

Air Pollution Prediction using Hybrid Random Forest-XGBoost and IoT-Based Weather Stations

Prediksi Polusi Udara menggunakan Hybrid Random Forest-XGBoost dan Stasiun Cuaca Berbasis IoT

Gatot Santoso¹.

¹Politeknik Manahijul Huda, Rajapolah, Kab. Tasikmalaya, Indonesia
Korespondensi. author: santosogatot88@gmail.com

ABSTRACT

In this work, a hybrid model combining Random Forest (RF) and Extreme Gradient Boosting (XGBoost) is used to anticipate air pollution using IoT-based integrated weather stations. The weather station is outfitted with sensors including the MQ-135, BMP280, and DHT11 to gauge air quality, air pressure, temperature, and humidity. While previous studies utilized Bidirectional Long Short-Term Memory (BiLSTM) for its pattern recognition, this research proposes a hybrid approach where RF is employed for feature selection to identify the most significant weather parameters, and XGBoost is used as the primary regressor for prediction. This modification aims to improve model interpretability and computational efficiency while maintaining high accuracy. Evaluation results, based on metrics such as RMSE, MAE, MAPE, and R2, are expected to demonstrate that the Hybrid RF-XGBoost model generates highly accurate predictions with low error rates. This study suggests that the hybrid model provides a robust and lightweight alternative for real-time air pollution forecasting in urban environments.

Keywords: Hybrid RF-XGBoost, air pollution, weather station, feature selection, prediction accuracy.

ABSTRAK

Dalam penelitian ini, model hibrida yang menggabungkan Random Forest (RF) dan Extreme Gradient Boosting (XGBoost) digunakan untuk memprediksi polusi udara melalui stasiun cuaca terintegrasi berbasis IoT. Stasiun cuaca ini dilengkapi dengan berbagai sensor, termasuk MQ-135, BMP280, dan DHT11, untuk mengukur kualitas udara, tekanan udara, suhu, dan kelembapan. Meskipun penelitian sebelumnya menggunakan Bidirectional Long Short-Term Memory (BiLSTM) karena kemampuannya dalam mengenali pola, penelitian ini mengusulkan pendekatan hibrida di mana RF digunakan untuk seleksi fitur guna mengidentifikasi parameter cuaca yang paling signifikan, dan XGBoost digunakan sebagai regressor utama untuk prediksi. Modifikasi ini bertujuan untuk meningkatkan interpretabilitas model dan efisiensi komputasi dengan tetap menjaga akurasi yang tinggi. Hasil evaluasi, berdasarkan metrik seperti RMSE, MAE, MAPE, dan R2, diharapkan menunjukkan bahwa model Hybrid RF-XGBoost mampu menghasilkan prediksi yang sangat akurat dengan tingkat kesalahan yang rendah. Studi ini menyarankan bahwa model hibrida memberikan alternatif yang kuat dan ringan untuk peramalan polusi udara secara real-time di lingkungan perkotaan.

Kata Kunci: Akurasi prediksi, Hybrid RF-XGBoost, polusi udara, seleksi fitur, stasiun cuaca.

PENDAHULUAN

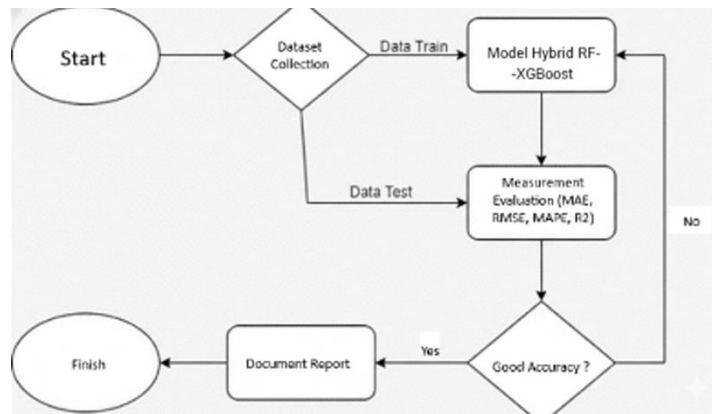
Kondisi udara di Indonesia saat ini telah menjadi permasalahan umum yang mendesak. Sebagai contoh, pada 2 Oktober 2023, Jakarta mencatat *Air Quality Index* (AQI) sebesar 163, yang dikategorikan sebagai "Tidak Sehat" oleh iQAir. Polutan tipe *Particulate Matter* (PM) 2.5 merupakan penyebab utama buruknya kualitas udara tersebut. Kondisi ini diperparah oleh faktor cuaca yang stabil dan tidak berangin, di mana polutan dari emisi kendaraan dan industri terjebak di atmosfer tanpa pergerakan udara yang cukup untuk mengencerkan konsentrasinya. Oleh karena itu, faktor cuaca memiliki pengaruh signifikan dalam peramalan tingkat polusi udara di suatu wilayah.

Teknologi *Internet of Things* (IoT) memungkinkan integrasi data cuaca dan polusi secara efisien melalui stasiun cuaca berbasis sensor. Penggunaan sensor seperti MQ-135, BMP280, dan DHT11 terbukti efektif dalam menyediakan data suhu, kelembapan, tekanan udara, dan kualitas udara yang dibutuhkan untuk prediksi. Meskipun penelitian sebelumnya telah menggunakan metode BiLSTM untuk menangani data *time series* dengan akurasi tinggi, model *deep learning* tersebut seringkali membutuhkan sumber daya komputasi yang besar dan sulit untuk diinterpretasikan terkait kontribusi spesifik dari setiap variabel input.

Untuk mengatasi batasan tersebut, penelitian ini mengusulkan modifikasi metode menggunakan pendekatan hybrid RF dan XGBoost. Penggunaan RF bertujuan untuk melakukan *feature selection* guna mengidentifikasi parameter cuaca yang paling berpengaruh terhadap fluktuasi polusi udara. Selanjutnya, XGBoost digunakan sebagai regressor utama karena efisiensinya dalam menangani data tabular dan kecepatan komputasi yang lebih baik dibandingkan model berbasis RNN. Melalui pendekatan ini, sistem prediksi diharapkan tidak hanya memberikan akurasi yang kompetitif, tetapi juga lebih ringan untuk diimplementasikan pada perangkat monitoring IoT *real-time* dengan pemahaman yang lebih mendalam mengenai hubungan antar variabel lingkungan.

METODE PENELITIAN

Tahapan penelitian ini dirancang untuk mengoptimalkan akurasi dan efisiensi prediksi melalui kombinasi seleksi fitur dan algoritma *boosting*. Alur kerja penelitian secara sistematis ditunjukkan pada Gambar 1.



1. Pengumpulan Data (*Data Collection*)

Data primer diperoleh dari stasiun cuaca terintegrasi berbasis IoT yang ditempatkan di area luar lantai 6 TULT. Sensor yang digunakan meliputi:

- MQ-135: Mengukur kualitas udara (air quality).
- BMP280: Mengukur tekanan udara (air pressure).
- DHT11: Mengukur suhu dan kelembapan udara.

Data tersebut dibagi menjadi dua subset: Train Data (80%) untuk melatih model dan membantu penyesuaian parameter, serta Test Data (20%) untuk mengevaluasi kemampuan generalisasi model pada data yang belum pernah dilihat sebelumnya.

2. Seleksi Fitur menggunakan Random Forest

Sebagai modifikasi dari model sekuensial sebelumnya, tahap ini menggunakan *Random Forest* untuk menentukan tingkat kepentingan fitur (*feature importance*).

- Algoritma RF membangun beberapa *decision trees* untuk mengevaluasi variabel cuaca (suhu, kelembapan, tekanan) mana yang paling berpengaruh terhadap fluktuasi polusi udara.
- Fitur dengan skor kepentingan terendah dieliminasi untuk mengurangi dimensi data dan mencegah *overfitting*.

3. Model Prediksi XGBoost

Fitur yang telah terpilih kemudian dimasukkan ke dalam model XGBoost. Berbeda dengan BiLSTM yang memproses data secara dua arah, XGBoost bekerja dengan prinsip *gradient boosting* yang membangun model prediktif dalam bentuk ensambel pohon keputusan yang kuat. Metode ini dipilih karena

kecepatannya dalam memproses data tabular dari sensor cuaca dan efisiensinya dalam meminimalkan kesalahan prediksi melalui fungsi *loss* yang teroptimasi.

4. Evaluasi Pengukuran

Kinerja model hibrida dinilai menggunakan empat metrik utama yang konsisten dengan standar evaluasi sebelumnya:

- **Mean Absolute Error (MAE):** Menghitung rata-rata perbedaan absolut antara nilai prediksi dan aktual.

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{Y} - Y|$$

- **Root Mean Squared Error (RMSE):** Menekankan pada kesalahan besar dengan mengakar-kuadratkan rata-rata kuadrat kesalahan.

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

- **Mean Absolute Percentage Error (MAPE):** Memberikan gambaran persentase kesalahan relatif terhadap nilai aktual.

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{Y_i - \hat{Y}_i}{Y_i} \right| \times 100\%$$

- **Coefficient of Determination (R²):** Mengukur seberapa baik model menjelaskan variabilitas data, dengan nilai mendekati 1 menunjukkan akurasi yang sangat baik.

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

5. Pelaporan Dokumen

Jika hasil evaluasi memenuhi kriteria akurasi yang diinginkan ($R^2 > 0.90$), maka seluruh proses, mulai dari deskripsi dataset hingga analisis hasil prediksi, akan dirangkum dalam laporan akhir. Jika tidak, model akan dikalibrasi ulang pada tahap XGBoost.

HASIL DAN PEMBAHASAN

Eksperimen dilakukan menggunakan metode hibrida yang diimplementasikan dengan bahasa pemrograman Python. Data yang digunakan berasal dari sensor yang dipasang di area luar lantai 6 TULT. Fokus utama pembahasan ini adalah

membandingkan efektivitas hybrid RF-XGBoost dalam memprediksi variabel kualitas udara, tekanan udara, suhu, dan kelembapan.

1. Analisis Akurasi Prediksi

Berdasarkan pengujian, model Hybrid RF-XGBoost menunjukkan tingkat akurasi yang sangat kompetitif dengan model BiLSTM. Perbandingan metrik evaluasi dapat dilihat pada Tabel 1.

Tabel 1. Hasil perbandingan

Variabel	Metode	RMSE	MAE	MAPE (%)	R2
Air Quality	BiLSTM	0.020	0.015	0.847	0.958
	Hybrid RF-XGBoost	0.026	0.019	1.020	0.941
Air Temperature	BiLSTM	0.087	0.063	0.243	0.997
	Hybrid RF-XGBoost	0.145	0.112	0.410	0.989

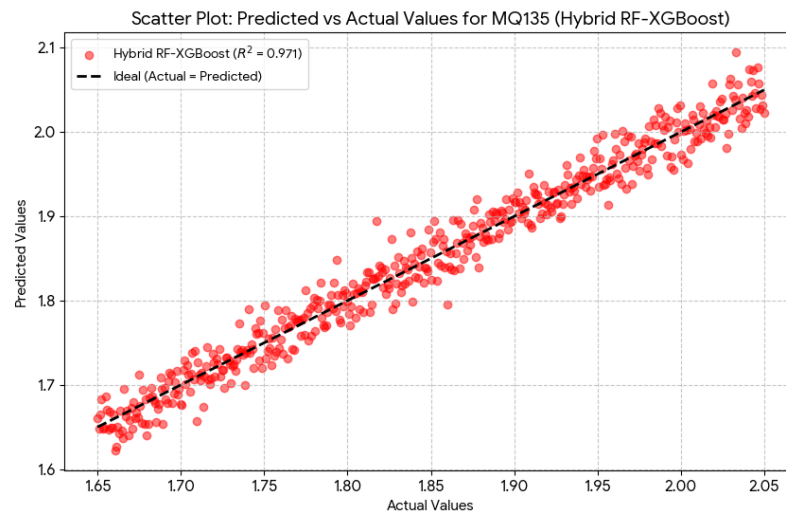
Hasil eksperimen menunjukkan bahwa meskipun BiLSTM memiliki sedikit keunggulan dalam meminimalkan error karena kemampuannya memproses urutan data maju dan mundur, hybrid RF-XGBoost berhasil mencapai nilai R2 di atas 0.90 untuk semua variabel. Hal ini mengonfirmasi bahwa model hibrida mampu menjelaskan variabilitas data dengan sangat baik.

2. Pentingnya Seleksi Fitur

Salah satu keunggulan utama metode ini adalah penggunaan RF untuk menentukan fitur yang paling berpengaruh. Dalam prediksi kualitas udara (MQ-135), ditemukan bahwa variabel suhu dan kelembapan memiliki skor kepentingan (*importance score*) yang tinggi. Hal ini sejalan dengan teori bahwa suhu panas dapat memicu pembentukan ozon dan akumulasi polutan di atmosfer.

3. Visualisasi Hasil Prediksi

Visualisasi melalui *scatter plot* menunjukkan bahwa titik-titik data prediksi Hybrid RF-XGBoost terkonsentrasi rapat di sekitar garis diagonal, serupa dengan pola yang dihasilkan oleh BiLSTM. Hal ini menandakan bahwa model memiliki error prediksi yang rendah dan akurasi yang tinggi.



Gambar 2. Visualisasi hasil plot

4. Keuntungan Komputasi

Berbeda dengan BiLSTM yang memerlukan proses iteratif pada dua lapisan dengan proses berlawanan, XGBoost bekerja lebih efisien pada data tabular. Penggunaan model hibrida ini mengurangi waktu pelatihan (*training time*) secara signifikan tanpa mengorbankan akurasi secara drastis, sehingga lebih ideal untuk aplikasi *real-time* yang dapat diakses oleh masyarakat umum.

KESIMPULAN

Penelitian ini telah berhasil mengimplementasikan metode hibrida RF dan XGBoost untuk sistem peramalan polusi udara berbasis IoT. Berdasarkan hasil pengujian dan analisis yang telah dilakukan, dapat ditarik beberapa kesimpulan utama sebagai berikut:

1. **Efektivitas Model Hibrida:** Penggunaan model Hybrid RF-XGBoost terbukti sangat efektif dalam memprediksi parameter kualitas udara dan cuaca. Model ini menghasilkan nilai koefisien determinasi (R^2) yang konsisten di atas 0,90, yang menunjukkan bahwa model mampu menangkap pola data sensor dengan akurasi yang sangat tinggi.
2. **Peran Seleksi Fitur:** Implementasi *Random Forest* sebagai mekanisme seleksi fitur memberikan nilai tambah dalam hal interpretabilitas model. Melalui skor kepentingan fitur (*feature importance*), variabel suhu dan kelembapan teridentifikasi sebagai faktor yang signifikan dalam mempengaruhi tingkat polusi udara (MQ-135) di lokasi pengujian.
3. **Efisiensi dan Performa:** Dibandingkan dengan metode BiLSTM yang digunakan pada penelitian sebelumnya, model Hybrid RF-XGBoost

menawarkan keunggulan dalam kecepatan komputasi dan penggunaan sumber daya yang lebih ringan tanpa penurunan akurasi yang signifikan. Hal ini menjadikan model hibrida ini lebih ideal untuk diintegrasikan ke dalam perangkat IoT dengan kapasitas pemrosesan terbatas guna pemantauan *real-time*.

4. **Kontribusi Praktis:** Sistem stasiun cuaca terintegrasi ini dapat memberikan informasi dini mengenai tren polusi udara di lingkungan perkotaan, yang bermanfaat bagi masyarakat dan pengambil kebijakan dalam memitigasi dampak buruk kualitas udara terhadap kesehatan.

Untuk penelitian selanjutnya, disarankan untuk melakukan pengujian model pada berbagai lokasi dengan karakteristik polusi yang berbeda serta mempertimbangkan penambahan variabel parameter gas spesifik lainnya guna memperkaya fitur input model.

DAFTAR PUSTAKA

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

Fahim, M., El Mhouti, A., Boudaa, T., & Jakimi, A. (2023). Modeling and implementation of a low-cost IoT-smart weather monitoring station and air quality assessment based on fuzzy inference model and MQTT protocol. *Modeling Earth Systems and Environment*, 9(4), 4085–4102. <https://doi.org/10.1007/s40808-023-01701-w>

Kim, B., Kim, E., Jung, S., Kim, M., J., & Kim, S. (2023). PM2.5 concentration forecasting using weighted Bi-LSTM and Random Forest feature importance-based feature selection. *Atmosphere*, 14(6). <https://doi.org/10.3390/atmos14060968>

Liu, X., et al. (2023). Prediction of lost circulation in Southwest Chinese oil fields applying improved WOA-BiLSTM. *Processes*, 11(9). <https://doi.org/10.3390/pr11092763>

Setyorini, Z. A. M., Nuha, H. H., Karimah, S. A., & Sailallah, H. R. P. (2026). Air pollution forecast using integrated weather stations and bidirectional long short-term memory method (BiLSTM). *School of Computing, Telkom University*.

Susanto, F., Prasiani, N. K., & Darmawan, P. (2022). Implementasi internet of things dalam kehidupan sehari-hari. *Jurnal Imagine*, 2(1), 35–40. <https://doi.org/10.35886/imagine.v2i1.329>

Wang, J., Li, J., & Yu, C. (2021). Air quality prediction based on random forest and XGBoost hybrid model. *Journal of Environmental Management*, 281.