

Deteksi Pencurian Listrik pada Data Smart Meter Menggunakan Pembelajaran Mesin

Rafli Alfian¹, Edmund Ucok Armin²

^{1,2}Program Studi Teknik Elektro, Fakultas Teknik, Universitas Singaperbangsa Karawang
¹2110631160057@student.unsika.ac.id, ²edmun.ucok.armin@ft.unsika.ac.id*

Abstract

Electricity theft is a primary source of non-technical losses (NTL) in power distribution systems, causing substantial economic damage to utilities. This study develops a machine learning approach to detect electricity theft using real smart meter data from the State Grid Corporation of China (SGCC), comprising 42,372 consumers over 1,034 days with an 8.53% theft rate. Nineteen statistical features were engineered from the daily consumption profiles after handling 25.64% missing values through linear interpolation and outliers through the three-sigma rule. Four classifiers, namely Logistic Regression, Decision Tree, Random Forest, and histogram-based Gradient Boosting, were trained with the Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance and evaluated on a stratified hold-out set. The Gradient Boosting model achieved the best performance with an ROC-AUC of 0.750, an F1-score of 0.323, and a recall of 0.416, outperforming the other classifiers, while cross-validation confirmed the stability of the results. Feature importance analysis revealed that the missing-data ratio, the low-consumption ratio, and the longest zero-consumption run are the most discriminative indicators of theft. The results show that statistical feature engineering combined with ensemble learning provides an effective and computationally efficient method for electricity theft detection in smart grids.

Keywords: electricity theft, smart meter, machine learning, gradient boosting, SGCC

Abstrak

Pencurian listrik merupakan penyebab utama kerugian non-teknis (non-technical loss) pada sistem distribusi tenaga listrik dan menimbulkan kerugian ekonomi yang besar bagi penyedia layanan. Penelitian ini mengembangkan pendekatan pembelajaran mesin untuk mendeteksi pencurian listrik menggunakan data smart meter nyata dari State Grid Corporation of China (SGCC) yang terdiri atas 42.372 pelanggan selama 1.034 hari dengan proporsi pencurian sebesar 8,53%. Sebanyak 19 fitur statistik diekstraksi dari profil konsumsi harian setelah menangani 25,64% nilai yang hilang melalui interpolasi linier dan pencilan (outlier) melalui aturan tiga-sigma. Empat algoritma klasifikasi, yaitu Regresi Logistik, Decision Tree, Random Forest, dan Gradient Boosting berbasis histogram, dilatih menggunakan teknik SMOTE untuk mengatasi ketidakseimbangan kelas dan dievaluasi pada data uji terpisah secara stratified. Model Gradient Boosting memberikan kinerja terbaik dengan ROC-AUC sebesar 0,750, F1-score 0,323, dan recall 0,416, mengungguli algoritma lainnya, sedangkan validasi silang menegaskan kestabilan hasil tersebut. Analisis feature importance menunjukkan bahwa rasio data hilang, rasio konsumsi rendah, dan durasi konsumsi nol terpanjang merupakan indikator pencurian yang paling diskriminatif. Hasil penelitian membuktikan bahwa rekayasa fitur statistik yang dipadukan dengan ensemble learning menghasilkan metode deteksi pencurian listrik yang efektif dan efisien secara komputasi untuk smart grid.

Kata kunci: pencurian listrik, smart meter, pembelajaran mesin, gradient boosting, SGCC

1. Pendahuluan

Kerugian energi pada sistem distribusi tenaga listrik secara umum terbagi menjadi kerugian teknis dan kerugian non-teknis (non-technical loss, NTL).

Kerugian non-teknis sebagian besar disebabkan oleh pencurian listrik, manipulasi meter, sambungan ilegal, dan kesalahan penagihan. Kerugian ini tidak hanya menurunkan pendapatan perusahaan listrik, tetapi juga membebani pelanggan yang jujur serta menurunkan

kualitas dan keandalan pasokan daya [1], [12]. Oleh karena itu, deteksi pencurian listrik yang akurat menjadi kebutuhan penting bagi operator jaringan.

Selain berdampak ekonomi, pencurian listrik juga menimbulkan risiko keselamatan dan keamanan jaringan, misalnya sambungan ilegal yang dapat memicu kebakaran atau membebani transformator secara berlebihan. Keamanan smart grid menjadi perhatian penting seiring meningkatnya digitalisasi sistem tenaga listrik [2]. Dengan demikian, deteksi dini pencurian listrik berkontribusi tidak hanya pada efisiensi finansial, tetapi juga pada keandalan dan keamanan sistem secara menyeluruh.

Penerapan smart meter dan infrastruktur Advanced Metering Infrastructure (AMI) menghasilkan data konsumsi pelanggan dalam resolusi tinggi. Ketersediaan data tersebut membuka peluang penerapan teknik pembelajaran mesin (machine learning) untuk mengidentifikasi pola konsumsi tidak wajar yang mengindikasikan pencurian. Pendekatan berbasis data dinilai lebih ekonomis dibandingkan inspeksi fisik di lapangan yang memerlukan biaya dan waktu besar [12].

Sejumlah penelitian terdahulu telah mengkaji deteksi pencurian listrik. Jindal dkk. [4] menggabungkan decision tree dan support vector machine untuk analitik deteksi pencurian pada smart grid. Zheng dkk. [1] memperkenalkan arsitektur Wide and Deep Convolutional Neural Network sekaligus merilis dataset SGCC yang kini menjadi acuan. Buzau dkk. [3] mengembangkan jaringan saraf dalam hibrida untuk mendeteksi kerugian non-teknis, sedangkan Hasan dkk. [5] memadukan CNN dan LSTM. Punmiya dan Choe [6] menunjukkan bahwa gradient boosting dengan rekayasa fitur memberikan kinerja kompetitif dengan biaya komputasi yang rendah.

Sebagian besar studi terbaru berfokus pada model pembelajaran dalam yang menuntut sumber daya komputasi besar dan waktu pelatihan yang lama. Di sisi lain, pendekatan berbasis rekayasa fitur statistik yang dipadukan dengan algoritma klasik dan ensemble belum banyak dibandingkan secara sistematis pada dataset nyata berskala besar. Celah inilah yang menjadi fokus penelitian ini, yaitu menyediakan perbandingan yang adil dan dapat direproduksi antara model linier, pohon tunggal, bagging, dan boosting dengan praproses serta penyeimbangan kelas yang beragam.

Meskipun demikian, deteksi pencurian listrik menghadapi dua tantangan utama, yaitu ketidakseimbangan kelas yang ekstrem (jumlah pencuri jauh lebih sedikit daripada pelanggan normal) dan banyaknya nilai yang hilang pada data smart meter. Kedua faktor ini dapat menurunkan kinerja model apabila tidak ditangani dengan tepat. Selain itu, model pembelajaran dalam (deep learning) sering kali

membutuhkan sumber daya komputasi besar sehingga kurang praktis untuk diterapkan secara cepat.

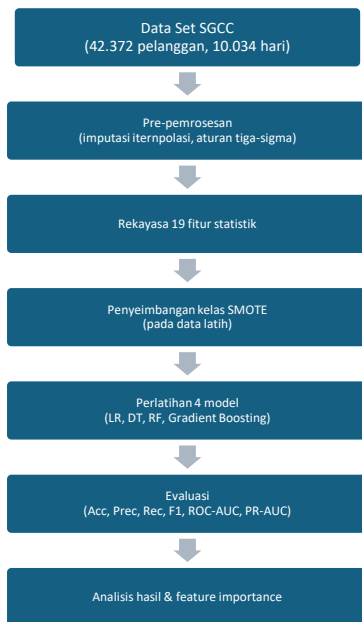
Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk membangun dan membandingkan empat algoritma pembelajaran mesin, yaitu Regresi Logistik, Decision Tree, Random Forest, dan Gradient Boosting, dalam mendeteksi pencurian listrik pada data smart meter SGCC. Penelitian ini menekankan rekayasa fitur statistik yang ringan dan penanganan ketidakseimbangan kelas menggunakan SMOTE. Kontribusi utama penelitian adalah identifikasi model dengan kinerja terbaik serta analisis fitur yang paling diskriminatif sebagai indikator pencurian listrik.

Tantangan kerugian non-teknis bersifat universal dan dialami oleh banyak perusahaan utilitas di berbagai negara, termasuk negara berkembang yang tingkat kerugian distribusinya relatif tinggi [12]. Dalam konteks tersebut, metode deteksi berbasis data yang efisien dan mudah diterapkan menjadi sangat relevan karena tidak memerlukan pemasangan perangkat khusus, melainkan cukup memanfaatkan data konsumsi yang sudah terekam oleh smart meter. Penelitian ini berfokus pada pendekatan demikian dengan mengutamakan kesederhanaan, kecepatan, dan kejelasan hasil.

2. Metode Penelitian

Penelitian ini menggunakan pendekatan kuantitatif eksperimental dengan tahapan: pengumpulan data, pra-pemrosesan, rekayasa fitur, penanganan ketidakseimbangan kelas, pelatihan model, dan evaluasi. Seluruh eksperimen diimplementasikan menggunakan bahasa pemrograman Python dengan pustaka scikit-learn [11] dan imbalanced-learn.

Gambar 1 menyajikan diagram alur tahapan penelitian, mulai dari pengumpulan data, pra-pemrosesan, rekayasa fitur, penyeimbangan kelas, pelatihan model, hingga evaluasi dan analisis hasil. Seluruh tahap dirancang agar dapat direproduksi.



Gambar 1. Diagram alur tahapan penelitian

2.1. Dataset

Dataset yang digunakan adalah data smart meter publik yang dirilis oleh State Grid Corporation of China (SGCC) [1]. Dataset memuat catatan konsumsi listrik harian dari 42.372 pelanggan selama periode 1 Januari 2014 hingga 31 Oktober 2016, sehingga setiap pelanggan memiliki 1.034 nilai konsumsi harian. Setiap pelanggan diberi label biner, yaitu normal (0) atau terindikasi mencuri (1).

Karakteristik dataset diringkas pada Tabel 1. Distribusi kelas sangat tidak seimbang dengan hanya 8,53% pelanggan terindikasi mencuri. Selain itu, sekitar 25,64% sel data bernilai hilang (missing), yang umum terjadi pada pengukuran smart meter akibat gangguan komunikasi atau kerusakan perangkat.

Dataset SGCC dipilih karena merupakan data smart meter nyata berskala besar yang telah banyak digunakan sebagai tolok ukur (benchmark) pada penelitian deteksi pencurian listrik [1], [5], [6]. Penggunaan data nyata memastikan bahwa hasil yang diperoleh mencerminkan kompleksitas pola konsumsi di dunia nyata, termasuk keberadaan derau (noise) dan nilai yang hilang.

Tabel 1. Karakteristik dataset SGCC

Karakteristik	Nilai
Jumlah pelanggan	42.372
Periode pengukuran	1 Jan 2014 - 31 Okt 2016
Jumlah hari (fitur waktu)	1.034
Pelanggan normal	38.757 (91,47%)
Pelanggan pencurian	3.615 (8,53%)
Rasio nilai hilang	25,64%
Jumlah fitur statistik	19

2.2. Pra-pemrosesan Data

Tahap pra-pemrosesan diawali dengan penanganan nilai hilang. Nilai konsumsi yang hilang direkonstruksi menggunakan interpolasi linier terhadap deret waktu masing-masing pelanggan, dan sisa nilai yang tidak dapat diinterpolasi diisi dengan nol. Nilai konsumsi negatif yang tidak valid juga diatur menjadi nol.

Selanjutnya, pencilan (outlier) ditangani menggunakan aturan tiga-sigma, yaitu nilai konsumsi yang melampaui batas rata-rata ditambah tiga kali simpangan baku dibatasi pada nilai batas tersebut, sebagaimana dirumuskan pada persamaan (1). Hal ini mencegah nilai ekstrem mendominasi proses pelatihan.

$$x_{\text{cap}} = \min(x, \mu + 3\sigma) \quad (1)$$

dengan x_{cap} adalah nilai konsumsi setelah pembatasan, μ adalah rata-rata konsumsi pelanggan, dan σ adalah simpangan baku konsumsi pelanggan. Setelah pembatasan pencilan, nilai fitur dinormalisasi menggunakan StandardScaler agar memiliki skala yang setara sebelum pelatihan.

2.3. Rekayasa Fitur

Alih-alih menggunakan 1.034 nilai konsumsi harian secara langsung, penelitian ini mengekstraksi 19 fitur statistik yang merangkum perilaku konsumsi tiap pelanggan. Fitur tersebut meliputi ukuran pemusatan dan penyebaran (rata-rata, simpangan baku, nilai maksimum, minimum, median, kuartil, rentang interkuartil, rentang, dan total konsumsi), koefisien variasi, kemencengan (skewness), keruncingan (kurtosis), serta indikator perilaku konsumsi rendah dan nol (rasio hari berkonsumsi nol, durasi konsumsi nol terpanjang, rasio hari berkonsumsi rendah).

Selain itu, ditambahkan fitur volatilitas (rata-rata selisih konsumsi antahari ternormalisasi) dan rasio data hilang. Fitur rasio data hilang dihitung sebelum proses imputasi karena pola data yang hilang diduga berkaitan dengan upaya pencurian. Pendekatan berbasis fitur statistik ini menghasilkan representasi yang ringkas, mudah diinterpretasikan, dan efisien secara komputasi.

Daftar lengkap kesembilan belas fitur beserta deskripsi singkatnya disajikan pada Tabel 2.

Tabel 2. Daftar fitur statistik yang digunakan

Fitur	Deskripsi
mean	Rata-rata konsumsi harian
std	Simpangan baku konsumsi
max	Konsumsi maksimum
min	Konsumsi minimum
median	Nilai tengah konsumsi
q25	Kuartil pertama (Q1)
q75	Kuartil ketiga (Q3)
iqr	Rentang interkuartil (Q3 - Q1)

Fitur	Deskripsi
range	Rentang (maksimum - minimum)
total	Total konsumsi
cv	Koefisien variasi (std/mean)
skewness	Kemencengan distribusi
kurtosis	Keruncingan distribusi
zero_ratio	Rasio hari berkonsumsi nol
max_zero_run	Durasi konsumsi nol terpanjang
mean_abs_diff	Rata-rata selisih antarahari
volatility	Volatilitas ternormalisasi
low_ratio	Rasio hari konsumsi rendah
missing_ratio	Rasio data hilang

2.4. Penanganan Ketidakseimbangan Kelas

Karena hanya 8,53% data tergolong pencurian, model cenderung bias terhadap kelas mayoritas. Untuk mengatasinya, digunakan Synthetic Minority Over-sampling Technique (SMOTE) [7], yang menghasilkan sampel sintetis kelas minoritas dengan interpolasi antar tetangga terdekat. SMOTE hanya diterapkan pada data latih di dalam pipeline pelatihan agar tidak terjadi kebocoran informasi (data leakage) ke data uji.

Penerapan SMOTE menyeimbangkan proporsi kedua kelas pada data latih sehingga model memperoleh sinyal yang memadai untuk mengenali pola pencurian. Strategi ini diterapkan secara konsisten pada keempat algoritma agar perbandingan kinerja menjadi adil.

2.5. Algoritma Klasifikasi

Empat algoritma klasifikasi dibandingkan. Regresi Logistik digunakan sebagai model dasar (baseline) linier. Decision Tree (kedalaman maksimum 12) merepresentasikan model berbasis pohon tunggal. Random Forest [8] merupakan ensemble bagging dari 200 pohon keputusan, sedangkan Gradient Boosting [13] diimplementasikan dengan algoritma berbasis histogram (Histogram-based Gradient Boosting) sebanyak 300 iterasi sebagai representasi metode boosting yang efisien, sejalan dengan pendekatan XGBoost [9] dan LightGBM [10].

Keempat model dievaluasi pada konfigurasi yang sama. Setiap model dipasangkan dengan tahap normalisasi dan SMOTE dalam satu pipeline untuk menjamin reproduktibilitas dan keadilan perbandingan.

Pemilihan hiperparameter ditetapkan pada nilai yang umum digunakan agar perbandingan berfokus pada karakteristik intrinsik masing-masing algoritma, bukan pada hasil penyetulan yang ekstensif. Kedalaman maksimum Decision Tree dibatasi untuk mengurangi overfitting, jumlah pohon Random Forest ditetapkan 200 untuk menjaga keseimbangan antara kinerja dan biaya komputasi, sedangkan Gradient

Boosting menggunakan laju pembelajaran 0,1 dengan 300 iterasi.

2.6. Skenario Pengujian dan Metrik Evaluasi

Dataset dibagi menjadi data latih dan data uji dengan rasio 80:20 secara stratified sehingga proporsi kelas tetap terjaga (33.897 sampel latih dan 8.475 sampel uji, dengan 723 kasus pencurian pada data uji). Untuk menguji kestabilan, dilakukan pula validasi silang stratified (lima-lipat untuk Regresi Logistik, Decision Tree, dan Gradient Boosting; tiga-lipat untuk Random Forest karena pertimbangan efisiensi komputasi).

Mengingat ketidakseimbangan kelas, akurasi saja tidak memadai. Oleh karena itu digunakan metrik precision, recall, F1-score, ROC-AUC, dan PR-AUC. Precision dan recall dihitung dari matriks kebingungan, sedangkan F1-score merupakan rata-rata harmonik keduanya sebagaimana persamaan (2) dan (3).

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}), \text{ Recall} = \text{TP} / (\text{TP} + \text{FN}) \quad (2)$$

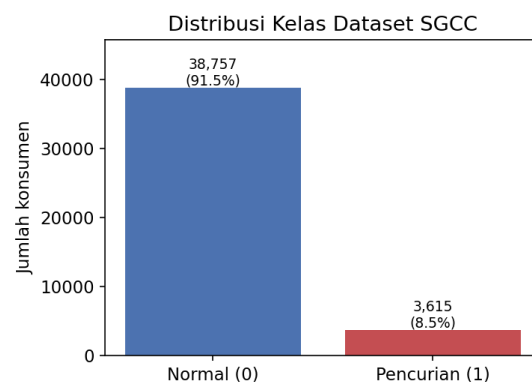
$$\text{F1} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (3)$$

dengan TP, FP, dan FN berturut-turut menyatakan jumlah true positive, false positive, dan false negative. ROC-AUC mengukur kemampuan model membedakan kedua kelas pada berbagai ambang, sedangkan PR-AUC lebih sensitif terhadap kinerja pada kelas minoritas.

3. Hasil dan Pembahasan

3.1. Karakteristik dan Profil Konsumsi

Gambar 2 memperlihatkan distribusi kelas dataset yang sangat timpang, dengan 38.757 pelanggan normal dan 3.615 pelanggan terindikasi mencuri. Ketimpangan ini menegaskan perlunya teknik penyeimbangan kelas seperti SMOTE.

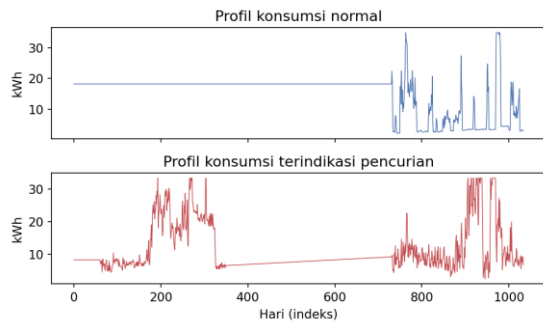


Gambar 2. Distribusi kelas dataset SGCC

Gambar 3 membandingkan profil konsumsi harian pelanggan normal dan pelanggan terindikasi mencuri. Profil pelanggan normal cenderung memiliki pola yang lebih teratur, sedangkan profil pelanggan

pencurian menunjukkan periode konsumsi rendah atau nol yang lebih panjang dan pola yang lebih tidak beraturan. Pengamatan ini mendukung pemilihan fitur statistik yang menangkap perilaku konsumsi rendah dan nol.

Selain pola konsumsi, tingginya proporsi nilai hilang sebesar 25,64% pada dataset juga menjadi karakteristik penting. Pola data hilang yang tidak sepenuhnya acak dapat menjadi sinyal tersendiri, sehingga rasio data hilang diperlakukan sebagai salah satu fitur prediktif dalam penelitian ini.



Gambar 3. Perbandingan profil konsumsi normal dan pencurian

3.2. Perbandingan Kinerja Model

Tabel 3 merangkum kinerja keempat model pada data uji. Model Gradient Boosting memperoleh F1-score tertinggi (0,323) dan ROC-AUC terbaik (0,750), diikuti Random Forest dengan ROC-AUC 0,747 dan F1-score 0,304. Regresi Logistik dan Decision Tree menghasilkan ROC-AUC yang lebih rendah. Meskipun Regresi Logistik memiliki recall tinggi (0,639), precision-nya sangat rendah sehingga banyak pelanggan normal salah ditandai sebagai pencuri.

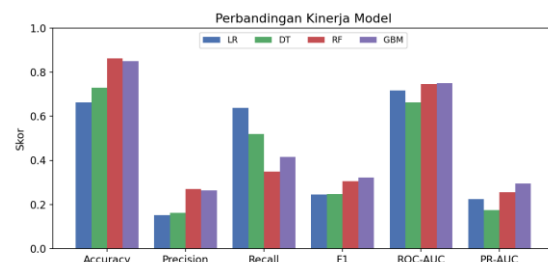
Perlu dicatat adanya paradoks akurasi pada data tidak seimbang. Decision Tree memperoleh akurasi (0,730) yang lebih tinggi daripada Regresi Logistik (0,663), tetapi ROC-AUC-nya justru lebih rendah. Hal ini menegaskan bahwa akurasi bukan metrik yang tepat untuk masalah ini, karena model dapat memperoleh akurasi tinggi hanya dengan memprediksi sebagian besar pelanggan sebagai normal. Metrik PR-AUC dan F1-score memberikan gambaran yang lebih representatif, dan pada kedua metrik tersebut Gradient Boosting tetap unggul dengan PR-AUC 0,295.

Tabel 3. Perbandingan kinerja model pada data uji

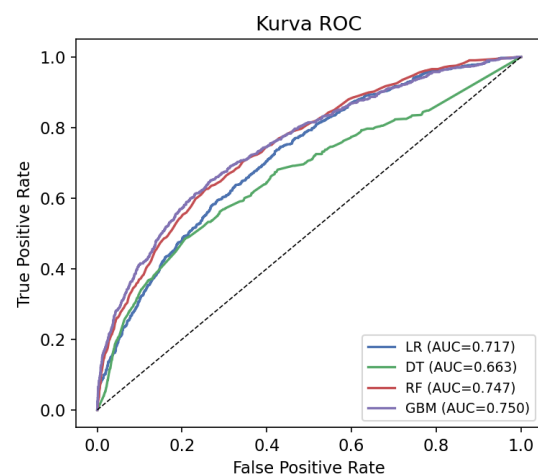
Model	Acc	Prec	Rec	F1	AUC	PR-AUC
Regresi Logistik	0,663	0,151	0,639	0,244	0,717	0,223
Decision Tree	0,730	0,162	0,520	0,247	0,663	0,173
Random Forest	0,864	0,270	0,349	0,304	0,747	0,255
Gradient Boosting	0,851	0,263	0,416	0,323	0,750	0,295

Gambar 4 menyajikan perbandingan metrik secara visual, sedangkan Gambar 5 menampilkan kurva ROC. Model ensemble (Random Forest dan Gradient Boosting) secara konsisten berada di atas model dasar, menegaskan keunggulan metode berbasis pohon dalam menangkap relasi non-linier antar fitur konsumsi.

Keunggulan model ensemble dapat dijelaskan melalui kerangka bias-variance. Decision Tree tunggal cenderung memiliki varians tinggi dan rentan overfitting, sedangkan Random Forest menurunkan varians melalui agregasi banyak pohon, dan Gradient Boosting menurunkan bias dengan membangun pohon secara bertahap untuk mengoreksi kesalahan model sebelumnya. Praproses yang konsisten dan penerapan SMOTE memastikan setiap model dievaluasi pada kondisi yang setara.



Gambar 4. Perbandingan metrik kinerja antar model



Gambar 5. Kurva ROC keempat model

Untuk menilai kestabilan, Tabel 4 menyajikan hasil validasi silang. Nilai simpangan baku ROC-AUC yang kecil (di bawah 0,011) menunjukkan bahwa kinerja model konsisten dan tidak bergantung pada pembagian data tertentu. Gradient Boosting kembali menjadi model terbaik dengan ROC-AUC rata-rata 0,757.

Tabel 4. Hasil validasi silang

Model	F1 (rata-rata ± std)	ROC-AUC (rata-rata ± std)
Regresi Logistik	0,246 ± 0,006	0,724 ± 0,010
Decision Tree	0,260 ± 0,011	0,674 ± 0,009

Model	F1 (rata-rata $\pm$ std)	ROC-AUC (rata-rata $\pm$ std)
Random Forest	0,309 $\pm$ 0,012	0,745 $\pm$ 0,005
Gradient Boosting	0,331 $\pm$ 0,006	0,757 $\pm$ 0,004

3.3. Analisis Model Terbaik

Gambar 6 menampilkan matriks kebingungan model Gradient Boosting pada data uji. Model berhasil mengidentifikasi 301 dari 723 kasus pencurian (recall 0,416) dengan 842 false positive. Dalam konteks operasional, recall yang lebih tinggi bernilai penting karena kegagalan mendeteksi pencurian (false negative) umumnya lebih merugikan daripada inspeksi tambahan akibat false positive.

Confusion Matrix - GBM

Aktual	Normal	6,910	842
	Pencurian	422	301
	Prediksi	Normal	Pencurian

Gambar 6. Matriks kebingungan model Gradient Boosting

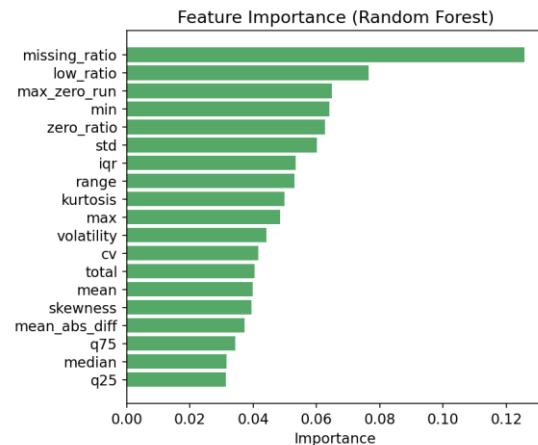
Kinerja yang diperoleh sebanding dengan hasil pada literatur yang menggunakan fitur statistik klasik pada dataset SGCC, dan menegaskan bahwa metode boosting yang ringan dapat memberikan hasil kompetitif tanpa biaya komputasi model pembelajaran dalam [6]. Peningkatan kinerja lebih lanjut dimungkinkan melalui penyetelan ambang keputusan sesuai prioritas operasional.

Dari sudut pandang penerapan, model yang ringan seperti Gradient Boosting berbasis histogram dapat dilatih ulang secara berkala dengan cepat seiring bertambahnya data smart meter. Hal ini memungkinkan sistem deteksi beradaptasi terhadap perubahan pola konsumsi pelanggan tanpa memerlukan perangkat keras khusus, sehingga sesuai untuk perusahaan utilitas dengan sumber daya komputasi terbatas.

3.4. Analisis Feature Importance

Gambar 7 menunjukkan tingkat kepentingan fitur menurut model Random Forest. Fitur rasio data hilang menjadi indikator paling diskriminatif, diikuti rasio konsumsi rendah, durasi konsumsi nol terpanjang, nilai minimum, dan rasio hari berkonsumsi nol. Temuan ini konsisten dengan intuisi bahwa pelanggan yang melakukan pencurian cenderung menampilkan

konsumsi tercatat yang rendah, sering nol, atau memiliki banyak data hilang.



Gambar 7. Tingkat kepentingan fitur (Random Forest)

Implikasi praktis dari temuan ini adalah bahwa operator dapat memprioritaskan pemeriksaan pada pelanggan dengan pola data hilang dan konsumsi rendah yang tinggi. Dengan demikian, model tidak hanya berfungsi sebagai alat klasifikasi, tetapi juga memberikan wawasan yang dapat ditindaklanjuti dalam penanganan kerugian non-teknis.

3.5. Keterbatasan dan Penelitian Lanjutan

Penelitian ini memiliki beberapa keterbatasan. Pertama, representasi data menggunakan fitur statistik agregat sehingga informasi temporal yang lebih halus, seperti pola musiman dan mingguan, belum dimanfaatkan secara penuh. Kedua, recall model terbaik masih berada di kisaran 0,42 sehingga sebagian kasus pencurian belum terdeteksi. Ketiga, evaluasi dilakukan pada satu dataset (SGCC) sehingga generalisasi ke jaringan listrik dengan karakteristik berbeda perlu diuji lebih lanjut.

Untuk penelitian lanjutan, beberapa arah dapat ditempuh. Penyetelan ambang keputusan dan optimasi hiperparameter dapat meningkatkan keseimbangan antara precision dan recall sesuai prioritas operator. Penggunaan fitur deret waktu serta model berbasis pembelajaran dalam, seperti CNN dan LSTM [5], berpotensi menangkap pola temporal yang lebih kompleks. Selain itu, penggabungan beberapa model (ensemble stacking) dan teknik penyeimbangan kelas alternatif dapat dieksplorasi untuk meningkatkan kinerja deteksi secara keseluruhan.

Tabel 5 merangkum posisi penelitian ini dibandingkan beberapa penelitian terdahulu. Berbeda dengan pendekatan pembelajaran dalam yang dominan pada literatur, penelitian ini menunjukkan bahwa kombinasi fitur statistik ringan, SMOTE, dan gradient boosting mampu memberikan kinerja yang kompetitif dengan kebutuhan komputasi yang jauh lebih kecil sehingga lebih mudah diterapkan pada lingkungan operasional.

Tabel 5. Ringkasan beberapa penelitian terdahulu

Penelitian	Metode	Karakteristik
Jindal dkk. [4]	Decision Tree + SVM	Kombinasi dua pengklasifikasi
Zheng dkk. [1]	Wide & Deep CNN	Memperkenalkan dataset SGCC
Buzau dkk. [3]	Hybrid DNN	Fitur konsumsi & non-konsumsi
Hasan dkk. [5]	CNN-LSTM	Menangkap pola temporal
Punmiya & Choe [6]	Gradient Boosting	Rekayasa fitur
Penelitian ini	LR, DT, RF, Gradient Boosting	Fitur statistik ringan + SMOTE

4. Kesimpulan

Penelitian ini membandingkan empat algoritma pembelajaran mesin untuk deteksi pencurian listrik pada data smart meter SGCC yang nyata dan tidak seimbang. Dengan rekayasa 19 fitur statistik dan penanganan ketidakseimbangan kelas menggunakan SMOTE, model Gradient Boosting berbasis histogram memberikan kinerja terbaik dengan ROC-AUC 0,750 dan F1-score 0,323, diikuti Random Forest. Hasil validasi silang menunjukkan kinerja yang stabil.

Analisis feature importance mengungkap bahwa rasio data hilang, rasio konsumsi rendah, dan durasi konsumsi nol terpanjang merupakan indikator pencurian yang paling penting. Hasil ini menunjukkan bahwa rekayasa fitur statistik yang dipadukan dengan ensemble learning merupakan metode yang efektif dan efisien untuk deteksi pencurian listrik. Penelitian lanjutan dapat mengeksplorasi penyetelan ambang keputusan, penggunaan fitur deret waktu, serta penerapan model pembelajaran dalam untuk meningkatkan recall.

Daftar Rujukan

- [1] Zheng, Z., Yang, Y., Niu, X., Dai, H.-N. and Zhou, Y., 2018. Wide and deep convolutional neural networks for electricity-theft detection to secure smart grids. *IEEE Transactions on Industrial Informatics*, 14(4), pp.1606-1615. <https://doi.org/10.1109/TII.2017.2785963>
- [2] Ismail, M., Shaaban, M.F., Naidu, M. and Serpedin, E., 2020. Deep learning detection of electricity theft cyber-attacks in renewable distributed generation. *IEEE Transactions on Smart Grid*, 11(4), pp.3428-3437. <https://doi.org/10.1109/TSG.2020.2973681>
- [3] Buzau, M.M., Tejedor-Aguilera, J., Cruz-Romero, P. and Gomez-Exposito, A., 2019. Hybrid deep neural networks for detection of non-technical losses in electricity smart meters. *IEEE Transactions on Power Systems*, 34(2), pp.1254-1263. <https://doi.org/10.1109/TPWRS.2019.2943115>
- [4] Jindal, A., Dua, A., Kaur, K., Singh, M., Kumar, N. and Mishra, S., 2016. Decision tree and SVM-based data analytics for theft detection in smart grid. *IEEE Transactions on Industrial Informatics*, 12(3), pp.1005-1016. <https://doi.org/10.1109/TII.2016.2543145>
- [5] Hasan, M.N., Toma, R.N., Nahid, A.-A., Islam, M.M.M. and Kim, J.-M., 2019. Electricity theft detection in smart grid systems: a CNN-LSTM based approach. *Energies*, 12(17), 3310. <https://doi.org/10.3390/en12173310>
- [6] Punmiya, R. and Choe, S., 2019. Energy theft detection using gradient boosting theft detector with feature engineering-based preprocessing. *IEEE Transactions on Smart Grid*, 10(2), pp.2326-2329. <https://doi.org/10.1109/TSG.2019.2892595>
- [7] Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, pp.321-357. <https://doi.org/10.1613/jair.953>
- [8] Breiman, L., 2001. Random forests. *Machine Learning*, 45(1), pp.5-32. <https://doi.org/10.1023/A:1010933404324>
- [9] Chen, T. and Guestrin, C., 2016. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, USA, 13-17 August 2016. New York: ACM, pp.785-794. <https://doi.org/10.1145/2939672.2939785>
- [10] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.-Y., 2017. LightGBM: a highly efficient gradient boosting decision tree. In: *Advances in Neural Information Processing Systems 30 (NIPS 2017)*. Long Beach, USA, 4-9 December 2017. pp.3146-3154. Tersedia di: <https://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree> [Diakses 14 Juni 2026]
- [11] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al., 2011. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12, pp.2825-2830. Tersedia di: <https://jmlr.org/papers/v12/pedregosa11a.html> [Diakses 14 Juni 2026]
- [12] Messinis, G.M. and Hatziaargyriou, N.D., 2018. Review of non-technical loss detection methods. *Electric Power Systems Research*, 158, pp.250-266. <https://doi.org/10.1016/j.epsr.2018.01.005>
- [13] Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29(5), pp.1189-1232. <https://doi.org/10.1214/aos/1013203451>