(2026), 4 (2): 283-292 **Scientica** 3021-8209

Jurnal Ilmiah Sain dan Teknologi

PERBANDINGAN MODEL *XGBOOST, LSTM,* DAN *ARIMAX* UNTUK MEMPREDIKSI PM2.5 DI JAKARTA

Dinah Ratulugina [1], Munawar [2]

[12]Jurusan Teknik Informatika, Fakultas Ilmu Komputer, Universitas Esa Unggul

Email : lugina10@gmail.com

Abstract

The increase in air pollution, especially PM2.5, has a serious impact on health and air quality, so accurate predictions are needed to support mitigation policies. This study aims to predict PM2.5 concentrations in Jakarta using three model approaches, namely XGBoost, LSTM, and ARIMAX. The results showed that ARIMAX provided the best performance (MAE 12.37, RMSE 17.60, and R^2 0.66), compared to XGBoost (MAE 12.89, RMSE 20.45, R^2 0.54), and LSTM (MAE 15.13, RMSE 22.32, R^2 0.46). Analysis of feature importance, permutation feature importance, and regression coefficients indicated that the variables PM10, PM2.5_lag, rainfall, average temperature, wind direction, and wind speed were the dominant factors influencing PM2.5 concentrations with a significance p-value <0.05. Meanwhile, the variables relative humidity, ozone, nitrogen dioxide, sunshine duration, and maximum temperature tended to have a lower or statistically insignificant influence. Time series models with exogenous variables, particularly ARIMAX, proved most effective in capturing the dynamics of PM2.5 concentrations and identifying the main meteorological factors and pollutants contributing to air pollution in Jakarta.

Article History

Submitted: 18 April 2026

Accepted: 21 April 2026

Published: 22 April 2026

Key Words

PM2.5, Air Quality Prediction, Air Pollution, Climate Factors, Jakarta

Abstrak

Peningkatan polusi udara, khususnya PM2.5 berdampak serius pada kesehatan dan kualitas udara sehingga prediksi akurat dibutuhkan untuk mendukung kebijakan mitigasi. Penelitian ini bertujuan untuk memprediksi konsentrasi PM2.5 di Jakarta menggunakan tiga pendekatan model yaitu *XGBoost*, *LSTM*, dan *ARIMAX*. Hasil penelitian menunjukkan bahwa *ARIMAX* memberikan performa terbaik (*MAE* 12.37, *RMSE* 17.60, dan R^2 0.66), dibandingkan *XGBoost* (*MAE* 12.89, *RMSE* 20.45, R^2 0.54), dan *LSTM* (*MAE* 15.13, *RMSE* 22.32, R^2 0.46). Analisis *feature importance, permutation feature importance,* dan koefisien regresi mengindikasikan bahwa variabel PM10, PM2.5_*lag*, curah hujan, suhu rata-rata, arah angin, dan kecepatan angin merupakan faktor dominan yang memengaruhi konsentrasi PM2.5 dengan signifikansi *p-value* < 0.05. Sementara itu, variabel kelembapan relatif, ozon, nitrogen dioksida, lama penyinaran matahari, serta suhu maksimum cenderung memiliki pengaruh yang lebih rendah atau tidak signifikan secara statistik. Model *time series* dengan variabel eksogen, khususnya *ARIMAX*, terbukti paling efektif menangkap dinamika konsentrasi PM2.5 dan mengidentifikasi faktor meteorologi serta polutan utama penyumbang polusi udara di Jakarta.

Sejarah Artikel

Submitted: 18 April 2026

Accepted: 21 April 2026

Published: 22 April 2026

Kata Kunci

PM2.5, Prediksi Kualitas Udara, Polusi Udara, Faktor Iklim, Jakarta

1. PENDAHULUAN

Salah satu komponen utama dari polusi udara yang paling berdampak terhadap kesehatan manusia adalah partikel halus berukuran ≤2.5 mikrometer, yang dikenal sebagai PM2.5 (Bhattarai et al., 2024). Partikel halus ini membawa senyawa berbahaya yang dapat menembus jauh ke dalam sistem paru-paru dan aliran darah, memicu penyakit kardiovaskular, pernapasan, kanker paru, stroke, dan kelahiran prematur (Gao et al., 2023). Menurut WHO (2021), sekitar 90% populasi dunia pada tahun 2019 terpapar udara dengan konsentrasi PM2.5 melebihi ambang batas aman, menyebabkan sekitar 7 juta kematian dini setiap tahun.

(2026), 4 (2): 283-292 **Scientica** 3021-8209
Jurnal Ilmiah Sain dan Teknologi

Gambar 1 Tren Konsentrasi PM2.5 di Jakarta (Centre for Research on Energy and Clean Air (CREA), 2024)

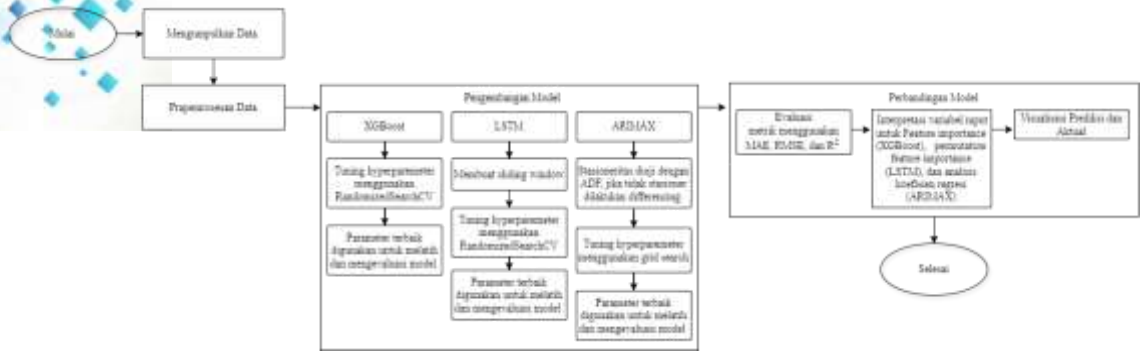
Gambar 1 menunjukkan tren konsentrasi PM2.5 di Jakarta selama lima tahun terakhir (2019–2023). Terlihat adanya peningkatan signifikan setiap tahun pada periode April hingga Agustus, yang bertepatan dengan musim kemarau berdasarkan laporan BMKG, dengan puncaknya umumnya terjadi pada bulan Agustus. Pola ini mengindikasikan adanya pengaruh musiman yang kemungkinan besar dipicu oleh faktor klimatologis seperti suhu, curah hujan, kelembapan, serta arah dan kecepatan angin.Polusi udara di Jakarta terutama berasal dari sektor transportasi, industri, dan aktivitas manusia, dengan kendaraan bermotor menjadi kontributor utama PM2.5 (Toharudin et al., 2023).

Faktor iklim seperti suhu, kelembapan, curah hujan, dan kecepatan angin turut memengaruhi konsentrasi PM2.5. Suhu tinggi mempercepat pembentukan partikel halus, sementara hujan dan angin membantu menurunkannya melalui *deposisi* dan *dispersi*. Kondisi stagnan, seperti angin lemah dan kekeringan, mendukung akumulasi polutan, termasuk dari kebakaran hutan. Perubahan cuaca dapat menjelaskan hingga 50% variasi harian PM2.5 (Bhattarai et al., 2024). Studi juga menunjukkan PM2.5 sangat berkorelasi dengan PM10 dan *CO*, serta cukup signifikan dengan NO_2, namun lemah dengan SO_2 dan O_3. Pemahaman hubungan ini penting untuk merancang sistem prediksi kualitas udara dan peringatan dini yang lebih efektif (Ismail et al., 2024).

Berbagai pendekatan telah digunakan untuk memprediksi konsentrasi PM2.5. Model statistik klasik seperti *ARIMA* dan *ARIMAX* efektif dalam menangkap pola musiman dan tren jangka panjang(Siddique et al., 2025), sementara model multivariat seperti VAR digunakan untuk menganalisis keterkaitan antar variabel secara simultan, meskipun terbatas pada hubungan linier dan memerlukan asumsi stasioneritas(Jayadri et al., 2024). Di sisi lain, model *machine learning* seperti *XGBoost* dan *LSTM* lebih fleksibel dalam menangani hubungan *non-linear* tanpa asumsi distribusi data, namun bersifat *black-box* dan membutuhkan penyetelan parameter yang cermat(Li et al., 2022).

Penelitian ini bertujuan untuk menganalisis dan memprediksi pengaruh faktor-faktor iklim serta polutan lain terhadap konsentrasi PM2.5 di Jakarta. Fokus utama terletak pada pemahaman hubungan antara variabel iklim seperti suhu, kelembapan, curah hujan, penyinaran matahari, dan arah angin, serta polutan seperti PM10, *CO*, O_3, NO_2, dan SO_2 terhadap peningkatan atau penurunan kadar PM2.5. Hasil penelitian ini diharapkan dapat menjadi dasar penyusunan kebijakan pengendalian PM2.5 yang lebih tepat sasaran dan mendukung upaya perlindungan kesehatan masyarakat.

2. METODE PENELITIAN



Gambar 2 Metodologi Penelitian

2.1 Pengumpulan Data

Sumber data polusi udara yang digunakan dalam penelitian ini diperoleh dari situs web Kaggle *https://www.kaggle.com/datasets/senadu34/air-quality-index-in-jakarta-2010-2021* (Taufiq Pohan, 2023). Dataset yang digunakan berisi data polusi udara Kota Jakarta, yang dipantau mulai dari 1 Januari 2019 hingga 30 November 2023. Dataset ini mencakup variabel-variabel seperti stasiun pemantauan, Partikulat Matter 10 (PM10), Partikulat Matter 2.5 (PM2.5), Sulfur Dioksida (SO_2), Karbon Monoksida (CO), Ozon (O_3), dan Nitrogen Dioksida (NO_2). Data meteorologi juga dimanfaatkan untuk menganalisis tingkat kualitas udara. Dataset diperoleh melalui situs resmi *https://dataonline.bmkg.go.id/dataonline-home* (Direktorat Data dan Komputasi - BMKG, 2024), yang menyediakan data historis serta prakiraan cuaca. Variabel meteorologi yang digunakan meliputi temperatur minimum (Tn), temperatur maksimum (Tx), temperatur rata-rata (Tavg), kelembapan rata-rata (RH_avg), curah hujan (RR), durasi penyinaran matahari (ss), kecepatan angin maksimum (ff_x), arah angin saat kecepatan maksimum (ddd_x), kecepatan angin rata-rata (ff_avg), dan arah angin dominan (ddd_car).

2.2 Prapemrosesan Data

Prapemrosesan data dilakukan untuk memastikan kualitas dan konsistensi dataset sebelum proses pelatihan model. Langkah awal penghapusan data duplikat serta klasifikasi kolom berdasarkan tipe data. Kolom bertipe tanggal dikonversi ke *datetime64*, nilai numerik ke *float64*, sedangkan variabel kategorikal ke *category*. Penanganan nilai hilang dilakukan melalui imputasi menggunakan nilai rata-rata, sedangkan *outlier* diidentifikasi dan ditangani dengan metode *Interquartile Range (IQR)*, lalu digantikan dengan nilai median. Transformasi juga diterapkan pada variabel arah angin yang dikonversi ke bentuk derajat, kemudian di *cyclic encoding*, menghasilkan dua fitur baru yaitu *wind_direction_sin* dan *wind_direction_cos*. Untuk mempertimbangkan autokorelasi dalam data deret waktu, ditambahkan fitur *lag* berupa nilai PM2.5 pada hari sebelumnya. Fitur yang digunakan meliputi variabel iklim dan fitur musiman yang telah dijelaskan sebelumnya, yaitu PM10, CO, O_3, NO_2, Tn, Tx, Tavg, RH_avg, RR, ss, ddd_x, ff_avg, *wind_direction_sin*, dan PM2.5_*lag*. Sementara itu, variabel SO_2, ff_x, dan *wind_direction_cos* tidak disertakan karena memiliki nilai korelasi yang rendah atau mendekati nol terhadap PM2.5. Kemudian dataset dibagi menjadi tiga bagian, data pelatihan (70%), validasi (10%), dan pengujian (20%). Terakhir *normalisasi* menggunakan *MinMaxScaler*, untuk memastikan data berada dalam rentang yang optimal bagi model *LSTM*.

2.3 Pengembangan Model

2.3.1 Model *XGBoost*

Extreme Gradient Boosting (*XGBoost*) adalah model *non-linier* yang mampu menangkap hubungan kompleks antara fitur, tidak seperti model *linier* yang mengasumsikan *linearitas* antara fitur input dan variabel target (Mandvi et al., 2024). Untuk meningkatkan performa, penyetelan hiperparameter dilakukan menggunakan metode *RandomizedSearchCV* dengan ruang pencarian meliputi *max_depth, learning_rate, n_estimators, subsample, colsample_bytree, gamma, min_child_weight, reg_alpha, dan reg_lambda.* Persamaan dalam *XGBoost* ditunjukkan pada Pers. (1).

$$Obj(\Theta) = \sum_{i=1}^{n} L\,(y_i,\hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (1)$$

Struktur pohon dan bobot pada setiap daun ditentukan melalui optimasi fungsi objektif *Obj(Θ)* yang mencakup fungsi *error L(Θ)* untuk mengukur selisih prediksi dan data aktual, serta fungsi regulasi *Ω(Θ)* untuk mengendalikan kompleksitas model. Selama pelatihan, pohon dibangun hingga mencapai jumlah tertentu, kemudian diperbarui berdasarkan nilai optimal fungsi objektif. Prediksi akhir kualitas udara diperoleh dari penjumlahan nilai prediksi seluruh subpohon.

2.3.2 Model *LSTM*

Long Short-Term Memory (*LSTM*) merupakan salah satu varian *Recurrent Neural Networks (RNN)* yang mampu mempertahankan informasi jangka panjang melalui mekanisme sel memori (Drewil & Al-Bahadili, 2022). Arsitektur jaringan saraf dibangun menggunakan model *Sequential* yang terdiri atas lapisan input, *LSTM*, *dense*, dan *dropout*. Proses *tuning hyperparameter* dilakukan menggunakan metode *RandomizedSearchCV* dengan ruang pencarian mencakup jumlah *unit*, tingkat *dropout*, *learning rate*, *batch size*, dan jumlah *epoch*. Proses optimasi model menggunakan algoritma *Adam optimizer*, dan diterapkan juga *EarlyStopping* untuk menghentikan proses pelatihan secara otomatis. Persamaan yang digunakan dalam algoritma *LSTM* ditunjukkan pada Pers. (2) - (7).

$$f_t = \sigma\big(w_f * [h_t - 1, x_t] + b_f\big) \quad (2)$$

$$i_t = \sigma(w_i[h_t - 1, x_t] + b_i) \quad (3)$$

$$\tilde{C_t} = \tanh\,(w_c * [h_t - 1, x_t] + b_c) \quad (4)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C_t} \quad (5)$$

$$O_t = \sigma(w_0 * [h_{t-1}, x_1] + b_o) \quad (6)$$

$$h_t = O_t * \tanh\,(C_t) \quad (7)$$

Di mana *ft* adalah *forget gate*, *σ* merupakan fungsi *sigmoid*, *wf* adalah bobot, h_{t-1} adalah output dari blok sebelumnya, x_t adalah vektor input, dan *bf* adalah *bias*. Simbol ∗ menunjukkan perkalian elemen demi elemen, sementara *Ct* merepresentasikan *cell state*, h_t adalah *hidden state*, dan *Ot* adalah *output gate* (Drewil & Al-Bahadili, 2022).

2.3.3 Model *ARIMAX*

Model *ARIMAX* merupakan pengembangan dari *ARIMA* yang tidak hanya menganalisis data historis dan pola musiman, tetapi juga menggabungkan variabel eksogen untuk meningkatkan akurasi prediksi. *ARIMA* efektif menangkap dinamika jangka pendek melalui tren dan musiman, sementara *ARIMAX* memperluas kemampuan ini dengan memasukkan faktor luar (Siddique et al., 2025, Wang et al., 2023). Stasioneritas pada deret waktu berarti

data memiliki rata-rata, varians, serta autokorelasi yang konstan sepanjang waktu. Untuk menguji stasioneritas, penelitian ini menggunakan *Augmented Dickey-Fuller Test (ADF Test)* (Gopu et al., 2021). Persamaan yang digunakan dalam algoritma *ARIMAX* ditunjukkan pada Pers. (8).

$$Y_t = \emptyset_0 + \sum_{i=1}^{p} \emptyset_i Y_{t-i} + \sum_{j=1}^{q} \beta_j X_{t-j} + \sum_{k=1}^{r} \theta_k \in_{t-k} + \in_t \quad (8)$$

Variabel dependen pada waktu ke-*t* dilambangkan sebagai Y_t, dengan konstanta $\emptyset_0$. Komponen *AR* ($\emptyset_i$) mempresentasikan pengaruh nilai masa lalu Y_t. Sedangkan variabel eksogen X_{t-j}dikaitkan dengan koefisien β_j. Komponen *MA* (θ_k) menunjukkan pengaruh *error* sebelumnya ($\in_t$) terhadap nilai sekarang. Parameter *p, q, dan r* masing-masing menunjukkan jumlah *lag AR, differencing*, dan *MA*, sementara *r* untuk eksogen. *Tuning* parameter dilakukan secara manual dengan *grid search* untuk menentukan kombinasi (*p, d, q*) terbaik berdasarkan nilai *Mean Squared Error (MSE)* terendah. Rentang nilai *p* dan *q* ditetapkan pada 0 hingga 3, sedangkan nilai *d* ditentukan melalui uji stasioneritas. Jika data sudah stasioner maka *d* = 0, namun jika tidak, *d* disesuaikan sesuai tingkat *differencing* yang diperlukan.

2.4 Perbandingan Model

Evaluasi metrik digunakan untuk menilai efisiensi dan keandalan kinerja setiap model. Metrik yang digunakan dalam penelitian ini antara lain *Mean Absolute Error (MAE), Root Mean Squared Error (RMSE),* dan *Koefisien Determinasi* (R^2)

$$MAE = \frac{1}{n}\sum_{i\equiv 1}^{n} |y_i - \hat{y}_i| \quad (9)$$

Di mana *y* merupakan nilai aktual, *y* ̂adalah nilai yang diprediksi, dan *n* menunjukkan jumlah keseluruhan observasi (Mandvi et al., 2024)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \quad (10)$$

Di mana Dalam konteks evaluasi model, *yi* merepresentasikan nilai aktual dari observasi ke-*i*, sementara *y î* menunjukkan nilai prediksi dari observasi ke- *i*. Selain itu, *n* mengacu pada jumlah total titik data yang dianalisis (Rad et al., 2024).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad (11)$$

Nilai R^2 berkisar dari 0 hingga 1, di mana $R^2 = 0$ menunjukkan bahwa model tidak dapat menjelaskan variansi dalam data, sedangkan $R^2 = 1$ menunjukkan bahwa model sangat sesuai dengan data dan mampu menjelaskan variabilitas sepenuhnya (Rad et al., 2024).

Interpretasi variabel input juga dilakukan untuk menilai kemampuan masing-masing model dalam mengidentifikasi faktor iklim dan polutan udara yang paling berpengaruh

terhadap konsentrasi PM2.5. Pada *XGBoost* digunakan *feature importance*, *LSTM* digunakan *permutation feature importance*, sedangkan *ARIMAX* digunakan analisis koefisien regresi.

1) Visualisasi perbandingan antara nilai aktual dan nilai prediksi turut digunakan untuk menilai sejauh mana model mampu menangkap pola temporal dalam data.

3. Hasil dan pembahasan

3.1 Model *XGBoost*

Proses pelatihan model *XGBoost* dilakukan dengan *tuning hyperparameter* menggunakan Teknik *RandomizedSearchCV*. Berdasarkan hasil pencarian, diperoleh kombinasi *hyperparameter* terbaik yaitu: *subsample* = 0.6, *max_depth* = 3, *learning_rate* = 0.05, *n_estimators* = 100, *colsample_bytree* = 0.8, *min_child_weight* = 3, *gamma* = 5, *reg_alpha* = 0, dan *reg_lambda* = 1. Kombinasi ini menunjukkan bahwa model bekerja optimal dengan kedalaman pohon yang relatif rendah (*max_depth* = 3) dan *learning_rate* kecil (0.05), sehingga mencegah *overfitting* sekaligus menjaga stabilitas proses pembelajaran. Dengan konfigurasi tersebut, model menghasilkan performa yang cukup baik, ditunjukkan oleh nilai evaluasi metrik *MAE* = 12.89, *RMSE* = 20.45, dan R^2 = 0.54. Nilai R^2 di atas 0.5 menunjukkan bahwa model mampu menjelaskan sekitar 54% variasi konsentrasi PM2.5, meskipun masih terdapat faktor lain di luar model yang memengaruhi variasi data.

3.2 Model *LSTM*

Proses pelatihan model *LSTM* dilakukan dengan *tuning hyperparameter* menggunakan teknik *RandomizedSearchCV*. Berdasarkan hasil pencarian, diperoleh kombinasi *hyperparameter* terbaik, yaitu jumlah unit *LSTM* (*model_units*) = 128, *learning rate* = 0.0005, *dropout rate* = 0.2, jumlah *epochs* = 50, dan *batch size* = 32. Dengan konfigurasi tersebut, model *LSTM* menghasilkan nilai evaluasi *MAE* sebesar 15.13, *RMSE* sebesar 22.32, dan R^2 sebesar 0.46.

3.3 Model *ARIMAX*

Langkah pertama dilakukan pengujian stasioneritas pada data target menggunakan uji *Augmented Dickey-Fuller (ADF)*. Hasil pengujian menunjukkan bahwa data target sudah stasioner sehingga nilai *d* ditetapkan sebesar 0. Selanjutnya, rentang nilai *p* dan *q* ditentukan pada 0, 1, 2, dan 3. Proses pemilihan parameter *ARIMAX* dilakukan melalui *grid search* menggunakan data pelatihan dan validasi. Dari hasil percobaan, kombinasi parameter terbaik diperoleh pada orde (0,0,0) dengan nilai *Mean Squared Error (MSE)* sebesar 341,38.

3.4 Perbandingan Model

Table 1 Perbandingan evaluasi model

Metrik	*XGBoost*	*LSTM*	*ARIMAX*
MAE	12.89	15.13	12.37
RMSE	20.45	22.32	17.60
R^2	0.54	0.46	0.66

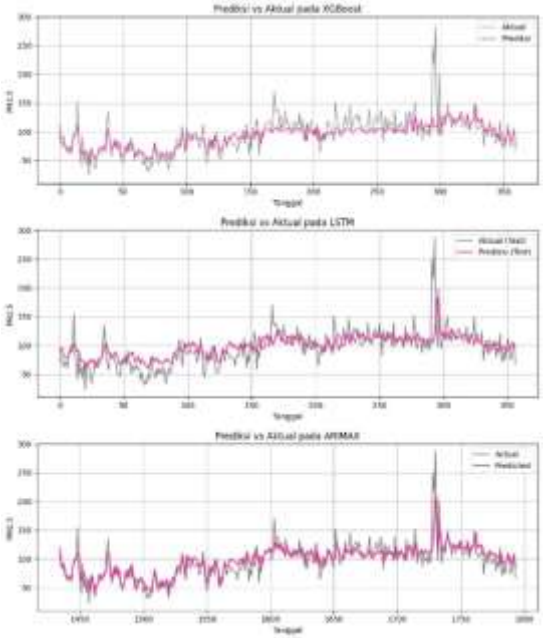
Dapat dilihat bahwa model *ARIMAX* menunjukkan performa prediksi paling baik dalam memodelkan konsentrasi PM2.5 berdasarkan variabel iklim dan polutan lain. Model *ARIMAX* mencatat nilai *MAE* sebesar 12.37 dan *RMSE* sebesar 17.60, yang berarti kesalahan rata-rata absolut dan kuadrat akar rata-rata kesalahan prediksi *ARIMAX* lebih kecil dibandingkan *XGBoost* dan *LSTM*. Selain itu, nilai R^2 sebesar 0.66 menunjukkan bahwa sekitar 66% variabilitas data PM2.5 dapat dijelaskan oleh model *ARIMAX*, menjadikannya model dengan kemampuan prediktif tertinggi di antara ketiganya. Model *XGBoost* menempati posisi tengah, dengan *MAE* sebesar 12.89, *RMSE* sebesar 20.45, dan R^2 sebesar 0.54, menunjukkan performa yang masih lebih baik dari *LSTM* namun belum melampaui *ARIMAX*. Sedangkan model *LSTM* memberikan hasil yang paling rendah, dengan *MAE* sebesar 15.13, *RMSE* sebesar 22.32, dan

R^2 sebesar 0.46, menandakan bahwa *LSTM* hanya mampu menjelaskan sekitar 46% variabilitas data.

Table 2 Perbandingan interpretasi variabel input

Fitur	***Feature Importance* (*XGBoost*)**	***Permutation Feature Importance* (*LSTM*)**	**Regresi (*ARIMAX*)**	
			Koefisien	***p-value***
PM10	12831.59	7.33	0.56	0.000
PM2.5_*lag*	11615.84	215.00	0.36	0.000
ddd_x	9905.31	26.38	-0.02	0.000
wind_direction_sin	4609.56	70.66	2.98	0.008
RR	4227.93	-2.02	-0.18	0.000
RH_avg	4174.81	12.07	0.03	0.811
O_3	3896.02	-0.46	-0.01	0.301
Tavg	3542.52	-4.77	4.36	0.000
NO_2	3272.38	-0.45	0.04	0.207
Tn	2799.55	4.39	-2.00	0.011
ss	2477.60	6.68	-0.16	0.453
CO	2424.89	-3.74	0.15	0.028
Tx	2364.04	14.01	-1.24	0.104
ff_avg	2236.97	48.01	-2.67	0.001

Hasil analisis perbandingan antara *XGBoost*, *LSTM*, dan *ARIMAX* menunjukkan bahwa ketiga model memiliki kecenderungan yang konsisten dalam menempatkan PM2.5_*lag* dan PM10 sebagai variabel paling berpengaruh. Pada *XGBoost*, kedua variabel tersebut memperoleh skor gain tertinggi (>11.000), sementara pada *LSTM*, PM2.5_*lag* tercatat menghasilkan kenaikan *error* terbesar (ΔMSE = 215), diikuti PM10 (ΔMSE = 7.33). Hal ini diperkuat oleh *ARIMAX*, di mana kedua variabel memiliki koefisien signifikan ($p < 0.01$). Informasi historis polutan terbukti menjadi faktor paling dominan dalam memprediksi konsentrasi PM2.5. Selain polutan historis, faktor meteorologi juga memberikan kontribusi penting. Pada *LSTM*, *wind_direction_sin* (ΔMSE = 70.66), ddd_x (ΔMSE = 26.39), serta ff_avg (ΔMSE = 48.01) menunjukkan pengaruh yang besar terhadap performa prediksi, meskipun pada *XGBoost* nilainya relatif moderat. *ARIMAX* juga mengonfirmasi signifikansi variabel ini ($p < 0.05$). Hal ini menegaskan bahwa arah dan kecepatan angin memainkan peran penting dalam dispersi polutan, sesuai dengan teori *dispersi* atmosfer. Sebaliknya, variabel Tavg memperlihatkan hasil yang kontradiktif antar model. Pada *ARIMAX*, Tavg berpengaruh positif dan signifikan (coef = 4.361, $p < 0.01$), menunjukkan bahwa peningkatan suhu rata-rata cenderung meningkatkan konsentrasi PM2.5.



Gambar 3 Perbandingan grafik prediksi vs aktual

Berdasarkan grafik perbandingan prediksi dengan data aktual, ketiga model *XGBoost*, *LSTM*, dan *ARIMAX* sama-sama mampu mengikuti tren umum konsentrasi PM2.5. Namun, terdapat perbedaan pada tingkat kedekatan prediksi terhadap nilai aktual. Model *ARIMAX* menunjukkan pola prediksi yang paling rapat menempel pada data aktual, termasuk pada fluktuasi sedang hingga tinggi, meskipun sesekali terdapat keterlambatan respons pada lonjakan ekstrem. Model *XGBoost* mampu menangkap tren dasar dengan cukup baik, tetapi cenderung menghasilkan prediksi yang lebih halus dan kurang akurat pada titik ekstrem. Sementara itu, model *LSTM* relatif stabil dalam memprediksi pola temporal, namun seringkali meredam lonjakan sehingga nilai prediksinya lebih rendah dibandingkan aktual.

4. KESIMPULAN

Penelitian ini bertujuan untuk memprediksi konsentrasi PM2.5 di Jakarta dengan menganalisis pengaruh faktor iklim dan polutan lain menggunakan tiga pendekatan model, yaitu *XGBoost*, *LSTM*, dan *ARIMAX*. Hasil evaluasi menunjukkan bahwa model *ARIMAX* memberikan performa terbaik dibandingkan dua model lainnya, dengan nilai *MAE* sebesar 12.37, *RMSE* sebesar 17.60, serta R^2 sebesar 0.66. Nilai koefisien determinasi tersebut mengindikasikan bahwa model *ARIMAX* mampu menjelaskan sekitar 66% variasi konsentrasi PM2.5 berdasarkan variabel prediktor yang digunakan. Sementara itu, model *XGBoost* menghasilkan nilai R^2 sebesar 0.54, yang berarti cukup baik dalam memprediksi, namun tidak seakurat *ARIMAX*. Model *LSTM* memiliki performa paling rendah dengan R^2 sebesar 0.46, yang menunjukkan keterbatasan dalam menangkap pola data polusi udara pada penelitian ini.

Selain itu, analisis variabel input menunjukkan bahwa PM10, suhu rata-rata (Tavg), serta arah angin memiliki pengaruh yang signifikan terhadap konsentrasi PM2.5, sedangkan variabel lain seperti *CO* dan curah hujan berkontribusi lebih rendah. Dengan demikian, faktor meteorologi dan polutan tertentu terbukti berperan penting dalam peningkatan maupun penurunan konsentrasi PM2.5.

Secara keseluruhan, penelitian ini memberikan gambaran bahwa pemilihan model prediksi dan pemahaman terhadap faktor iklim maupun polutan menjadi kunci dalam upaya memonitor serta mengendalikan kualitas udara. Hasil ini dapat menjadi dasar pengambilan keputusan bagi pemerintah daerah dalam menetapkan kebijakan lingkungan yang lebih tepat sasaran, khususnya dalam menghadapi tantangan polusi udara di perkotaan seperti Jakarta.

DAFTAR PUSTAKA

Bhattarai, H., Tai, A. P. K., Val Martin, M., & Yung, D. H. Y. (2024). Responses of fine particulate matter (PM2.5) air quality to future climate, land use, and emission changes: Insights from modeling across shared socioeconomic pathways. *Science of The Total Environment*, *948*, 174611. https://doi.org/10.1016/J.SCITOTENV.2024.174611

Centre for Research on Energy and Clean Air (CREA). (2024). *Indonesia's air quality: Decline in 2023 due to lack of intervention and El Niño. What about 2024?* https://energyandcleanair.org/wp/wp-content/uploads/2024/04/EN-CREA_ID-AQ-decline-in-2023-due-to-lack-of-intervention-and-El-Nino.-What-about-2024.pdf

Chen, H., Guan, M., & Li, H. (2021). Air Quality Prediction Based on Integrated Dual *LSTM* Model. *IEEE Access*, *9*, 93285–93297. https://doi.org/10.1109/ACCESS.2021.3093430

Direktorat Data dan Komputasi - BMKG. (2024). *Data Online*. BMKG. https://dataonline.bmkg.go.id/dataonline-home

Drewil, G. I., & Al-Bahadili, R. J. (2022). Air pollution prediction using *LSTM* deep learning and metaheuristics algorithms. *Measurement: Sensors*, *24*. https://doi.org/10.1016/j.measen.2022.100546

Gao, J., He, S., & Wang, L. (2023). County level study of the interaction effect of PM2.5 and climate sustainability on mortality in China. *Front. Public Health*. https://doi.org/https://doi.org/10.3389/fpubh.2022.1036272

Gopu, P., Panda, R. R., & Nagwani, N. K. (2021). Time Series Analysis Using *ARIMA* Model for Air Pollution Prediction in Hyderabad City of India. In *Advances in Intelligent Systems and Computing* (pp. 47–56). https://doi.org/10.1007/978-981-33-6912-2_5

Ismail, N. A., Abdullah, S., Mansor, A. A., Ahmad, A. N., Ahmed, A. N., & Ismail, M. (2024). Trend and Interrelationship of PM2.5, Gaseous Pollutants and Meteorological Factors in Kuala Terengganu, Malaysia. *International Journal of Design and Nature and Ecodynamics*, *19*(4), 1251–1260. https://doi.org/10.18280/ijdne.190416

Jayadri, B. L., Pangastuti, M., Farhan, M., & Kartiasih, F. (2024). Determinants of PM2.5 Concentration in DKI Jakarta Province: A VAR Model Approach. *Inferensi*, *7*(1), 27. https://doi.org/10.12962/j27213862.v7i1.19843

Li, X., Huo, H., & Liu, Z. (2022). Analysis and prediction of PM2.5 concentration based on *LSTM-XGBoost*-SVR model. *Research Square*. https://doi.org/10.21203/rs.3.rs-2158285/v1

Mandvi, Patel, P. K., & Singh, H. K. (2024). Performance analysis of machine learning models for AQI prediction in Gorakhpur City: a critical study. *Environmental Monitoring and Assessment*, *196*(10). https://doi.org/10.1007/s10661-024-13107-x

Rad, A. K., Razmi, S. O., Nematollahi, M. J., Naghipour, A., Golkar, F., & Mahmoudi, M. (2024). Machine learning models for predicting interactions between air pollutants in Tehran Megacity, Iran. *Alexandria Engineering Journal*, *104*, 464–479. https://doi.org/10.1016/j.aej.2024.08.023

Siddique, M. A. B., Mahalder, B., Haque, M. M., & Ahammad, S. K. S. (2025). Impact of climatic and water quality parameters on Tilapia (Oreochromis niloticus) broodfish growth: Integrating *ARIMA* and *ARIMAX* for precise modeling and forecasting. *PLoS ONE*, *20*(3 March). https://doi.org/10.1371/journal.pone.0313846

Taufiq Pohan. (2023). *Air Quality Index in Jakarta*. Kaggle. https://www.kaggle.com/datasets/senadu34/air-quality-index-in-jakarta-2010-2021

Toharudin, T., Caraka, R. E., Pratiwi, I. R., Kim, Y., Gio, P. U., Sakti, A. D., Noh, M., Nugraha, F. A. L., Pontoh, R. S., Putri, T. H., Azzahra, T. S., Cerelia, J. J., Darmawan, G., & Pardamean, B. (2023). Boosting Algorithm to Handle Unbalanced Classification of PM2.5Concentration Levels by Observing Meteorological Parameters in Jakarta-

Indonesia Using AdaBoost, *XGBoost*, CatBoost, and LightGBM. *IEEE Access*, *11*, 35680–35696. https://doi.org/10.1109/ACCESS.2023.3265019

Wang, Y., Gao, C., Zhao, T., Jiao, H., Liao, Y., Hu, Z., & Wang, L. (2023). A comparative study of three models to analyze the impact of air pollutants on the number of pulmonary tuberculosis cases in Urumqi, Xinjiang. *PLoS ONE*, *18*(1 January). https://doi.org/10.1371/journal.pone.0277314