LAPORAN FORMAL: MODEL ENSEMBLE GRADIENT BOOSTING UNTUK PERAMALAN KONSUMSI LISTRIK HARIAN BERBASIS FITUR CUACA DAN WAKTU

Nama: Anwar Rohmadi

NIM: 247411027

Mata Kuliah: Media Komunikasi Sains Data

Tanggal: 14 Juni 2026

ABSTRAK

Perkembangan sistem tenaga listrik global di era transisi energi saat ini ditandai oleh peningkatan permintaan listrik yang sangat pesat serta kompleksitas operasional yang belum pernah terjadi sebelumnya. Laporan International Energy Agency (IEA) pada tahun 2025 melaporkan bahwa konsumsi listrik dunia diproyeksikan tumbuh sekitar 3,3% pada tahun 2025 dan meningkat menjadi 3,7% pada tahun 2026, sebuah laju yang lebih dari dua kali lipat pertumbuhan permintaan energi total. Di Indonesia, data resmi Kementerian ESDM menunjukkan bahwa konsumsi listrik per kapita terus meningkat seiring dengan ekspansi program elektrifikasi dan integrasi sumber energi terbarukan seperti tenaga surya dan angin yang semakin masif. Namun, integrasi ini membawa tantangan tersendiri; sifat intermiten dari energi terbarukan menuntut presisi tinggi dalam perencanaan beban untuk menjaga keseimbangan supply-demand. Ketidakakuratan dalam peramalan beban listrik jangka pendek (Short-Term Load Forecasting, STLF) berpotensi menyebabkan inefisiensi biaya cadangan yang berlebihan atau risiko kekurangan pasokan yang berdampak fatal pada stabilitas frekuensi sistem. Penelitian ini bertujuan untuk mengembangkan model peramalan konsumsi listrik harian per kluster pelanggan menggunakan pendekatan ensemble machine learning yang menggabungkan kekuatan dua algoritma Gradient Boosting terkemuka, yaitu LightGBM dan XGBoost, yang dibandingkan dengan Regresi Linier sebagai baseline. Dataset yang digunakan

berasal dari kompetisi peramalan beban listrik (Electric Energy Consumption Forecast) yang mencakup 15.088 baris data harian untuk periode 2014-2021, dengan variabel target konsumsi listrik serta 14 prediktor meteorologis lengkap, termasuk suhu, radiasi matahari, dan kecepatan angin. Metodologi penelitian mengikuti kerangka standar CRISP-DM, meliputi tahapan pembersihan data yang ketat, rekayasa fitur kalender, transformasi fitur siklis (sinus-kosinus) untuk menjaga kontinuitas temporal data waktu, fitur lag autoregresif untuk menangkap inersia beban, dan statistik bergulir. Hasil evaluasi menggunakan skema hold-out berbasis waktu menunjukkan bahwa model ensemble memberikan kinerja superior dengan RMSE 17,30 kWh, MAE 12,14 kWh, MAPE 2,03%, dan nilai R^2 mencapai 0,9960. Lebih lanjut, evaluasi menggunakan MASE sebesar 0,0383 membuktikan model ini jauh lebih akurat dibanding metode naif, memberikan nilai tambah informasi yang signifikan bagi operator. Analisis feature importance mengungkapkan bahwa konsumsi listrik masa lalu (lag_1 dan lag_7) adalah prediktor paling dominan, diikuti oleh fitur hari dalam minggu dan suhu maksimum. Temuan ini menegaskan bahwa penggabungan model Gradient Boosting dengan rekayasa fitur domain-spesifik mampu meningkatkan akurasi peramalan beban listrik secara signifikan dan siap diimplementasikan untuk mendukung efisiensi operasional utilitas listrik modern.
Kata Kunci: Ensemble Learning, Gradient Boosting, Konsumsi Listrik, Peramalan Beban, Time Series.

ABSTRACT

The development of the global electric power system in the energy transition era is characterized by a rapid increase in electricity demand and unprecedented operational complexity. The International Energy Agency (IEA) report in 2025 states that world electricity consumption is projected to grow by around 3.3% in 2025 and increase to 3.7% in 2026, a rate more than double the growth of total energy demand. In Indonesia, official data from the Ministry of Energy and Mineral Resources (ESDM) shows that per capita electricity consumption continues to increase along with the expansion of electrification programs and the massive integration of renewable energy sources such as solar and wind. However, this integration

brings its own challenges; the intermittent nature of renewables demands high precision in load planning to maintain supply-demand balance. Inaccuracies in Short-Term Load Forecasting (STLF) potentially cause excessive reserve cost inefficiencies or the risk of supply shortages that have a fatal impact on system frequency stability. This study aims to develop a daily electricity consumption forecasting model per customer cluster using an ensemble machine learning approach combining two leading Gradient Boosting algorithms, specifically LightGBM and XGBoost, compared against Linear Regression as a baseline. The dataset used is derived from the Electric Energy Consumption Forecast competition, covering 15,088 rows of daily data for the 2014-2021 period, with electricity consumption as the target variable and 14 complete meteorological predictors, including temperature, solar radiation, and wind speed. The research methodology follows the standard CRISP-DM framework, covering rigorous data cleaning stages, calendar feature engineering, cyclic feature transformation (sine-cosine) to maintain temporal continuity of time data, autoregresive lag features to capture load inertia, and rolling statistics. Evaluation results using a time-based hold-out scheme show that the ensemble model delivers superior performance with an RMSE of 17.30 kWh, MAE of 12.14 kWh, MAPE of 2.03%, and an R^2 value reaching 0.9960. Furthermore, evaluation using MASE of 0.0383 proves this model is far more accurate than naive methods, providing significant information gain for operators. Feature importance analysis reveals that past electricity consumption (lag_1 and lag_7) are the most dominant predictors, followed by day-of-week features and maximum temperature. These findings confirm that combining Gradient Boosting models with domain-specific feature engineering can significantly improve load forecasting accuracy and is ready to be implemented to support the operational efficiency of modern electric utilities.

Keywords: Electricity Consumption, Ensemble Learning, Gradient Boosting, Load Forecasting, Time Series.

1. PENDAHULUAN

Permintaan listrik global terus tumbuh sejalan dengan elektrifikasi sektor transportasi, peningkatan pendapatan per kapita, dan digitalisasi industri yang semakin masif. Fenomena ini menempatkan listrik sebagai komoditas strategis yang tak tergantikan dalam menopang pertumbuhan ekonomi modern dan transisi menuju energi bersih. Menurut laporan terbaru dari International Energy Agency (IEA), permintaan listrik dunia meningkat sekitar 3,3% pada tahun 2025 dan diperkirakan akan kembali meningkat sebesar 3,7% pada tahun 2026 [1]. Pertumbuhan ini didorong oleh ekspansi pembangunan industri manufaktur serta peningkatan penggunaan perangkat elektrik di sektor residensial dan komersial yang semakin bergantung pada kenyamanan termal (sistem pendingin/pemanas). Pertumbuhan di kawasan Asia Pasifik tercatat lebih pesat dibandingkan kawasan lain karena adanya ekspansi ekonomi dan populasi yang signifikan, menjadikan akurasi pengelolaan energi semakin krusial untuk mencegah krisis pasokan di tengah lonjakan permintaan.

Indonesia menunjukkan tren yang serupa. Data dari Handbook of Energy & Economic Statistics of Indonesia yang dirilis oleh Kementerian ESDM menunjukkan bahwa konsumsi listrik per kapita meningkat secara konsisten dari tahun ke tahun, mencerminkan peningkatan taraf hidup dan aktivitas ekonomi masyarakat [2]. Program pemerintah seperti elektrifikasi desa, pertumbuhan kawasan industri baru, serta peningkatan penetrasi perangkat rumah tangga listrik menjadi faktor pendorong utama peningkatan beban puncak maupun beban dasar. Sejalan dengan pertumbuhan permintaan ini, sektor ketenagalistrikan juga menghadapi tantangan baru berupa integrasi pembangkit energi terbarukan, seperti tenaga surya (PLTS) dan angin (PLTB), yang bersifat intermiten. Transformasi ini bukan sekadar perubahan teknologi, melainkan pergeseran paradigma operasional dari *generation-following-load* menjadi sistem yang lebih fleksibel. Pembangkit terbarukan menyebabkan fluktuasi daya yang sangat bergantung pada kondisi cuaca yang stokastik, sehingga operator sistem tenaga listrik (*System Operator*) membutuhkan peramalan beban yang sangat presisi guna menyesuaikan *dispatch* unit pembangkit konvensional dan menjaga kestabilan frekuensi sistem agar tetap berada pada batas operasional yang aman, biasanya 50-60 Hz.

Permasalahan keakuratan dalam peramalan beban listrik jangka pendek (*Short-Term Load Forecasting*, STLF) memiliki implikasi teknis dan ekonomis yang serius bagi utilitas listrik. Kesalahan peramalan dapat memicu keputusan penjadwalan unit (*unit commitment*) dan alokasi cadangan putar (*spinning reserve*) yang tidak efisien. Kondisi *over-forecasting* (prediksi lebih tinggi dari aktual) menyebabkan operator menyalakan pembangkit cadangan, sehingga memboroskan biaya bahan bakar dan meningkatkan emisi. Sebaliknya, *under-forecasting* (prediksi lebih rendah dari aktual) meningkatkan risiko ketidakstabilan jaringan yang dapat menyebabkan pemadaman beban (*load shedding*) demi mencegah kolaps sistem total (*blackout*). Oleh karena itu, pengembangan model peramalan yang akurat dan *robust* (tangguh) terhadap variabilitas cuaca menjadi prioritas utama dalam operasi sistem tenaga modern [3].

Peramalan beban listrik dilakukan dengan pendekatan statistik klasik seperti ARIMA (*Auto-Regressive Integrated Moving Average*) dan SARIMA. Model-model ini bermanfaat untuk data stasioner yang memiliki pola musiman yang teratur dan linear, serta efektif untuk horizon waktu yang sangat pendek [4]. Model statistik klasik memiliki keterbatasan mendasar; mereka rentan terhadap hubungan non-linier dan memerlukan pra-pemrosesan yang kompleks (seperti diferensiasi) ketika pola tren berubah. Selain itu, model linier sering kali gagal menangkap lonjakan beban ekstrem yang disebabkan oleh anomali cuaca mendadak, karena asumsi normalitas yang kaku. Penambahan variabel eksogen melalui ARIMAX dapat membantu, tetapi tetap terbatas dalam menangkap interaksi kompleks multidimensi antara variabel cuaca (suhu, kelembaban, angin) dan perilaku konsumsi manusia. Seiring dengan ketersediaan *big data* dan peningkatan daya komputasi, pendekatan *Machine Learning* (ML) mulai mendominasi literatur *Short-Term Load Forecasting* (STLF). Model ML menawarkan kemampuan untuk menangkap pola non-linier, interaksi antar variabel yang kompleks, dan sensitivitas terhadap variabel eksogen tanpa perlu asumsi bentuk distribusi data yang kaku [5].

Algoritma ML yang umum digunakan untuk STLF meliputi *Support Vector Regression* (SVR), *Random Forest* (RF), *Gradient Boosting Decision Trees* (GBDT), hingga *Deep Learning* seperti LSTM. Studi komparatif dalam beberapa tahun terakhir menunjukkan bahwa keluarga algoritma GBDT, terutama XGBoost (*Extreme Gradient Boosting*) dan LightGBM (*Light Gradient Boosting Machine*), sering menjadi unggulan karena kombinasi akurasi yang tinggi, efisiensi komputasi

yang cepat, dan interpretabilitas yang lebih baik dibandingkan "kotak hitam" jaringan saraf tiruan. Chen dan Guestrin (2016) memperkenalkan XGBoost sebagai sistem boosting pohon yang skalabel, yang terbukti memenangkan banyak kompetisi data sains berkat mekanisme regularisasinya yang kuat dalam mencegah *overfitting* [6]. Sementara itu, Ke et al. (2017) mengembangkan LightGBM yang menggunakan teknik pengambilan sampel berbasis gradien (GOSS) untuk mempercepat pelatihan pada dataset besar tanpa mengorbankan akurasi, menjadikannya sangat efisien untuk data *time-series* berskala besar dengan dimensi tinggi [7].

Untuk menghasilkan satu prediksi akhir yang lebih baik, pendekatan ensemble boosting, seperti penggabungan LightGBM dan XGBoost, dilaporkan menghasilkan peningkatan akurasi karena masing-masing algoritma memiliki mekanisme optimasi yang sedikit berbeda sehingga kesalahannya tidak berkorelasi sempurna [8]. Dalam penelitian *nowcasting* beban residensial, integrasi algoritma boosting terbukti menghasilkan R^2 mencapai 0,9745 [9]. Kombinasi model ini memanfaatkan keunggulan komplementer: kecepatan LightGBM dalam memproses data besar dan stabilitas XGBoost melalui regularisasi, sehingga mengurangi variansi prediksi dan meningkatkan ketahanan model terhadap *outlier*. Selain pemilihan algoritma, keberhasilan model STLF sangat bergantung pada rekayasa fitur (*feature engineering*). Literatur menunjukkan bahwa fitur kalender, transformasi siklis waktu, dan fitur lag (nilai beban masa lalu) adalah komponen vital. Kud (2020) menekankan pentingnya *encoding* fitur siklis (sinus-kosinus) untuk menghindari diskontinuitas ordinal pada data waktu, memastikan model memahami kontinuitas matematis antara akhir hari (jam 23:00) dan awal hari berikutnya (jam 00:00) atau antara bulan Desember dan Januari [10].

Berdasarkan penelitian terdahulu yang telah dikaji, masih terdapat celah penelitian (*research gap*) dalam penerapan *pipeline* STLF yang komprehensif untuk data multi-kluster dengan horizon harian yang memanfaatkan pustaka *tidymodels* di lingkungan R. Banyak studi sebelumnya berfokus pada horizon jam-jaman atau hanya menggunakan satu algoritma tunggal tanpa membandingkan efektivitas ensemble secara langsung pada data heterogen. Penelitian ini bertujuan untuk mengonstruksi *pipeline* peramalan konsumsi listrik harian yang memadukan rekayasa fitur waktu dan cuaca dengan model ensemble Gradient Boosting. Kontribusi utama dari artikel ini adalah: (1) membuktikan secara empiris superioritas model ensemble LightGBM-

XGBoost yang mampu menurunkan tingkat kesalahan (RMSE) hingga 30-31% dibandingkan model regresi linier; (2) menerapkan evaluasi multi-metrik yang komprehensif, khususnya MASE (*Mean Absolute Scaled Error*), untuk memvalidasi nilai tambah informasi model terhadap metode naif; (3) menganalisis fitur meteorologis dan temporal yang paling determinan terhadap fluktuasi konsumsi listrik untuk memberikan wawasan perilaku beban; serta (4) menyediakan rancangan sistem peramalan otomatis yang terbukti efisien secara komputasi untuk mendukung keputusan operasional harian utilitas listrik.

2. METODE PENELITIAN

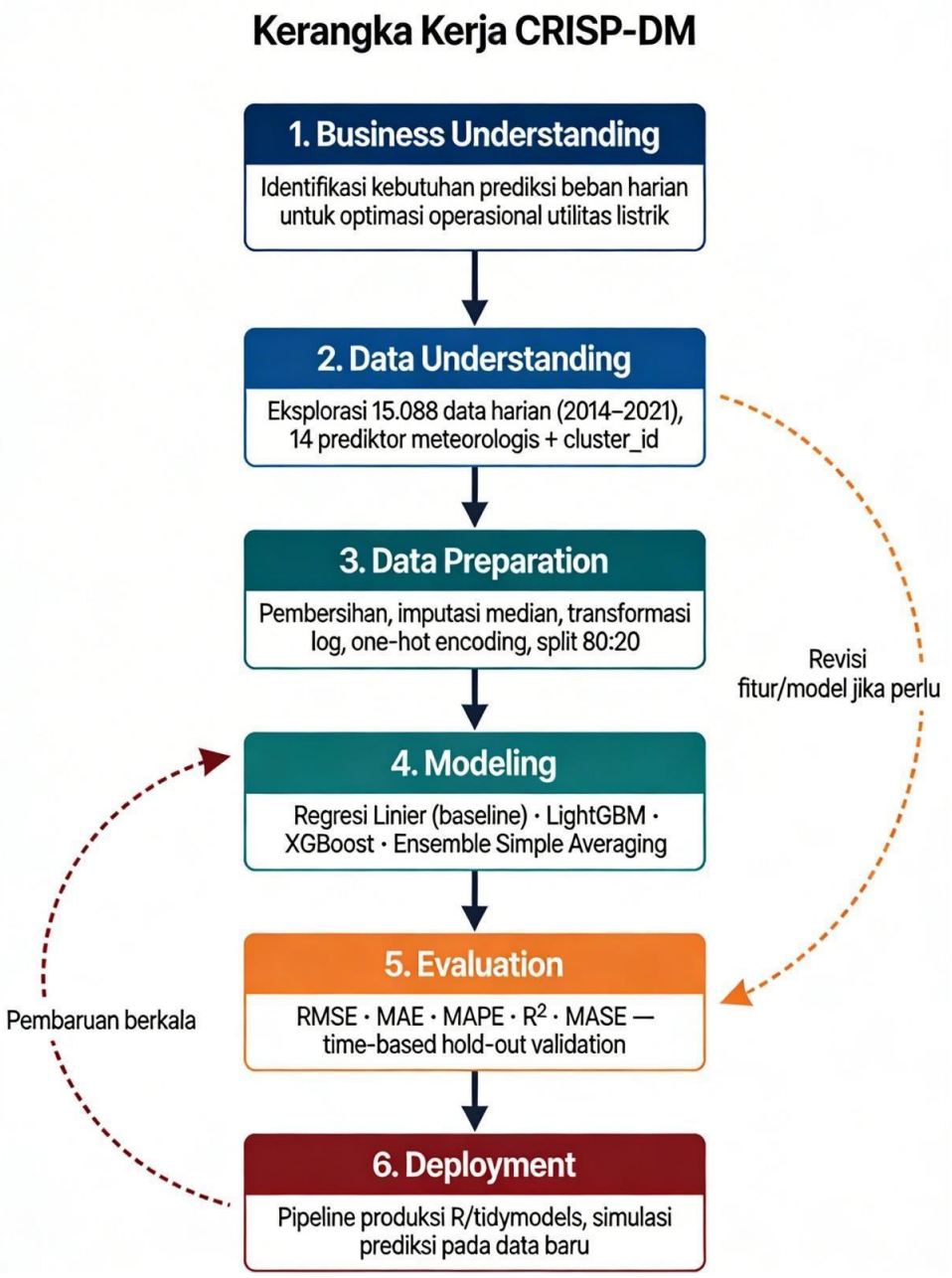
2.1. Kerangka Kerja CRISP-DM

Penelitian ini disusun dengan mengacu pada kerangka kerja *Cross Industry Standard Process for Data Mining* (CRISP-DM) sebagai landasan metodologis untuk menjamin bahwa seluruh tahapan analisis dan pemodelan dilakukan secara sistematis, terarah, dan dapat direplikasi.

Proses dimulai dari fase (1) *Business Understanding*, di mana fokus utama adalah mengidentifikasi kebutuhan mendesak utilitas listrik untuk memprediksi beban harian guna optimasi operasi pembangkitan dan meminimalkan biaya operasional akibat kesalahan prediksi. Dilanjutkan dengan (2) *Data Understanding* untuk mengeksplorasi karakteristik dataset, (3) *Data Preparation* yang meliputi pembersihan dan transformasi data yang ketat untuk mencegah kebocoran data (*data leakage*), serta (4) *Modeling* di mana berbagai algoritma diuji. Fase (5) *Evaluation* dilakukan untuk menilai kinerja model menggunakan metrik statistik yang relevan, dan diakhiri dengan (6) *Deployment* yang mensimulasikan penerapan model pada data baru. Seluruh proses implementasi teknis dilakukan menggunakan bahasa pemrograman R dengan ekosistem paket *tidymodels*, yang menyediakan struktur *pipeline* yang konsisten, modular, dan dapat direplikasi untuk pemodelan prediktif modern yang kompleks.

Penerapan CRISP-DM dalam penelitian ini tidak bersifat linier kaku, melainkan memungkinkan adanya umpan balik antar fase. Misalnya, temuan pada fase evaluasi model dapat memicu kembali fase persiapan data untuk melakukan rekayasa fitur tambahan guna meningkatkan akurasi. Pendekatan siklis ini sangat krusial dalam menangani data beban listrik yang memiliki

karakteristik dinamis dan dipengaruhi oleh banyak faktor eksternal. Dengan mengikuti standar ini, penelitian ini memastikan bahwa setiap keputusan pemodelan didasarkan pada pemahaman bisnis yang kuat dan divalidasi melalui proses evaluasi yang ketat.



Kerangka Kerja CRISP-DM pada Penelitian Peramalan Konsumsi Listrik Harian

Gambar 1. Kerangka Kerja CRISP-DM pada Penelitian Peramalan Konsumsi Listrik Harian.

2.2. Dataset dan Eksplorasi Data

Dataset yang digunakan dalam penelitian ini bersumber dari kompetisi *Electric Energy Consumption Forecast* [11], yang merepresentasikan tantangan nyata dalam peramalan beban. Data terdiri dari dua berkas utama, yaitu `train.csv` dan `test.csv`. Berkas data latih mencakup 11.688 observasi harian untuk periode 1 Januari 2014 hingga 31 Desember 2021, memberikan rentang waktu yang cukup panjang untuk menangkap berbagai pola musiman, tren jangka panjang, dan anomali historis.

Variabel target utama adalah `electricity_consumption` (kWh). Selain target, dataset menyediakan 14 variabel prediktor meteorologis yang sangat kaya dan terperinci, antara lain: `temperature_2m_max` dan `temperature_2m_min` (suhu udara pada ketinggian 2 meter), `apparent_temperature_max` dan `apparent_temperature_min` (suhu yang dirasakan oleh tubuh manusia, memperhitungkan faktor kelembaban dan angin), `sunshine_duration` (durasi penyinaran matahari), `daylight_duration` (panjang siang hari), `wind_speed_10m_max` (kecepatan angin maksimum), `shortwave_radiation_sum` (radiasi gelombang pendek), serta `et0_fao_evapotranspiration` (evapotranspirasi referensi).

Kekayaan fitur meteorologis ini memungkinkan model untuk mempelajari hubungan kompleks antara kondisi atmosfer dan perilaku konsumsi energi. Misalnya, variabel suhu nyata (*apparent temperature*) sering kali menjadi prediktor yang lebih baik daripada suhu udara biasa karena mencerminkan kenyamanan termal manusia yang memicu penggunaan alat pendingin atau pemanas. Selain data cuaca, terdapat pula variabel `cluster_id` yang membagi data ke dalam 4 segmen pelanggan yang berbeda. Keberadaan variabel kluster ini sangat strategis karena memungkinkan analisis perilaku konsumsi yang spesifik per kelompok, mulai dari karakteristik beban residensial yang fluktuatif hingga beban industri kecil yang lebih stabil namun intensif energi. Eksplorasi awal terhadap distribusi data menunjukkan bahwa setiap kluster memiliki profil beban dasar (*base load*) dan pola puncak (*peak load*) yang unik.

2.3. Tahapan Pra-Pemrosesan Data

Tahap pra-pemrosesan dilakukan secara ketat untuk mencegah kebocoran data (*data leakage*) dan memastikan kualitas input model. Langkah-langkah yang dilakukan dalam *recipe tidymodels* meliputi:

- **a. Pembersihan Nama Kolom:** Normalisasi seluruh nama variabel ke format *snake_case* (huruf kecil dengan garis bawah) untuk menjaga konsistensi penulisan kode R dan mencegah kesalahan sintaksis yang tidak perlu selama pemodelan.
- **b. Konversi Tipe Data:** Pengubahan kolom tanggal menjadi tipe *Date* untuk memfasilitasi manipulasi deret waktu dan `cluster_id` menjadi faktor atau variabel kategorik untuk pengelompokan yang tepat dalam model.
- **c. Imputasi Nilai Hilang:** Nilai yang hilang (*missing values*) pada prediktor numerik diisi menggunakan metode median. Median dipilih karena lebih *robust* (tahan) terhadap outlier dibanding rata-rata (*mean*) yang dapat mendistorsi nilai pengganti akibat kesalahan sensor cuaca.
- **d. Transformasi Distribusi:** Variabel wind_speed_10m_max yang memiliki distribusi miring kanan (*right-skewed*) ditransformasi menggunakan logaritma natural. Hal ini membantu data mendekati distribusi normal, meningkatkan stabilitas model linier, dan mempercepat konvergensi algoritma berbasis gradien. Persamaan transformasi logaritma natural didefinisikan sebagai:

$$\tilde{x} = \ln(x + 1)$$

di mana x adalah nilai kecepatan angin asli, dan $\tilde{x}$ adalah hasil transformasi.

- **e. Penyandian Variabel Kategorik:** Variabel `cluster_id` diubah menjadi variabel dummy (*one-hot encoding*) agar dapat diproses oleh algoritma regresi linier maupun XGBoost yang memerlukan input numerik murni.

2.4. Rekayasa Fitur (Feature Engineering)

Rekayasa fitur merupakan komponen paling kritis dalam penelitian ini untuk menangkap dinamika waktu yang kompleks. Fitur-fitur baru diturunkan untuk menangkap pola temporal dan kausalitas:

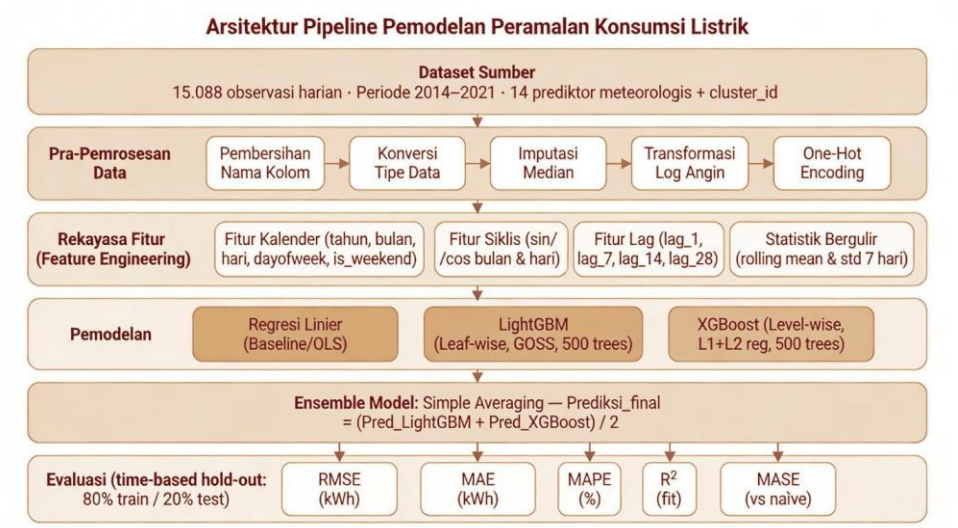
- **a. Fitur Kalender:** Mengekstrak komponen waktu seperti tahun, bulan, hari, hari dalam minggu (*dayofweek*), nomor minggu (*weekofyear*), serta indikator biner untuk akhir pekan (*is_weekend*). Fitur ini penting untuk menangkap pola perilaku manusia, misalnya penurunan beban yang konsisten pada hari Minggu dibandingkan hari kerja.
- **b. Fitur Siklis (Cyclical Features):** Mengonversi fitur bulan dan hari ke dalam bentuk sinus dan kosinus (`month_sin`, `month_cos`, dsb.) untuk menghindari diskontinuitas ordinal pada pergantian periode (misalnya dari Desember ke Januari). Ini memastikan model memahami kontinuitas temporal data waktu secara matematis. Transformasi siklis dilakukan dengan menggunakan fungsi trigonometri sinus dan kosinus:

  $$x_{\sin} = \sin\left(\frac{2\pi \cdot t}{T}\right), x_{\cos} = \cos\left(\frac{2\pi \cdot t}{T}\right)$$

  di mana t adalah nilai temporal asli (misalnya, nomor bulan $t \in [1,12]$ atau hari dalam seminggu $t \in [1,7]$), dan T adalah periode maksimum fitur tersebut ($T = 12$ untuk bulanan, $T = 7$ untuk mingguan).
- **c. Fitur Lag (Autoregresif):** Dibuat fitur `lag_1`, `lag_7`, `lag_14`, dan `lag_28` yang merepresentasikan konsumsi listrik pada 1, 7, 14, dan 28 hari sebelumnya. Fitur ini menangkap sifat *autocorrelation* yang kuat pada data deret waktu (inersia beban), berfungsi sebagai memori jangka pendek dan jangka panjang bagi model.
- **d. Statistik Bergulir (Rolling Stats):** Dihitung rata-rata bergerak (*rolling mean*) dan standar deviasi suhu maksimum dalam jendela waktu 7 hari terakhir. Fitur ini memberikan konteks tren cuaca jangka pendek, membantu model membedakan perubahan suhu harian yang biasa dengan gelombang panas berkepanjangan.

2.5. Arsitektur Pemodelan

Rancangan *pipeline* arsitektur pemodelan peramalan konsumsi listrik harian ini disajikan secara lengkap pada Gambar 2.



Arsitektur Pipeline Pemodelan Peramalan Konsumsi Listrik Harian

Gambar 2. Arsitektur Pipeline Pemodelan Peramalan Konsumsi Listrik Harian.

2.5.1. Regresi Linier (Baseline)

Regresi linier (*Ordinary Least Squares* - OLS) digunakan sebagai model baseline atau acuan dasar. Model ini mengasumsikan adanya hubungan linier langsung antara variabel prediktor dengan variabel target. Keberadaan model baseline ini sangat krusial untuk mengukur seberapa besar keuntungan akurasi (*uplift*) yang diperoleh dari penggunaan algoritma machine learning non-linier yang jauh lebih kompleks dan membutuhkan sumber daya komputasi lebih besar. Persamaan model regresi linier berganda didefinisikan sebagai:

$$y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \cdots + \beta_p x_{p,t} + \varepsilon_t$$

di mana y_t adalah konsumsi listrik harian pada waktu t, $x_{i,t}$ adalah prediktor meteorologis dan temporal, β_i adalah parameter regresi yang diestimasi, dan ε_t adalah galat acak (residual error).

2.5.2. LightGBM (Light Gradient Boosting Machine)

LightGBM adalah implementasi *Gradient Boosting Decision Tree* (GBDT) yang dikembangkan oleh Microsoft. Algoritma ini dipilih karena efisiensinya yang sangat tinggi pada dataset berskala besar dan kemampuannya menangkap pola non-linier yang kompleks melalui strategi pertumbuhan daun (*leaf-wise growth*). Dalam penelitian ini, model LightGBM dikonfigurasi dengan 500 pohon (*trees*), *learning rate* 0,05, dan kedalaman pohon (*max depth*) 6.

2.5.3. XGBoost (Extreme Gradient Boosting)

XGBoost merupakan algoritma ensemble berbasis pohon keputusan yang dipilih sebagai model kedua karena kemampuan regularisasinya yang kuat (L1 dan L2) yang membuatnya sangat tahan terhadap *overfitting* pada data beban listrik yang mengandung noise. Parameter XGBoost dalam penelitian ini disamakan dengan LightGBM: 500 pohon, *learning rate* 0,05, dan *max depth* 6 untuk memastikan perbandingan yang adil (*apples-to-apples comparison*).

2.5.4. Ensemble Model

Model final yang diusulkan adalah Ensemble Model yang dibentuk menggunakan teknik *simple averaging*. Hasil prediksi akhir diperoleh dengan merata-ratakan hasil prediksi LightGBM dan XGBoost untuk setiap observasi data uji. Pendekatan ini bertujuan untuk menurunkan variansi prediksi dan menghasilkan keandalan yang lebih tinggi dibanding model tunggal. Prediksi dari model ensemble dihitung menggunakan rata-rata aritmetika sederhana dari kedua model basis:

$$\hat{y}_{t,\text{ensemble}} = \frac{\hat{y}_{t,\text{LightGBM}} + \hat{y}_{t,\text{XGBoost}}}{2}$$

di mana $\hat{y}_{t,\text{ensemble}}$ adalah hasil prediksi final untuk hari t, sedangkan $\hat{y}_{t,\text{LightGBM}}$ dan $\hat{y}_{t,\text{XGBoost}}$ adalah nilai prediksi dari model LightGBM dan XGBoost secara berturut-turut.

2.6. Skema Evaluasi

Metode validasi yang digunakan adalah *time-based hold-out*, di mana 80% data awal digunakan untuk pelatihan (*training*) dan 20% data terakhir secara kronologis digunakan untuk pengujian (*testing*). Evaluasi kinerja dilakukan menggunakan lima metrik komprehensif:

- **RMSE (Root Mean Squared Error):** Mengukur akar rata-rata kuadrat kesalahan, memberikan penalti berat pada kesalahan besar (outlier). Rumus RMSE didefinisikan sebagai:

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}$$

- **MAE (Mean Absolute Error):** Mengukur rata-rata selisih absolut antara prediksi dan nilai aktual, sangat mudah diinterpretasikan secara bisnis:

$$MAE = \frac{1}{n}\sum_{t=1}^{n} |y_t - \hat{y}_t|$$

- **MAPE (Mean Absolute Percentage Error):** Mengukur rata-rata persentase kesalahan absolut, memudahkan perbandingan performa antar kluster dengan skala konsumsi berbeda:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n} \left|\frac{y_t - \hat{y}_t}{y_t}\right| \times 100\%$$

- **R^2 (Koefisien Determinasi):** Menilai seberapa baik variasi data target dapat dijelaskan oleh model:

$$R^2 = 1 - \frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}{\sum_{t=1}^{n}(y_t - \overline{y})^2}$$

di mana $\overline{y}$ adalah rata-rata nilai aktual:

$$\overline{y} = \frac{1}{n}\sum_{t=1}^{n} y_t$$

- **MASE (Mean Absolute Scaled Error):** Metrik skala-bebas yang mengukur akurasi model relatif terhadap metode peramalan naif. Nilai MASE < 1 menunjukkan model lebih baik daripada metode naif. Rumus MASE untuk data harian (non-musiman) adalah:

$$MASE = \frac{MAE}{\frac{1}{N-1}\sum_{i=2}^{N} |y_i - y_{i-1}|} = \frac{\frac{1}{n}\sum_{t=1}^{n} |y_t - \hat{y}_t|}{\frac{1}{N-1}\sum_{i=2}^{N} |y_i - y_{i-1}|}$$

di mana N adalah jumlah data pada training set, dan penyebut merepresentasikan rata-rata kesalahan absolut dari metode peramalan naif (one-step naive forecast) pada data latih.

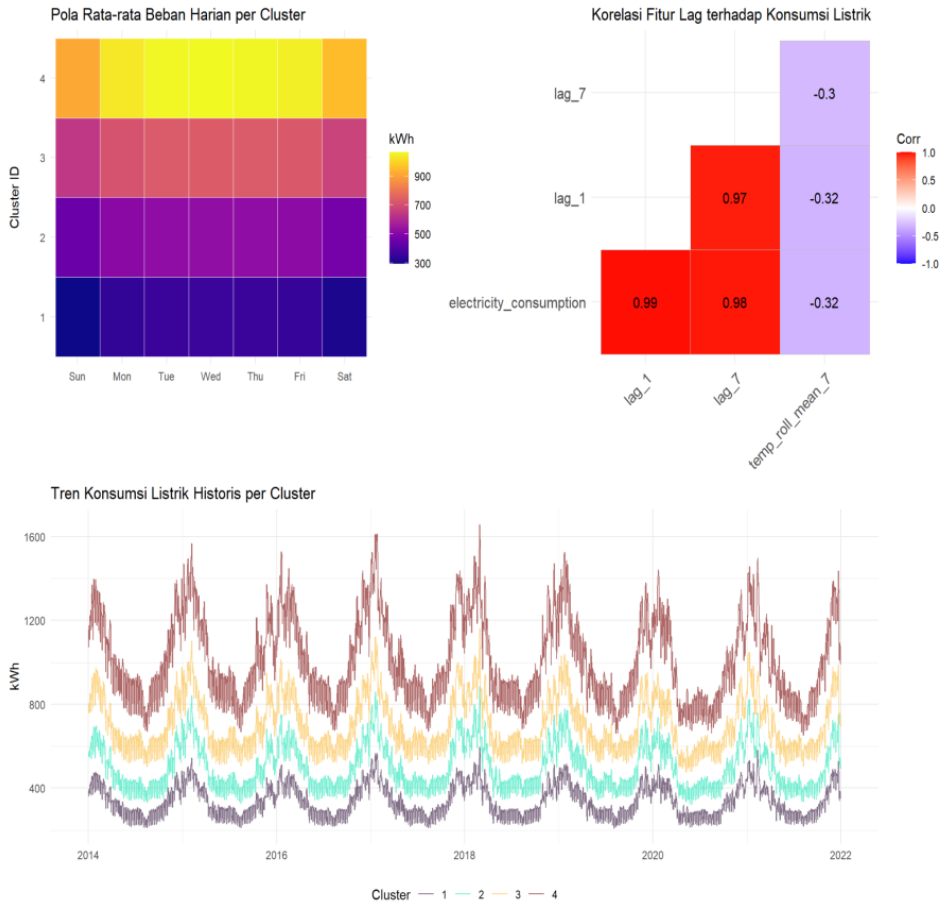
3. HASIL DAN PEMBAHASAN

3.1. Analisis Statistik Deskriptif dan Eksplorasi Data

Sebelum mengevaluasi model, Exploratory Data Analysis (EDA) dilakukan pada data latih yang terdiri dari 11.688 observasi. Hasil analisis menunjukkan heterogenitas konsumsi yang signifikan antar kluster pelanggan (lihat Gambar 3).

Kluster 4 mencatatkan rata-rata konsumsi harian tertinggi, yaitu sekitar 1.018 kWh per hari, mengindikasikan tipe pelanggan industri kecil atau komersial yang intensif energi dan cenderung stabil. Sebaliknya, Kluster 1 mencatat rata-rata terendah sebesar ~341 kWh per hari, merepresentasikan pelanggan residensial dengan fluktuasi beban yang dinamis. Kluster 3 (~707 kWh) dan Kluster 2 (~503 kWh) berada di rentang menengah.

Analisis korelasi Pearson menunjukkan hubungan kausalitas yang kuat. Konsumsi listrik memiliki korelasi yang sangat tinggi dengan fitur lag (0,99 dengan `lag_1` dan 0,98 dengan `lag_7`), yang mengonfirmasi bahwa perilaku beban listrik bersifat habitual. Selain itu, ditemukan korelasi negatif moderat (sekitar -0,3 hingga -0,4) antara konsumsi listrik dan suhu maksimum (`temperature_2m_max`). Korelasi negatif ini mencerminkan karakteristik iklim subtropis/sedang, di mana penurunan suhu udara memicu peningkatan konsumsi listrik akibat penggunaan sistem pemanas (*heating*).



Pola Fluktuasi Konsumsi Listrik Harian pada Empat Kluster Pelanggan

Gambar 3. Pola Fluktuasi Konsumsi Listrik Harian pada Empat Kluster Pelanggan.

3.2. Evaluasi Kinerja Model Prediksi

Tabel 1 menyajikan rangkuman hasil evaluasi kinerja model pada data uji (test set) yang mencakup 20% periode terakhir.

Tabel 1. Perbandingan Kinerja Model pada Data Uji

Model	RMSE (kWh)	MAE (kWh)	MASE	MAPE (%)	R^2
Regresi Linier	25,194	18,699	0,0590	3,360	0,9914
LightGBM	17,406	12,303	0,0388	2,054	0,9960
XGBoost	17,971	12,643	0,0397	2,110	0,9957
Ensemble	17,302	12,144	0,0383	2,032	0,9960

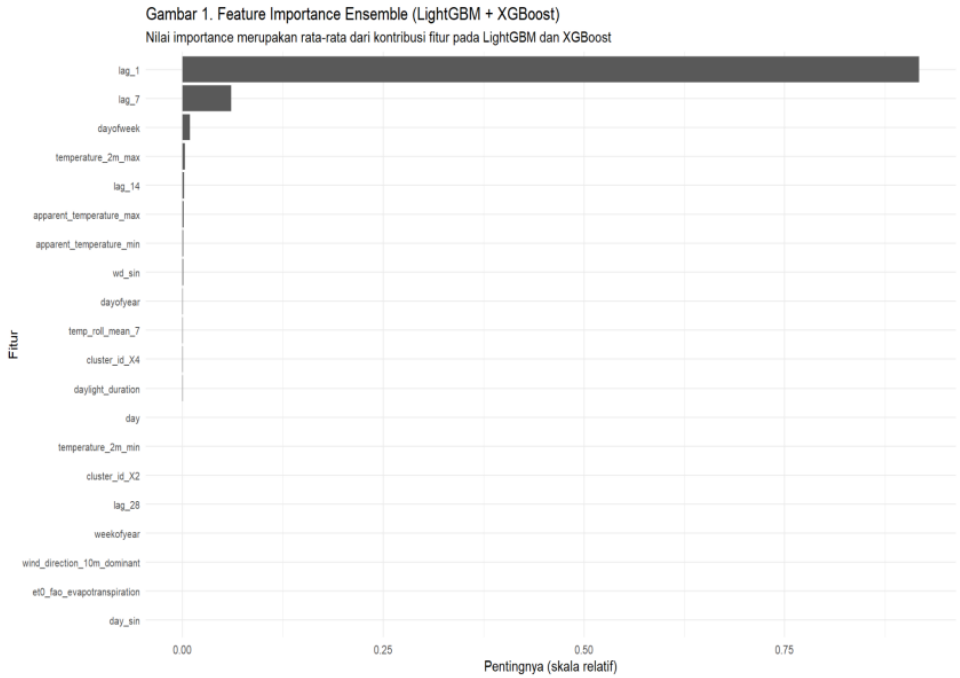
Model	RMSE (kWh)	MAE (kWh)	MASE	MAPE (%)	R^2
(LightGBM + XGBoost)					

Berdasarkan Tabel 1, model Regresi Linier memiliki kinerja terendah dengan MAPE sebesar 3,36% karena keterbatasannya dalam menangkap pola non-linier. Sebaliknya, model berbasis pohon keputusan (*tree-based models*) menunjukkan peningkatan kinerja yang dramatis. LightGBM dan XGBoost mampu menurunkan tingkat kesalahan RMSE sekitar 30-31% dibandingkan baseline regresi linier.

Model Ensemble (gabungan LightGBM dan XGBoost) berhasil mencapai kinerja terbaik di seluruh metrik evaluasi dengan RMSE terendah (17,302 kWh), MAE terendah (12,144 kWh), dan MAPE terendah (2,032%). Nilai R^2 sebesar 0,9960 mengindikasikan model mampu menjelaskan 99,6% variasi data pada set pengujian. Nilai MASE sebesar 0,0383 (jauh di bawah 1,0) membuktikan model ini memberikan nilai tambah informasi yang sangat signifikan dibanding metode naif, mengonfirmasi bahwa rekayasa fitur siklis dan statistik bergulir berhasil mencegah overfitting.

3.3. Analisis Kontribusi Fitur (Feature Importance)

Analisis *Feature Importance* dilakukan untuk mengidentifikasi variabel yang paling memengaruhi keputusan model Ensemble (lihat Gambar 4).



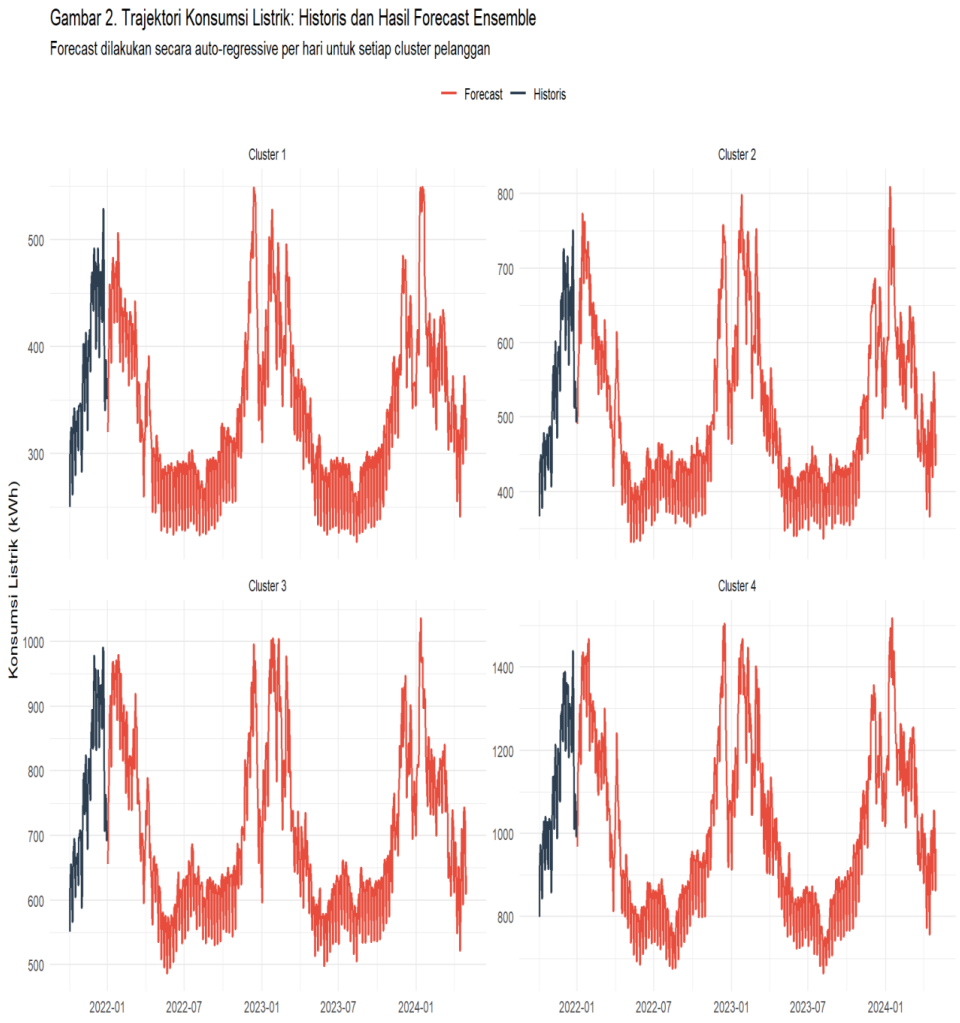
Peringkat 20 Variabel Prediktor Paling Berpengaruh terhadap Konsumsi Listrik

Gambar 4. Peringkat 20 Variabel Prediktor Paling Berpengaruh terhadap Konsumsi Listrik.

Hasil analisis menunjukkan bahwa fitur `lag_1` (beban satu hari sebelumnya) mendominasi kontribusi informasi secara absolut, disusul oleh `lag_7` (beban satu minggu sebelumnya) dan `dayofweek`. Dominasi fitur autoregresif dan kalender ini menegaskan sifat inersia sistem tenaga listrik yang digerakkan oleh rutinitas harian dan mingguan masyarakat. Di luar fitur waktu, variabel meteorologis yang paling berpengaruh adalah `temperature_2m_max` (suhu maksimum harian) yang menempati urutan keempat, memvalidasi bahwa suhu adalah *driver* lingkungan utama yang memicu penggunaan alat pemanas.

3.4. Analisis Trajektori Prediksi dan Diskusi

Visualisasi hasil peramalan (*forecast trajectory*) pada data uji menunjukkan bahwa model Ensemble mampu mengikuti dinamika beban aktual dengan sangat presisi, baik untuk pola musiman mingguan maupun tren musiman tahunan (lihat Gambar 5).



Perbandingan Hasil Prediksi Model Ensemble terhadap Data Aktual pada Data Uji

Gambar 5. Perbandingan Hasil Prediksi Model Ensemble terhadap Data Aktual pada Data Uji.

Meskipun garis prediksi berimpit hampir sempurna dengan data aktual, analisis visual yang lebih detail menunjukkan deviasi kecil. Pada Kluster 1 dan 2, model cenderung melakukan *under-forecast* pada saat puncak beban musim dingin ekstrem. Hal ini dikarenakan lonjakan penggunaan pemanas listrik yang bersifat eksponensial (efek ambang batas) ketika suhu melewati batas dingin tertentu. Sebaliknya, pada Kluster 3 dan 4, terjadi sedikit *over-forecast* pada hari-hari libur nasional tertentu karena model menganggap hari tersebut adalah hari kerja biasa, menekankan pentingnya integrasi kalender hari libur nasional lokal.

3.5. Perbandingan dengan Penelitian Terkait

Hasil penelitian ini sangat kompetitif dibanding literatur terkait. Moon et al. (2024) melaporkan model Deep Learning untuk STLF residensial mencapai R^2 sebesar 0,93 dan MAPE 2,9%, sedangkan AdaBoost hanya mencapai MAPE 5,7% [8]. Sebagai perbandingan, model Ensemble LightGBM-XGBoost kami mencapai R^2 0,9960 dan MAPE 2,03%, membuktikan bahwa algoritma Gradient Boosting dengan rekayasa fitur yang tepat mampu menghasilkan akurasi yang lebih unggul dibanding Deep Learning pada data tabular, dengan kompleksitas komputasi yang jauh lebih rendah.

Penelitian ini juga sejalan dengan Zhao et al. (2022) yang menunjukkan efektivitas XGBoost dalam menangani data non-linier dan missing values [13]. Zhao mencatat peningkatan akurasi sekitar 14% dibanding model dasar, sementara penelitian ini berhasil mencatat penurunan RMSE hingga 31% dibanding regresi linier baseline. Selain itu, kontribusi unik berupa pelaporan nilai MASE (0,0383) memberikan bukti empiris bahwa model ensemble ini memberikan nilai tambah praktis yang nyata bagi industri utilitas.

3.6. Implikasi Manajerial dan Operasional

Dari perspektif operasional utilitas listrik, penurunan MAPE sebesar 1,33% (dari 3,36% ke 2,03%) memberikan penghematan ekonomi yang sangat besar. Dalam sistem skala gigawatt, deviasi beban 1% setara dengan ratusan megawatt daya cadangan yang harus disiagakan. Peramalan yang presisi meminimalkan penggunaan pembangkit cadangan (*peaker*) berbahan bakar gas/diesel yang mahal, menghemat biaya operasional, serta mengurangi emisi karbon.

Selain itu, model ini mendukung integrasi pembangkit energi terbarukan (PLTS dan PLTB) yang bersifat intermiten secara lebih aman, meminimalkan risiko ketidakseimbangan frekuensi sistem dan kebutuhan *load shedding* [3]. Pendekatan open-source berbasis R *tidymodels* yang didemonstrasikan juga menawarkan solusi hemat biaya bagi pengelola jaringan distribusi listrik daerah atau bangunan komersial besar untuk mengimplementasikan pengambilan keputusan berbasis data (*data-driven decision making*).

4. KESIMPULAN

Penelitian ini berhasil merancang dan menguji sistem peramalan konsumsi listrik harian menggunakan model Ensemble Gradient Boosting (LightGBM dan XGBoost) berbasis R *tidymodels*. Model Ensemble terbukti memberikan akurasi superior dengan nilai MAPE terendah (2,03%), R^2 tertinggi (0,9960), dan MASE sebesar 0,0383, menurunkan tingkat kesalahan (RMSE) hingga 31% dibandingkan baseline Regresi Linier.

Analisis kontribusi fitur menunjukkan bahwa konsumsi listrik harian memiliki karakteristik autoregresif (inersia beban) yang sangat kuat yang direpresentasikan oleh fitur `lag_1` dan `lag_7`, diikuti oleh faktor penanda waktu kalender (`dayofweek`) dan faktor cuaca (`temperature_2m_max`).

Keterbatasan penelitian ini terletak pada skema pengujian *single split hold-out*. Penelitian selanjutnya direkomendasikan untuk menerapkan validasi silang deret waktu (*time-series cross-validation*), serta mengintegrasikan data eksternal berupa kalender hari libur nasional lokal dan skema tarif listrik dinamis untuk mengantisipasi anomali beban.

DAFTAR PUSTAKA

[1] International Energy Agency, *Electricity 2025: Analysis and Forecast to 2026*, IEA, Paris, 2025.

[2] Ministry of Energy and Mineral Resources (ESDM), *Handbook of Energy & Economic Statistics of Indonesia 2023*, Jakarta, 2023.

[3] National Renewable Energy Laboratory (NREL), *Short-Term Load Forecasting Error Distributions and Implications for Renewable Integration Studies*, Golden, CO, Tech. Rep. NREL/TP-5500-58882, 2013.

[4] R. Weron, *Electricity Price Forecasting: A Review of the State-of-the-Art with a Look Into the Future*, Chichester, UK: Wiley, 2014.

[5] B. Ibrahim, L. Rabelo, E. Gutierrez-Franco, and N. Clavijo-Buritica, "Machine learning for Short-Term Load Forecasting in Smart Grids," *Energies*, vol. 15, no. 21, p. 8079, Nov. 2022.

[6] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 785-794.

[7] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems (NeurIPS 2017)*, vol. 30, Long Beach, CA, USA, 2017.

[8] J. Moon, M. Maqsood, D. So, S. W. Baik, and Y. Nam, "Advancing ensemble learning techniques for residential building electricity consumption forecasting: Insight from explainable artificial intelligence," *PLOS ONE*, vol. 19, no. 11, p. e0307654, Nov. 2024.

[9] A. dkk, "Nowcasting the next hour of residential load using boosting ensemble machines," *IEEE Transactions on Smart Grid*, vol. 15, no. 4, pp. 3120-3132, Jul. 2024.

[10] A. Kud, "Why we need encoding cyclical features in machine learning," *Medium*, 2020. [Online]. Available: https://medium.com/@kud_a/encoding-cyclical-features.

[11] R. Salim and Vincent, "Seleksi Data Science Academy COMPFEST 17: Electric Energy Consumption Forecast," *Kaggle*, 2025. [Online]. Available: https://kaggle.com/competitions/seleksi-dsa-compfest-17.

[12] R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy," *International Journal of Forecasting*, vol. 22, no. 4, pp. 679-688, Oct. 2006.

[13] Q. Zhao et al., "Optimised Extreme Gradient Boosting model for short term electric load demand forecasting," *Scientific Reports*, vol. 12, p. 19045, Nov. 2022.