Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1730−1742
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8655

Prediksi dan Optimalisasi Konsumsi Energi Smart Atmospheric Water Generator (SAWG) Menggunakan XGBoost Regression

Halim Jayakusuma Wiradinata*, Heru Agus Santoso

Fakultas Ilmu Komputer, Program Studi Teknik Informatika, Universitas Dian Nuswantoro, Semarang, Indonesia
Email: [1,*]111202214044@mhs.dinus.ac.id, [2]heru.agus.santoso@dsn.dinus.ac.id
Email Penulis Korespondensi: 111202214044@mhs.dinus.ac.id
Submitted: **07/11/2025**; Accepted: **09/12/2025**; Published**: 11/12/2025**

Abstrak−Penurunan ketersediaan air bersih mendorong pemanfaatan Smart Atmospheric Water Generator (SAWG) sebagai alternatif penyediaan air, namun konsumsi listriknya berfluktuasi mengikuti suhu, kelembapan, dan pola operasi. Penelitian ini membangun model prediksi konsumsi energi (kWh) pada SAWG menggunakan Extreme Gradient Boosting (XGBoost) dan mendemonstrasikan skema optimasi operasional berbasis prediksi untuk penjadwalan hemat energi. Data pencatatan SAWG (1.601 baris, 9 variabel) dipraolah melalui penanganan nilai hilang, konversi numerik, serta deteksi noise/outlier hingga diperoleh 1.313 baris layak olah. Fitur yang digunakan mencakup parameter lingkungan dan sinyal kelistrikan, ditambah fitur waktu berupa jam-ke (hour of day), hari-dalam-minggu (day of week), dan bulan (month). Pemodelan dilakukan dengan pembagian deret waktu secara kronologis (split 80:20 sebagai konfigurasi utama dan 60:40 sebagai uji ketahanan), Time Series Cross-Validation pada blok data latih, serta penyetelan hiperparameter menggunakan GridSearchCV. Evaluasi pada hold-out test set menunjukkan bahwa kinerja model masih terbatas pada konteks deret waktu (R^2 bernilai negatif dan kesalahan prediksi berada pada orde puluhan kWh), namun model tetap mampu memeringkat jam-jam dengan konsumsi energi terprediksi relatif rendah. Pada tahap optimasi operasional berbasis prediksi, keluaran model per jam ($\hat{y}_t$) digunakan oleh Greedy Scheduler untuk memilih H = 8 jam operasi dengan energi terprediksi minimum. Dibandingkan jadwal naive (total energi terprediksi 47,493 kWh), skenario greedy menghasilkan sekitar 43,134 kWh dengan estimasi penghematan ±9,18%. Hasil ini menunjukkan bahwa, meskipun akurasi model masih perlu ditingkatkan, penjadwalan berbasis prediksi berpotensi menurunkan konsumsi energi tanpa mengubah arsitektur perangkat dan dapat dikembangkan sebagai komponen decision-support pada operasi SAWG.

Kata Kunci: Smart AWG; Konsumsi Energi; XGBoost Regression; Pemodelan Prediktif; Optimasi Operasional

Abstract−The decreasing availability of clean water has motivated the use of Smart Atmospheric Water Generator (SAWG) systems as an alternative water source, but their electrical energy consumption fluctuates with ambient conditions and operating patterns. This study develops a predictive model of SAWG energy consumption (kWh) using Extreme Gradient Boosting (XGBoost) and demonstrates a prediction-based operational optimization scheme for energy-efficient scheduling. The SAWG logging dataset (1,601 rows, 9 variables) is preprocessed through missing-value handling, numeric conversion, and noise/outlier detection, resulting in 1,313 usable records. The feature set includes environmental parameters, electrical signals, and time features: hour of day, day of week, and month. Modeling employs chronological time-based splits (80:20 as the main configuration and 60:40 as a robustness check), *Time Series Cross-Validation* on the training block, and hyperparameter tuning via GridSearchCV. Evaluation on the hold-out test sets shows that the model's performance in a strict time-series setting remains limited: for the 80:20 split, the test results are approximately MAE = 23.16 kWh, MSE = 648.93 kWh^2, and $R^2 = -0.22$, while for the 60:40 split they are MAE = 27.21 kWh, MSE = 932.17 kWh^2, and $R^2 = -1.75$. Although the model cannot yet explain the overall variance of energy consumption satisfactorily, it can still be used to rank hours by predicted energy. In the prediction-based operational optimization stage, hourly model outputs $\hat{y}_t$are fed into a *Greedy Scheduler* that selects H = 8 operating hours with the lowest predicted energy. Compared with a naive schedule, which yields a total predicted energy of 47.493 kWh over the simulation horizon, the greedy schedule achieves 43.134 kWh, corresponding to an estimated saving of about 9.18%. These results indicate that prediction-based scheduling can reduce SAWG energy consumption without modifying the device hardware and can be further developed as a decision-support component for SAWG operation.

Keywords: Smart AWG; Energy Prediction; XGBoost Regression; Operational Optimization.

1. PENDAHULUAN

Krisis ketersediaan air bersih terus menguat akibat musim kering berulang, pertumbuhan populasi, dan degradasi kualitas sumber air permukaan. Dalam konteks penelitian ini, istilah Smart Atmospheric Water Generator (SAWG) merujuk pada AWG yang telah dilengkapi sensor lingkungan dan kelistrikan, unit kendali mikroprosesor, serta modul komunikasi Internet of Things (IoT) sehingga mampu melakukan pencatatan (logging) otomatis, pemantauan jarak jauh, dan menjadi basis pengambilan keputusan operasional berbasis data [4], [11], [12]. Atmospheric water generator (AWG) sebagai perangkat yang mengekstraksi uap air ambien menjadi air muncul sebagai opsi strategis karena tidak bergantung pada pasokan air permukaan maupun air tanah [1]–[3]. Tantangan utama yang kerap menghambat adopsi AWG adalah konsumsi listrik yang relatif tinggi serta berfluktuasi mengikuti dinamika lingkungan seperti suhu, kelembapan relatif, dan pola operasi perangkat. Ketidakpastian beban energi ini memperbesar biaya operasional dan menurunkan keberlanjutan sistem. Penelitian ini merespons persoalan tersebut dengan menawarkan pendekatan prediktif yang empiris, sistematis, dan berbasis teori untuk memperkirakan konsumsi energi (kWh) pada tingkat perangkat SAWG menggunakan XGBoost regression. Dengan model yang presisi dan mudah direplikasi, operator dapat mengambil keputusan berbasis data misalnya penjadwalan, penetapan set-point, dan pengaturan jam operasi agar biaya energi turun tanpa mengorbankan produksi air.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1730−1742
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8655

Secara state of the art, penelitian *AWG* berkembang pada dua arus besar yang saling melengkapi. Arus pertama menekankan kemajuan material dan proses penjerapan/pelepasan uap air serta rancangan sistem pengambil air udara (*atmospheric water harvesting/AWH*). Ulasan mengenai pendekatan pasif *water harvesting from air* merangkum mekanisme kondensasi dan sorpsi beserta prospeknya [2], sementara tinjauan komprehensif terbaru memetakan lanskap teknologi *AWH* lintas pendekatan kondensasi, desikan, membran dan menyoroti isu integrasi energi dalam berbagai iklim [3]. Rancangan *all-weather* berdaya surya ditawarkan untuk mencapai efisiensi menyeluruh antara input energi dan keluaran air [5]. Arus kedua berfokus pada evaluasi kinerja yang terstandardisasi: usulan indeks evaluasi global untuk solusi hibrida *HVAC–AWG* menegaskan perlunya metrik komprehensif yang menyatukan aspek termal dan konsumsi energi [6]. Di sisi lain, benchmark produksi *AWG* di Amerika Serikat memperlihatkan variabilitas performa lintas perangkat serta menegaskan kebutuhan metrik yang seragam untuk perbandingan yang adil [7]. Studi pada iklim panas-lembap menunjukkan relasi erat antara produksi air dan konsumsi listrik sehingga strategi operasi adaptif menjadi krusial [8], sedangkan optimasi siklus adsorpsi–desorpsi pada wilayah kering memperlihatkan potensi penghematan energi dengan kebijakan operasi yang kontekstual [9]. Hubungan empiris antara energi dan keluaran *AWG* juga telah ditelaah pada skala eksperimental [10]. Sejalan dengan transformasi digital, ekosistem *Internet of Things (IoT)* untuk *AWG* menghadirkan data logging waktu nyata untuk pemantauan perangkat dan kualitas air [11], dan secara umum smart technology/machine learning semakin diadopsi dalam pemantauan dan pengolahan air [12]. Arah baru berbasis machine learning bahkan telah mendorong prediksi panen air pada membran deposisi ion [13] serta *computational screening material metal-organic frameworks (MOF)* untuk memilih kandidat *AWH* yang unggul [15]. Di sisi kebijakan dan implementasi, penilaian potensi *AWG* di wilayah kering dan rancangan untuk daerah berkelembapan rendah–menengah menunjukkan konteks sosial-teknis yang luas yang perlu diakomodasi [16], [17], sementara buku dan monograf terbaru menyediakan landasan konseptual dan rekayasa untuk pengembangan *AWG* modern [18], [19]. Rangkaian state of the art ini menegaskan satu benang merah: isu konsumsi energi pada *AWG* khususnya prediksi beban pada tingkat perangkat masih memerlukan pendekatan analitik yang logis, relevan, dan dapat dijalankan secara rutin pada lingkungan operasional *SAWG*.

Pada saat yang sama, *XGBoost* menempati posisi penting dalam pemodelan konsumsi atau beban energi karena kemampuannya menangkap non-linearitas, interaksi fitur, dan ketahanan terhadap skala fitur yang beragam. Tinjauan mutakhir di domain energi menekankan praktik deret-waktu yang kuat, mulai dari beban kota hingga jaringan daya skala besar, dengan pemilihan fitur cuaca/kalender, skema validasi yang ketat, serta tolok ukur yang transparan [20]. Secara khusus, *XGBoost* untuk *short-term load forecasting* menunjukkan kinerja kompetitif pada power *grid big data* [21], dan *XGBoost quantile regression* pada kinerja energi bangunan memungkinkan prediksi berikut ketidakpastian yang relevan ketika keputusan operasional membawa konsekuensi biaya [22]. Integrasi ensemble kooperatif dan *SHapley Additive exPlanations (SHAP)* meningkatkan akurasi sekaligus memberikan feature attribution yang dapat dijelaskan [23]. Sinergi XGBoost dengan metaheuristics memperlihatkan penguatan performa pada beban gedung [24], sedangkan boosting ensemble efektif untuk nowcasting beban residensial ber-horizon sangat pendek [25]. Pada sistem energi terbarukan, *XGBoost-based featurization* bermanfaat sebagai tahap prapemrosesan bagi model deep learning hilir [26]; di ranah kendaraan listrik, studi komparatif menegaskan generalitas pendekatan *machine learning* untuk memetakan konsumsi energi pada domain berbeda [27]. Telaah komparatif yang lebih luas antara metode klasik, abu-abu, kabur, dan cerdas menempatkan *XGBoost* dalam lanskap metodologis yang kuat dan berbasis teori, sehingga pemilihan model dapat dipertanggungjawabkan secara ilmiah [28].

Berdasarkan celah tersebut, penelitian ini menawarkan kontribusi: (i) menyusun model prediksi konsumsi energi (kWh) *SAWG* berbasis *XGBoost regression* yang logis, empiris, dan replikabel; (ii) melaksanakan evaluasi menyeluruh menggunakan cross-validation dan hold-out test dengan metrik *mean absolute error (MAE)*, *root mean squared error (RMSE)*, dan *coefficient of determination (R^2)*; (iii) menyediakan interpretabilitas melalui *feature importance* serta *SHAP* bila relevan untuk mengidentifikasi faktor penggerak konsumsi energi seperti suhu, kelembapan relatif, sinyal kelistrikan, dan pola waktu; serta (iv) mendemonstrasikan Optimasi Operasional Berbasis Prediksi yaitu pemanfaatan keluaran model ($\hat{y}_t$) untuk penjadwalan harian menggunakan *Greedy Scheduler* pada kuota operasi tertentu (misalnya $H = 8$jam) sebagai *day-ahead scheduling*. Dengan memilih jam operasi pada periode ber-$\hat{y}_t$ terendah, pendekatan ini bertujuan menurunkan energi terprediksi dibanding jadwal statis baseline dan berfungsi sebagai komponen decision-support yang mudah diintegrasikan ke alur operasional *SAWG* tanpa perubahan arsitektur perangkat.

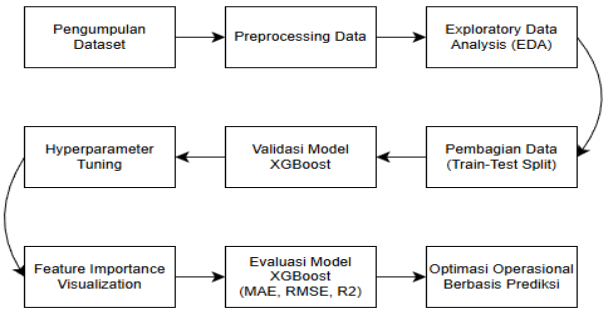
2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Untuk mencapai tujuan penelitian ini, dilakukan serangkaian tahapan yang tersusun secara sistematis mulai dari pengumpulan data pemantauan *Smart Atmospheric Water Generator* (SAWG), analisis dan pembersihan data dari nilai hilang maupun noise, eksplorasi awal melalui *exploratory data analysis* (EDA), hingga pembentukan fitur lingkungan, kelistrikan, dan waktu. Tahap berikutnya adalah pemisahan data latih uji berbasis waktu, pembangunan model *XGBoost regression*, serta validasi dan penyetelan hiperparameter untuk memperoleh konfigurasi terbaik. Seluruh rangkaian tersebut dirancang mengikuti kaidah pemodelan deret waktu untuk konsumsi energi sehingga alur data dan tujuan analitis pada tiap tahap tetap konsisten. Selain itu, keluaran model prediksi diintegrasikan ke dalam

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1730−1742
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8655

skenario optimasi operasional berbasis penjadwalan untuk menilai potensi penghematan energi secara kuantitatif. Alur lengkap tahapan penelitian ditunjukkan pada Gambar 1.



Gambar 1. Tahapan Penelitian

2.2 Pengumpulan Data

Pengumpulan data pada penelitian ini dilakukan dengan memanfaatkan sumber data primer berupa rekaman pemantauan (*logging*) perangkat *Smart Atmospheric Water Generator (SAWG)* yang tersimpan berupa dataset excel. Data tersebut diambil dari sistem pemantauan *SAWG* yang merekam kondisi lingkungan dan parameter operasi perangkat secara berkala. Proses pengambilan data dilakukan dengan mengekspor data hasil pencatatan dari *logger* ke format tabular agar dapat diolah lebih lanjut pada lingkungan komputasi. Dataset yang dikumpulkan memuat satu kolom waktu (*timestamp*) sebagai penanda urutan pengukuran dan beberapa kolom variabel pengukuran, berupa suhu udara (Suhu (AC)), kelembapan relatif (Kelembaban (%)), kualitas udara/partikulat (Kualitas Udara (µg/m³)), level air tangki (Level Air (%)), serta parameter kelistrikan berupa tegangan (Tegangan (V)), arus (Arus (A)), dan daya sesaat (Daya (W)). Kolom Energi (kWh) ditetapkan sebagai atribut target karena penelitian ini berfokus pada pemodelan dan prediksi konsumsi energi SAWG. Seluruh kolom tersebut dipertahankan karena mewakili dua kelompok informasi yang diperlukan model, yaitu kondisi lingkungan yang memengaruhi kemampuan kondensasi atau panen uap air dan kondisi kelistrikan yang secara langsung berhubungan dengan beban energi perangkat[1],[2]. Pengumpulan dataset dilakukan pada satu periode operasi kontinu (satu rentang waktu pemantauan) agar pola perubahan energi dapat diamati dalam urutan waktu yang konsisten. Dataset yang digunakan sebagai sampel penelitian dapat dilihat pada Tabel 1.

Tabel 1. Dataset Penelitian

Variabel	Satuan	Peran Dalam Model
Suhu	°C	Input
Kelembapan	%	Input
Kualitas Udara	µg/m³	Input
Level Air	%	Input
Tegangan	V	Input
Arus	A	Input
Daya	W	Input
Energi	kWh	Target Output

2.3 Preprocessing Data

Preprocessing Data dilakukan agar dataset yang telah dikumpulkan berada dalam kondisi siap olah sebelum digunakan pada pemodelan *XGBoost regression*. Langkah ini diperlukan karena data hasil pencatatan sensor dan *logger* kerap memuat nilai hilang, format numerik yang tidak seragam, maupun pembacaan ekstrem yang perlu ditinjau terlebih dahulu [4], [11], [12]. Pertama, kolom waktu *(timestamp)* dikonversi ke format tanggal waktu dan seluruh data diurutkan secara kronologis. Pengurutan ini penting karena beberapa studi kinerja *atmospheric water generator* menunjukkan bahwa perubahan suhu dan kelembapan dari waktu ke waktu berpengaruh terhadap energi yang dibutuhkan untuk proses kondensasi [6]–[9]. Setelah urutan waktu seragam, seluruh kolom untuk mencegah kegagalan pelatihan model. Kedua, dilakukan penanganan nilai hilang. Nilai kosong yang muncul sebagai celah pendek dapat diisi dengan interpolasi linier atau forward filling, sedangkan hilang dalam rentang panjang dibuang agar tidak menimbulkan bias pada model. Pendekatan ini sejalan dengan praktik pembersihan data sensor pada sistem air berbasis *IoT* dan *machine learning* [4], [11], [12]. Ketiga, dilakukan pemeriksaan noise pada kolom numerik, yaitu mendeteksi sel yang seharusnya berisi angka tetapi tercampur karakter lain nilai seperti ini dikonversi ke numerik atau ditandai kembali sebagai hilang dan mengikuti prosedur penanganan sebelumnya. Keempat, dilakukan identifikasi outlier menggunakan metode rentang antar-kuartil (IQR). Pada variabel daya dan kelistrikan, outlier tidak langsung dihapus karena lonjakan sesaat dapat terjadi pada operasi AWG, sebagaimana dilaporkan pada analisis empiris konsumsi energi dan optimasi siklus AWG [8]–[10]. Nilai ekstrem yang masih masuk akal secara fisis dipertahankan, sedangkan nilai yang jelas tidak realistis dapat dipangkas. Hasil *preprocessing* ini menghasilkan dataset yang bersih,

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1730−1742
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8655

konsisten, dan terdokumentasi, sehingga dapat digunakan pada tahap pemisahan data dan pemodelan energi menggunakan gradient boosting sebagaimana dipraktikkan pada studi peramalan energi terkini [20]–[23].

2.4 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) dilakukan untuk memperoleh gambaran awal mengenai pola, sebaran, serta hubungan antarfitur pada dataset *SAWG* sebelum proses pemodelan. Tahap ini berfungsi sebagai jembatan antara *preprocessing data* dan pemilihan model, karena *EDA* memungkinkan peneliti memastikan bahwa variabel-variabel yang akan dimasukkan ke dalam *XGBoost regression* memang informatif terhadap target Energi (kWh) dan tidak mengandung penyimpangan data yang tidak terdeteksi sebelumnya [4], [11], [12]. Pertama, dilakukan statistik deskriptif terhadap seluruh variabel numerik yaitu Suhu (AC), Kelembaban, Kualitas Udara, Level Air, Tegangan, Arus, Daya, serta Energi untuk melihat rentang nilai minimum–maksimum, kecenderungan pusat, dan penyebaran data. Langkah ini penting karena beberapa studi kinerja *atmospheric water generator* menunjukkan bahwa perubahan suhu dan kelembapan pada rentang operasi tertentu akan langsung tercermin pada kebutuhan energi dan produksi air [6]–[9]. Dengan statistik deskriptif, peneliti dapat memastikan bahwa data yang dimiliki masih berada dalam rentang fisis yang wajar bagi operasi *SAWG*. Kedua, dilakukan visualisasi sebaran dan deret waktu pada variabel lingkungan dan kelistrikan untuk melihat pola harian atau periodik. Deret waktu ini membantu mendeteksi adanya lonjakan konsumsi daya yang mungkin berkaitan dengan siklus kompresi, defrost, atau perubahan kondisi udara masuk fenomena yang juga diamati pada analisis empiris konsumsi energi *AWG* di iklim panas maupun lembap [8], [10]. Jika ditemukan pola yang berulang, temuan tersebut dapat dijadikan dasar penambahan fitur waktu pada tahap pemodelan, sejauh diperbolehkan dalam ruang lingkup penelitian. Ketiga, dilakukan analisis korelasi menggunakan koefisien korelasi Pearson atau Spearman antara variabel masukan dan target Energi (kWh). Tujuannya adalah mengidentifikasi fitur mana yang mempunyai kedekatan hubungan paling kuat terhadap target dan mengamati kemungkinan multikolinearitas di antara fitur-fitur kelistrikan. Analisis seperti ini lazim digunakan pada studi peramalan energi dan *gradient boosting* agar pemodelan yang dihasilkan selaras dengan perilaku data sensor yang sesungguhnya [20]–[23]. Korelasi yang tinggi pada fitur tertentu misalnya Daya atau Arus mengindikasikan bahwa fitur tersebut harus diprioritaskan dalam pemodelan, sedangkan korelasi rendah pada sebagian fitur lingkungan perlu dijelaskan sebagai karakteristik set data, bukan kelemahan metode. Keempat, hasil EDA didokumentasikan dalam bentuk tabel ringkasan, grafik sebaran, dan *correlation matrix* untuk memudahkan penjelasan pada bagian hasil dan pembahasan. Dengan adanya EDA yang terdokumentasi, alasan pemilihan fitur dan metrik evaluasi pada tahap pemodelan dapat ditelusuri kembali, sekaligus menunjukkan bahwa penelitian ini mengikuti alur analisis data yang sistematis sebagaimana direkomendasikan pada penelitian IoT dan smart water terbaru [4], [11], [12].

2.5 Pembagian Data

Pada tahap ini, himpunan data yang telah melalui preprocessing dan exploratory data analysis (EDA) dipisahkan menjadi vektor fitur X dan target y. Vektor fitur X mencakup parameter lingkungan dan operasi perangkat, yakni Suhu, Kelembaban, Kualitas Udara, Level Air, Tegangan, Arus, Daya, serta tiga fitur waktu yang diekstraksi dari *timestamp* yang terdiri dari hour, dayofweek, dan month, sedangkan Energi ditetapkan sebagai atribut target y. Mengingat data bersifat deret waktu, pembagian data latih uji dilakukan secara kronologis dengan terlebih dahulu mengurutkan seluruh baris berdasarkan *timestamp*. Penelitian ini menerapkan dua skema hold-out, yakni 80:20 sebagai konfigurasi utama dan 60:40 sebagai *robustness check*. Pada skema 80:20, sekitar 80% sampel awal (±1050 baris) digunakan sebagai data latih dan 20% sampel terakhir ±263 baris digunakan sebagai data uji; sedangkan pada skema 60:40, sekitar 60% sampel awal (±787 baris) menjadi data latih dan 40% sampel terakhir (±526 baris) menjadi data uji. Dengan pendekatan ini, data uji selalu merepresentasikan periode waktu yang lebih baru dibanding data latih sehingga mencegah kebocoran informasi masa depan atau *data leakage* dan menghasilkan evaluasi yang lebih realistis untuk skenario prediksi operasi SAWG ke depan. Seluruh proses pelatihan model, validasi silang, dan penyetelan hiperparameter dilaksanakan secara eksklusif pada blok data latih, sementara blok data uji dipertahankan terpisah sebagai *hold-out test set* independen untuk penilaian akhir. Skema pemisahan deret waktu seperti ini selaras dengan praktik pemodelan konsumsi atau beban energi berbasis gradient boosting yang menekankan pentingnya hold-out set kronologis guna menilai kemampuan generalisasi model pada data operasi mendatang [20]–[23].

2.6 Validasi Model (Time Series Cross-Validation)

Validasi model dilakukan untuk memastikan bahwa kinerja XGBoost regression tidak hanya baik pada satu skema pembagian data saja, tetapi juga stabil ketika deret waktu dilipat ke dalam beberapa blok pelatihan–validasi yang berurutan. Pada penelitian ini digunakan skema *Time Series Cross-Validation* dengan 5 lipatan, di mana blok data latih hasil pemisahan berbasis waktu dibagi lagi secara kronologis: pada setiap putaran, segmen data paling belakang berperan sebagai data validasi, sedangkan seluruh segmen sebelumnya digunakan sebagai data pelatihan. Urutan waktu selalu dipertahankan sehingga tidak ada sampel dari masa depan yang ikut dilatih, sehingga risiko kebocoran informasi atau biasa disebut *data leakage* dapat diminimalkan. Metrik kinerja seperti R^2, MAE, dan MSE dihitung pada setiap lipatan dan kemudian dirata-ratakan untuk memperoleh estimasi performa yang lebih stabil serta tidak bergantung pada satu rentang waktu saja. Pendekatan *Time Series Cross-Validation* ini lazim digunakan pada

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1730−1742
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8655

pemodelan beban dan konsumsi energi berbasis gradient boosting karena lebih sesuai dengan sifat kronologis data dan mampu memberikan gambaran kemampuan generalisasi model pada periode operasi selanjutnya [20]–[23].

2.7 Hyperparameter Tuning

Untuk memperoleh konfigurasi model XGBoost regression yang lebih optimal, dilakukan proses penyetelan hiperparameter (*hyperparameter tuning*) pada blok data latih hasil pemisahan berbasis waktu (time-based split). Proses ini menggunakan skema *Time Series Cross-Validation* 5 lipatan sebagaimana dijelaskan pada Subbab 2.6, sehingga setiap kombinasi hiperparameter dievaluasi pada beberapa rentang waktu yang berbeda tanpa melanggar urutan kronologis. Hiperparameter yang disetel antara lain jumlah pohon keputusan (*n_estimators*), kedalaman maksimum pohon (*max_depth*), laju pembelajaran (*learning_rate*), proporsi sampel yang digunakan pada setiap pohon (*subsample*), serta proporsi fitur yang diambil per pohon (*colsample_bytree*). Pencarian kombinasi terbaik dilakukan menggunakan prosedur pencarian grid (*Grid Search*) dengan ruang pencarian yang dibatasi pada nilai-nilai yang umum digunakan dalam pemodelan beban dan konsumsi energi. Rata-rata nilai R^2, MAE, dan MSE pada lipatan validasi digunakan sebagai kriteria seleksi, dan konfigurasi yang memberikan kompromi terbaik antara kesalahan yang rendah dan stabilitas antar-lipatan dipilih sebagai model akhir yang kemudian dievaluasi pada *hold-out set* dan menjadi dasar simulasi optimasi operasional [20]–[23].

2.8 Feature Importance

Untuk menafsirkan determinan utama konsumsi energi SAWG, dilakukan analisis *feature importance* pada model XGBoost hasil penyetelan parameter. Dua metrik dilaporkan. Pertama, *F-score (weight)*, yakni frekuensi sebuah fitur dipilih sebagai pemisah atau sering disebut *split* pada keseluruhan pohon ansambel. Kedua, *gain*, yaitu peningkatan rata-rata kualitas pemisahan (penurunan *loss*) yang dikaitkan dengan sebuah fitur pada saat fitur tersebut digunakan. Secara metodologis, pelaporan *feature importance* merupakan praktik yang lazim dalam pemodelan konsumsi energi berbasis *gradient boosting* karena menyediakan jejak penjelasan yang ringkas dan langsung dapat ditindaklanjuti untuk pengambilan keputusan operasional [20], [21], [22], [23]. Hasil *feature importance* kemudian dipakai untuk mengurutkan variabel yang paling berkontribusi terhadap variasi Energi (kWh) misalnya sinyal kelistrikan (Daya, Arus), diikuti penanda kondisi lingkungan seperti suhu dan kelembaban serta fitur waktu sehingga penajaman strategi kendali dapat diarahkan pada prediktor yang paling dominan. Apabila diperlukan tingkat penjelasan yang lebih mendalam pada tingkat observasi, analisis dapat dilanjutkan menggunakan pendekatan *post-hoc explainability* seperti *SHapley Additive exPlanations (SHAP)* untuk menilai arah dan besar pengaruh masing-masing fitur terhadap prediksi individual, sebagaimana diterapkan pada studi terkini di ranah *energy forecasting* dan sistem cerdas [23], [26], [27]. Temuan ini menjadi landasan pada bagian pembahasan untuk merekomendasikan fitur mana yang perlu dipantau, dikendalikan, atau dioptimasi terlebih dahulu dalam operasi SAWG.

2.9 Evaluasi Model XGBoost

Evaluasi dilakukan untuk menilai sejauh mana model XGBoost regression mampu memprediksi konsumsi energi SAWG secara akurat dan konsisten pada data yang tidak pernah dilihat sebelumnya. Pengujian dilaksanakan pada *hold-out test set* yang telah dipisahkan sejak tahap pembagian data secara kronologis (time-based split), sehingga nilai kinerja yang diperoleh merefleksikan kemampuan generalisasi model terhadap periode operasi yang lebih baru, bukan sekadar kecocokan pada data latih. Model yang diuji merupakan konfigurasi terbaik hasil penyetelan hiperparameter dengan *Time Series Cross-Validation* pada blok data latih, sebagaimana diuraikan pada Subbab 2.7. Pemilihan metrik mengikuti praktik evaluasi pada pemodelan konsumsi/beban energi berbasis *gradient boosting* dan peramalan deret waktu di domain energi [20], [21], [23], [28]. Tiga metrik regresi digunakan, yakni *Mean Absolute Error* (MAE), *Mean Squared Error* (MSE), dan *Coefficient of Determination* (R^2). Kinerja akhir dilaporkan pada dua skema pembagian, yaitu 80:20 (konfigurasi utama) dan 60:40 (uji ketahanan), untuk menilai stabilitas hasil di bawah variasi proporsi data latih–uji. Pada masing-masing skema, hasil model dasar (*baseline*) dibandingkan dengan model pasca *hyperparameter tuning* guna menilai dampak penyetelan parameter terhadap nilai MAE, MSE, dan R^2, baik ketika terjadi perbaikan maupun ketika kinerja justru menurun akibat konfigurasi yang kurang sesuai dengan karakteristik data. Dengan demikian, evaluasi ini tidak hanya menunjukkan angka kinerja akhir, tetapi juga memberikan gambaran yang lebih lengkap mengenai kemampuan generalisasi model XGBoost dan keterbatasannya pada konteks deret waktu konsumsi energi SAWG.

2.9.1 Mean Absolute Error (MAE)

MAE digunakan untuk mengukur rata-rata besarnya kesalahan prediksi tanpa mempertimbangkan arah kesalahan. Metrik ini dinyatakan dalam satuan yang sama dengan variabel target Energi, sehingga mudah diinterpretasikan dalam konteks operasional *SAWG*. Nilai MAE yang lebih kecil menunjukkan bahwa secara rata-rata model menghasilkan prediksi yang lebih mendekati nilai aktual. Berikut merupakan persamaan matematika dari *Mean Absolute Error*.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|yi - \hat{y}i| \quad (1)$$

2.9.2 Root Mean Squared Error (RMSE)

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1730−1742
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8655

RMSE digunakan untuk menilai seberapa besar kesalahan kuadrat rata-rata yang dibuat oleh model. Berbeda dengan *MAE*, *RMSE* memberikan penalti yang lebih besar terhadap kesalahan yang besar sehingga lebih sensitif terhadap deviasi ekstrem. Dalam pemodelan energi, hal ini penting karena konsumsi dapat meningkat tajam pada kondisi operasi tertentu. Berikut merupakan persamaan matematika dari *Root Mean Squared Error*.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(yi - \hat{y}i)^2} \quad (2)$$

2.9.3 Coefficient of Determination (R^2)

R^2 digunakan untuk mengukur proporsi variasi data aktual yang dapat dijelaskan oleh model. Nilai R^2 berada pada rentang 0 sampai 1, dengan nilai yang lebih mendekati 1 menunjukkan bahwa model mampu menjelaskan variasi konsumsi energi dengan lebih baik. Nilai R^2 yang tinggi biasanya menjadi indikator bahwa pemilihan fitur dan parameter model sudah sesuai. Berikut merupakan persamaan matematika dari koefisien determinasi atau R^2.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad (3)$$

Ketiga metrik tersebut digunakan secara bersamaan agar diperoleh gambaran evaluasi yang komprehensif: MAE untuk melihat kesalahan rata-rata dalam satuan kWh, RMSE untuk mendeteksi adanya kesalahan besar, dan R^2 untuk menilai kemampuan model dalam menjelaskan variasi konsumsi energi SAWG.

2.10 Optimasi Operasional Berbasis Prediksi

Sejalan dengan tujuan penelitian untuk mendukung pengambilan keputusan berbasis data, model prediktif XGBoost yang telah dievaluasi dimanfaatkan untuk merancang dan mensimulasikan skenario optimasi operasional berbasis prediksi. Tahap ini mendemonstrasikan bagaimana model peramalan dapat digunakan secara praktis untuk pengendalian beban (load control) dan penjadwalan operasi (operational scheduling) pada perangkat SAWG [7], [15], [18]. Metode optimisasi yang diterapkan adalah Greedy Scheduler harian. Pendekatan ini menggunakan prediksi konsumsi energi per jam dari model XGBoost sebagai kriteria utama seleksi jam operasi. Pemanfaatan model ML misalnya XGBoost sebagai *surrogate* untuk strategi penjadwalan dan manajemen energi merupakan praktik yang umum pada domain Energy Management Systems (EMS), khususnya saat ruang kebijakan besar namun evaluasi cepat diperlukan [7], [12], [16]. Alur metodologis simulasi optimisasi adalah sebagai berikut:

a. Prediksi Konsumsi Energi Per Jam. Data bertimestamp hasil praproses di-*resample* ke interval 1 jam (*resample* 'H'). Model XGBoost final (θ*, hasil tuning pada Subbab 2.9) digunakan untuk memprediksi energi $\hat{y}$(kWh) pada setiap jam berdasarkan fitur lingkungan dan operasional rata-rata pada jam tersebut [7], [15].
b. Penentuan Kuota Operasi (H). Ditetapkan kuota jam operasi harian H(mis. $H = 8$jam) yang merepresentasikan total durasi target perangkat diizinkan beroperasi per hari, konsisten dengan batasan operasional SAWG [12], [18].
c. Skenario Baseline (Naive). Sebagai pembanding, didefinisikan skenario naif: perangkat beroperasi pada Hjam kerja standar pertama dalam hari (mis. 08:00–16:00) tanpa mempertimbangkan prediksi $\hat{y}$[12].
d. Skenario Optimal (Greedy). Diterapkan algoritma greedy yang memilih Hslot jam dengan $\hat{y}$terendah dari seluruh slot harian yang tersedia. Karena evaluasi $\hat{y}$berbasis model sangat cepat, heuristik greedy efektif sebagai kebijakan *myopic* yang hemat komputasi untuk meminimalkan energi kumulatif [12], [16].
e. Analisis Penghematan. Total konsumsi energi kumulatif untuk skenario naif ($E_{\text{naïve}}$)dan skenario greedy (E_{optimal})dibandingkan. Persentase penghematan dihitung sebagai

$$\Delta(\%) = 100 \times \frac{E_{\text{naïve}} - E_{\text{optimal}}}{E_{\text{naïve}}} \quad (4)$$

3. HASIL DAN PEMBAHASAN

3.1 Dataset

Dataset yang dianalisis pada penelitian ini merupakan hasil pencatatan (logging) perangkat Smart Atmospheric Water Generator (SAWG) yang merekam kondisi lingkungan dan parameter kelistrikan secara berkala dalam satu periode operasi. Dataset tersebut terdiri atas 1601 baris dan 9 kolom, sehingga secara kuantitatif dinilai cukup representatif untuk pemodelan regresi pada tingkat perangkat. Struktur variabelnya mencakup satu variabel waktu (timestamp) sebagai penanda urutan pencatatan; empat variabel lingkungan yaitu Suhu (AC) (suhu udara, °C), Kelembaban (%) (kelembapan relatif, %), Kualitas Udara (µg/m³) (konsentrasi partikel polutan), dan Level Air (%) (indikasi kondisi air/tangki); serta empat variabel kelistrikan/operasional yaitu Tegangan (V), Arus (A), Daya (W), dan Energi (kWh). Variabel Energi (kWh) ditetapkan sebagai variabel target karena penelitian ini berfokus pada pemodelan dan prediksi konsumsi energi SAWG berdasarkan kondisi lingkungan dan sinyal kelistrikan. Pemeriksaan awal menunjukkan bahwa terdapat sejumlah nilai hilang (NaN), khususnya pada kolom kelistrikan seperti Daya (W) dan Tegangan (V), sehingga diperlukan tahapan pra-pemrosesan pada bab berikutnya. Variabel waktu telah berhasil dikonversi ke format datetime sehingga dapat digunakan untuk analisis berbasis waktu, dan nilai-nilai awal memperlihatkan variasi yang

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1730−1742
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8655

realistis untuk operasi SAWG di iklim panas dan lembap yang dimana suhu sekitar 25–38 °C, kelembapan 39–55%, kualitas udara 0–31 µg/m³.

3.2 Preprocessing Data

Tahap *preprocessing* menghasilkan perubahan yang cukup signifikan pada kualitas dan jumlah data. Pertama, dilakukan koreksi skala pada variabel kelistrikan, yakni Arus dan Energi, karena hasil inspeksi menunjukkan kedua variabel tersebut tercatat 1.000 kali lebih besar dari rentang yang wajar untuk operasi *SAWG*. Setelah koreksi skala diterapkan, seluruh kolom diperiksa nilai hilangnya. Hasilnya, nilai kosong hanya muncul pada kolom-kolom kelistrikan: Tegangan sebanyak 130 baris, Arus 104 baris, Daya 155 baris, dan Energi 107 baris, sedangkan kolom lingkungan yakni Suhu, Kelembaban, Kualitas Udara, dan Level Air tercatat lengkap. Seluruh baris yang mengandung nilai hilang atau nilai non-numerik pada kolom numerik kemudian dihapus sehingga ukuran data berkurang dari 1601 baris menjadi 1313 baris dengan jumlah kolom tetap 9. Pengurangan 288 baris ini menunjukkan bahwa pembersihan terutama menyasar data kelistrikan yang tidak lengkap.

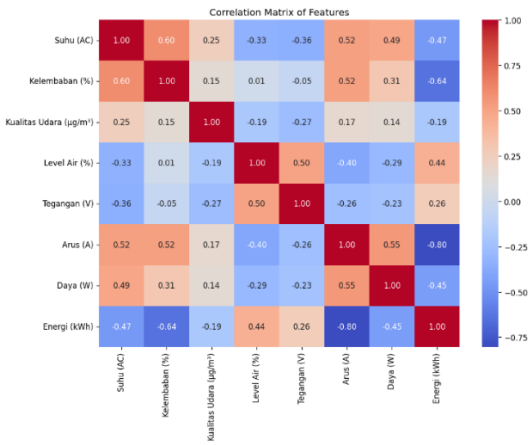
Setelah pembersihan nilai hilang, dilakukan identifikasi *outlier* menggunakan metode rentang antar-kuartil (IQR). Hasilnya menunjukkan: Suhu (AC) memiliki 1 *outlier* dengan batas wajar 22,450–34,850 °C; Kelembaban (%) memiliki 6 *outlier* pada 25,000–73,000%; Kualitas Udara (µg/m³) memiliki 28 *outlier* pada −1,500 hingga 2,500 µg/m³; Level Air (%) tidak memiliki *outlier* (batas −30,000–130,000%); Tegangan (V) memiliki 2 *outlier* pada 214,100–234,900 V; Arus (A) memiliki 1 *outlier* pada −2,613 hingga 4,358 A; Daya (W) memiliki 240 *outlier* pada 188,250–494,250 W; dan Energi (kWh) tidak memiliki *outlier* (batas −81,420–154,745 kWh). Mengingat operasi SAWG dapat mengalami lonjakan daya sesaat, *outlier* pada Daya (W) tidak langsung dihapus dan didokumentasikan sebagai bagian dari dinamika operasi, sedangkan nilai yang jelas tidak realistis ditangani pada tahap pembersihan. Ringkasan tersebut mendasari dataset akhir yang disajikan pada Tabel 2.

Tabel 2. Ringkasan outlier

Variabel	Jumlah outlier (n)	Batas bawah	Batas atas	Catatan
Suhu (AC) (°C)	1	22.45	34.85	Sedikit *outlier*; rentang wajar sempit
Kelembaban (%)	6	25	73	Sebaran moderat
Kualitas Udara (µg/m³)	28	−1.500	2.5	Nilai negatif menunjukkan batas statistik IQR
Level Air (%)	0	−30.000	130	Tidak ada *outlier*
Tegangan (V)	2	214.1	234.9	Variasi kecil di sekitar suplai
Arus (A)	1	−2.613	4.358	Nilai bawah negatif berasal dari batas IQR
Daya (W)	240	188.25	494.25	*Outlier* terbanyak; terkait dinamika beban
Energi (kWh)	0	−81.420	154.745	Tidak ada *outlier* setelah koreksi skala

3.3 Exploratory Data Analysis (EDA)

Setelah dilakukan preprocessing, langkah selanjutnya adalah melakukan *Exploratory Data Analysis* (EDA) untuk memahami hubungan antar variabel dan karakteristik distribusi data sebelum pemodelan XGBoost regression. Koefisien korelasi Pearson antar variabel numerik dihitung dan divisualisasikan dalam bentuk *heatmap* pada Gambar 2.

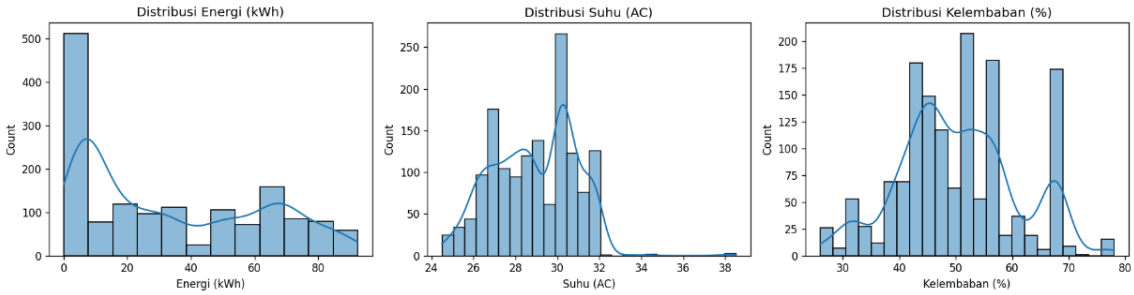


Gambar 2. Heatmap korelasi antarvariabel

Distribusi variabel utama kemudian dianalisis menggunakan histogram yang ditampilkan pada Gambar 3. Distribusi variabel utama. Gambar 3 menunjukkan bahwa distribusi Energi (kWh) cenderung tidak simetris dan

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1730−1742
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8655

mengindikasikan adanya beberapa kelompok nilai misalnya pada rentang energi rendah hingga menengah, yang dapat merefleksikan perbedaan mode operasi seperti idle, beban parsial, dan beban penuh. Distribusi Suhu (AC) berada pada kisaran suhu ruang yang wajar untuk lingkungan tropis, sedangkan Kelembaban (%) tersebar pada rentang yang relevan dengan operasi *atmospheric water generator*. Secara keseluruhan, EDA melalui Gambar 2 dan Gambar 3 mengonfirmasi bahwa yang pertama; data memiliki variasi yang cukup untuk pelatihan model *machine learning*, lalu kedua; terdapat pola hubungan yang dapat dimanfaatkan oleh model, dan terakhir yaitu terdapat anomali korelasi pada variabel Energi (kWh) yang diperlakukan sebagai keterbatasan data dan menjadi pertimbangan penting dalam interpretasi hasil pemodelan.



Gambar 3. Distribusi Variabel Utama

3.4 Pembagian Data

Pada tahap ini, skema pembagian data latih dan data uji yang telah dijelaskan pada Subbab 2.5 diimplementasikan pada keseluruhan deret waktu yang telah melalui proses preprocessing dan EDA. Seluruh observasi terlebih dahulu diurutkan berdasarkan waktu pencatatan sehingga urutan kronologis tetap terjaga. Dari deret waktu tersebut kemudian dibentuk dua skenario pembagian, yaitu skema 80:20 sebagai konfigurasi utama dan skema 60:40 sebagai pembanding. Pada skema 80:20, sekitar delapan puluh persen observasi pada bagian awal deret digunakan sebagai data latih, sedangkan dua puluh persen observasi terakhir digunakan sebagai data uji. Skema ini mensimulasikan kondisi nyata ketika model dilatih menggunakan data historis dan kemudian dievaluasi pada periode operasi yang lebih baru.

Selanjutnya, skema 60:40 diterapkan dengan prinsip yang sama, di mana sekitar enam puluh persen observasi awal berperan sebagai data latih dan empat puluh persen observasi terakhir sebagai data uji. Skenario kedua ini digunakan untuk melihat konsistensi kinerja model ketika proporsi data latih dikurangi dan porsi data uji diperbesar. Pada kedua skema, data uji dipertahankan sebagai *hold-out set* independen yang tidak dilibatkan dalam proses pelatihan, validasi silang, maupun penyetelan hiperparameter. Dengan demikian, evaluasi kinerja model yang disajikan pada subbab berikutnya benar-benar merefleksikan kemampuan model dalam melakukan prediksi pada data SAWG yang belum pernah "dilihat" sebelumnya dan bebas dari kebocoran informasi antarperiode waktu.

3.5 Validasi Model (Time Series Cross-Validation)

Setelah model XGBoost regression dibangun pada data latih hasil pembagian deret waktu (skema 80:20 dan 60:40), dilakukan proses validasi menggunakan *Time Series Cross-Validation* 5 lipatan pada blok data latih masing-masing skema. Pada setiap lipatan, sejumlah segmen awal deret waktu digunakan sebagai data pelatihan, sedangkan segmen berikutnya yang lebih baru digunakan sebagai data validasi. Dengan cara ini, urutan waktu tetap dipertahankan dan tidak ada data dari masa depan yang ikut dilatih, sehingga risiko *data leakage* dapat diminimalkan. Metrik yang dievaluasi pada setiap lipatan meliputi R^2, *Mean Absolute Error* (MAE), dan *Mean Squared Error* (MSE), yang kemudian dirata-ratakan untuk memperoleh gambaran umum kinerja model pada berbagai rentang waktu di dalam data latih.

Hasil validasi menunjukkan bahwa nilai MAE dan MSE antar-lipatan berada pada orde yang sebanding dengan evaluasi pada *hold-out set*, sementara nilai R^2 cenderung rendah bahkan bernilai negatif pada beberapa lipatan. Pola ini mengindikasikan bahwa, meskipun model mampu mengikuti sebagian variasi pola konsumsi energi SAWG, kemampuan penjelasan variansi total masih terbatas dan prediksi sering menyimpang cukup jauh dari nilai aktual pada periode tertentu. Selain itu, perbedaan kinerja antara skema 80:20 dan 60:40 tidak menunjukkan perbaikan signifikan ketika proporsi data latih diperbesar, sehingga menguatkan indikasi bahwa keterbatasan bukan hanya berasal dari jumlah data, tetapi juga dari karakteristik dan kualitas dataset yang telah dibahas pada subbab EDA. Secara keseluruhan, validasi dengan *Time Series Cross-Validation* ini memberikan gambaran yang lebih realistis tentang performa model XGBoost pada skenario deret waktu dan menjadi dasar untuk interpretasi hasil pada data uji independen di subbab berikutnya.

3.6 Hyperparameter Tuning

Tahap *hyperparameter tuning* dilakukan untuk menyempurnakan kinerja model dasar yang sebelumnya telah divalidasi menggunakan *Time Series Cross-Validation* pada data latih. Penyetelan dilakukan dengan bantuan GridSearchCV yang dikombinasikan dengan skema validasi silang 5-lipatan berbasis deret waktu, sehingga setiap

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1730−1742
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8655

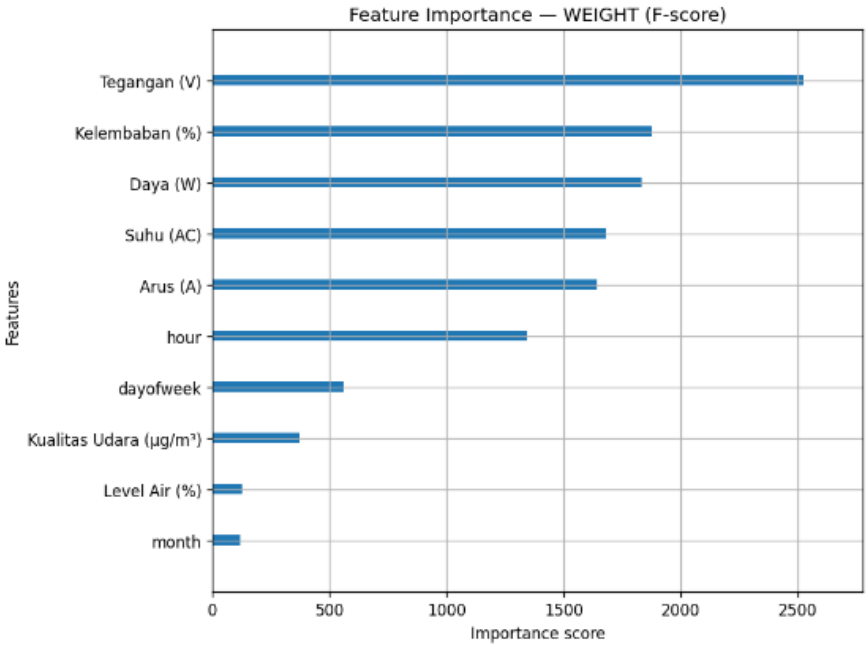
kombinasi parameter diuji secara konsisten pada beberapa segmen waktu yang berbeda tanpa melanggar urutan kronologis. Ruang pencarian yang digunakan mencakup parameter-parameter utama XGBoost, yaitu jumlah pohon (*n_estimators* = 500 dan 1000), kedalaman maksimum pohon (*max_depth* = 5 dan 7), laju pembelajaran (*learning_rate* = 0,05 dan 0,1), serta rasio pengambilan sampel baris dan fitur (*subsample* = 0,8 dan *colsample_bytree* = 0,8). Kombinasi ini dipilih karena berpengaruh langsung terhadap kemampuan model dalam menyeimbangkan kompleksitas dan generalisasi pada pemodelan konsumsi energi berbasis gradient boosting.
Hasil pencarian menunjukkan bahwa konfigurasi terbaik diperoleh pada parameter :

a. n_estimators : 1000
b. max_depth : 7
c. learning_rate : 0.1
d. subsample : 0.8
e. colsample_bytree : 0.8

Dengan kombinasi tersebut, model menghasilkan skor R^2 terbaik sebesar 0,9919 pada proses pencarian berbasis *Time Series Cross-Validation*. Nilai ini sedikit lebih tinggi daripada rata-rata R^2 pada validasi model dasar (0,9886), yang menunjukkan bahwa penyetelan hiperparameter berhasil meningkatkan kemampuan model dalam menjelaskan variasi konsumsi energi SAWG pada data latih. Peningkatan ini juga mengindikasikan bahwa model merespons positif terhadap pengaturan kedalaman pohon yang lebih tinggi dan laju pembelajaran yang tidak terlalu kecil, sehingga pola nonlinier antarfitur khususnya antara variabel kelistrikan dan informasi waktu dapat ditangkap dengan lebih baik. Konfigurasi terbaik inilah yang selanjutnya digunakan sebagai model final pada tahap evaluasi terhadap *hold-out test set* dan sebagai dasar simulasi optimasi operasional pada subbab berikutnya.

3.7 Feature Importance

Analisis *feature importance* dilakukan pada model XGBoost terbaik untuk mengetahui variabel mana yang paling berkontribusi terhadap prediksi Energi (kWh). Gambar 4 menampilkan *feature importance* berdasarkan metrik WEIGHT (F-score), yaitu seberapa sering suatu fitur digunakan sebagai pemisah simpul pada seluruh pohon dalam ensemble. Terlihat bahwa Tegangan (V) merupakan fitur yang paling sering muncul, diikuti oleh Kelembaban (%), Daya (W), Suhu (AC), dan Arus (A). Fitur waktu hour menempati posisi menengah, sedangkan dayofweek, Kualitas Udara (µg/m³), Level Air (%), dan month relatif jarang digunakan. Pola ini mengindikasikan bahwa dari sisi frekuensi pemakaian, sinyal kelistrikan dan kondisi udara (tegangan, kelembaban, daya, suhu, arus) menjadi informasi utama yang dipakai model untuk membentuk pemisahan keputusan, sementara informasi waktu dan beberapa variabel lingkungan lain lebih berperan sebagai pelengkap.

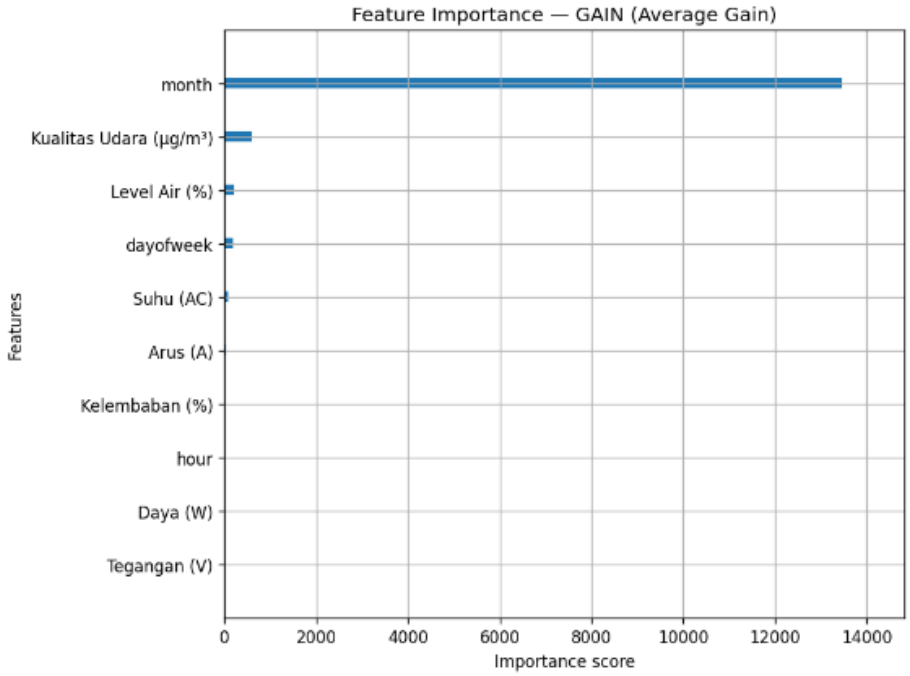


Gambar 4. Feature Importance F-score/weight

Gambar 5 menunjukkan *feature importance* berdasarkan metrik GAIN (average gain), yaitu rata-rata penurunan kesalahan (loss) ketika suatu fitur digunakan sebagai pemisah simpul. Berbeda dengan metrik WEIGHT, pada metrik GAIN fitur month justru mendominasi dengan nilai yang jauh lebih besar dibanding fitur lain, diikuti oleh Kualitas Udara (µg/m³) dan Level Air (%). Fitur dayofweek dan Suhu (AC) memberikan kontribusi sedang, sedangkan Arus (A), Kelembaban (%), hour, Daya (W), dan Tegangan (V) memiliki nilai gain yang relatif kecil. Hal ini menunjukkan bahwa meskipun fitur kelistrikan sering digunakan pada banyak simpul, pemisahan yang memberikan penurunan kesalahan terbesar justru banyak melibatkan informasi bulan, kualitas udara, dan level air. Dengan kata

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1730−1742
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8655

lain, aspek musiman/waktu (month) dan kondisi lingkungan tertentu berpotensi memicu perubahan pola konsumsi energi yang cukup tajam pada sebagian titik data.



Gambar 5. Feature Importance by Gain

Perbedaan pola antara metrik WEIGHT dan GAIN ini menggarisbawahi bahwa kontribusi fitur tidak hanya ditentukan oleh seberapa sering fitur tersebut digunakan, tetapi juga seberapa besar dampak setiap pemisahan terhadap penurunan kesalahan model. Secara umum, hasil *feature importance* mengonfirmasi bahwa konsumsi energi SAWG dipengaruhi oleh kombinasi faktor kelistrikan (tegangan, daya, arus), kondisi lingkungan (kelembaban, suhu, kualitas udara, level air), serta informasi waktu (jam, hari, bulan). Namun dominasi fitur month pada metrik GAIN dan temuan anomali pada analisis korelasi sebelumnya menunjukkan bahwa interpretasi kausal perlu dilakukan dengan hati-hati, karena masih terdapat keterbatasan kualitas dan cakupan dataset yang bisa memengaruhi cara model mempelajari pola konsumsi energi.

3.8 Evaluasi Model XGBoost

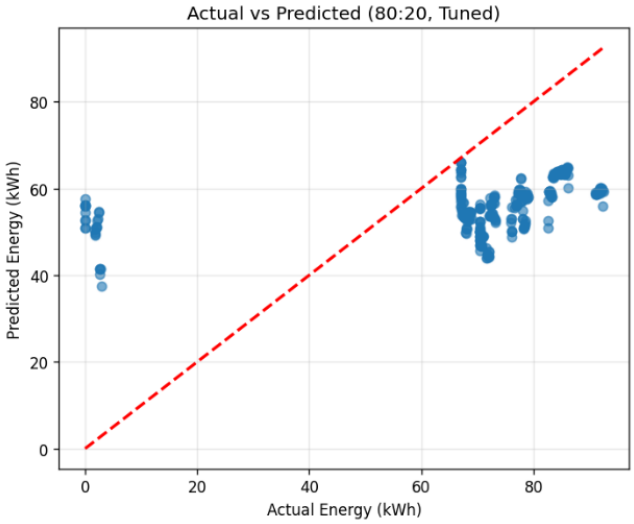
Kinerja akhir model XGBoost regression dievaluasi pada *hold-out test set* untuk dua skema pembagian deret waktu, yaitu split 80:20 sebagai konfigurasi utama dan split 60:40 sebagai pembanding. Tabel 3. Performa model XGBoost pada data uji merangkum nilai *Mean Absolute Error* (MAE), *Mean Squared Error* (MSE), dan koefisien determinasi (R^2) untuk kedua skema tersebut. Pada split 80:20, model menghasilkan MAE sebesar sekitar 23,16 kWh, MSE sekitar 648,93 kWh^2, dan R^2 sebesar −0,22. Sementara itu, pada split 60:40, nilai MAE meningkat menjadi sekitar 27,21 kWh, MSE menjadi sekitar 932,17 kWh^2, dan R^2 turun hingga −1,75. Nilai R^2 yang negatif pada kedua skema menunjukkan bahwa dalam pengujian yang lebih ketat berbasis waktu ini, model belum mampu mengungguli prediksi sederhana berbasis rata-rata energi, sehingga kemampuan penjelasan variansi total konsumsi energi SAWG masih terbatas.

Tabel 3. Performa Model XGBoost

Split	MAE	MSE	R2
80:20	23.1609	648.9291	-0.2189
60:40	27.2137	932.1677	-1.7509

Dibandingkan antara kedua skema, split 80:20 memberikan kesalahan yang lebih kecil dan nilai R^2 yang kurang negatif dibanding split 60:40, yang mengindikasikan bahwa proporsi data latih yang lebih besar sedikit membantu model mengenali pola historis sebelum diuji pada periode waktu yang lebih baru. Namun, peningkatan tersebut belum cukup untuk menghasilkan kinerja yang memuaskan, terlebih dengan sebaran titik pada grafik Actual vs Predicted (Gambar 6) yang masih jauh dari garis diagonal $y = x$. Kombinasi nilai MAE dan MSE yang relatif tinggi, R^2 yang negatif, serta visualisasi yang menunjukkan penyimpangan prediksi pada banyak pengamatan menegaskan bahwa model XGBoost dalam konfigurasi ini masih menghadapi keterbatasan dalam memodelkan pola konsumsi energi SAWG pada konteks deret waktu, sehingga hasil evaluasi perlu dibaca dengan mempertimbangkan keterbatasan dataset yang telah dibahas pada subbab sebelumnya.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1730−1742
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8655

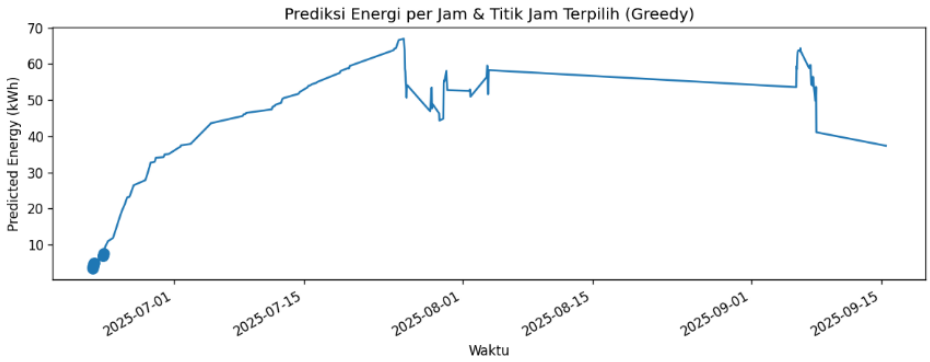


Gambar 6. Perbandingan nilai aktual dan nilai prediksi

3.9 Hasil Simulasi Optimalisasi Operasional Berbasis Prediksi

Pada tahap ini, model XGBoost final digunakan sebagai dasar simulasi penjadwalan operasional SAWG dengan pendekatan *Greedy Scheduler*. Deret prediksi konsumsi energi per jam $\hat{y}_t$ yang dihasilkan model untuk horizon waktu tertentu dimanfaatkan untuk membandingkan dua skenario: (1) skenario *naive* (baseline), di mana perangkat beroperasi pada delapan jam kerja berturut-turut tanpa mempertimbangkan variasi beban terprediksi; dan (2) skenario optimal (*greedy*), di mana delapan slot jam dipilih dari seluruh horizon dengan mengutamakan nilai $\hat{y}_t$ terendah hingga total durasi operasi mencapai 8 jam.

Hasil simulasi diringkas pada Gambar 7 dan Tabel 4. Gambar 7 menampilkan profil prediksi energi per jam beserta penandaan titik-titik jam yang dipilih oleh algoritma *greedy*. Titik terpilih cenderung berada pada bagian kurva dengan nilai prediksi yang relatif rendah. Tabel 4 menunjukkan delapan timestamp rekomendasi beserta nilai $\hat{y}_t$-nya, misalnya 22 Juni 2025 pukul 07.00–12.00 dengan rentang prediksi sekitar 3,61–5,03 kWh serta beberapa jam pada 23 Juni 2025 pukul 09.00–11.00 dengan prediksi sekitar 6,89–7,57 kWh. Pada horizon simulasi ini, total energi terprediksi untuk jadwal *naive* adalah sekitar 47,493 kWh, sedangkan jadwal hasil seleksi *greedy* menghasilkan total energi terprediksi sekitar 43,134 kWh, sehingga diperoleh estimasi penghematan sekitar 9,18% terhadap konsumsi energi pada horizon yang sama.



Gambar 7. Perbandingan nilai aktual dan nilai prediksi

Meskipun besarnya penghematan masih terbatas dan sangat bergantung pada akurasi model serta pola beban pada dataset yang digunakan, hasil ini menunjukkan bahwa keluaran model prediksi dapat dimanfaatkan untuk menyusun jadwal operasi yang sedikit lebih hemat dibanding pola jam kerja statis. Dengan kata lain, bahkan pada kondisi di mana kemampuan prediksi model belum optimal, integrasi prediksi energi dengan skema penjadwalan *greedy* sudah dapat berperan sebagai langkah awal menuju pengelolaan operasi SAWG yang lebih efisien dan adaptif terhadap variasi beban.

Tabel 4. Jam operasi rekomendasi dan $\hat{y}_t$(kWh)

Timestamp	y_hat
6/22/2025 07:00	3.609
6/22/2025 08:00	3.832
6/22/2025 09:00	4.239

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1730−1742
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8655

Timestamp	y_hat
6/22/2025 10:00	4.862
6/22/2025 12:00	5.031
6/23/2025 09:00	6.892
6/23/2025 10:00	7.102
6/23/2025 11:00	7.568

4. KESIMPULAN

Secara keseluruhan, penelitian ini telah menyusun dan menguji sebuah rantai pemodelan konsumsi energi SAWG berbasis XGBoost regression dengan tahapan yang jelas dan dapat direplikasi. Alur yang dibangun mencakup pembersihan data dan penanganan nilai tidak wajar, pembentukan fitur waktu (jam, hari, bulan), serta pembagian data deret waktu secara kronologis menjadi dua skema utama, yaitu split 80:20 dan 60:40. Pada blok data latih, model divalidasi dan disetel menggunakan *Time Series Cross-Validation* serta prosedur *hyperparameter tuning*, kemudian dianalisis lebih lanjut melalui *feature importance* untuk melihat kontribusi masing-masing variabel. Di bawah skema evaluasi deret waktu yang lebih ketat ini, kinerja model pada data uji masih terbatas: split 80:20 menghasilkan MAE ≈ 23,16 kWh, MSE ≈ 648,93 kWh², dan $R^2 \approx -0{,}22$, sedangkan split 60:40 memberikan MAE ≈ 27,21 kWh, MSE ≈ 932,17 kWh², dan $R^2 \approx -1{,}75$. Hasil tersebut mengindikasikan bahwa model baru mampu mengikuti sebagian pola konsumsi energi, namun belum memadai untuk menjelaskan variansi keseluruhan secara deret waktu. Analisis *feature importance* menunjukkan bahwa kombinasi variabel kelistrikan (tegangan, daya, arus), kondisi lingkungan (kelembaban, suhu, kualitas udara, level air), serta informasi waktu masih menjadi faktor dominan, sekaligus menegaskan adanya keterbatasan dataset yang memengaruhi cara model belajar. Studi ini memperlihatkan bagaimana keluaran model prediksi dapat diintegrasikan ke dalam mekanisme penjadwalan operasional berbasis data. Keluaran model ($\hat{y}_t$) dimanfaatkan dalam skema penjadwalan harian menggunakan Greedy Scheduler, yang membandingkan jadwal operasi naive dengan jadwal hasil seleksi jam-jam berenergi terprediksi rendah. Pada horizon simulasi, jadwal naive menghasilkan total energi terprediksi 47,493 kWh, sedangkan jadwal greedy menghasilkan 43,134 kWh sehingga memberikan estimasi penghematan sekitar 9,18%. Penghematan ini masih bersifat prediktif dan sangat bergantung pada kualitas data serta akurasi model, sementara berbagai kendala praktis seperti target produksi air minimum, biaya transisi ON/OFF, dan maintenance window belum dimasukkan ke dalam formulasi. Oleh karena itu, penelitian lanjutan perlu diarahkan pada peningkatan kualitas dan cakupan data (lintas perangkat, musim, dan lokasi), pengayaan fitur dengan informasi meteorologi dan parameter operasi tambahan, eksplorasi model deret waktu yang lebih khusus, serta pengembangan skema optimasi terbatasi yang merepresentasikan kondisi operasi nyata. Dengan penguatan tersebut, pendekatan pemodelan dan penjadwalan berbasis prediksi yang dikembangkan dalam studi ini berpotensi berkembang menjadi sistem pendukung keputusan yang lebih matang untuk meningkatkan efisiensi energi SAWG di lapangan.

REFERENCES

[1] N. Khan, S. Khan, Q. Khorajiya, J. Sairan, and Prof. M. A. Gulbarga, "Atmospheric Water Generator," International Journal for Research in Applied Science and Engineering Technology, vol. 10, no. 4, pp. 929–935, Apr. 2022, doi: 10.22214/ijraset.2022.41406.

[2] X. Liu, D. Beysens, and T. Bourouina, "Water Harvesting from Air: Current Passive Approaches and Outlook," *ACS Materials Letters*, vol. 4, no. 5, pp. 1003–1024, May 2022, doi: 10.1021/acsmaterialslett.1c00850.

[3] J. Wang, Z. Yang, Z. Li, H. Fu, and J. Chen, "Comprehensive review on atmospheric water harvesting technologies," *Journal of Water Process Engineering*, vol. 69, p. 106836, Jan. 2025, doi: 10.1016/j.jwpe.2024.106836.

[4] A. Bhardwaj et al., "Smart IoT and Machine Learning-based Framework for Water Quality Assessment and Device Component Monitoring," Environmental Science and Pollution Research, vol. 29, no. 30, pp. 46018–46036, Jun. 2022, doi: 10.1007/s11356-022-19014-3.

[5] P. Wang, J. Xu, Z. Bai, R. Wang, and T. Li, "Designing next-generation all-weather and efficient atmospheric water harvesting powered by solar energy," *Energy & Environmental Science*, vol. 18, no. 14, pp. 7005–7022, 2025, doi: 10.1039/D5EE01454A.

[6] L. Cattani, P. Cattani, R. Figoni, and A. Magrini, "Performance Assessment of Atmospheric Water Generators: A Review of Evaluation Tools and Proposal for a Novel Advanced Global Evaluation Index for HVAC–AWG Hybrid Solutions," *Applied Sciences*, vol. 14, no. 24, p. 11793, Dec. 2024, doi: 10.3390/app142411793.

[7] E. Sadowski, E. Mbonimpa, and C. M. Chini, "Benchmarks of production for atmospheric water generators in the United States," *PLOS Water*, vol. 2, no. 6, p. e0000133, Jun. 2023, doi: 10.1371/journal.pwat.0000133.

[8] F. Faraz Ahmad, C. Ghenai, M. al Bardan, M. Bourgon, and A. Shanableh, "Performance analysis of atmospheric water generator under hot and humid climate conditions: Drinkable water production and system energy consumption," *Case Studies in Chemical and Environmental Engineering*, vol. 6, p. 100270, Dec. 2022, doi: 10.1016/j.cscee.2022.100270.

[9] I. I. El-Sharkawy, S. Haridy, M. Hassan, A. Radwan, and M. M. Abd-Elhady, "Optimization of atmospheric water harvesting cycles for sustainable water supply in arid regions," *International Journal of Thermofluids*, vol. 24, p. 100977, Nov. 2024, doi: 10.1016/j.ijft.2024.100977.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1730−1742
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8655

[10] B. Asiabanpour, N. Ownby, M. Summers, and F. Moghimi, "Atmospheric Water Generation and Energy Consumption: An Empirical Analysis," in *2019 IEEE Texas Power and Energy Conference (TPEC)*, IEEE, Mar. 2019, pp. 1–6. doi: 10.1109/TPEC.2019.8662164.
[11] F. Gaggini, R. N. F. Crespo, M. Calò, V. M. Gentile, A. Macii, and E. Patti, "An Internet of Things Ecosystem for Atmospheric Water Generators," *IEEE Access*, vol. 13, pp. 77565–77581, Apr. 2025, doi: 10.1109/ACCESS.2025.3562742.
[12] M. Lowe, R. Qin, and X. Mao, "A Review on Machine Learning, Artificial Intelligence, and Smart Technology in Water Treatment and Monitoring," Water, vol. 14, no. 9, p. 1384, Apr. 2022, doi: 10.3390/w14091384.
[13] G. Barletta, S. Moitra, S. Derrible, A. Mathew, A. M. Nair, and C. M. Megaridis, "Exploring machine learning models to predict atmospheric water harvesting with an ion deposition membrane," *Journal of Water Process Engineering*, vol. 72, p. 107476, Apr. 2025, doi: 10.1016/j.jwpe.2025.107476.
[14] M. S. Islam and B. Asiabanpour, "Machine Learning-Based Optimization of Surface Temperature in TPMS Geometries for Atmospheric Water Generation Systems," in *Proc. IISE Annual Conference & Expo 2025*, Atlanta, GA, USA, 2025, doi: 10.5281/zenodo.15644127.
[15] L. Li, Z. Shi, H. Liang, J. Liu, and Z. Qiao, "Machine Learning-Assisted Computational Screening of Metal-Organic Frameworks for Atmospheric Water Harvesting," *Nanomaterials*, vol. 12, no. 1, p. 159, Jan. 2022, doi: 10.3390/nano12010159.
[16] C. Cuevas, A. Cendoya, D. Sacasas, and M. Pezo, "Evaluation of the Potential of Atmospheric Water Generators to Mitigate Water Scarcity in Northern Chile," *Processes*, vol. 13, no. 9, p. 3003, Sep. 2025, doi: 10.3390/pr13093003.
[17] C. Rincón, J. Alencastre, E. Barrantes, and C. Carbajal, "Atmospheric water generator: design for rural zones with low and medium humidity level," *South Florida Journal of Development*, vol. 5, no. 7, p. e4081, Jul. 2024, doi: 10.46932/sfjdv5n7-002.
[18] E. Fosso-Kankeu, A. Al Alili, H. Mittal, and B. Mamba, Eds., *Atmospheric Water Harvesting: Development and Challenges*. Cham: Springer, 2023, doi: 10.1007/978-3-031-21746-3.
[19] S. D. Deshmukh and A. V. Deshmukh, *Atmospheric Water Generation (AWG): Engineering Principles, Technologies and Applications*. Lulu.com, 2025.
[20] N. Maleki, O. Lundström, A. Musaddiq, J. Jeansson, T. Olsson, and F. Ahlgren, "Future energy insights: Time-series and deep learning models for city load forecasting," *Applied Energy*, vol. 374, p. 124067, Nov. 2024, doi: 10.1016/j.apenergy.2024.124067.
[21] L. Ren, L. Zhang, H. Wang, and Q. Guo, "An Extreme Gradient Boosting Algorithm for Short-Term Load Forecasting Using Power Grid Big Data," 2019, pp. 479–490. doi: 10.1007/978-981-13-2288-4_46.
[22] S. G. Kangalli Uyar, B. K. Ozbay, and B. Dal, "Interpretable building energy performance prediction using XGBoost Quantile Regression," *Energy and Buildings*, vol. 340, p. 115815, Aug. 2025, doi: 10.1016/j.enbuild.2025.115815.
[23] E. L. Alba, G. A. Oliveira, M. H. D. M. Ribeiro, and É. O. Rodrigues, "Electricity Consumption Forecasting: An Approach Using Cooperative Ensemble Learning with SHapley Additive exPlanations," *Forecasting*, vol. 6, no. 3, pp. 839–863, Sep. 2024, doi: 10.3390/forecast6030042.
[24] Z. Mustaffa and M. H. Sulaiman, "Advanced forecasting of building energy loads with XGBoost and metaheuristic algorithms integration," *Energy Storage and Saving*, Aug. 2025, doi: 10.1016/j.enss.2025.03.005.
[25] A. Muqtadir, B. Li, Z. Ying, C. Songsong, and S. N. Kazmi, "Nowcasting the next hour of residential load using boosting ensemble machines," *Scientific Reports*, vol. 15, no. 1, p. 7157, Feb. 2025, doi: 10.1038/s41598-025-91767-6.
[26] H. Abbasimehr, R. Paki, and A. Bahrini, "A novel XGBoost-based featurization approach to forecast renewable energy consumption with deep learning models," *Sustainable Computing: Informatics and Systems*, vol. 38, p. 100863, Apr. 2023, doi: 10.1016/j.suscom.2023.100863.
[27] I. Hussain, K. B. Ching, C. Uttraphan, K. G. Tay, and A. Noor, "Evaluating machine learning algorithms for energy consumption prediction in electric vehicles: A comparative study," *Scientific Reports*, vol. 15, no. 1, p. 16124, May 2025, doi: 10.1038/s41598-025-94946-7.
[28] A. T. Mustafa and O. S. A.-D. Al-Yozbaky, "Forecasting Energy Demand and Generation Using Time Series Models: A Comparative Analysis of Classical, Grey, Fuzzy, and Intelligent Approaches," *Franklin Open*, p. 100350, Aug. 2025, doi: 10.1016/j.fraope.2025.100350.