

Analisis Empiris Algoritma Gradient Boosting untuk Deteksi Fraud pada Ketidakseimbangan Kelas Ekstrem

Fahmi Izhari*¹; Rusnai Rahayu²; Hasan al Fachri Satia Simbolon³; Ade Linhar P⁴; Rafi Septiawan Putra⁵

^{1,2,3,4,5}Teknologi Informasi, UIN Syekh Ali Hasan Ahmad Addary
Padangsidempuan

fahmi_izhari@uinsyahada.ac.id¹; rusnairahayu@uinsyahada.ac.id²;
hasanalfachri@uinsyahada.ac.id³; adelinhar@uinsyahada.ac.id⁴;
rafiputra95@uinsyahada.ac.id⁵

Abstract

Detecting fraudulent financial transactions requires robust machine learning models capable of handling highly imbalanced tabular data. In real-world scenarios, the ratio of legitimate to fraudulent transactions often exceeds 1:500, rendering standard accuracy metrics deceptive due to the accuracy paradox. This study presents a rigorous empirical benchmarking of three state-of-the-art Gradient Boosting Decision Tree (GBDT) algorithms—XGBoost, CatBoost, and LightGBM—on the standard ULB Credit Card Fraud dataset comprising 284,807 transactions. To address the extreme class imbalance (0.17% fraud rate), mathematical penalty adjustments, specifically `scale_pos_weight` and `balanced_auto_class_weights`, were integrated into the objective functions of the models. Experimental results demonstrate that XGBoost achieved the highest discriminative power with an ROC-AUC of 0.9812 and a Recall of 0.8673, while maintaining the lowest computational latency of 3.92 seconds. CatBoost and LightGBM also exhibited exceptional recall capabilities, exceeding 0.88, validating the efficacy of algorithmic weight penalization over traditional resampling techniques. The analysis confirms that XGBoost provides the optimal trade-off between predictive robustness and computational efficiency for realtime, large-scale financial fraud detection systems.

Keywords: *fraud detection; gradient boosting; class imbalance; XGBoost; machine learning.*

Abstrak

Mendeteksi transaksi keuangan yang curang memerlukan model *machine learning* yang kuat dan mampu menangani data tabular yang sangat tidak seimbang. Dalam skenario dunia nyata, rasio transaksi legitimasi terhadap penipuan sering kali melebihi 1:500, menjadikan metrik akurasi standar menyesatkan akibat adanya paradoks akurasi. Penelitian ini menyajikan *benchmarking* empiris yang ketat dari tiga algoritma *Gradient Boosting Decision Tree* (GBDT) terdepan XGBoost, CatBoost, dan LightGBM pada dataset standar ULB Credit Card Fraud yang terdiri dari 284.807 transaksi. Untuk mengatasi ketidakseimbangan kelas yang ekstrem (tingkat fraud 0,17%), penyesuaian penalti matematis, khususnya `scale_pos_weight` dan `balanced_auto_class_weights`, diintegrasikan ke dalam fungsi objektif model. Hasil eksperimen menunjukkan bahwa XGBoost mencapai daya diskriminatif tertinggi dengan ROC-AUC sebesar 0,9812 dan Recall sebesar 0,8673, sembari

mempertahankan latensi komputasi terendah yaitu 3,92 detik. CatBoost dan LightGBM juga menunjukkan kapabilitas *recall* yang luar biasa, melebihi 0,88, memvalidasi kemanjuran penaltis bobot algoritmik dibandingkan teknik *resampling* tradisional. Analisis ini mengonfirmasi bahwa XGBoost memberikan *trade-off* optimal antara ketahanan prediktif dan efisiensi komputasi untuk sistem deteksi fraud keuangan berskala besar secara *realtime*.

Kata Kunci: deteksi *fraud*; *gradient boosting*; ketidakseimbangan kelas; XGBoost; *machine learning*.

PENDAHULUAN

Deteksi penipuan keuangan merepresentasikan tantangan kritis dalam sains data modern, yang dicirikan oleh volume transaksi yang masif, fitur tabular berdimensi tinggi, dan ketidakseimbangan kelas yang parah (Hilal dkk., 2021; Yang dkk., 2026). Biaya dari *false negative* (kegagalan mendeteksi penipuan) secara signifikan lebih besar daripada *false positive*, sehingga mengharuskan algoritma untuk memaksimalkan *recall* sembari mempertahankan skalabilitas komputasi. Kondisi tersebut menjadikan deteksi *fraud* berbeda dari klasifikasi konvensional karena keberhasilan model tidak cukup ditentukan oleh jumlah prediksi yang benar secara keseluruhan, tetapi terutama oleh kemampuannya mengenali transaksi fraud yang jumlahnya relatif sangat sedikit.

Selama satu dekade terakhir, *Gradient Boosting Decision Trees* (GBDT), khususnya XGBoost, CatBoost, dan LightGBM, telah menetapkan standar *state of the art* untuk data tabular. Algoritma-algoritma ini secara iteratif membangun ansambel pohon keputusan untuk meminimalkan fungsi kerugian yang dapat didiferensiasi, secara efektif menangkap interaksi non-linear yang kompleks antar fitur transaksi (Faqih & Sugihartono, 2025; Rizky dkk., 2022; Saputra dkk., 2025). Kemampuan tersebut membuat GBDT banyak digunakan pada permasalahan klasifikasi yang memiliki hubungan kompleks antarvariabel. Namun, performa algoritma tetap dipengaruhi oleh distribusi kelas yang digunakan selama proses pembelajaran.

Meskipun memiliki keunggulan prediktif, model GBDT menghadapi kesulitan matematis saat berhadapan dengan distribusi kelas yang sangat timpang (Pristyanto dkk., 2023). Pada dataset keuangan dunia nyata, transaksi *fraud* sering kali hanya mencakup kurang dari 1% dari total data (Anh Luong & Xie, 2026). Ketidakseimbangan tersebut menyebabkan kontribusi kelas mayoritas jauh lebih

dominan dalam proses optimasi model. Akibatnya, model dapat memperoleh performa keseluruhan yang tinggi meskipun masih gagal mendeteksi sebagian besar transaksi fraud yang menjadi perhatian utama sistem.

Dalam kondisi ini, model naif yang memprediksi semua transaksi sebagai normal akan tetap menghasilkan akurasi di atas 99%, sebuah fenomena yang dikenal sebagai Paradoks Akurasi (*Accuracy Paradox*) (Martínez dkk., 2025). Oleh karena itu, metrik evaluasi standar menjadi tidak relevan, dan algoritma harus dimodifikasi secara matematis untuk memberikan penalti yang lebih berat pada kesalahan klasifikasi kelas minoritas (Kraiem dkk., 2021). Pendekatan tradisional seperti *oversampling* (SMOTE) atau *undersampling* sering kali menimbulkan *overfitting* atau menghilangkan informasi berharga dari kelas mayoritas (Zhao dkk., 2026). Kondisi ini menunjukkan bahwa penanganan ketidakseimbangan kelas perlu mempertimbangkan karakteristik data sekaligus mekanisme pembelajaran algoritma.

Sebagai alternatif, modifikasi langsung pada fungsi objektif algoritma GBDT melalui penaltis bobot kelas menawarkan solusi yang lebih elegan secara matematis. XGBoost dan LightGBM memanfaatkan parameter *scale_pos_weight* untuk menyeimbangkan gradien, sementara CatBoost menggunakan mekanisme *ordered boosting* dan *auto_class_weights* untuk memitigasi bias prediksi (Daswati dkk., 2026; Hancock & Khoshgoftaar, 2020; Nugraha & Syarif, 2023). Pendekatan ini memungkinkan seluruh data tetap digunakan dalam proses pembelajaran, tetapi kesalahan terhadap kelas *fraud* diberikan pengaruh yang lebih besar terhadap proses optimasi. Dengan demikian, perubahan komposisi data akibat *resampling* dapat dihindari sekaligus mempertahankan informasi transaksi mayoritas.

Meskipun ketiga algoritma ini telah divalidasi secara luas, perbandingan empiris yang ketat mengenai *trade off* antara metrik *Recall*, *Precision*, dan biaya komputasi pada dataset dengan rasio ketidakseimbangan ekstrem (di atas 1:500) masih perlu dieksplorasi lebih mendalam. Perbedaan mekanisme internal XGBoost, CatBoost, dan LightGBM memungkinkan munculnya karakteristik performa yang berbeda ketika diterapkan pada kondisi ketidakseimbangan ekstrem. Oleh sebab itu, evaluasi tidak hanya perlu mempertimbangkan kemampuan model dalam

mendeteksi kelas *fraud*, tetapi juga hubungan antara kemampuan tersebut dengan *Precision*, *F1-Score*, *ROC-AUC*, dan waktu komputasi.

Penelitian ini menjembatani kesenjangan tersebut dengan menyediakan analisis komparatif antara XGBoost, CatBoost, dan LightGBM pada dataset standar ULB *Credit Card Fraud* yang memiliki rasio ketidakseimbangan 1:577. Penggunaan dataset dengan proporsi *fraud* yang sangat rendah memberikan lingkungan pengujian yang sesuai untuk mengevaluasi kemampuan algoritma dalam mengenali kelas minoritas tanpa mengubah struktur data melalui teknik *resampling*. Dengan kondisi eksperimen yang sama, perbedaan kinerja ketiga algoritma dapat dianalisis berdasarkan kemampuan prediktif dan efisiensi komputasinya.

Kontribusi utama dari penelitian ini adalah evaluasi mendalam tentang bagaimana penalitis bobot matematis memaksa batas keputusan (*decision boundary*) model untuk menangkap pola *fraud* yang jarang, serta analisis latensi komputasi untuk menentukan kelayakan penerapannya pada sistem pemrosesan transaksi *realtime*. Dengan demikian, penelitian tidak hanya membandingkan nilai metrik masing-masing algoritma, tetapi juga memberikan gambaran mengenai *trade off* antara kemampuan mendeteksi *fraud* dan kebutuhan komputasi. Hasil analisis tersebut menjadi dasar untuk memahami karakteristik GBDT dalam menghadapi ketidakseimbangan kelas yang ekstrem.

METODE PENELITIAN

Penelitian ini menggunakan dataset standar ULB *Credit Card Fraud Detection* yang berisi 284.807 transaksi kartu kredit, di mana hanya 492 transaksi (0,17%) yang berlabel *fraud*, menghasilkan rasio ketidakseimbangan kelas yang ekstrem yaitu 1:577. Fitur V1 hingga V28 merupakan atribut numerik hasil transformasi *Principal Component Analysis* (PCA) (Le Borgne dkk., 2022). Untuk memastikan konsistensi matematis dan konvergensi yang stabil pada algoritma berbasis pohon keputusan, dua fitur non PCA, yaitu *Time* dan *Amount*, distandarisasi menggunakan normalisasi *Z-score*. Dataset kemudian dibagi menjadi 80% data pelatihan dan 20% data pengujian menggunakan teknik *stratified*

sampling untuk mempertahankan rasio ketidakseimbangan asli pada kedua himpunan data.

Tahap pra-pemrosesan dilakukan sebelum proses pembentukan model untuk memastikan data berada dalam kondisi yang sesuai dengan rancangan eksperimen. Penggunaan *stratified sampling* menjadi penting karena pembagian secara acak tanpa mempertimbangkan distribusi kelas berpotensi menghasilkan proporsi kelas yang berbeda antara data pelatihan dan pengujian. Dengan mempertahankan distribusi kelas, evaluasi ketiga algoritma dapat dilakukan pada kondisi yang relatif setara. Selain itu, penggunaan data asli tanpa teknik *resampling* memungkinkan penelitian mengevaluasi pengaruh penaltis bobot kelas secara langsung terhadap proses pembelajaran model.

Tiga algoritma *Gradient Boosting Decision Tree* (GBDT) terdepan, yaitu XGBoost, CatBoost, dan LightGBM, dievaluasi secara komparatif. Meskipun ketiga algoritma ini memiliki mekanisme internal yang berbeda XGBoost menggunakan aproksimasi histogram tingkat kedua, CatBoost mengandalkan *ordered boosting* untuk mencegah bias prediksi, dan LightGBM memanfaatkan *Gradient-based One-Side Sampling* (GOSS) untuk efisiensi memori. Perbedaan mekanisme tersebut memungkinkan penelitian mengamati karakteristik masing-masing algoritma ketika menghadapi distribusi kelas yang sangat timpang.

Ketiganya menghadapi tantangan matematis yang sama saat berhadapan dengan distribusi kelas yang sangat timpang. Dalam kondisi standar, fungsi kerugian akan didominasi oleh gradien dari kelas mayoritas (normal), sehingga model akan mengabaikan kelas minoritas (*fraud*). Permasalahan tersebut dapat menyebabkan model lebih banyak mengoptimalkan kesalahan pada transaksi normal karena jumlah observasinya jauh lebih besar. Oleh karena itu, diperlukan mekanisme yang dapat meningkatkan kontribusi kelas *fraud* selama proses optimasi tanpa mengubah jumlah observasi dalam dataset.

Untuk mengatasi hal tersebut tanpa melakukan *resampling* data yang berisiko menghilangkan informasi, penelitian ini memodifikasi fungsi objektif standar menjadi Fungsi Kerugian Berbobot (*Weighted Log Loss*) (Kumbhar dkk., 2025). Modifikasi matematis ini didefinisikan secara formal dalam Persamaan 1:

$$\mathcal{L}(\theta) = \sum_{i=1}^N w_{y_i} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \quad (1)$$

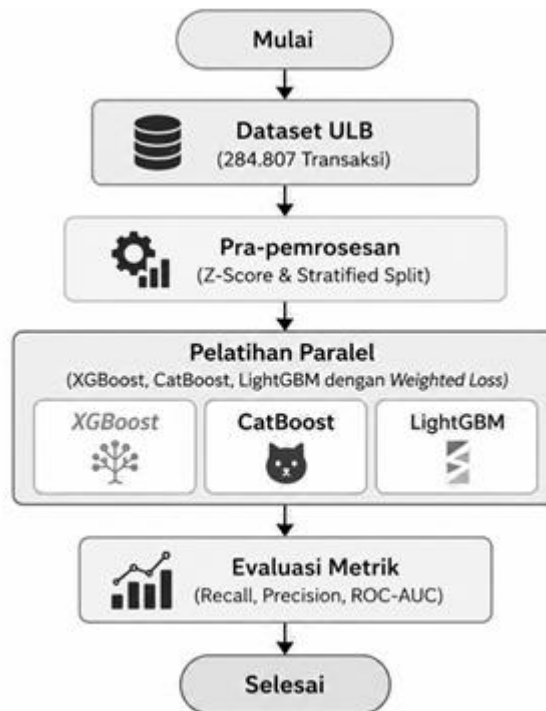
di mana $\mathcal{L}(\theta)$ merepresentasikan fungsi kerugian total yang akan diminimalkan oleh model, N adalah jumlah total sampel dalam satu *batch*, y_i adalah label kelas yang sebenarnya (0 untuk normal, 1 untuk *fraud*), dan $\hat{y}_i$ adalah probabilitas prediksi yang dihasilkan oleh model. Variabel kunci dalam persamaan ini adalah w_{y_i} , yang merupakan bobot kelas (*class weight*) untuk sampel ke- i . Pemberian bobot tersebut memungkinkan kesalahan pada kelas minoritas memberikan kontribusi yang lebih besar terhadap fungsi kerugian dibandingkan kesalahan pada kelas mayoritas.

Secara konseptual, pembobotan tersebut mengubah kontribusi masing-masing observasi terhadap proses optimasi tanpa mengubah jumlah sampel yang digunakan. Ketika bobot kelas *fraud* diberikan lebih besar, kesalahan klasifikasi terhadap transaksi *fraud* memperoleh penalti yang lebih tinggi. Mekanisme ini diharapkan mendorong model membentuk *decision boundary* yang lebih sensitif terhadap karakteristik kelas minoritas. Konsekuensinya kemudian diamati melalui perubahan *Recall*, *Precision*, dan *F1-Score* pada tahap evaluasi.

Selanjutnya, ketiga algoritma dilatih menggunakan data pelatihan yang telah diproses dengan mekanisme pembobotan kelas. Data pengujian tidak digunakan dalam proses pembentukan model dan hanya digunakan untuk mengukur kemampuan generalisasi. Pendekatan ini memungkinkan perbandingan dilakukan secara konsisten karena XGBoost, CatBoost, dan LightGBM dievaluasi pada data pengujian yang sama. Dengan demikian, perbedaan kinerja yang diperoleh dapat dikaitkan dengan karakteristik algoritma dan mekanisme penanganan ketidakseimbangan kelas yang digunakan.

Evaluasi kinerja dilakukan dengan mempertimbangkan metrik yang relevan terhadap karakteristik dataset yang sangat tidak seimbang. Metrik yang digunakan meliputi *Accuracy*, *Precision*, *Recall*, *F1-Score*, dan *ROC-AUC*. Selain kinerja prediktif, penelitian juga mengukur waktu pelatihan sebagai indikator efisiensi komputasi. Penggunaan beberapa metrik secara bersamaan diperlukan karena nilai *Accuracy* yang tinggi tidak secara langsung menunjukkan kemampuan model dalam mendeteksi transaksi *fraud* pada kondisi ketidakseimbangan ekstrem.

Secara komputasional, seluruh tahapan penelitian dirancang secara sistematis untuk memastikan reproduisibilitas dan validitas empiris. Alur kerja komprehensif mulai dari akuisisi data, pra-pemrosesan, inialisasi model dengan penyesuaian bobot, pelatihan paralel, hingga evaluasi metrik digambarkan dalam Gambar 1.



Gambar 1. Alur Kerja Kerangka *Benchmarking* Deteksi *Fraud*

Alur tersebut menunjukkan bahwa penelitian dilakukan melalui tahapan yang berurutan, dimulai dari dataset ULB, dilanjutkan dengan pra-pemrosesan, pembentukan dan pelatihan tiga model GBDT, kemudian diakhiri dengan evaluasi kinerja. Struktur tersebut memastikan bahwa proses benchmarking dilakukan menggunakan prosedur yang seragam sehingga hasil XGBoost, CatBoost, dan LightGBM dapat dibandingkan berdasarkan kinerja prediktif maupun efisiensi komputasi.

HASIL DAN PEMBAHASAN

3.1. Kinerja Prediktif dan Fenomena Paradoks Akurasi

Evaluasi empiris pada dataset ULB yang memiliki ketidakseimbangan kelas ekstrem (rasio 1:577) mengungkapkan temuan kritis mengenai metrik evaluasi. Seperti yang disajikan pada Tabel 1, ketiga algoritma *Gradient Boosting* melaporkan skor Akurasi yang sangat tinggi, berkisar antara 0,9331 hingga 0,9968.

Namun, dalam konteks data yang sangat timpang, metrik Akurasi mengalami fenomena *Accuracy Paradox*, di mana model naif yang memprediksi seluruh transaksi sebagai "normal" secara matematis tetap akan menghasilkan akurasi di atas 99,8%. Oleh karena itu, evaluasi kinerja secara ketat difokuskan pada metrik *Recall*, *Precision*, *F1-Score*, dan *ROC-AUC*.

Hasil tersebut menunjukkan bahwa tingginya Akurasi tidak secara langsung mencerminkan kemampuan model dalam mendeteksi transaksi *fraud*. Perbedaan jumlah antara kelas normal dan *fraud* menyebabkan kontribusi kelas mayoritas sangat dominan terhadap nilai Akurasi. Dengan demikian, model dapat menghasilkan nilai Akurasi yang tinggi meskipun masih terdapat transaksi *fraud* yang tidak berhasil dikenali. Kondisi ini memperkuat alasan penggunaan metrik yang secara khusus dapat menggambarkan kemampuan model dalam mengenali kelas minoritas.

Integrasi fungsi kerugian berbobot terbukti secara signifikan mengubah perilaku konvergensi model. Ketiga algoritma berhasil mencapai *Recall* di atas 0,86, yang mengindikasikan bahwa model secara efektif mendeteksi lebih dari 86% transaksi *fraud* yang sebenarnya. Secara matematis, penalti asimetris yang diberikan pada *false negative* memaksa pohon keputusan untuk memperluas batas keputusannya (*decision boundary*) ke dalam wilayah kelas minoritas. Perubahan tersebut menunjukkan bahwa pemberian bobot kelas mampu meningkatkan perhatian model terhadap transaksi yang sebelumnya berpotensi tertutupi oleh dominasi kelas mayoritas.

Konsekuensi dari ekspansi batas ini adalah peningkatan *false positive* yang tajam, yang terefleksi pada rendahnya skor *Precision* (0,0226 hingga 0,3346). Dalam domain deteksi *fraud* keuangan, *trade-off* ini sangat rasional dan dapat diterima secara industri, mengingat biaya finansial akibat *false negative* (kegagalan mendeteksi penipuan) jauh lebih besar dibandingkan biaya operasional untuk meninjau *false positive* (transaksi sah yang tertahan).

Tabel 1. Perbandingan Kinerja dan Efisiensi Komputasi pada Dataset *Fraud* ULB

Model	Akurasi	Presisi	Recall	F1-Score	ROC-AUC	Waktu Latihan (s)
XGBoost	0,9968	0,3346	0,8673	0,4830	0,9812	3,92

CatBoost	0,9919	0,1611	0,8878	0,2727	0,9755	7,61
LightGBM	0,9331	0,0226	0,8980	0,0441	0,9373	4,85

Berdasarkan Tabel 1, XGBoost menghasilkan Akurasi sebesar 0,9968, Presisi 0,3346, *Recall* 0,8673, *F1-Score* 0,4830, dan *ROC-AUC* 0,9812 dengan waktu latihan 3,92 detik. CatBoost menghasilkan *Recall* yang lebih tinggi, yaitu 0,8878, tetapi memiliki Presisi 0,1611 dan *F1-Score* 0,2727 dengan waktu latihan 7,61 detik. Sementara itu, LightGBM memperoleh *Recall* tertinggi sebesar 0,8980, tetapi menghasilkan Presisi 0,0226 dan *F1-Score* 0,0441. Hasil tersebut menunjukkan bahwa peningkatan *Recall* tidak selalu diikuti peningkatan keseimbangan antara *Precision* dan *Recall*.

Perbedaan tersebut penting karena *F1-Score* memberikan gambaran keseimbangan antara kemampuan model menemukan transaksi *fraud* dan ketepatan prediksi *fraud*. XGBoost memperoleh *F1-Score* sebesar 0,4830, lebih tinggi dibandingkan CatBoost sebesar 0,2727 dan LightGBM sebesar 0,0441. Dengan demikian, hasil pengujian memperlihatkan bahwa ketiga algoritma memberikan karakteristik yang berbeda dalam merespons ketidakseimbangan kelas, meskipun seluruhnya menggunakan mekanisme pembobotan untuk meningkatkan perhatian terhadap kelas minoritas.

3.2. Analisis Komparatif Algoritma dan Efisiensi Komputasi

Dari perspektif komputasi dan daya diskriminatif, XGBoost muncul sebagai algoritma yang paling optimal. XGBoost berhasil mencapai *ROC-AUC* tertinggi sebesar 0,9812 dan *F1-Score* terbaik (0,4830) dengan latensi komputasi tercepat, yaitu hanya 3,92 detik. Keunggulan ini didukung oleh mekanisme aproksimasi histogram tingkat kedua yang secara efisien memetakan ruang fitur numerik hasil PCA tanpa memerlukan komputasi *split* yang mahal secara komputasi.

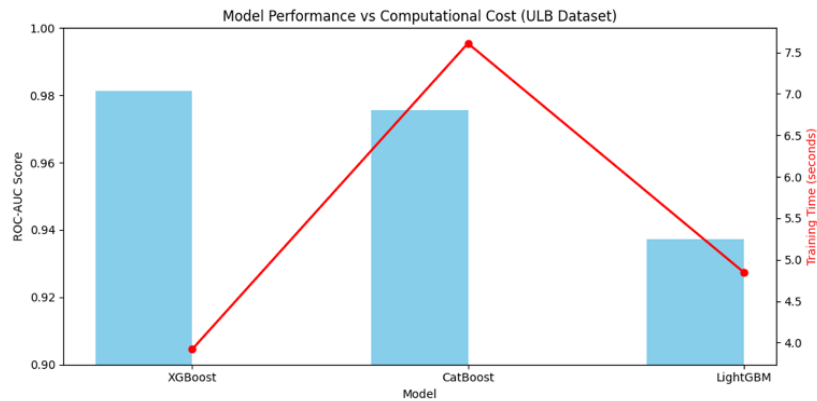
Nilai *ROC-AUC* tersebut menunjukkan kemampuan XGBoost dalam membedakan kelas *fraud* dan kelas normal pada berbagai kemungkinan ambang keputusan. Sementara itu, nilai *F1-Score* yang lebih tinggi menunjukkan bahwa model tidak hanya mampu mempertahankan *Recall*, tetapi juga menghasilkan tingkat Presisi yang relatif lebih baik dibandingkan dua algoritma lainnya.

Kombinasi kedua karakteristik tersebut menjadi penting pada dataset dengan ketidakseimbangan kelas 1:577 karena evaluasi hanya berdasarkan *Recall* dapat menghasilkan interpretasi yang tidak lengkap.

Di sisi lain, CatBoost menunjukkan kapabilitas *Recall* tertinggi (0,8878), yang membuktikan bahwa mekanisme *ordered boosting*-nya sangat tangguh dalam menangkap pola anomali tanpa mengalami bias prediksi. Namun, keunggulan ini dibayar dengan waktu pelatihan yang hampir dua kali lipat lebih lama (7,61 detik) dibandingkan XGBoost. Selain itu, nilai Presisi CatBoost sebesar 0,1611 menyebabkan *F1-Score* yang diperoleh hanya mencapai 0,2727. Hasil ini menunjukkan bahwa peningkatan *Recall* belum secara otomatis menghasilkan keseimbangan performa klasifikasi yang lebih baik.

Sementara itu, LightGBM, meskipun memanfaatkan *Gradient-based One-Side Sampling* (GOSS) untuk menyelesaikan pelatihan dengan cepat (4,85 detik), mengalami degradasi performa yang signifikan pada *Precision* (0,0226). Hal ini menyebabkan *F1-Score* LightGBM anjlok ke angka 0,0441, mengindikasikan bahwa model menghasilkan terlalu banyak *false positive* dibandingkan *true positive* dalam upayanya mengejar *Recall*. Dengan kata lain, kemampuan LightGBM dalam menangkap transaksi *fraud* yang lebih tinggi belum diikuti oleh kemampuan mempertahankan ketepatan prediksi pada kelas tersebut.

Perbandingan tersebut menunjukkan bahwa efisiensi komputasi perlu dianalisis bersama dengan kinerja prediktif. Waktu pelatihan XGBoost sebesar 3,92 detik merupakan yang terendah, diikuti LightGBM sebesar 4,85 detik dan CatBoost sebesar 7,61 detik. Sementara itu, nilai *ROC-AUC* masing-masing sebesar 0,9812, 0,9755, dan 0,9373 menunjukkan adanya perbedaan daya diskriminatif antaralgoritma. Dengan demikian, hasil eksperimen memberikan gambaran bahwa setiap algoritma memiliki karakteristik yang berbeda antara kemampuan mendeteksi kelas minoritas dan kebutuhan komputasinya.



Gambar 2. Analisis Komparatif Kinerja Prediktif (ROC-AUC) terhadap Biaya Komputasi pada Dataset ULB.

Gambar 2 memperlihatkan hubungan antara *ROC-AUC* sebagai representasi daya diskriminatif model dan waktu pelatihan sebagai representasi biaya komputasi. Visualisasi tersebut memperkuat hasil pada Tabel 1 bahwa perbedaan kinerja tidak hanya terlihat pada metrik klasifikasi, tetapi juga pada kebutuhan waktu komputasi masing-masing algoritma. Posisi setiap model dalam grafik memberikan gambaran mengenai hubungan antara kemampuan diskriminatif dan efisiensi proses pelatihan.

Secara keseluruhan, hasil empiris ini mengonfirmasi bahwa untuk sistem deteksi *fraud* berskala besar yang menuntut pemrosesan *realtime*, XGBoost menawarkan keseimbangan terbaik (*optimal trade off*). Algoritma ini tidak hanya memaksimalkan daya diskriminatif model terhadap kelas minoritas melalui penaltis bobot matematis, tetapi juga mempertahankan efisiensi komputasi yang superior dibandingkan CatBoost dan LightGBM.

KESIMPULAN

Penelitian ini berhasil melakukan *benchmarking* empiris terhadap tiga algoritma *Gradient Boosting* (XGBoost, CatBoost, LightGBM) dalam mendeteksi penipuan kartu kredit pada kondisi ketidakseimbangan kelas yang ekstrem (1:577). Temuan menunjukkan bahwa modifikasi matematis pada fungsi objektif melalui penaltis bobot kelas jauh lebih unggul dibandingkan teknik *resampling* data, berhasil memaksa model mencapai *Recall* di atas 86% tanpa mengorbankan integritas data mayoritas. XGBoost terbukti sebagai algoritma paling optimal, mencapai *ROC-AUC* tertinggi sebesar 0,9812 dengan latensi komputasi tercepat

(3,92 detik), menjadikannya sangat ideal untuk sistem pemrosesan transaksi *realtime*. Penelitian selanjutnya akan mengeksplorasi integrasi algoritma ini dengan *Federated Learning* untuk mendeteksi fraud lintas institusi keuangan tanpa melanggar privasi data nasabah.

REFERENCES

- Anh Luong, N. D., & Xie, S. (2026). Explainable ensemble machine learning for financial transaction fraud detection: Insights from XGBoost and deep neural networks. *The Journal of Finance and Data Science*, 12, 100195. <https://doi.org/https://doi.org/10.1016/j.jfds.2026.100195>
- Daswati, O., Wijayanto, H., & Afendi, F. (2026). Comparative Study Of Lightgbm, Catboost, And Random Forest In Modeling Public Complaints Classification. *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, 20, 2535–2548. <https://doi.org/10.30598/barekengvol20iss3pp2535-2548>
- Faqih, A., & Sugihartono, T. (2025). Perbandingan Algoritma XGBoost dan LSTM dalam Prediksi Harga Saham Tesla Menggunakan Data Tahun 2025. *Jurnal Pendidikan dan Teknologi Indonesia*, 5(6), 1563–1573. <https://doi.org/10.52436/1.jpti.836>
- Hancock, J., & Khoshgoftaar, T. (2020). CatBoost for big data: an interdisciplinary review. *Journal of Big Data*, 7. <https://doi.org/10.1186/s40537-020-00369-8>
- Hilal, W., Gadsden, S., & Yawney, J. (2021). Financial Fraud: A Review of Anomaly Detection Techniques and Recent Advances. *Expert Systems with Applications*, 193, 116429. <https://doi.org/10.1016/j.eswa.2021.116429>
- Kraiem, M. S., Sánchez-Hernández, F., & Moreno-García, M. N. (2021). Selecting the Suitable Resampling Strategy for Imbalanced Data Classification Regarding Dataset Properties. An Approach Based on Association Models. *Applied Sciences*, 11(18). <https://doi.org/10.3390/app11188546>
- Kumbhar, M., Bandaru, S., & Karlsson, A. (2025). Imbalanced data oversampling through subspace optimization with Bayesian reinforcement. *Artificial Intelligence Review*, 59(1), 1. <https://doi.org/10.1007/s10462-025-11417-1>

- Le Borgne, Y.-A., Siblini, W., Lebichot, B., & Bontempi, G. (2022). *Reproducible Machine Learning for Credit Card Fraud detection: Practical handbook*.
- Martínez, Ó., Sánchez, P., & Alcaraz, E. (2025). *A Review of Machine Learning and Deep Learning Approaches for Fraud Detection Across Financial and Supply Chain Domains*. <https://doi.org/10.21203/rs.3.rs-7893661/v1>
- Nugraha, W., & Syarif, M. (2023). Teknik Weighting untuk Mengatasi Ketidakseimbangan Kelas Pada Prediksi Churn Menggunakan XGBoost, LightGBM, dan CatBoost. *Techno.Com*, 22, 97–108. <https://doi.org/10.33633/tc.v22i1.7191>
- Pristyanto, Y., Mukarabiman, Z., & Nugraha, A. (2023). Extreme Gradient Boosting Algorithm to Improve Machine Learning Model Performance on Multiclass Imbalanced Dataset. *JOIV: International Journal on Informatics Visualization*, 7, 710–715. <https://doi.org/10.30630/joiv.7.3.1102>
- Rizky, P., Hirzi, R., & Hidayaturrohmah, U. (2022). Perbandingan Metode LightGBM dan XGBoost dalam Menangani Data dengan Kelas Tidak Seimbang. *J Statistika: Jurnal Ilmiah Teori dan Aplikasi Statistika*, 15, 228–236. <https://doi.org/10.36456/jstat.vol15.no2.a5548>
- Saputra, G. E., Swari, M. H. P., & Nurlaili, A. L. (2025). Implementasi Algoritma XGBoost, CatBoost, dan LGBM untuk Klasifikasi Pencemaran Udara. *Jiip - Jurnal Ilmiah Ilmu Pendidikan*, 8(12), 14135–14139. <https://doi.org/10.54371/jiip.v8i12.10102>
- Yang, H., Shukur, Z., & Sahran, S. (2026). A Review of Artificial Intelligence for Financial Fraud Detection. *Applied Sciences*, 16(4). <https://doi.org/10.3390/app16041931>
- Zhao, X., Wang, S., Qiu, S., Suo, S., & Lu, L. (2026). Optimizing the Undersampling Technique by Feature Weighting for Software Defect Prediction. *International Journal of Computational Intelligence Systems*, 19(1), 118. <https://doi.org/10.1007/s44196-026-01174-6>