

Deteksi Penipuan Kartu Kredit Melalui Pendekatan Stacking Ensemble Berbasis Gradient Boosting dengan Optimasi Hyperparameter *Credit Card Fraud Detection Using a Gradient Boosting Stacking Ensemble with Hyperparameter Optimization*

Shafa Syahida^{*1}, Donna Nur Tamara², Diah Septiani³

^{1,2,3} Program Studi S1 Sains Data, Telkom University, Purwokerto, Jawa Tengah

e-mail: ^{*1}shafasyahida@student.telkomuniversity.ac.id,

^{*2}donnanur@student.telkomuniversity.ac.id,

^{*3}diahseptiani@telkomuniversity.ac.id

Abstrak - Deteksi fraud kartu kredit merupakan tantangan penting karena data transaksi memiliki ketidakseimbangan kelas yang tinggi. Penelitian ini membangun model Stacking Ensemble berbasis Gradient Boosting dengan XGBoost, LightGBM, dan CatBoost sebagai base learner, serta Logistic Regression sebagai meta-learner. Tahapan penelitian meliputi preprocessing menggunakan RobustScaler, feature selection berbasis SHAP, penanganan ketidakseimbangan kelas menggunakan SMOTE, dan optimasi hyperparameter menggunakan Optuna. Hasil penelitian menunjukkan bahwa Stacking Ensemble menghasilkan performa terbaik dengan precision 0,7798, recall 0,8673, F1-Score 0,8213, ROC-AUC 0,9789, dan PR-AUC 0,8534. Hasil ini menunjukkan bahwa penggabungan beberapa model gradient boosting mampu menghasilkan keseimbangan precision dan recall yang lebih baik dibandingkan model individual.

Kata kunci – Deteksi fraud kartu kredit; logistic regression; optimasi hyperparameter; SMOTE; stacking ensemble

Abstract - Credit card fraud detection is a challenging task due to severe class imbalance in transaction data. This study developed a Gradient Boosting-based Stacking Ensemble model using XGBoost, LightGBM, and CatBoost as base learners, with Logistic Regression as the meta-learner. The research stages included preprocessing using RobustScaler, SHAP-based feature selection, class imbalance handling using SMOTE, and hyperparameter optimization using Optuna. The results showed that the Stacking Ensemble achieved the best performance with a precision of 0.7798, recall of 0.8673, F1-Score of 0.8213, ROC-AUC of 0.9789, and PR-AUC of 0.8534. These results indicated that combining multiple gradient boosting models provided a better balance between precision and recall than individual models.

Keywords – Credit card fraud detection; hyperparameter tuning; logistic regression; SMOTE; stacking ensemble

I. PENDAHULUAN

Penggunaan kartu kredit sebagai alat pembayaran terus mengalami pertumbuhan yang signifikan seiring dengan meningkatnya volume transaksi digital di seluruh dunia. Pada tahun 2024, total volume transaksi kartu kredit, debit, dan prepaid secara global mencapai \$51,92 triliun [1]. Namun, pertumbuhan tersebut diikuti oleh meningkatnya ancaman penipuan kartu kredit yang menyebabkan kerugian finansial besar bagi individu maupun institusi keuangan. Berdasarkan laporan Nilson Report [2], kerugian akibat fraud kartu kredit

secara global mencapai \$33,83 miliar pada tahun 2023. Lebih lanjut, laporan yang sama memproyeksikan bahwa kerugian kumulatif akibat fraud kartu kredit akan mencapai \$403,88 miliar dalam satu dekade ke depan.

Ancaman fraud kartu kredit mendorong institusi keuangan untuk mengembangkan sistem deteksi yang lebih canggih dan adaptif. *Machine learning* telah menjadi pendekatan yang banyak digunakan karena kemampuannya dalam mengungkap pola kompleks dan *nonlinier* yang sulit diidentifikasi oleh metode berbasis aturan konvensional [3]. Di antara berbagai algoritma *machine learning*, model berbasis *gradient boosting* seperti XGBoost, LightGBM, dan CatBoost telah menunjukkan performa unggul dalam berbagai tugas klasifikasi, termasuk deteksi fraud [4]. Kemampuan ketiganya dalam menangani data berdimensi tinggi, toleransi terhadap nilai pencilan, serta mekanisme regularisasi bawaan menjadikan model-model ini pilihan yang relevan untuk kasus *fraud detection*.

Salah satu tantangan utama dalam deteksi fraud kartu kredit adalah ketidakseimbangan kelas (*class imbalance*) yang ekstrem pada data transaksi, di mana transaksi fraud hanya mewakili sebagian kecil dari keseluruhan data. Pada dataset European Credit Card yang banyak digunakan dalam penelitian, hanya 0,17% dari total transaksi merupakan transaksi *fraud* [5]. Kondisi ini menyebabkan model *machine learning* cenderung bias terhadap kelas mayoritas sehingga gagal mendeteksi transaksi fraud secara efektif [3]. Untuk mengatasi permasalahan tersebut, teknik *Synthetic Minority Oversampling Technique* (SMOTE) banyak diterapkan untuk menghasilkan sampel sintesis dari kelas minoritas sehingga distribusi kelas menjadi lebih seimbang [5].

Selain tantangan ketidakseimbangan kelas, performa model tunggal seringkali tidak cukup optimal untuk menangani kompleksitas pola *fraud*. Pendekatan *stacking ensemble* telah terbukti mampu meningkatkan performa klasifikasi dengan cara menggabungkan prediksi beberapa model dasar melalui *meta-learner* [5]. Penelitian Abdelghafour et al. [6] menunjukkan bahwa model *stacking* dengan XGBoost, LightGBM, dan CatBoost sebagai *base learner* serta *Logistic Regression* sebagai *meta-learner* mampu menghasilkan ROC-AUC sebesar 0,9887 pada dataset yang sama. Meskipun demikian, studi tersebut belum menerapkan SMOTE secara benar dalam proses validasi silang, sehingga berpotensi menghasilkan *data leakage* yang membuat estimasi performa menjadi terlalu optimis.

Masalah *data leakage* merupakan isu metodologis yang sering ditemukan dalam penelitian deteksi *fraud*. Hayat et al. [7] mengidentifikasi bahwa salah satu penyebab utamanya adalah penerapan teknik *scaling* dan *oversampling* sebelum pembagian data latih dan data uji, sehingga informasi dari data uji bocor ke proses pelatihan model. Studi tersebut menunjukkan bahwa bahkan model sederhana dapat mencapai hasil yang tampak sangat baik apabila protokol evaluasi yang digunakan tidak tepat. Senada dengan hal tersebut, Kabane [8] secara spesifik membuktikan bahwa penerapan SMOTE sebelum *train-test split* pada dataset kartu kredit Kaggle dapat menghasilkan estimasi performa yang menyesatkan. Oleh karena itu, penerapan *pipeline preprocessing* yang benar di mana pembagian data dilakukan terlebih dahulu sebelum *scaling* maupun *oversampling* menjadi aspek kritis yang harus diperhatikan.

Di samping permasalahan metodologis tersebut, optimasi *hyperparameter* juga memainkan peran penting dalam memaksimalkan performa model. Pendekatan tuning tradisional seperti *grid search* dan *random search* memiliki keterbatasan efisiensi pada ruang pencarian yang besar

[9]. Optuna, sebagai *framework* optimasi berbasis *Tree-structured Parzen Estimator* (TPE), menawarkan strategi pencarian yang lebih efisien dan adaptif dibandingkan pendekatan konvensional [10]. Selain optimasi *hyperparameter*, pemilihan fitur yang relevan juga berkontribusi pada performa dan efisiensi model. Metode SHAP (*SHapley Additive Explanations*) memungkinkan penghitungan kontribusi setiap fitur terhadap prediksi model secara konsisten, sehingga dapat digunakan sebagai dasar pemilihan fitur yang lebih dapat diinterpretasi dibandingkan metode berbasis statistik konvensional [11].

Berdasarkan latar belakang tersebut, penelitian ini mengusulkan pendekatan *Stacking Ensemble* Berbasis *Gradient Boosting* dengan Optimasi *Hyperparameter* untuk deteksi penipuan kartu kredit. *Pipeline* yang diusulkan menerapkan *train-test split* sebelum proses *preprocessing* apapun, *RobustScaler* yang hanya di-fit pada data latih, SMOTE yang diterapkan di dalam setiap fold validasi silang untuk menghindari *data leakage*, pemilihan fitur berbasis SHAP, serta *tuning hyperparameter* XGBoost, LightGBM, dan CatBoost menggunakan Optuna.

II. TINJAUAN TEORI

2.1 Ensemble Learning

Ensemble learning merupakan pendekatan *machine learning* yang menggabungkan beberapa model untuk menghasilkan performa prediksi yang lebih baik dibandingkan model tunggal. Prinsip utama *ensemble learning* adalah memanfaatkan keunggulan dari beberapa model sehingga kesalahan yang dihasilkan oleh satu model dapat dikompensasi oleh model lainnya. Pendekatan ini banyak digunakan pada permasalahan klasifikasi karena mampu meningkatkan akurasi, stabilitas, dan kemampuan generalisasi model [12].

2.2 Stacking Ensemble

Stacking Ensemble merupakan metode ensemble yang menggabungkan beberapa model dasar (*base learner*) dan sebuah model tingkat kedua (*meta-learner*) untuk menghasilkan prediksi yang lebih baik dibandingkan model individual [6]. Pada metode *stacking*, prediksi yang dihasilkan oleh beberapa *base learner* digunakan sebagai masukan bagi *meta-learner*. Untuk mengurangi risiko *overfitting*, proses pembentukan data pelatihan *meta-learner* umumnya dilakukan menggunakan pendekatan *Out-of-Fold* (OOF) *Prediction* melalui *cross validation* [6].

2.3 Gradient Boosting Models

a. XGBoost

XGBoost (*Extreme Gradient Boosting*) merupakan algoritma *ensemble* berbasis *gradient boosting* yang membangun model secara bertahap melalui penambahan pohon keputusan untuk memperbaiki kesalahan prediksi model sebelumnya. XGBoost menggabungkan fungsi *loss* dan regularisasi sehingga mampu meningkatkan kemampuan generalisasi model sekaligus mengurangi risiko *overfitting* [13].

Fungsi objektif XGBoost dapat dirumuskan sebagai berikut [13]:

$$Obj^{(t)} = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^t \Omega(f_k) \dots \dots \dots (1)$$

dengan $Obj^{(t)}$ merupakan fungsi objektif pada iterasi ke- t , y_i merupakan nilai aktual, $\hat{y}_i^{(t)}$ merupakan hasil prediksi pada iterasi ke- t , l merupakan fungsi *loss*, f_k merupakan pohon keputusan ke- k , dan Ω merupakan fungsi regularisasi.

b. LightGBM

LightGBM merupakan algoritma *gradient boosting* yang dikembangkan untuk meningkatkan efisiensi komputasi dan penggunaan memori pada dataset berukuran besar. LightGBM menggunakan pendekatan *histogram-based learning* dan *leaf-wise tree growth* sehingga mampu menghasilkan proses pelatihan yang lebih cepat dibandingkan metode *boosting* konvensional [14].

Fungsi objektif LightGBM pada iterasi ke- t dapat dituliskan sebagai berikut [15]:

$$Obj^{(t)} = \sum_{i=1}^n \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \dots \dots \dots (2)$$

dengan g_i merupakan gradien pertama, h_i merupakan gradien kedua (Hessian), $f_t(x_i)$ merupakan nilai prediksi pohon pada iterasi ke- t , dan Ω merupakan fungsi regularisasi.

c. CatBoost

CatBoost merupakan algoritma *gradient boosting* yang dirancang untuk mengurangi *prediction shift* dan *overfitting* melalui mekanisme *ordered boosting* [16].

Proses pembaruan prediksi pada CatBoost dapat dituliskan sebagai berikut [26]:

$$F_t(x) = F_{t-1}(x) + \eta h_t(x) \dots \dots \dots (3)$$

dengan $F_t(x)$ merupakan prediksi model pada iterasi ke- t , $F_{t-1}(x)$ merupakan prediksi model sebelumnya, η merupakan learning rate, dan $h_t(x)$ merupakan *weak learner* pada iterasi ke- t .

2.4 Optimasi Hyperparameter Menggunakan Optuna

Optuna merupakan framework optimasi *hyperparameter* yang menggunakan algoritma *Tree-structured Parzen Estimator* (TPE). TPE merupakan pendekatan *Bayesian Optimization* yang membangun distribusi probabilitas berdasarkan hasil evaluasi trial sebelumnya, kemudian menggunakan distribusi tersebut untuk menentukan kandidat *hyperparameter* berikutnya yang berpotensi memberikan performa lebih baik [10].

2.5 RobustScaler dan Penanganan Ketidakseimbangan Kelas

RobustScaler merupakan metode normalisasi yang menggunakan median dan rentang interkuartil (*Interquartile Range/IQR*) sehingga lebih tahan terhadap keberadaan *outlier* dibandingkan metode *scaling* berbasis rata-rata dan standar deviasi [17]. *Synthetic Minority Oversampling Technique* (SMOTE) merupakan metode *oversampling* yang menghasilkan sampel sintesis baru pada kelas minoritas dengan melakukan interpolasi antara suatu sampel dan tetangga terdekatnya (*nearest neighbors*).

Dibandingkan metode *oversampling* konvensional yang hanya menggandakan data, SMOTE mampu menghasilkan variasi data baru sehingga membantu model mempelajari karakteristik kelas minoritas dengan lebih baik [7].

2.6 Metrik Evaluasi

Precision digunakan untuk mengukur proporsi prediksi positif yang benar terhadap seluruh prediksi positif yang dihasilkan model [15].

$$Precision = \frac{TP}{TP + FP} \dots \dots \dots (4)$$

Recall digunakan untuk mengukur kemampuan model dalam menemukan seluruh data positif yang sebenarnya [15].

$$Recall = \frac{TP}{TP + FN} \dots \dots \dots (5)$$

F1-Score merupakan rata-rata harmonik antara *Precision* dan *Recall* sehingga mampu memberikan keseimbangan antara kedua metrik tersebut [15].

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \dots \dots \dots (6)$$

ROC-AUC (*Receiver Operating Characteristic–Area Under the Curve*) digunakan untuk mengukur kemampuan model dalam membedakan kelas positif dan negatif pada berbagai nilai *threshold* klasifikasi dengan *TPR(t)* merupakan nilai *True Positive Rate* pada *threshold t*, *FPR(t)* merupakan nilai *False Positive Rate* pada *threshold t*, *t* merupakan *threshold* klasifikasi, dan $\int$ menunjukkan perhitungan luas area di bawah kurva ROC [15].

$$ROC-AUC = \int_0^1 TPR(t) d(FPR(t)) \dots \dots \dots (7)$$

PR-AUC (*Precision-Recall Area Under the Curve*) mengukur luas area di bawah kurva *Precision-Recall* dan sering digunakan pada dataset dengan ketidakseimbangan kelas yang tinggi dengan *Precision(r)* merupakan nilai *precision* pada tingkat *recall r*, *r* merupakan nilai *recall*, dan $\int$ menunjukkan perhitungan luas area di bawah kurva *Precision-Recall* [3].

$$PR-AUC = \int_0^1 Precision(r) dr \dots \dots \dots (8)$$

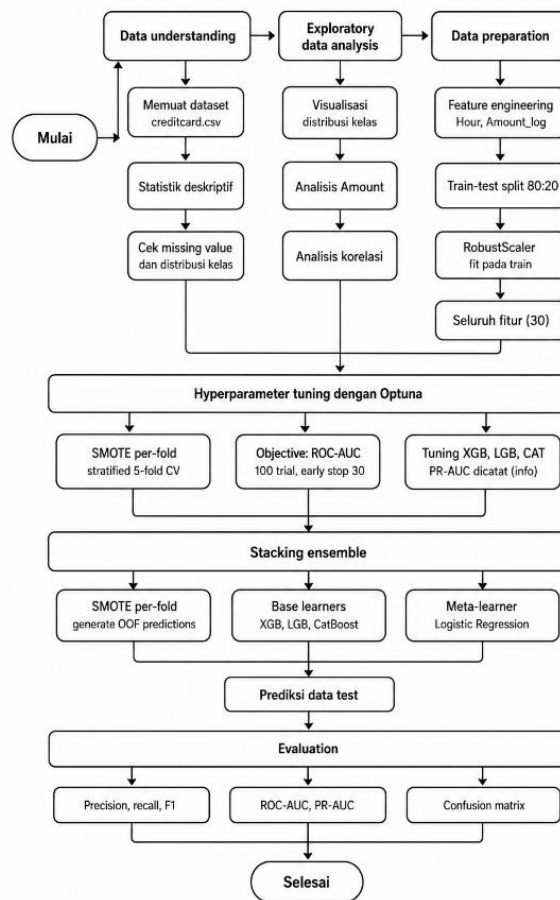
III. METODE

3.1 Dataset

Penelitian ini menggunakan dataset Credit Card Fraud Detection yang tersedia secara publik melalui platform Kaggle. Dataset berisi 284.807 transaksi kartu kredit dengan 492 transaksi *fraud* dan 284.315 transaksi *non-fraud*. Dataset terdiri atas 31 atribut yang mencakup fitur V1 hingga V28 hasil transformasi *Principal Component Analysis* (PCA), atribut *Time*, atribut *Amount*, dan variabel target *Class*. Variabel target bernilai 0 untuk transaksi *non-fraud* dan 1 untuk transaksi *fraud* [3][11].

3.2 Tahapan Penelitian

Penelitian dilakukan melalui beberapa tahapan yang meliputi data *understanding*, *exploratory data analysis* (EDA), *preprocessing data*, optimasi *hyperparameter*, pembentukan model *stacking ensemble*, dan evaluasi model. Alur penelitian ditunjukkan pada Gambar 1.



Gambar 1. Kerangka Kerja Implementasi

Setelah memahami karakteristik data melalui tahap data *understanding* dan EDA, dilakukan *preprocessing* yang meliputi *feature engineering*, *train-test split*, *feature scaling* menggunakan *RobustScaler*, dan *feature selection* menggunakan SHAP. Selanjutnya dilakukan optimasi *hyperparameter* menggunakan Optuna pada masing-masing model *base learner*, yaitu XGBoost, LightGBM, dan CatBoost. Model yang telah dioptimasi kemudian digabungkan menggunakan metode *stacking ensemble* dengan *Logistic Regression* sebagai *meta-learner*. Tahap akhir dilakukan evaluasi menggunakan metrik *Precision*, *Recall*, *F1-Score*, *ROC-AUC*, dan *PR-AUC*.

3.3 Preprocessing Data

Tahap *preprocessing* dilakukan untuk mempersiapkan data sebelum proses pelatihan model. *Feature engineering* dilakukan dengan mengubah atribut *Time* menjadi atribut *Hour* serta mentransformasi atribut *Amount* menggunakan fungsi logaritma natural sehingga menghasilkan atribut *Amount_log*. Setelah itu dataset dibagi menggunakan *stratified train-test split* dengan rasio 80:20 untuk mempertahankan distribusi kelas *fraud* dan *non-fraud*.

Data latih kemudian dinormalisasi menggunakan *RobustScaler* yang hanya di-fit pada

training set untuk menghindari *data leakage*. Selanjutnya dilakukan *feature selection* menggunakan SHAP untuk mengevaluasi kontribusi masing-masing fitur terhadap prediksi model. Kandidat jumlah fitur dievaluasi menggunakan *5-Fold Cross Validation* berdasarkan nilai *F1-Score* dan kombinasi fitur terbaik digunakan pada tahap pemodelan berikutnya.

3.4 Optimasi Hyperparameter

Optimasi *hyperparameter* dilakukan menggunakan Optuna dengan algoritma *Tree-structured Parzen Estimator* (TPE). Proses optimasi menggunakan *5-Fold Stratified Cross Validation* dengan metrik objektif berupa nilai ROC-AUC rata-rata. Setiap model diberikan maksimum 100 trial dan menerapkan *early stopping* apabila tidak terjadi peningkatan performa selama 30 trial berturut-turut.

Search space *hyperparameter* yang digunakan untuk XGBoost, LightGBM, dan CatBoost ditunjukkan pada Tabel 1.

Tabel 1. Search Space Hyperparameter Model

Model	Hyperparameter
XGBoost	learning_rate [0.01, 0.30], n_estimators [50, 1000], max_depth [3, 10], min_child_weight [1, 10], gamma [0.0, 5.0], subsample [0.5, 1.0], colsample_bytree [0.3, 1.0], reg_alpha [0.0, 1.0], reg_lambda [1.0, 100.0], scale_pos_weight {10, 15, 20}
LightGBM	n_estimators [50, 200], learning_rate [0.01, 0.30], num_leaves [2, 50], max_depth [3, 12], min_data_in_leaf [10, 100], min_child_weight [0.001, 1.0], subsample [0.6, 1.0], colsample_bytree [0.6, 1.0], min_child_samples [5, 100], reg_alpha [0.001, 10], reg_lambda [0.001, 10], min_split_gain [0.001, 1.0], bagging_fraction [0.6, 1.0], bagging_freq [1, 10]
CatBoost	iterations [200, 1000], depth [3, 6], learning_rate [0.01, 0.20], l2_leaf_reg [3.0, 10.0], min_data_in_leaf [5, 30], random_strength [1.0, 10.0], model_size_reg [0.0, 1.0], rsm [0.7, 1.0]

3.5 Pemodelan dan Evaluasi

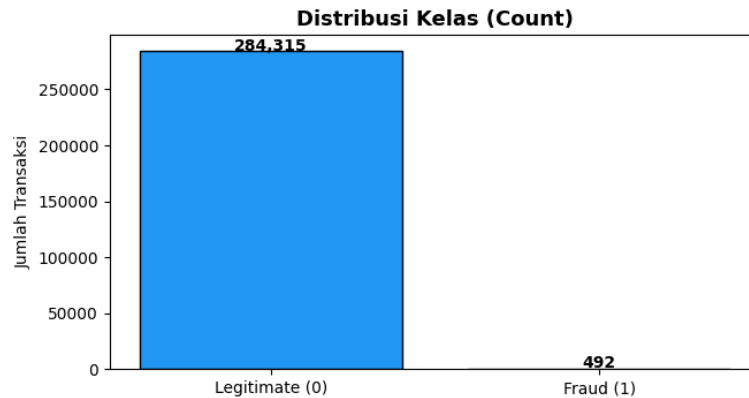
Pemodelan dilakukan menggunakan tiga algoritma *gradient boosting*, yaitu XGBoost, LightGBM, dan CatBoost sebagai *base learner*. Ketiga model terlebih dahulu dioptimasi menggunakan Optuna, kemudian digunakan untuk membentuk prediksi *Out-of-Fold* (OOF) melalui *5-Fold Stratified Cross Validation*. Prediksi probabilitas yang dihasilkan digunakan sebagai masukan bagi *Logistic Regression* sebagai *meta-learner*. Untuk mengatasi ketidakseimbangan kelas, SMOTE diterapkan hanya pada data *training* di setiap *fold* validasi silang sehingga tidak terjadi *data leakage*. Model yang telah dilatih kemudian dievaluasi pada *test set* menggunakan metrik *Precision*, *Recall*, *F1-Score*, ROC-AUC, dan PR-AUC.

IV. HASIL DAN PEMBAHASAN

4.1 Hasil Preprocessing dan Feature Selection

Analisis distribusi kelas menunjukkan bahwa proporsi transaksi *fraud* hanya sebesar 0,1727% dari keseluruhan data. Kondisi tersebut mengindikasikan bahwa dataset

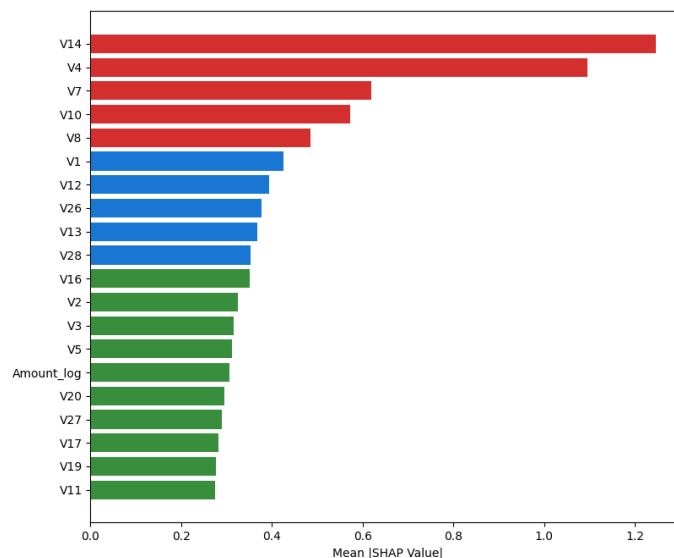
termasuk ke dalam kategori *highly imbalanced* dataset sehingga diperlukan strategi penanganan ketidakseimbangan kelas pada tahap pemodelan.



Gambar 2. Distribusi Kelas *Fraud* dan *Non-Fraud*

Selain ketidakseimbangan kelas, atribut *Amount* menunjukkan distribusi yang sangat condong ke kanan (*right-skewed distribution*). Oleh karena itu dilakukan transformasi logaritmik menggunakan fungsi $\log_{1p}$ untuk mengurangi pengaruh nilai transaksi yang sangat besar sehingga proses pembelajaran model menjadi lebih stabil.

Selanjutnya dilakukan analisis kontribusi fitur menggunakan SHAP (*SHapley Additive Explanations*) untuk mengidentifikasi fitur yang paling berpengaruh terhadap prediksi model. Hasil pemeringkatan fitur berdasarkan nilai SHAP ditunjukkan pada Gambar 3.



Gambar 3. SHAP *Feature Importance*

Berdasarkan Gambar 3, fitur V14, V4, V7, V10, dan V8 merupakan fitur yang memiliki kontribusi terbesar terhadap prediksi model. Dominasi fitur-fitur tersebut menunjukkan bahwa beberapa komponen hasil transformasi PCA memiliki kemampuan yang lebih baik dalam membedakan transaksi *fraud* dan *non-fraud* dibandingkan fitur lainnya. Selain itu, fitur *Amount_log* juga menunjukkan kontribusi yang cukup tinggi sehingga informasi nominal transaksi tetap berperan dalam proses klasifikasi.

Untuk menentukan jumlah fitur yang digunakan pada tahap pemodelan, dilakukan evaluasi beberapa kandidat jumlah fitur menggunakan *5-Fold Cross Validation*. Hasil evaluasi ditunjukkan pada Tabel 2.

Tabel 2. Evaluasi Pemilihan Jumlah Fitur

Jumlah Fitur	Mean F1 Score
Top-10	0,5694
Top-17	0,7540
Top-20	0,7519
Seluruh Fitur (30)	0,7765

Berdasarkan Tabel 2, penggunaan seluruh 30 fitur menghasilkan nilai *F1-Score* tertinggi sebesar 0,7765. Hasil ini menunjukkan bahwa pengurangan jumlah fitur tidak memberikan peningkatan performa pada dataset yang digunakan. Meskipun beberapa fitur memiliki kontribusi yang lebih kecil berdasarkan SHAP, informasi yang dikandungnya tetap membantu model dalam mengenali pola *fraud* yang kompleks. Oleh karena itu, seluruh fitur dipertahankan pada tahap pemodelan.

4.2 Hasil Evaluasi Model

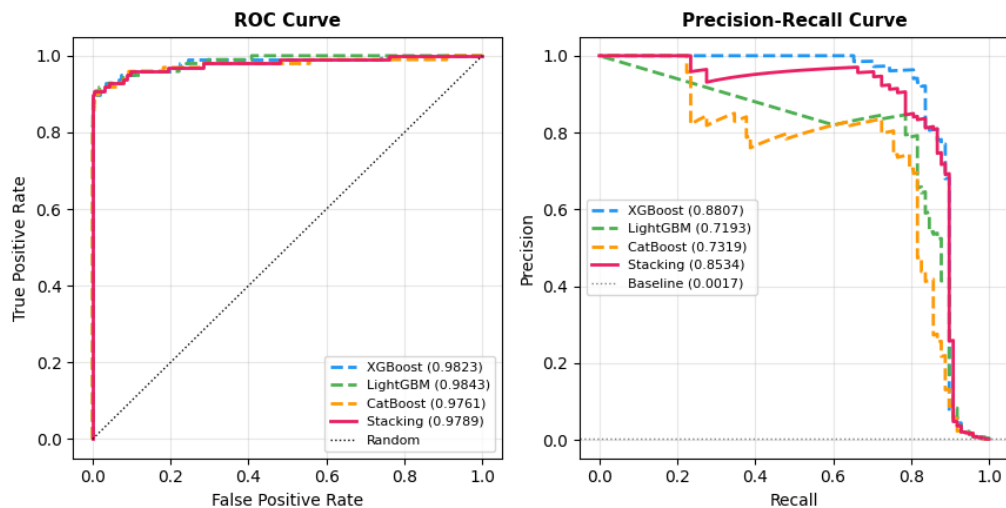
Evaluasi dilakukan terhadap tiga model individual, yaitu XGBoost, LightGBM, dan CatBoost, serta model usulan berupa *Stacking Ensemble* yang menggabungkan ketiga model tersebut melalui *Logistic Regression* sebagai *meta-learner*. Hasil evaluasi seluruh model ditunjukkan pada Tabel 3.

Tabel 3. Perbandingan Kinerja Seluruh Model

Model	Precision	Recall	F1-Score	ROC-AUC	PR-AUC
XGBoost	0,6241	0,8980	0,7364	0,9823	0,8807
LightGBM	0,2391	0,8980	0,3777	0,9843	0,7193
CatBoost	0,0714	0,9082	0,1324	0,9761	0,7319
Stacking Ensemble	0,7798	0,8673	0,8213	0,9789	0,8534

Berdasarkan Tabel 3, seluruh model menghasilkan nilai ROC-AUC di atas 0,97 yang menunjukkan kemampuan diskriminasi yang sangat baik dalam membedakan transaksi *fraud* dan *non-fraud*. LightGBM menghasilkan nilai ROC-AUC tertinggi sebesar 0,9843, sedangkan XGBoost memperoleh nilai PR-AUC tertinggi sebesar 0,8807. Namun demikian, pada kasus *fraud detection* dengan ketidakseimbangan kelas yang sangat tinggi, evaluasi tidak dapat hanya bergantung pada satu metrik. Oleh karena itu, *F1-Score* digunakan sebagai indikator utama karena mampu merepresentasikan keseimbangan antara *precision* dan *recall*.

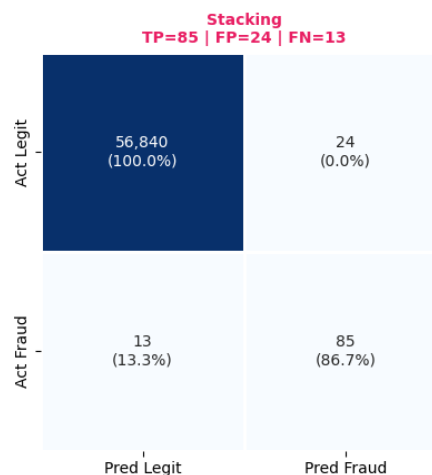
Hasil pengujian menunjukkan bahwa *Stacking Ensemble* menghasilkan *F1-Score* tertinggi sebesar 0,8213 dengan *precision* sebesar 0,7798 dan *recall* sebesar 0,8673. Nilai tersebut lebih tinggi dibandingkan seluruh model individual yang digunakan sebagai baseline. Hasil ini menunjukkan bahwa penggabungan beberapa model *gradient boosting* mampu menghasilkan representasi pola *fraud* yang lebih baik dibandingkan penggunaan model tunggal.



Gambar 4. Perbandingan ROC Curve dan Precision-Recall Curve Seluruh Model

Berdasarkan Gambar 4, seluruh model menghasilkan kurva ROC yang mendekati sudut kiri atas grafik dengan nilai ROC-AUC di atas 0,97, sehingga menunjukkan kemampuan yang sangat baik dalam membedakan transaksi *fraud* dan *non-fraud* pada berbagai *threshold* klasifikasi. LightGBM memperoleh ROC-AUC tertinggi sebesar 0,9843, diikuti XGBoost sebesar 0,9823, Stacking Ensemble sebesar 0,9789, dan CatBoost sebesar 0,9761. Namun, karena dataset memiliki ketidakseimbangan kelas yang sangat tinggi, *Precision-Recall Curve* menjadi indikator yang lebih relevan. Pada grafik tersebut, XGBoost menghasilkan PR-AUC tertinggi sebesar 0,8807, diikuti *Stacking Ensemble* sebesar 0,8534, yang menunjukkan bahwa kedua model mampu mempertahankan *precision* yang relatif baik ketika *recall* meningkat. Meskipun demikian, *Stacking Ensemble* tetap dipilih sebagai model terbaik karena menghasilkan *F1-Score* tertinggi, sehingga menunjukkan keseimbangan paling baik antara *precision* dan *recall* dibandingkan model lainnya.

Untuk memperoleh gambaran yang lebih rinci mengenai karakteristik prediksi model terbaik, dilakukan analisis menggunakan confusion matrix sebagaimana ditunjukkan pada Gambar 5.



Gambar 5. Confusion Matrix Stacking Ensemble

Berdasarkan Gambar 5, *Stacking Ensemble* berhasil mendeteksi 85 transaksi *fraud* dan melewati 13 transaksi *fraud*. Model juga menghasilkan 24 transaksi *non-fraud* yang salah diklasifikasikan sebagai *fraud*. Jumlah *false positive* yang relatif kecil menunjukkan bahwa model mampu mengurangi kesalahan klasifikasi pada transaksi normal, sementara jumlah *true positive* yang tinggi menunjukkan bahwa model tetap mampu mengenali sebagian besar transaksi *fraud*. Hasil ini memperlihatkan bahwa *Stacking Ensemble* memiliki keseimbangan yang baik antara kemampuan mendeteksi *fraud* dan kemampuan meminimalkan alarm palsu.

Secara keseluruhan, hasil penelitian menunjukkan bahwa pendekatan *Stacking Ensemble* berbasis *Gradient Boosting* berhasil meningkatkan keseimbangan antara *precision* dan *recall* pada permasalahan deteksi *fraud* kartu kredit yang memiliki ketidakseimbangan kelas tinggi. Meskipun XGBoost menghasilkan PR-AUC tertinggi dan LightGBM menghasilkan ROC-AUC tertinggi, *Stacking Ensemble* memperoleh *F1-Score* tertinggi sebesar 0,8213 sehingga dipilih sebagai model terbaik pada penelitian ini.

V. KESIMPULAN DAN SARAN

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa pendekatan *Stacking Ensemble* berbasis *Gradient Boosting* berhasil diterapkan untuk mendeteksi transaksi *fraud* kartu kredit pada dataset yang memiliki ketidakseimbangan kelas yang tinggi. Proses pengembangan model dilakukan melalui tahapan *preprocessing* yang meliputi *feature engineering*, *stratified train-test split*, *RobustScaler*, *feature selection* berbasis SHAP, penanganan ketidakseimbangan kelas menggunakan SMOTE, serta optimasi *hyperparameter* menggunakan Optuna. Hasil analisis SHAP menunjukkan bahwa fitur V14, V4, V7, V10, dan V8 merupakan fitur yang paling berkontribusi terhadap prediksi model. Selain itu, hasil evaluasi menunjukkan bahwa penggunaan seluruh fitur memberikan performa yang lebih baik dibandingkan penggunaan sebagian fitur hasil seleksi.

Hasil eksperimen menunjukkan bahwa setiap model *baseline* memiliki karakteristik yang berbeda dalam mendeteksi transaksi *fraud*. XGBoost menghasilkan PR-AUC tertinggi sebesar 0,8807, LightGBM memperoleh ROC-AUC tertinggi sebesar 0,9843, sedangkan CatBoost menghasilkan *recall* tertinggi sebesar 0,9082. Namun demikian, model *Stacking Ensemble* mampu memberikan keseimbangan terbaik antara *precision* dan *recall* dengan memperoleh *precision* sebesar 0,7798, *recall* sebesar 0,8673, *F1-Score* sebesar 0,8213, ROC-AUC sebesar 0,9789, dan PR-AUC sebesar 0,8534. Hasil tersebut menunjukkan bahwa penggabungan beberapa model *gradient boosting* melalui *Logistic Regression* sebagai *meta-learner* mampu menghasilkan performa yang lebih baik dibandingkan penggunaan model individual.

Penelitian selanjutnya dapat mengeksplorasi metode *ensemble* lain seperti *blending* atau *weighted voting*, menggunakan *meta-learner* yang berbeda, membandingkan berbagai teknik penanganan ketidakseimbangan kelas, serta menguji model pada dataset *fraud* yang lebih beragam. Selain itu, optimasi *threshold* klasifikasi dan penerapan metode *Explainable Artificial Intelligence* (XAI) pada model akhir juga dapat dilakukan untuk meningkatkan interpretabilitas dan kesesuaian model pada implementasi nyata.

VI. DAFTAR PUSTAKA

- [1] The Nilson Report, "Global Card Fraud Losses at \$33 Billion," GlobeNewswire. Accessed: Jun. 17, 2026. [Online]. Available: <https://finance.yahoo.com/news/global-card-fraud-losses-33-170800960.html>
- [2] The Nilson Report, "Payment Card Fraud Losses Approach \$34 Billion," GlobeNewswire. Accessed: Jun. 17, 2026. [Online]. Available: <https://markets.financialcontent.com/bpas/article/gnwcq-2025-1-6-payment-card-fraud-losses-approach-34-billion>
- [3] T. Albalawi and S. Dardouri, "Enhancing credit card fraud detection using traditional and deep learning models with class imbalance mitigation," *Front. Artif. Intell.*, vol. 8, 2025, doi: 10.3389/frai.2025.1643292.
- [4] J. Liu, Y. Xiao, and L. Tan, "Application of Machine Learning Model in Fraud Identification: A Comparative Study of CatBoost, XGBoost and LightGBM," 2025, doi: 10.54254/2755-2721/119/2025.21637.
- [5] F. Almalki and M. Masud, "Financial Fraud Detection Using Explainable AI and Stacking Ensemble Methods," May 2025, [Online]. Available: <http://arxiv.org/abs/2505.10050>
- [6] E. B. Abdelghafour, C. Mohamed, A. Noura, and B. Abdelhamid, "Enhancing Credit Card Fraud Detection Using a Stacking Model Approach and Hyperparameter Optimization," 2024. [Online]. Available: www.ijacsa.thesai.org
- [7] K. Hayat and B. Magnier, "Data Leakage and Deceptive Performance: A Critical Examination of Credit Card Fraud Detection Methodologies," *Mathematics*, vol. 13, no. 16, Aug. 2025, doi: 10.3390/math13162563.
- [8] S. Kabane, "Impact of Sampling Techniques and Data Leakage on XGBoost Performance in Credit Card Fraud Detection," Dec. 2024, [Online]. Available: <http://arxiv.org/abs/2412.07437>
- [9] E. A. L. M. Btoush, X. Zhou, R. Gururajan, K. C. Chan, and O. Alsodi, "Enhancing credit card fraud detection with a stacking-based hybrid machine learning approach," *PeerJ Comput. Sci.*, vol. 11, 2025, doi: 10.7717/peerj-cs.3007.
- [10] J. P. Lai, Y. L. Lin, H. C. Lin, C. Y. Shih, Y. P. Wang, and P. F. Pai, "Tree-Based Machine Learning Models with Optuna in Predicting Impedance Values for Circuit Analysis," *Micromachines (Basel)*, vol. 14, no. 2, Feb. 2023, doi: 10.3390/mi14020265.
- [11] N. Baisholan, J. E. Dietz, S. Gnatyuk, M. Turdalyuly, E. T. Matson, and K. Baisholanova, "FraudX AI: An Interpretable Machine Learning Framework for Credit Card Fraud Detection on Imbalanced Datasets," *Computers*, vol. 14, no. 4, Apr. 2025, doi: 10.3390/computers14040120.
- [12] R. K. Gupta et al., "Enhanced framework for credit card fraud detection using robust feature selection and a stacking ensemble model approach," *Results in Engineering*, vol. 26, Jun. 2025, doi: 10.1016/j.rineng.2025.105084.
- [13] A. Mehdiyari, A. Chehri, A. Jakimi, and R. Saadane, "Hyperparameter Optimization with Genetic Algorithms and XGBoost: A Step Forward in Smart Grid Fraud Detection," *Sensors*, vol. 24, no. 4, Feb. 2024, doi: 10.3390/s24041230.
- [14] H. Z. Shao, X. Liu, D. S. Zong, and Q. J. Song, "Optimization of diabetes prediction methods based on combinatorial balancing algorithm," Dec. 01, 2024, Springer Nature. doi: 10.1038/s41387-024-00324-z.
- [15] X. Zhao, Y. Liu, and Q. Zhao, "Improved LightGBM for Extremely Imbalanced Data and Application to Credit Card Fraud Detection," 2024, doi: 10.1109/ACCESS.2023.1120000.
- [16] L. Wang, C. Zhang, W. Wang, T. Deng, T. Ma, and P. Shuai, "Slope Stability Assessment Using an Optuna-TPE-Optimized CatBoost Model," *Eng.*, vol. 6, no. 8, Aug. 2025, doi: 10.3390/eng6080185.
- [17] J. M. H. Pinheiro et al., "The Impact of Feature Scaling in Machine Learning: Effects on Regression and Classification Tasks," *IEEE Access*, vol. 13, pp. 199903–199931, 2025, doi: 10.1109/ACCESS.2025.3635541.