
Recall	98,78%
F1-Score	15,06%
ROC-AUC	99,91%

4.6 Analisis ROC Curve dan ROC-AUC

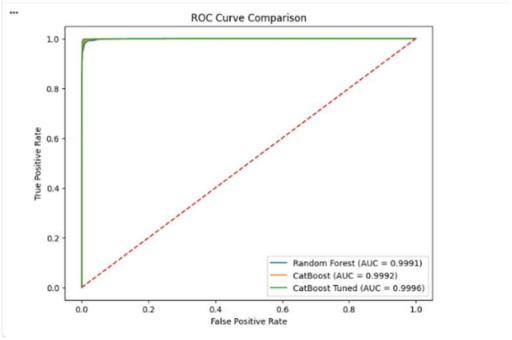
Evaluasi ROC-AUC dilakukan untuk mengetahui kemampuan masing-masing model dalam membedakan transaksi fraud dan non-fraud pada berbagai *classification threshold*. ROC Curve menggambarkan hubungan antara True Positive Rate dan False Positive Rate.

Berdasarkan hasil visualisasi ROC Curve, ketiga model menghasilkan kurva yang berada sangat dekat dengan area kiri atas grafik. Nilai ROC-AUC yang diperoleh adalah:

- Random Forest = 99,91%
- CatBoost Baseline = 99,92%
- CatBoost Tuned = 99,96%

CatBoost Tuned menghasilkan nilai ROC-AUC tertinggi. Hasil tersebut menunjukkan bahwa setelah dilakukan *hyperparameter tuning*, kemampuan model dalam memberikan pemisahan atau *ranking* antara transaksi fraud dan non-fraud mengalami peningkatan.

Namun, selisih ROC-AUC ketiga model relatif kecil. Oleh karena itu, ROC-AUC saja tidak cukup untuk menjelaskan perbedaan performa model pada kelas fraud. Hal tersebut menjadi alasan penggunaan PR-AUC dan analisis Confusion Matrix sebagai evaluasi tambahan.



Gambar 2. ROC Curve Comparison

Gambar menunjukkan bahwa ketiga kurva berada sangat dekat dengan nilai TPR = 1 pada tingkat FPR yang relatif rendah. Hal tersebut konsisten dengan nilai ROC-AUC yang sangat tinggi.

4.7Analisis Precision-Recall Curve dan PR-AUC

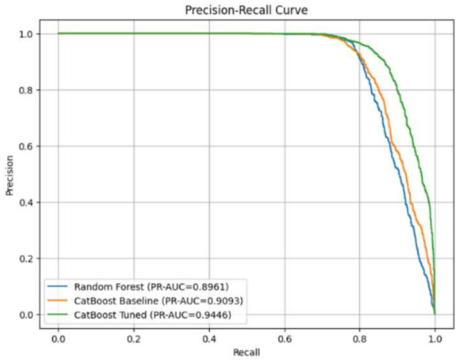
Selain ROC-AUC, penelitian menggunakan Precision-Recall Curve dan PR-AUC karena dataset PaySim memiliki ketidakseimbangan kelas yang sangat tinggi.

Hasil PR-AUC adalah:

- Random Forest = 89,61%
- CatBoost Baseline = 90,93%
- CatBoost Tuned = 94,46%

CatBoost Tuned menghasilkan PR-AUC tertinggi dibandingkan kedua model lainnya. Nilai tersebut menunjukkan bahwa CatBoost Tuned memiliki kemampuan yang lebih baik dalam mempertahankan hubungan antara Precision dan Recall pada berbagai *classification threshold*.

Peningkatan PR-AUC dari 90,93% pada CatBoost Baseline menjadi 94,46% pada CatBoost Tuned menunjukkan bahwa peningkatan performa setelah tuning tidak hanya terlihat pada ROC-AUC, tetapi juga pada evaluasi yang lebih berfokus terhadap kelas positif.



Gambar 3. Precision-Recall Curve Comparison

Berdasarkan grafik, kurva CatBoost Tuned berada relatif lebih tinggi dibandingkan Random Forest dan CatBoost Baseline, terutama ketika nilai Recall berada pada rentang tinggi. Hal tersebut menunjukkan kemampuan CatBoost Tuned dalam mempertahankan Precision yang lebih baik ketika model berusaha mendeteksi lebih banyak transaksi fraud.

4.8 Analisis Potensi Information Leakage

Berdasarkan hasil evaluasi awal, Random Forest dan CatBoost menghasilkan nilai ROC-AUC yang sangat tinggi, yaitu berada pada kisaran 0,99. Nilai tersebut menimbulkan pertanyaan mengenai kemungkinan adanya *information leakage*, yaitu kondisi ketika model memperoleh informasi yang secara konseptual belum seharusnya tersedia pada saat prediksi sehingga dapat menghasilkan performa yang terlalu optimis. Oleh karena itu, dilakukan beberapa pengujian tambahan untuk mengetahui apakah nilai ROC-AUC yang tinggi tersebut didominasi oleh atribut tertentu atau merupakan karakteristik dari dataset PaySim.

4.8.1 Penghapusan Fitur *newbalanceOrig* dan *newbalanceDest*

Pengujian pertama dilakukan dengan menghapus atribut newbalanceOrig dan newbalanceDest. Kedua atribut tersebut merepresentasikan saldo setelah transaksi diproses. Dalam konteks deteksi fraud secara *real-time*, kedua informasi tersebut secara konseptual dapat dipertanyakan ketersediaannya pada saat keputusan prediksi dilakukan sehingga diuji sebagai fitur yang berpotensi menimbulkan *information leakage*.

Setelah kedua fitur tersebut dihapus, ketiga model dilatih kembali menggunakan konfigurasi yang sama. Hasil pengujian ditunjukkan pada Tabel berikut.

Tabel 5 Hasil Pengujian Setelah Penghapusan newbalanceOrig dan newbalanceDest

Model	**ROC-AUC**
Random Forest	0,9980
CatBoost	0,9977
CatBoost Tuned	0,9992

4.8.2 Penghapusan fitur IsFlaggedFraud

Pengujian kedua dilakukan dengan menghapus atribut isFlaggedFraud. Atribut tersebut merupakan indikator yang berkaitan dengan penandaan transaksi yang dianggap mencurigakan sehingga diuji secara terpisah untuk mengetahui pengaruhnya terhadap performa model.

Setelah atribut tersebut dihapus, ketiga model dilatih dan dievaluasi kembali. Hasil pengujian ditunjukkan pada Tabel berikut.

Tabel 6 Penghapusan Fitur isFlaggedFraud

Model	**ROC-AUC**

Random Forest	0,9981
CatBoost	0,9978
CatBoost Tuned	0,9993

4.8.3 Single-Feature Evaluation

Untuk memperoleh pemeriksaan tambahan, dilakukan single-feature evaluation, yaitu pengujian dengan menggunakan satu fitur pada setiap proses pelatihan. Pengujian ini bertujuan untuk mengetahui apakah terdapat satu atribut tertentu yang secara individual mampu menghasilkan performa klasifikasi yang sangat tinggi dan berpotensi menjadi sumber *information leakage*.

Pada Random Forest, hasil pengujian menunjukkan nilai ROC-AUC tertinggi diperoleh oleh oldbalanceOrg sebesar 0,8688, diikuti oleh type sebesar 0,8529, step sebesar 0,8524, amount sebesar 0,8181, dan oldbalanceDest sebesar 0,6341.

Sementara itu, pada CatBoost, nilai ROC-AUC tertinggi diperoleh oleh fitur type sebesar 0,8529, diikuti oldbalanceOrg sebesar 0,8492, step sebesar 0,8410, amount sebesar 0,7873, dan oldbalanceDest sebesar 0,6337.

Hasil tersebut menunjukkan bahwa tidak terdapat satu fitur yang secara individual menghasilkan ROC-AUC mendekati 1,00. Nilai tertinggi pada pengujian single-feature hanya mencapai 0,8688 pada fitur *oldbalanceOrg*. Oleh karena itu, performa tinggi yang diperoleh model multivariat lebih tepat dipahami sebagai hasil dari kombinasi beberapa atribut yang saling melengkapi dalam proses klasifikasi, bukan karena satu fitur tertentu saja.

4.8.4 Final Result

Berdasarkan seluruh pengujian yang dilakukan, yaitu penghapusan newbalanceOrig, newbalanceDest, dan isFlaggedFraud, evaluasi ulang model, serta *single-feature evaluation*, tidak ditemukan indikasi kuat bahwa tingginya nilai ROC-AUC pada penelitian ini didominasi oleh satu atribut tertentu.

Penghapusan fitur yang diduga berpotensi menimbulkan *information leakage* tidak menyebabkan penurunan ROC-AUC secara drastis, sedangkan pengujian *single-feature* menunjukkan bahwa tidak terdapat satu fitur yang secara individual mampu menghasilkan performa mendekati model multivariat. Dengan demikian, hasil pengujian mendukung interpretasi bahwa performa tinggi diperoleh dari kombinasi informasi antarfitur dan karakteristik dataset PaySim.

Meskipun demikian, hasil tersebut tidak digunakan sebagai bukti mutlak bahwa seluruh bentuk *information leakage* telah dieliminasi. Secara khusus, atribut newbalanceOrig dan newbalanceDest merepresentasikan kondisi saldo setelah transaksi sehingga penggunaannya perlu dikaitkan dengan definisi waktu prediksi apabila model akan diterapkan pada sistem deteksi fraud secara *real-time*.

5. KESIMPULAN DAN SARAN

Berdasarkan hasil penelitian, algoritma CatBoost dan Random Forest mampu mendeteksi transaksi fraud pada dataset PaySim yang memiliki ketidakseimbangan kelas. Random Forest menghasilkan Recall tertinggi sebesar 98,78%, tetapi memiliki Precision yang rendah sebesar 8,15% dan menghasilkan *False Positive* sebanyak 18.284 transaksi. CatBoost Baseline memberikan keseimbangan yang lebih baik dengan Precision 79,19%, Recall 85,27%, dan F1-Score 82,12%. Setelah dilakukan *hyperparameter tuning*, CatBoost Tuned meningkatkan Recall menjadi 92,08%, ROC-AUC 99,96%, dan PR-AUC 94,46%, serta menurunkan *False Negative* dari 242 menjadi 130 transaksi. Dengan demikian, CatBoost Tuned memiliki kemampuan deteksi fraud yang lebih kuat,

khususnya dalam mengurangi transaksi fraud yang tidak terdeteksi, meskipun terjadi penurunan Precision dan F1-Score.

Analisis potensi *information leakage* melalui penghapusan beberapa fitur dan *single-feature evaluation* tidak menunjukkan indikasi kuat bahwa performa tinggi model hanya disebabkan oleh satu fitur tertentu. Untuk penelitian selanjutnya, disarankan menerapkan temporal validation, melakukan *hyperparameter tuning* secara lebih sistematis, mengoptimalkan *classification threshold*, serta mengembangkan *feature engineering* berbasis pola transaksi historis. Pengujian pada dataset transaksi fraud lainnya juga diperlukan untuk mengetahui kemampuan generalisasi model pada kondisi yang lebih mendekati penerapan nyata.

DAFTAR REFERENSI

[1] Gomber, P., Kauffman, R. J., Parker, C., & Weber, B. W. (2018). On the Fintech Revolution: Interpreting the Forces of Innovation, Disruption, and Transformation in Financial Services. Journal of Management Information Systems, 35(1), 220–265.

[2] Arner, D. W., Barberis, J., & Buckley, R. P. (2016). The Evolution of Fintech: A New Post-Crisis Paradigm? Georgetown Journal of International Law, 47(4), 1271–1319.

[3] Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). A Comprehensive Survey of Data Mining-Based Fraud Detection Research. Artificial Intelligence Review, 34(1), 1–14.

[4] Dal Pozzolo, A., Caelen, O., Le Borgne, Y.-A., Waterschoot, S., & Bontempi, G. (2014). Learned Lessons in Credit Card Fraud Detection from a Practitioner Perspective. Expert Systems with Applications.

[5] Mitchell, T. M. (1997). Machine Learning. New York: McGraw-Hill.

[6] Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32.

[7] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased Boosting with Categorical Features. Advances in Neural Information Processing Systems (NeurIPS), 31.

[8] Lopez-Rojas, E., & Axelsson, S. (2016). PaySim: A Financial Mobile Money Simulator for Fraud Detection.

[9] Gupta, P., Varshney, A., Khan, M. R., Ahmed, R., Shuaib, M., & Alam, S. (2023). Unbalanced Credit Card Fraud Detection Data: A Machine Learning-Oriented Comparative Study of Balancing Techniques. Procedia Computer Science, 218, 2575–2584.

[10] Chawla, N. V. (2009). Data Mining for Imbalanced Datasets: An Overview. In Data Mining and Knowledge Discovery Handbook. Springer.

[11] He, H., & Garcia, E. A. (2009). Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284.

[12] Fawcett, T. (2006). An Introduction to ROC Analysis. Pattern Recognition Letters, 27(8), 861–874.

[13] Dorogush, A. V., Ershov, V., & Gulin, A. (2018). CatBoost: Gradient Boosting with Categorical Features Support. arXiv:1810.11363.

[14] Géron, A. (2022). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow (3rd ed.). O'Reilly Media.

[15] Ramineni, P., & Mastouri, R. (2025). Credit Card Fraud Detection Using CatBoost.

[16] CatBoost, XGBoost, and LightGBM: A Comparative Study for Financial Fraud Detection. (2025). Preprint.
[17] Assabil, J. J., & Obagbuwa, I. C. Credit Card Fraud Detection Using Machine Learning Algorithms: A Comparative Study of Six Models. International Journal of Intelligent Systems and Applications in Engineering.
[18] Sopiyan, M., Fauziah, F., & Wijaya, Y. F. Fraud Detection Using Random Forest, Logistic Regression, and Gradient Boosting. JUITA.
[19] Handayani, K. et al. Analisis Kinerja Logistic Regression dan Random Forest pada Deteksi Fraud E-Commerce Menggunakan SMOTE & PCA. Jurnal Sistem Informasi Akuntansi.
[20] Whitrow, C., Hand, D. J., Juszczak, P., Weston, D., & Adams, N. M. (2009). Transaction Aggregation as a Strategy for Credit Card Fraud Detection. Data Mining and Knowledge Discovery, 18(1), 30–55.
[21] Nguyen, N. et al. (2022). A Proposed Model for Card Fraud Detection Based on CatBoost and Deep Neural Network. IEEE Access.
[22] Almalki, F., & Masud, M. (2025). Financial Fraud Detection Using Explainable AI and Stacking Ensemble Methods.
[23] Wang, Y. (2025). A Data Balancing and Ensemble Learning Approach for Credit Card Fraud Detection. Preprint.
[24] Yu, S., Chang, V., Huỳnh, G. L., Jesus, V., & Luo, J. (2025). Advanced Supervised Machine Learning Algorithms in Credit Card Fraud Detection. SciTePress (Procedia).
[25] Optimizing Credit Card Fraud Detection with Random Forests and SMOTE. (2025). Scientific Reports.
[26] Phua, C. R. S. et al. (2025). Credit Card Fraud Detection Using Ensemble Variation: Logistic Regression, SVC, and Random Forest.
[27] Ibrahim, M. M., & Alfauzan, S. (2026). Analisis Kinerja Model Machine Learning untuk Mendeteksi Transaksi Fraud pada Sistem Pembayaran Online.
[28] Icha, N. N., Lopes, D. S. P., & Hanafi, M. (2024). Optimasi Deteksi Penipuan Kartu Kredit Menggunakan Regresi Logistik.
[29] Carcillo, F., Le Borgne, Y. A., Caelen, O., Bontempi, G., et al. (2021). Combining Unsupervised and Supervised Learning in Credit Card Fraud Detection. Information Sciences.
[30] Aggarwal, C. C. (2017). Outlier Analysis (2nd ed.). Springer.
[31] Chicco, D., & Jurman, G. (2020). The Advantages of the Matthews Correlation Coefficient (MCC) over F1 Score and Accuracy in Binary Classification Evaluation. BMC Genomics, 21(6).
[32] GSMA. (2024). State of the Industry Report on Mobile Money 2024.
[33] Worldpay. (2024). Global Payments Report 2024.
[34] Bank Indonesia. (2024). Statistik Sistem Pembayaran Indonesia 2024.