Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) DOI: 10.25126/jtiik.20241117640
Vol. 11, No. 1, Februari 2024, hlm. 125-134 p-ISSN: 2355-7699
Akreditasi KEMENRISTEKDIKTI, No. 36/E/KPT/2019 e-ISSN: 2528-6579

DETEKSI TRANSAKSI FRAUD KARTU KREDIT MENGGUNAKAN OVERSAMPLING ADASYN DAN SELEKSI FITUR SVM-RFECV

I Wayan Dharmana[*1], I Gede Aris Gunadi[2], Luh Joni Erawati Dewi[3]

[1,2,3] Universitas Pendidikan Ganesha, Singaraja
Email: [1]dharmana@undiksha.ac.id, [2]gedearisgunadi@undiksha.ac.id, [3]joni.erawati@undiksha.ac.id
[*]Penulis Korespondensi

(Naskah masuk: 16 Agustus 2023, diterima untuk diterbitkan: 23 Januari 2024)

Abstrak

Perkembangan kejahatan transaksi *fraud* kartu kredit memberikan dampak kerugian finansial bagi pemegang kartu. Pengembangan model deteksi transaksi *fraud* menggunakan *machine learning* telah dilakukan, namun memiliki beberapa tantangan meliputi ketidakseimbangan data serta dimensi dataset yang besar. Penelitian ini mengusulkan pendekatan pengembangan dengan seleksi fitur menggunakan SVM-RFECV dan metode *oversampling* dengan ADASYN. Pendekatan ini diharapkan mampu mengatasi permasalahan dimensi data serta ketidakseimbangan data yang terjadi. Seleksi fitur dengan SVM-RFECV menghasilkan variabel optimal pada rasio data latih 70% sejumlah 390 variabel, rasio data latih 80% sejumlah 400 variabel dan rasio data latih 90% sejumlah 390 variabel. Metode ADASYN telah memperbaiki ketidakseimbangan data dengan menghasilkan data sintetis berdasarkan rasio *oversampling* meliputi 100%, 50% dan 25%. Model yang menggunakan data hasil *oversampling* mengalami peningkatan kinerja AUC dan *recall*. Kinerja AUC tertinggi dihasilkan sejumlah 88,08% pada data latih 70%, *oversampling* 100% dan algoritma LGBM. Sedangkan, kinerja *recall* tertinggi sejumlah 83,08% dihasilkan saat menggunakan data latih 70%, *oversampling* 100% dengan algoritma AdaBoost. Berdasarkan pembahasan ini, maka dapat disimpulkan bahwa penggunaan *oversampling* dengan ADASYN dan seleksi fitur SVM-RFECV dapat dipertimbangkan untuk meningkatkan kinerja AUC dan *recall*.

Kata kunci: kartu kredit, *deteksi fraud, machine learning,* data tidak seimbang, seleksi fitur

DETECTION OF CREDIT CARD FRAUD TRANSACTION USING ADASYN OVERSAMPLING AND SVM-RFECV FEATURE SELECTION

Abstract

The growth of credit card fraud transaction crimes has a financial impact on cardholders. The development of credit card fraud transaction detection models using machine learning has been carried out, but there are several challenges, including data imbalance and large dataset dimensions. This research proposes a development approach with feature selection using SVM-RFECV and oversampling methods with ADASYN. This approach is expected to be able to overcome the problems of data dimensions and data imbalance. Feature selection with SVM-RFECV produced optimal variables at a training data ratio of 70% with 390 variables, a training data ratio of 80% with 400 variables, and a training data ratio of 90% with 390 variables. The ADASYN method has improved data imbalance by generating synthetic data based on the oversampling ratio, including 100%, 50%, and 25%. Models using oversampled data experienced improved AUC and recall performance. The highest AUC performance was produced at 88.08% on 70% training data, 100% oversampling, and the LGBM algorithm. Meanwhile, the highest recall performance of 83.08% was produced when using 70% training data, 100% oversampling with the AdaBoost algorithm. Based on this discussion, it can be concluded that the use of oversampling with ADASYN and feature selection with SVM-RFECV can be considered to improve AUC and recall performance.

Keywords: *credit card, fraud detection, machine learning, imbalanced dataset, feature selection*

1. PENDAHULUAN

Kejahatan penyalahgunaan kartu kredit untuk melakukan transaksi *fraud* memiliki dampak kerugian finansial dan terjadi cukup masif. Di India, telah terjadi kasus kebocoran informasi pribadi kartu kredit dari 70 Juta orang serta telah beredar pada situs Dark Web (Dubey, Mundhe and Kadam, 2020). Menurut laporan European Central Bank, nilai total transaksi *fraud* pada kartu yang diterbitkan pada

kawasan SEPA telah mencapai 1,87 Miliar Euro pada tahun 2019 dan sebagian besar terjadi pada kanal layanan digital (European Central Bank, 2021). Data lain dari Federal Trade Commission menyatakan bahwa pada tahun 2021 telah terjadi 88.354 transaksi *fraud* kartu kredit di Amerika Serikat dengan nilai transaksi mencapai 181 juta US Dollar (Federal Trade Commission, 2022). Data kasus-kasus ini memperlihatkan bahwa kasus *fraud* dalam kartu kredit perlu dilakukan pencegahan untuk mengurangi kerugian secara finansial.

Solusi pencegahan transaksi fraud menggunakan model *machine learning* perlu dikembangkan karena volume transaksi yang besar serta pola transaksi *fraud* yang dapat berubah sesuai perkembangan strategi kejahatan. Namun dalam pengembangan model prediksi ini terdapat beberapa masalah yang dihadapi yakni ketidakseimbangan rasio dataset *fraud* serta konsistensi kinerja model. Ketidakseimbangan data berkaitan dengan rasio antara transaksi *fraud* dan *not fraud* yang cukup jauh dan dapat mempengaruhi hasil prediksi. Dari segi performa, dataset yang tidak seimbang dapat menghasilkan akurasi yang tinggi. Namun hal ini bersifat bias dan hasil prediksinya condong ke kelas mayoritas. Selain itu, dimensi dataset yang besar menjadi suatu tantangan yang mempengaruhi komputasi. Sehingga dibutuhkan pendekatan untuk mengurangi dimensi tanpa menghilangkan informasi penting dalam suatu dataset.

Penelitian terkait deteksi transaksi *fraud* kartu kredit telah dilakukan oleh beberapa peneliti dengan melakukan pengembangan model menggunakan algoritma berbasis *supervised learning* seperti SVM, Naïve Bayes, Logistic Regression, Decision Tree, KNN, Random Forest, GBM, LightGBM, XGBoost, AdaBoost, serta CatBoost (Taha and Malebary, 2020) (Dileep, Navaneeth and Abhishek, 2021) (Sumanth et al., 2022) (Alfaiz and Fati, 2022) (Madhurya et al., 2022). Pendekatan untuk mengatasi data tidak seimbang pernah dilakukan dengan teknik *resampling* seperti Random Oversampling, SMOTE dan ADASYN oleh beberapa peneliti (Gupta et al., 2023) (Moreira et al., 2022) (Berkmans and Karthick, 2022) (Lu et al., 2020). Untuk pendekatan permasalahan terkait dimensi data yang besar juga pernah diusulkan menggunakan beberapa metode seperti *Pearson's Correlation Coefficient*, K-Best dan SVM-RFE (Mqadi, Naicker and Adeliyi, 2021) (Zhang et al., 2022) (Malik et al., 2022).

Penelitian oleh Berkmans dan Karthick memaparkan bahwa data yang tidak seimbang menjadi suatu tantangan yang signifikan untuk mengembangan model deteksi transaksi *fraud* kartu kredit. Jumlah sampel positif (transaksi *fraud*) yang minim mendorong adanya proses *oversampling* yang perlu dilakukan untuk menghasil lebih banyak data sintetis dari kelas minoritas. Dalam penelitian ini, metode berbasis SMOTE direkomendasikan untuk melakukan proses *resampling* dengan kinerja *recall* tertinggi sejumlah 81% dan precission sejumlah 86%. Dari aspek *future research*, pendekatan *resampling* lainnya perlu dilakukan kajian lebih lanjut dengan tujuan untuk meningkatkan kinerja dari *classifier* yang digunakan (Berkmans and Karthick, 2022). Hal ini sejalan dengan kekurangan metode SMOTE yang berpotensi membuat data sintetis yang *noisy* serta tumpang tindih (Grina, Elouedi and Lefevre, 2020) (Lu et al., 2020).

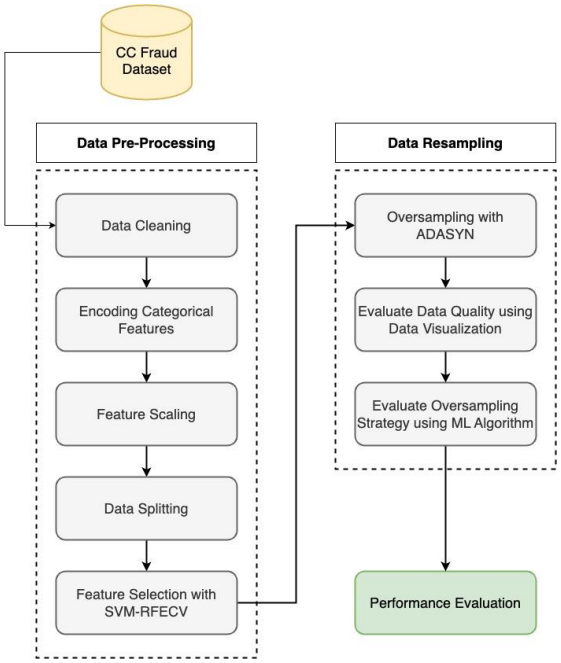
Penelitian oleh Malik dkk pernah memaparkan pendekatan seleksi fitur untuk mengatasi permasalahan dimensi dataset yang berpengaruh kepada sumber daya komputasi yang digunakan. Penelitian tersebut menggunakan metode seleksi fitur *correlation-based filter* serta SVM-RFE. Model prediksi yang menggunakan seleksi fitur dan *classifier* AdaBoost dan LGBM ini berhasil mendapatkan kinerja sejumlah 82% AUC. Dari segi aspek *future research*, terdapat pekerjaan lanjutan terkait penggunaan metode seleksi fitur lain dengan tujuan mendapatkan kinerja prediksi yang lebih optimal (Malik et al., 2022). Dalam penggunaan SVM-RFE terdapat keterbatasan dalam menentukan jumlah variabel yang optimal yang harus ditetapkan manual oleh peneliti. Sehingga metode ini hanya mampu memilih variabel sesuai dengan jumlah yang ditentukan.

Berdasarkan pemaparan tersebut, penelitian ini mengusulkan pendekatan deteksi transaksi *fraud* kartu kredit menggunakan pendekatan *oversampling* dengan metode ADASYN dan seleksi fitur dengan SVM-RFECV. Tujuan dari penelitian ini untuk mengusulkan suatu pendekatan alternatif pengembangan model deteksi transaksi *fraud* dengan memperbaiki distribusi data tidak seimbang dan permasalahan dimensi data dengan pendekatan seleksi fitur dengan kinerja yang lebih optimal. Keterbatasan dalam melakukan penentuan jumlah variabel paling optimal disolusikan dengan menggunakan SVM-RFECV yang memanfaatkan *cross validation* untuk memastikan jumlah variabel yang optimal dari beberapa subset yang diuji coba. Metode ADASYN digunakan dengan mempertimbangkan kemampuan membuat data sintetis berdasarkan distribusi tertimbang untuk sampel data berbeda dari kelas minoritas yang bergantung pada tingkat kesulitan sampel tersebut dipelajari.

Pembahasan dalam penelitian ini diharapkan dapat menjadi referensi bagi penyedia jasa keuangan khususnya penerbit kartu kredit untuk melakukan pengembangan sistem deteksi transaksi *fraud*. Selain itu, penelitian ini juga diharapkan memberikan kontribusi referensi terkait alternatif pendekatan pengembangan model deteksi *fraud* kartu kredit dengan *oversampling* ADASYN serta seleksi fitur SVM-RFECV.

2. METODE PENELITIAN

2.1. Rancangan Penelitian



Gambar 1. Alur Rancangan Penelitian

Berdasarkan ilustrasi alur penelitian pada Gambar 1, alur penelitian dapat dikelompokan sebagai berikut:

1. Pengumpulan Data
   Tahapan yang dilakukan untuk mengumpulkan dataset yang akan digunakan untuk proses pelatihan dan pengujian model prediksi transaksi *fraud*.
2. *Data Pre-Processing*
   Tahapan sebelum melakukan pengembangan model yang bertujuan untuk melakukan standarisasi data sesuai standar dataset untuk *machine learning*. Adapun tahapan data pre-processing meliputi *data cleaning, encoding categorical features, feature scaling, feature selection,* serta *data splitting.* Proses *data cleaning* dilakukan untuk mengolah dataset sehingga terhindar dari baris yang memiliki data kosong. Berikutnya, proses *encoding categorical features* dilakukan untuk mengkonversi kolom yang bersifat kategorikal menjadi numerik. Data juga dilakukan normalisasi menggunakan *MinMaxScaler* sehingga dataset akan memiliki rentang nilai 0 hingga 1. Berikutnya, proses *data splitting* dilakukan dengan membagi dataset menjadi beberapa komposisi data latih dan uji meliputi 70:30, 80:20 dan 90:10. Kemudian, masing-masing data latih dilakukan seleksi fitur untuk menemukan jumlah variabel yang optimal.
3. *Data Resampling*
   Tahapan data *resampling* bertujuan untuk memperbaiki distribusi data latih yang digunakan dalam proses deteksi transaksi fraud. Adapun tahapan ini meliputi proses *oversampling* dengan ADASYN, analisis kualitas data dengan visualisasi distribusi, serta uji klasifikasi.
4. Evaluasi Strategi *Oversampling*
   Dalam tahapan ini dilakukan evaluasi strategi *oversampling* yang paling optimal berdasarkan kinerja sesuai dengan metrik evaluasi AUC, *precision*, *recall* serta *f1-score*.

2.2. Teknik Pengumpulan Data

Teknik pengumpulan data yang dilakukan dalam penelitian ini meliputi studi literatur dan studi dokumentasi. Studi literatur dilakukan dengan mempelajari referensi dari artikel, buku serta sumber pustaka lainnya untuk mengetahui teori-teori dan penelitian sebelumnya yang relevan. Studi dokumentasi dilakukan dengan mengambil data riwayat transaksi kartu kredit yang terdapat pada dataset IEEE-CIS Fraud Detection yang sumber datanya disediakan Vesta Corporation dan dipublikasikan pada situs Kaggle.

Dataset yang digunakan merupakan riwayat transaksi yang meliputi tabel transaksi dan tabel identitas. Data yang digunakan merupakan data transaksi dengan tipe pembayaran kartu kredit yang berjumlah 148.986 baris data. Apabila kedua tabel digabungkan, maka dataset ini memiliki 432 variabel. Adapun variabel dataset yang digunakan dapat dilihat pada Tabel 1 dan Tabel 2.

Tabel 1. Variabel Tabel Transaksi

Variabel	Jenis	Keterangan
TransactionID	Numerik	ID Transaksi sebagai indeks
isFraud	Numerik	Target prediksi dengan nilai 0 (*Not Fraud*) dan 1 (*Fraud*)
TransactionDT	Numerik	*Timedelta* dari transaksi yang dilakukan
TransactionAmt	Numerik	Nominal transaksi dalam mata uang USD
ProductCD	Kategorikal	Kode produk setiap transaksi
card1-card6	Kategorikal	Informasi kartu seperti tipe kartu, kategori kartu, bank penerbit, negara, dan lainnya
addr1-addr2	Kategorikal	Alamat
dist1-dist2	Numerik	Jarak
P_emaildomain	Kategorikal	Domain email pembeli
R_emaildomain	Kategorikal	Domain email penerima
C1-C14	Numerik	Jumlah, seperti data jumlah alamat ditemukan yang diasosiasikan dengan kartu
D1-D15	Numerik	*Timedelta*, seperti jeda hari dengan transaksi sebelumnya
M1-M9	Kategorikal	*Match* yang berkaitan dengan tingkat kecocokan beberapa variabel seperti kecocokan dengan nama pemegang kartu dan alamat. Nilainya terdiri dari T (*True*) dan F (*False*)

V1-V339	Numerik	Variabel numerik khusus yang disediakan Vesta Corporation meliputi *ranking*, *counting*, serta variabel berkaitan lainnya

Tabel 2. Variabel Tabel Identitas

Variabel	Jenis	Keterangan
TransactionID	Numerik	ID Transaksi sebagai indeks
id_01-id_38	Numerik	Variabel numerik yang memuat informasi identitas meliputi informasi jaringan (IP, ISP, *Proxy*, dll) dan *digital signature* (*User Agent*, *Browser*, OS, *Version* dan lainnya). Variabel sudah dilakukan *masking* dalam bentuk format numerik.
DeviceType	Kategorikal	Informasi tipe perangkat seperti *mobile* atau *desktop*
DeviceInfo	Kategorikal	Informasi jenis perangkat yang digunakan

2.3. *Data Pre-Processing*

Tahapan *data pre-processing* dimulai dengan proses *data cleaning* yang dilakukan dengan melakukan penghapusan data yang memiliki nilai kosong pada kelompok variabel yang signifikan. Kelompok variabel setidaknya memiliki 10 variabel yang saling berkaitan seperti *counting*, *delta*, *v-variable*, serta *identity*. Apabila suatu data mengalami kekosongan nilai pada keseluruhan variabel pada suatu kelompok, maka data akan dihapuskan. Sedangkan apabila masih memenuhi ambang batas, maka akan dilakukan imputasi sesuai jenis variabel.

Berdasarkan hasil proses *data cleaning*, terdapat penghapusan data sejumlah 70.654 data. Apabila dilihat dari penyebab penghapusan, sebagian besar data transaksi tidak memiliki nilai pada data identitas. Maka dari itu, dalam penelitian ini akan fokus terhadap transaksi *fraud* yang memiliki identitas. Untuk data dengan *missing value* yang tersisa dilakukan imputasi dengan ketentuan *zero imputation* terhadap variabel numerik dan imputasi modus (*most frequent*) terhadap variabel kategorikal. Imputasi pada variabel numerik akan dilakukan dengan mengisi nilai 0. Sedangkan untuk variabel kategorikal akan dilakukan dengan mengisi nilai modus.

Data yang digunakan untuk pelatihan dan pengujian model prediksi harus dalam bentuk numerik. Hal ini karena *machine learning* hanya dapat mengolah data yang bersifat numerik. Oleh sebab itu, proses *encoding categorical features* perlu dilakukan untuk mengubah kolom kategorikal menjadi numerik. Dalam proses ini, teknik *label encoding* digunakan untuk mengubah variabel kategorikal menjadi nilai numerik. Setiap nilai unik pada variabel independen kategorikal akan diberikan nilai dengan rentang 0 hingga banyaknya data dikurangi satu. Adapun hasil transformasi dataset dapat dilihat pada Gambar 2.

Gambar 2. Hasil *Label Encoding* Kolom ProductCD

TransactionID	TransactionDT	TransactionAmt	ProductCD	card1	card2	card3	card4
2987004	0.000000	0.013009	0.25	0.200932	0.860972	0.381679	0.666667
2987010	0.000003	0.019781	0.00	0.891164	0.589615	0.129771	0.666667
2987017	0.000010	0.026089	0.25	0.603313	0.185930	0.381679	0.666667
2987022	0.000018	0.013009	0.25	0.041417	0.976549	0.381679	1.000000
2987038	0.000042	0.006469	0.75	0.256500	0.668342	0.381679	0.000000

Gambar 3. Hasil Normalisasi dengan *MinMaxScaler*

Dari segi skala data, dataset memiliki jangkauan data yang bervariasi antar fiturnya seperti TransactionAmt dan card4. TransactionAmt merupakan fitur yang merepresentasikan nominal transaksi pembayaran yang dilakukan menggunakan kartu kredit dengan rentang nilai antara 0,272 USD hingga 3.822,95 USD. Berbeda dengan variabel card4 yang merepresentasikan nilai biner dari penerbit kartu dengan merek Visa memiliki jangkauan data terbatas antara nilai 0 hingga 3. Apabila tidak dilakukan normalisasi, akan ada fitur yang memiliki kontribusi yang lebih dominan dibanding fitur lain dikarenakan jangkauan data yang berbeda. Proses *feature scaling* menggunakan MinMaxScaler dilakukan untuk normalisasi dataset yang digunakan. Adapun visualisasi hasil dataset yang telah dinormalisasi dapat dilihat pada Gambar 3.

Dataset yang sudah dilakukan normalisasi akan dibagi menjadi data latih dan data uji. Data latih digunakan dalam proses pelatihan model yang selanjutnya juga akan dibagi kembali menjadi data latih dan data validasi. Sedangkan data uji digunakan untuk mengevaluasi kemampuan model prediksi dalam melakukan prediksi terhadap data yang belum pernah digunakan saat proses pelatihan. Pembagian data latih dan data uji dilakukan dengan beberapa rasio data latih meliputi 70%, 80% dan 90%. Seluruh data latih mengalami ketidakseimbangan data antara kelas *fraud* dan *not fraud*.

2.4. *Data Resampling*

Proses *oversampling* dilakukan menggunakan ADASYN dengan pengaturan rasio *oversampling* yang dilakukan terhadap kelas minoritas. Hal ini untuk mengetahui keterkaitan antara rasio *oversampling* dengan kinerja prediksi. Dalam tahapan ini digunakan beberapa rasio *oversampling* meliputi 100%, 50% dan 25% dari kelas mayoritas. Konfigurasi *hyperparameter* yang digunakan untuk

melakukan *oversampling* dapat dilihat Tabel 3. Strategi rasio *oversampling* ini dikombinasikan beberapa variasi data latih yang dibagi pada tahapan *data splitting*. Hasil kombinasi rasio data latih dan rasio *oversampling* menghasilkan 12 kombinasi strategi. Termasuk dengan skenario saat data latih tidak dilakukan *oversampling*.

Tabel 3. *Hyperparameter* ADASYN

Hyperparameter	**Keterangan**	**Nilai**
sampling_strategy	Rasio *oversampling* yang menentukan jumlah data minoritas	1 = 100% 0.5 = 50% 0.25 = 25%
random_state	Konfigurasi nilai random yang digunakan dalam pembuatan data sintetis	0

Untuk memastikan kualitas data sintetis yang dihasilkan dengan *oversampling*, dilakukan tahapan analisis kualitas data. Distribusi antara data sebelum dan sesudah *oversampling* dilakukan perbandingan melalui visualisasi histogram. Histogram distribusi ini dibandingkan dari segi bentuk dan lokasi puncak. Apabila kedua histogram memiliki bentuk dan lokasi puncak yang sama, maka data sintetis yang dihasilkan memiliki distribusi yang sama pada variabel tersebut.

Data sintetis yang telah dilakukan analisis kualitas data akan dilanjut ke proses uji klasifikasi. Dalam proses ini menggunakan algoritma AdaBoost dan LGBM sebagai *classifier*. Apabila dikombinasikan dengan rasio data latih, rasio *oversampling* dan algoritma, maka terdapat 24 strategi yang dapat digunakan. Seluruh strategi ini akan dilakukan analisis untuk menemukan strategi paling optimal. Metrik evaluasi yang digunakan adalah AUC, *precision*, *recall* dan *f1-score*.

2.5. Evaluasi Strategi Oversampling

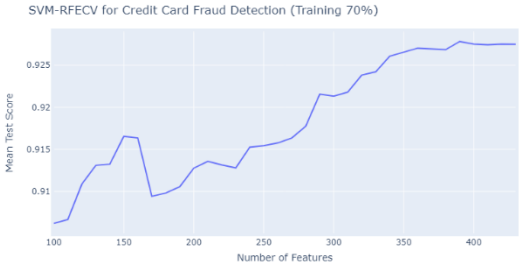
Untuk mengukur kinerja dari strategi yang digunakan, maka dilakukan evaluasi menggunakan metrik evaluasi yang umum digunakan. Hal ini untuk memastikan bahwa strategi yang digunakan mampu bekerja dengan baik untuk melakukan deteksi transaksi *fraud* dengan dataset dunia nyata. Dalam penelitian ini, ada beberapa metrik evaluasi yang digunakan meliputi AUC, *precission, recall* serta *f1-score*.

Dalam melakukan analisis strategi *oversampling* yang optimal, setiap metrik akan dilakukan analisis untuk mengetahui strategi yang paling optimal. Penelitian ini akan memberikan rekomendasi strategi yang optimal sesuai dengan metrik evaluasi yang akan digunakan dalam melakukan deteksi transaksi *fraud*. Analisis juga dilakukan dengan melakukan perbandingan *confusion matrix* untuk melihat kemampuan prediksi suatu model.

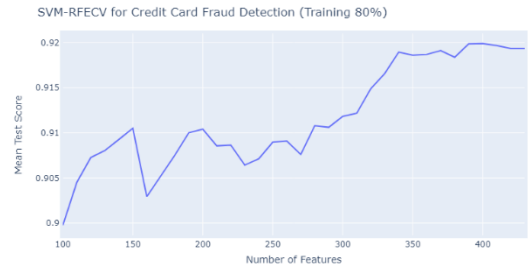
3. HASIL DAN PEMBAHASAN

3.1. Hasil Seleksi Fitur dengan SVM-RFECV

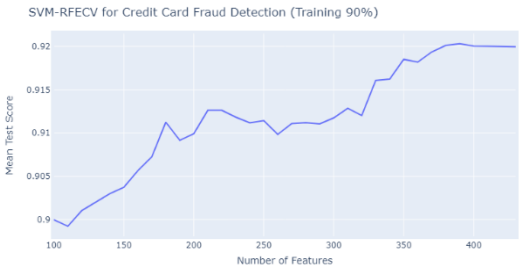
Metode SVM-RFECV berhasil mendapatkan sejumlah variabel optimal sesuai dengan data latih yang telah telah disiapkan. Saat menggunakan data latih dengan rasio 70%, didapatkan bahwa variabel yang optimal sejumlah 390 variabel. Untuk rasio data latih 80%, kinerja subset paling optimal ditemukan pada subset dengan variabel berjumlah 400. Sedangkan untuk rasio data latih 90%, didapatkan sejumlah 390 variabel optimal. Variabel optimal ditentukan berdasarkan kinerja *cross-validation* dengan menggunakan metrik AUC. Adapun visualisasi dari perbandingan kinerja subset sesuai rasio data latih 70%, 80% dan 90% dapat dilihat pada Gambar 4, Gambar 5 dan Gambar 6.



Gambar 4. Perbandingan Kinerja Subset Setelah Seleksi Fitur SVM-RFECV (70%)



Gambar 5. Perbandingan Kinerja Subset Setelah Seleksi Fitur SVM-RFECV (80%)



Gambar 6. Perbandingan Kinerja Subset Setelah Seleksi Fitur SVM-RFECV (90%)

3.2. Hasil *Oversampling* dengan ADASYN

3.2.1. Hasil *Oversampling* Data Latih 70%

Jumlah transaksi *fraud* yang belum dilakukan *oversampling* pada data latih 70% memiliki jumlah 6.192. *Oversampling* pada data latih 70% berhasil

melakukan penambahan data kelas minoritas menjadi 48.650 data pada rasio *oversampling* 100%, 24.117 data pada rasio *oversampling* 50% dan 12.571 data pada rasio *oversampling* 25%. Rincian distribusi hasil *oversampling* dapat dilihat pada Tabel 4.

Tabel 4. Rincian Distribusi Data *Oversampling* Data Latih 70%

	Jumlah Data	
	Fraud	***Not Fraud***
Tanpa *Oversampling*	48.650	6.192
Oversampling 100%	48.650	48.650
Oversampling 50%	48.650	24.117
Oversampling 25%	48.650	12.571

Selanjutnya, distribusi sebelum dan sesudah *oversampling* dilakukan analisis kualitas data dengan melakukan visualisasi menggunakan histogram. Berdasarkan analisis histogram dari sampel variabel yang digunakan, seluruh histogram memiliki bentuk dan lokasi puncak yang sama. Sehingga dapat disimpulkan bahwa data sintetis pada proses *oversampling* dengan data latih 70% memiliki distribusi yang baik pada seluruh rasio *oversampling*.

Uji klasifikasi pada data latih 70% dilakukan dengan menggunakan algoritma AdaBoost dan LGBM. Berdasarkan kinerja AUC, kinerja tertinggi didapatkan sejumlah 88,08% saat menggunakan strategi dengan rasio *oversampling* 100% dan algoritma LGBM. Apabila melihat kinerja *recall*, kinerja tertinggi sejumlah 83,08% didapatkan saat menggunakan algoritma AdaBoost dengan oversampling 100%. Sedangkan untuk kinerja *precision* dan f1 score, strategi yang menggunakan algoritma LGBM tanpa *oversampling* menjadi yang paling optimal. Rincian hasil uji klasifikasi dapat dilihat pada Tabel 7.

3.2.2. Hasil *Oversampling* Data Latih 80%

Transaksi *fraud* dalam dataset yang dilakukan *oversampling* terdapat sejumlah 7.670 data. *Oversampling* data latih 80% berhasil dilakukan dengan dengan menghasilkan distribusi jumlah transaksi sejumlah 54.657 data pada rasio *oversampling* 100%, 27.109 data pada rasio *oversampling* 50% dan 13.335 data pada rasio *oversampling* 25%. Rincian perbandingan distribusi hasil *oversampling* dapat dilihat pada Tabel 5. Analisis kualitas data sintetis yang dihasilkan melalui tahapan *oversampling* pada data latih 80% menghasilkan distribusi histogram dengan bentuk dan lokasi puncak yang sama. Maka dari itu, kualitas data yang dihasilkan pada distribusi data ini mampu menjadi representasi dari data asli.

Hasil uji klasifikasi pada data latih 80% memaparkan bahwa kinerja AUC meningkat setelah dilakukan *oversampling*. Kinerja AUC tertinggi didapatkan sejumlah 86,19% pada rasio *oversampling* 50% dengan menggunakan algoritma LGBM. Dari segi *precision*, kinerja paling optimal sejumlah 87,47% dihasilkan saat menggunakan algoritma LGBM tanpa *oversampling*. Metrik *recall* mengalami peningkatan setelah dilakukan *oversampling* dengan kinerja terbaik sejumlah 80,83% didapatkan saat menggunakan *oversampling* 100% dan algoritma AdaBoost. Untuk kinerja *f1 score* tertinggi didapatkan sejumlah 79,42% saat menggunakan algoritma LGBM dan *oversampling* 50%. Rincian hasil uji klasifikasi pada data latih 80% dapat dilihat pada Tabel 8.

Tabel 5. Rincian Distribusi Data *Oversampling* Data Latih 80%

	Jumlah Data	
	Fraud	***Not Fraud***
Tanpa *Oversampling*	55.007	7.670
Oversampling 100%	55.007	54.657
Oversampling 50%	55.007	27.109
Oversampling 25%	55.007	13.335

3.2.3. Hasil *Oversampling* Data Latih 90%

Data latih yang belum dilakukan *oversampling* dengan rasio data latih 90% memiliki transaksi *fraud* sejumlah 8.915 transaksi. Proses *oversampling* dilakukan dengan menggunakan rasio *oversampling* 100%, 50% dan 25% yang masing-masing menghasilkan transaksi *fraud* sejumlah 62.734 data, 30.036 data dan 15.048 data. Rincian distribusi data hasil *oversampling* data latih 90% dapat dilihat pada Tabel 6. Analisis kualitas data sintetis dilakukan pada data latih 90% yang menghasilkan sebuah kesimpulan bahwa seluruh data sintetis memiliki distribusi yang sama dengan distribusi data asli berdasarkan analisis dengan histogram. Hal ini berdasarkan analisis terhadap bentuk dan lokasi puncak histogram dan data sintetis yang sama dengan data asli.

Tabel 6. Rincian Distribusi Data *Oversampling* Data Latih 90%

	Jumlah Data	
	Fraud	***Not Fraud***
Tanpa *Oversampling*	61.597	8.915
Oversampling 100%	61.597	62.734
Oversampling 50%	61.597	30.036
Oversampling 25%	61.597	15.048

Uji klasifikasi yang dilakukan terhadap data latih dengan rasio 90% menghasilkan kinerja AUC tertinggi sejumlah 85,58%. Kinerja ini dihasilkan saat menggunakan data yang telah dilakukan *oversampling* sejumlah 100% dan menggunakan algoritma LGBM. Dari segi *precision*, kinerja tertinggi sejumlah 86,51% dihasilkan saat menggunakan algoritma LGBM tanpa *oversampling*. Kinerja *recall* mengalami peningkatan saat dilakukan prediksi menggunakan data yang telah dilakukan *oversampling* dengan kinerja tertinggi sejumlah 81,06%. Nilai tertinggi ini didapatkan saat menggunakan algoritma AdaBoost dengan rasio *oversampling* 100%. Terakhir, dari segi kinerja f1 score tertinggi sejumlah 78,27% dihasilkan saat menggunakan algoritma LGBM dengan rasio

oversampling sejumlah 25%. Rincian lengkap terkait hasil uji klasifikasi dapat dilihat pada Tabel 9.

Tabel 7. Hasil Uji Klasifikasi Data Latih 70%

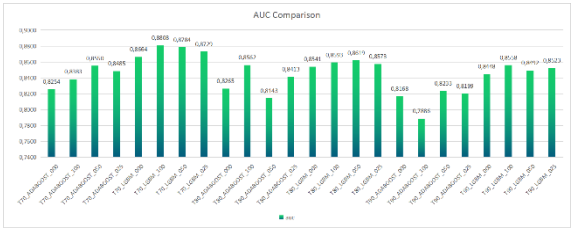
Model	**AUC**	***Precision***	***Recall***	***F1 Score***
AdaBoost – *Imbalanced Data*	0,8254	0,8482	0,6738	0,7510
AdaBoost – *Oversampling* 100%	0,8383	0,5062	0,8308	0,6291
AdaBoost – *Oversampling* 50%	0,8550	0,6921	0,7757	0,7315
AdaBoost – *Oversampling* 25%	0,8485	0,7555	0,7427	0,7491
LGBM – *Imbalanced Data*	0,8664	0,8912	0,7501	0,8146
LGBM – *Oversampling* 100%	0,8808	0,8225	0,7943	0,8082
LGBM – *Oversampling* 50%	0,8784	0,8454	0,7842	0,8136
LGBM – *Oversampling* 25%	0,8729	0,8550	0,7706	0,8106

Tabel 8. Hasil Uji Klasifikasi Data Latih 80%

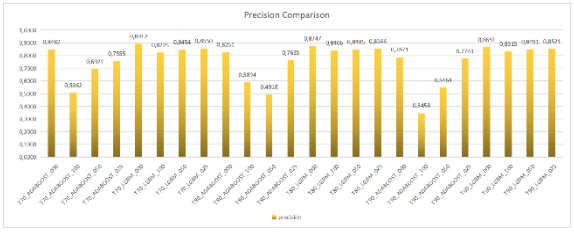
Model	**AUC**	***Precision***	***Recall***	***F1 Score***
AdaBoost – *Imbalanced Data*	0,8265	0,8261	0,6772	0,7443
AdaBoost – *Oversampling* 100%	0,8562	0,5894	0,8083	0,6817
AdaBoost – *Oversampling* 50%	0,8143	0,4918	0,7627	0,5980
AdaBoost – *Oversampling* 25%	0,8413	0,7635	0,7206	0,7414
LGBM – *Imbalanced Data*	0,8541	0,8747	0,7259	0,7934
LGBM – *Oversampling* 100%	0,8593	0,8406	0,7425	0,7885
LGBM – *Oversampling* 50%	0,8619	0,8485	0,7465	0,7942
LGBM – *Oversampling* 25%	0,8573	0,8544	0,7360	0,7908

Tabel 9. Hasil Uji Klasifikasi Data Latih 90%

Model	**AUC**	***Precision***	***Recall***	***F1 Score***
AdaBoost – *Imbalanced Data*	0,8168	0,7871	0,6609	0,7185
AdaBoost – *Oversampling* 100%	0,7886	0,3458	0,8106	0,4848
AdaBoost – *Oversampling* 50%	0,8233	0,5464	0,7401	0,6286
AdaBoost – *Oversampling* 25%	0,8199	0,7743	0,6696	0,7181
LGBM – *Imbalanced Data*	0,8448	0,8651	0,7063	0,7777
LGBM – *Oversampling* 100%	0,8558	0,8315	0,7343	0,7799
LGBM – *Oversampling* 50%	0,8492	0,8491	0,7179	0,7780
LGBM – *Oversampling* 25%	0,8523	0,8521	0,7237	0,7827

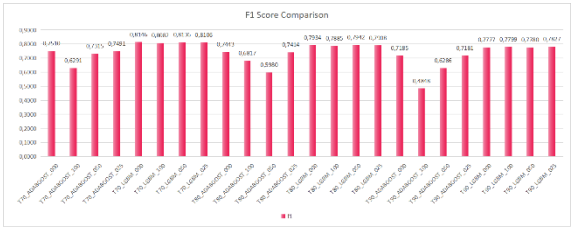


(a) Perbandingan Kinerja AUC



(b) Perbandingan Kinerja *Precision*

(c) Perbandingan Kinerja *Recall*



(d) Perbandingan Kinerja *F1 Score*

Gambar 7. Perbandingan Kinerja AUC, *Precision*, *Recall* dan *F1 Score*

(a) Sebelum *Oversampling* (b) Setelah *Oversampling*

Gambar 8. *Confusion Matrix* Strategi Data Latih 70% dengan Algoritma LGBM

3.3. Analisis Strategi *Oversampling*

Dalam penelitian ini terdapat 24 skenario strategi yang dapat digunakan untuk melakukan deteksi transaksi *fraud* kartu kredit yang merupakan hasil kombinasi antara rasio data latih, rasio data uji, rasio *oversampling* serta algoritma yang digunakan. Setiap strategi memiliki kinerja, keunggulan dan kelemahan masing-masing. Bagian ini membahas strategi yang optimal digunakan sesuai metrik yang dijadikan penilaian utama. Apabila menggunakan AUC sebagai metrik utama, model yang memiliki kinerja terbaik dari seluruh skenario adalah model yang menggunakan data latih 70%, data uji 30%, rasio *oversampling* 100% serta algoritma LGBM. Kinerja tertinggi didapatkan dengan nilai AUC 88,08%. Dari segi *precision*, strategi yang menggunakan data yang dilakukan *oversampling* cenderung mengalami penurunan kinerja. Hal berbeda terjadi apabila pengembangan model menjadi *recall* sebagai penilaian kinerja utama.

Kinerja *recall* cenderung meningkat apabila menggunakan data yang telah dilakukan *oversampling*. Kinerja tertinggi sejumlah 83,08% dihasilkan ketika menggunakan data latih 70%, data uji 30%, *oversampling* 100% dan algoritma AdaBoost. Kinerja *F1 Score* cenderung variatif sesuai dengan nilai kinerja *precision* serta *recall* yang dihasilkan masing-masing model prediksi. Hal ini karena *f1 score* merupakan harmonisasi dari kedua metrik ini. Kinerja *F1 Score* paling optimal didapatkan saat menggunakan data latih 70%, data uji 30%, tanpa *oversampling* dan menggunakan algoritma LGBM. Gambaran perbandingan kinerja seluruh skenario dapat dilihat pada Gambar 7.

Berdasarkan penjelasan sebelumnya, strategi yang menggunakan *oversampling* mampu meningkatkan kinerja AUC dan *recall*. Untuk memahami penyebab terjadinya peningkatan kinerja ini, maka akan dilakukan perbandingan menggunakan *confusion matrix* yang dapat dilihat pada Gambar 8. Dalam perbandingan tersebut terlihat bahwa terdapat peningkatan jumlah prediksi positif yang benar (*true positive*). Jumlah transaksi fraud yang diprediksi dengan benar mengalami peningkatan apabila dibandingkan dengan jumlah prediksi saat menggunakan data yang tidak seimbang yang awalnya berjumlah 2.819 menjadi 2.985 transaksi *fraud*. Berbeda dengan *precision* yang membandingkan jumlah prediksi positif yang benar. Kinerja *recall* membandingkan jumlah prediksi positif yang benar dengan jumlah aktual positif yang dalam hal ini merupakan jumlah aktual transaksi *fraud* sesuai data uji. Sehingga setiap adanya peningkatan jumlah transaksi *fraud* yang diprediksi dengan benar, maka akan menyebabkan peningkatan kinerja *recall*. Apabila melihat paparan tersebut sebelumnya, maka pemilihan strategi dalam melakukan deteksi transaksi *fraud* perlu menentukan metrik evaluasi yang digunakan agar strategi yang digunakan lebih optimal dan tepat sasaran.

4. KESIMPULAN DAN SARAN

Permasalahan transaksi *fraud* kartu kredit mendorong pengembangan model deteksi transaksi *fraud* yang mampu melakukan pencegahan terhadap terjadinya transaksi *fraud*. Model deteksi *fraud* memiliki beberapa tantangan seperti data tidak seimbang serta dimensi dataset yang cukup besar. Penelitian ini diharapkan mampu memberikan pendekatan alternatif untuk mengatasi data tidak seimbang dengan melakukan strategi *oversampling* yang optimal. Di sisi lain, permasalahan komputasi yang diakibatkan oleh dimensi dataset perlu diselesaikan dengan melakukan seleksi fitur tanpa menghilangkan variabel yang memberikan kontribusi optimal. Kinerja dari model deteksi juga diharapkan dapat mengalami peningkatan setelah dilakukan pendekatan alternatif dalam perbaikan data tidak seimbang serta dimensi dataset.

Penelitian ini telah melakukan pendekatan seleksi fitur dengan SVM-RFECV. Keterbatasan pada penelitian yang dilakukan oleh Malik dkk (2022) telah dilakukan pengembangan dengan menggunakan *cross validation* sehingga jumlah variabel yang optimal dapat ditentukan sesuai data latih yang digunakan. Seleksi fitur dalam penelitian ini menghasilkan variabel optimal sejumlah 390 variabel pada data latih 70%, 400 variabel optimal pada data latih 80% dan 390 variabel optimal pada data latih 90%. Variabel diseleksi berdasarkan nilai

kontribusi suatu variabel terhadap variabel target (*isFraud*). Dengan adanya seleksi fitur, maka variabel yang digunakan saat *oversampling*, pelatihan dan pengujian model menjadi lebih sedikit dan mempengaruhi sumber daya komputasi yang digunakan.

Pendekatan untuk mengatasi permasalahan data tidak seimbang dilakukan dengan melakukan *oversampling* terhadap data kelas minoritas menggunakan metode ADASYN. Metode ini juga mengatasi potensi adanya *overlapping* yang terjadi ketika menggunakan SMOTE dengan melakukan pembobotan dan fokus pada data yang sulit untuk dipelajari. ADASYN berhasil memperbaiki ketidakseimbangan data dengan membuat data sintetis sesuai rasio *oversampling* 100%, 50% dan 25%. Hasil analisis kualitas data berdasarkan distribusi data sintetis dengan histogram menunjukan bahwa data sintetis yang dihasilkan dengan ADASYN memiliki bentuk dan lokasi puncak yang sama. Data sintetis yang dihasilkan mampu menjadi representasi data transaksi *fraud* yang asli.

Hasil uji klasifikasi memaparkan bahwa model yang menggunakan data yang dilakukan *oversampling* mengalami peningkatan kinerja AUC dan Recall dibandingkan dengan model dengan data yang tidak dilakukan *oversampling*. Kinerja AUC tertinggi didapatkan sejumlah 88,08% saat menggunakan data latih 70%, rasio *oversampling* 100% dan algoritma LGBM. Sedangkan kinerja *recall* tertinggi didapatkan saat menggunakan data latih 70%, rasio *oversampling* 100% dan algoritma AdaBoost. *Recall* yang dihasilkan dalam strategi ini sejumlah 83,08%. Hasil menunjukan bahwa metode *oversampling* dengan ADASYN serta seleksi fitur dengan SVM-RFECV mampu menghasilkan kinerja yang optimal untuk melakukan deteksi transaksi *fraud* kartu kredit. Berdasarkan pemaparan tersebut, maka dapat disimpulkan bahwa metode ADASYN dan seleksi fitur SVM-RFECV dapat dipertimbangkan dalam melakukan deteksi transaksi *fraud* untuk meningkatkan kinerja AUC dan *recall*.

Dalam penelitian ini masih terdapat lingkup pekerjaan yang dapat dilakukan peningkatan dan pembahasan lanjutan pada penelitian berikutnya. Data transaksi *fraud* yang digunakan memiliki variabel terkait dengan waktu transaksi. Penelitian selanjutnya dapat melakukan analisis terkait dengan potensi terjadinya transaksi *fraud* yang berfokus pada waktu. Sehingga dapat diketahui keterkaitan antara waktu dengan potensi terjadinya transaksi *fraud* menggunakan pendekatan *time series*. Selain itu, terdapat suatu kondisi hasil model prediksi *machine learning* yang tidak mudah dipahami oleh manusia. Model *machine learning* tidak dapat menjelaskan penyebab suatu transaksi diklasifikasikan sebagai *fraud* atau *not fraud*. Maka dari itu, penelitian selanjutnya dapat fokus terkait dengan implementasi *explainable articial intelligence* (XAI) untuk memaparkan penyebab suatu transaksi dianggap sebagai transaksi *fraud* dengan visualisasi yang mudah dipahami oleh manusia.

DAFTAR PUSTAKA

ALFAIZ, N.S. and FATI, S.M., 2022. Enhanced Credit Card Fraud Detection Model Using Machine Learning. *Electronics*, 11(4), p.662. https://doi.org/10.3390/electronics11040662.

BERKMANS, T.J. and KARTHICK, S., 2022. Credit Card Fraud Detection with Data Sampling. In: *2022 International Conference on Power, Energy, Control and Transmission Systems (ICPECTS)*. IEEE. pp.1–6. https://doi.org/10.1109/ICPECTS56089.2022.10046729.

DILEEP, M.R., NAVANEETH, A. V and ABHISHEK, M., 2021. A Novel Approach for Credit Card Fraud Detection using Decision Tree and Random Forest Algorithms. In: *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*. IEEE. pp.1025–1028. https://doi.org/10.1109/ICICV50876.2021.9388431.

DUBEY, S.C., MUNDHE, K.S. and KADAM, A.A., 2020. Credit Card Fraud Detection using Artificial Neural Network and BackPropagation. In: *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*. IEEE. pp.268–273. https://doi.org/10.1109/ICICCS48265.2020.9120957.

EUROPEAN CENTRAL BANK, 2021. *Seventh report on card fraud*.

FEDERAL TRADE COMMISSION, 2022. *Consumer Sentinel Network Data Book 2021*. Washington, DC.

GRINA, F., ELOUEDI, Z. AND LEFEVRE, E., 2020. A Preprocessing Approach for Class-Imbalanced Data Using SMOTE and Belief Function Theory. pp.3–11. https://doi.org/10.1007/978-3-030-62365-4_1.

GUPTA, P., VARSHNEY, A., KHAN, M.R., AHMED, R., SHUAIB, M. and ALAM, S., 2023. Unbalanced Credit Card Fraud Detection Data: A Machine Learning-Oriented Comparative Study of Balancing Techniques. *Procedia Computer Science*, 218, pp.2575–2584. https://doi.org/10.1016/j.procs.2023.01.231.

LU, C., LIN, S., LIU, X. and SHI, H., 2020. Telecom Fraud Identification Based on ADASYN and Random Forest. In: *2020 5th International Conference on Computer and*

Communication Systems (ICCCS). IEEE. pp.447–452. https://doi.org/10.1109/ICCCS49078.2020.9118521.

MADHURYA, M.J., GURURAJ, H.L., SOUNDARYA, B.C., VIDYASHREE, K.P. and RAJENDRA, A.B., 2022. Exploratory analysis of credit card fraud detection using machine learning techniques. *Global Transitions Proceedings*, 3(1), pp.31–37. https://doi.org/10.1016/j.gltp.2022.04.006.

MALIK, E.F., KHAW, K.W., BELATON, B., WONG, W.P. and CHEW, X., 2022. Credit Card Fraud Detection Using a New Hybrid Machine Learning Architecture. *Mathematics*, 10(9), p.1480. https://doi.org/10.3390/math10091480.

MOREIRA, M.Â.L., JUNIOR, C. DE S.R., SILVA, D.F. DE L., DE CASTRO JUNIOR, M.A.P., COSTA, I.P. DE A., GOMES, C.F.S. and DOS SANTOS, M., 2022. Exploratory analysis and implementation of machine learning techniques for predictive assessment of fraud in banking systems. *Procedia Computer Science*, 214, pp.117–124. https://doi.org/10.1016/j.procs.2022.11.156.

MQADI, N.M., NAICKER, N. and ADELIYI, T., 2021. Solving Misclassification of the Credit Card Imbalance Problem Using Near Miss. *Mathematical Problems in Engineering*, 2021, pp.1–16. https://doi.org/10.1155/2021/7194728.

SUMANTH, C.H., KALYAN, P.P., RAVI, B. and BALASUBRAMANI., S., 2022. Analysis of Credit Card Fraud Detection using Machine Learning Techniques. In: *2022 7th International Conference on Communication and Electronics Systems (ICCES)*. IEEE. pp.1140–1144. https://doi.org/10.1109/ICCES54183.2022.9835751.

TAHA, A.A. and MALEBARY, S.J., 2020. An Intelligent Approach to Credit Card Fraud Detection Using an Optimized Light Gradient Boosting Machine. *IEEE Access*, 8, pp.25579–25587. https://doi.org/10.1109/ACCESS.2020.2971354.

ZHANG, A., YU, H., HUAN, Z., YANG, X., ZHENG, S. and GAO, S., 2022. SMOTE-RkNN: A hybrid re-sampling method based on SMOTE and reverse k-nearest neighbors. *Information Sciences*, 595, pp.70–88. https://doi.org/10.1016/j.ins.2022.02.038.