Terbit *online* pada laman : http://teknosi.fti.unand.ac.id/

Jurnal Nasional Teknologi dan Sistem Informasi

| ISSN (Print) 2460-3465 | ISSN (Online) 2476-8812 |

Artikel Penelitian

Analisis dan Prediksi Indeks Kualitas Udara Jakarta: Penerapan Algoritma XGBoost

Evandha Mustika Sari [a,*], *Cahya Sabila* [b], *Rifqi Fakhrizal Adam* [c], *Robert Kurniawan* [d]

[abc] *Program Studi D-IV Statistika, Politeknik Statistika STIS, Jl. Otto Iskandardinata No.64C, Jakarta 13330, Indonesia*
[d] *Program Studi D-IV Komputasi Statistika, Politeknik Statistika STIS, Jl. Otto Iskandardinata No.64C, Jakarta 13330, Indonesia*

INFORMASI ARTIKEL

Sejarah Artikel:
Diterima Redaksi: 14 Juli 2025
Revisi Akhir: 02 September 2025
Diterbitkan *Online*: 07 September 2025

KATA KUNCI

polusi udara,
XGBoost,
data mining,
prediksi

KORESPONDENSI

E-mail: 212212587@stis.ac.id

A B S T R A C T

Polusi udara merupakan masalah serius yang berdampak pada kesehatan dan kualitas hidup masyarakat di kota metropolitan seperti Jakarta. Untuk mengatasi tantangan ini, diperlukan metode prediksi kualitas udara yang akurat dan andal. *Extreme Gradient Boosting* (XGBoost) adalah algoritma *machine learning* yang unggul dalam menangani data non-linear dan kompleks, sehingga cocok digunakan untuk memodelkan kualitas udara. Penelitian ini bertujuan mengembangkan model prediksi kualitas udara di Jakarta menggunakan XGBoost dengan memanfaatkan data polutan pembangun Indeks Kualitas Udara (AQI) yang diperoleh melalui proses *data mining* menggunakan *Earth Engine Code Editor*. Evaluasi model dilakukan menggunakan metrik RMSE, MAE, R^2, dan RSE menunjukkan bahwa XGBoost memberikan performa prediksi yang sangat baik. Analisis *feature importance* mengidentifikasi bahwa SO_2, $PM_{2.5}$ dan PM_{10} sebagai faktor utama yang mempengaruhi kualitas udara di Jakarta. Hasil penelitian ini diharapkan dapat mendukung pemerintah dalam pengambilan kebijakan mitigasi polusi udara dan pengembangan sistem peringatan dini yang efektif untuk meningkatkan kualitas hidup masyarakat.

1. PENDAHULUAN

Sejak dimulainya revolusi industri, perkembangan pesat dalam bidang teknologi, energi, dan pembangunan sosial telah membawa banyak perubahan ke arah positif bagi kehidupan manusia. Namun, kemajuan ini juga menyebabkan dampak negatif terhadap lingkungan, salah satunya permasalahan polusi udara. Polusi udara merujuk pada perubahan kandungan udara yang disebabkan oleh adanya zat berbahaya, seperti partikel halus (PM), sulfur dioksida (SO_2), karbon monoksida (CO), nitrogen dioksida (NO_2), serta berbagai logam berat [1]

Polusi udara sampai saat ini masih menjadi ancaman serius yang terus berlanjut bagi kesehatan manusia dan keseimbangan ekosistem. Permasalahan lingkungan seperti hujan asam, kabut asap, penipisan lapisan ozon, dan pemanasan global merupakan akibat dari adanya polusi udara. Selain itu, Hoffman dalam [2] menyatakan bahwa paparan jangka panjang terhadap polusi udara dapat menyebabkan gangguan pernapasan, termasuk batuk, sesak dada, kanker paru-paru, serta penyakit pernapasan lainnya.
Indonesia menempati peringkat pertama sebagai negara dengan polusi udara tertinggi di Asia Tenggara. Per 23 Juni 2024, IQAir mencatat Jakarta sebagai kota yang memiliki tingkat polusi udara tertinggi ketiga di dunia [3]. Pemerintah Jakarta telah menerapkan berbagai kebijakan seperti menggalakkan penggunaan transportasi umum, membentuk tim pemantau kualitas udara, dan memberlakukan disinsentif pada tarif parkir, namun tingkat kualitas udara masih terus memburuk [4]. Oleh karena itu, pengembangan model prediksi kualitas udara yang akurat menjadi semakin penting. Dengan menggunakan metode prediksi yang akurat, pembuat kebijakan dapat melakukan penyesuaian dan intervensi yang tepat sasaran dalam tindakan pencegahan dan pengendalian polusi, perencanaan pengembangan kota, serta praktik manajemen perkotaan [5].

Perkembangan teknologi dalam pengumpulan dan pemrosesan data serta adanya integrasi berbagai disiplin ilmu menyebabkan *data mining* dan *machine learning* mulai digunakan untuk

https://doi.org/10.25077/TEKNOSI.v11i2.2025.161-169

analisis lingkungan. Hal ini memungkinkan untuk memperoleh informasi udara yang lebih akurat dan terkini sehingga dapat memberikan dasar yang kuat dalam penentuan kebijakan bagi pemerintah [6]. Adapun metode yang akan digunakan yaitu *Extreme Gradient Boosting Regressor* (XGBoost). Model XGBoost telah terbukti mampu meningkatkan akurasi prediksi indeks kualitas udara secara signifikan pada data spasial dan temporal di beberapa kota besar [7]. Algoritma yang digunakan pada model XGBoost berupa pohon keputusan berulang yang sebelumnya terdiri dari beberapa pohon keputusan [6].

Pada penelitian sebelumnya [8], prediksi AQI Jakarta dengan ARIMA menghasilkan MAPE dengan rentang 8% sampai 42% pada keenam variabel penyusun AQI yaitu yaitu $PM_{2.5}$, PM_{10}, SO_2, CO, NO_2, dan O_3. Kemudian penelitian lainnya [9] yaitu prediksi AQI di Beijing menggunakan metode ARIMA dan *Neural Network Model*, berkesimpulan bahwa metode ARIMA hanya mampu memprediksi data berdasarkan data historis sedangkan neural network kemampuan adaptasi diri yang baik, dan hasil prakiraan relatif baik dibuktikan dengan nilai MAPE metode *neural network* yang lebih kecil daripada metode ARIMA yaitu sebesar 12,81%. Hasil penelitian [10] juga menunjukkan bahwa XGBoost dan *Support Vector Regression* (SVR) memberikan performa pemantauan kualitas udara yang unggul, khususnya dalam menangani data yang bersifat *autocorrelated*.

Namun, meskipun model *machine learning* telah terbukti unggul dalam adaptasi dan akurasi, penerapan khusus dari algoritma XGBoost masih jarang diterapkan di Jakarta. Penerapan model XGBoost penting dilakukan karena memiliki keunggulan dalam menangani data non-linear dan kompleks [11]. Adapun penelitian terkait yang menerapkan XGBoost di Jakarta hanya terbatas pada satu atau beberapa polutan saja. Dengan demikian, dibutuhkan pemanfaatan data multi-polutan terbaru dalam bentuk data harian agar dapat memberikan gambaran kualitas udara secara *real-time.* Dengan demikian, penelitian ini diharapkan memiliki kontribusi signifikan dalam memberikan model prediksi AQI yang akurat dan adaptif terhadap data polutan harian di Jakarta.

Dalam penelitian ini, model dapat dibangun berdasarkan data rata-rata harian dan data konsentrasi enam polutan, yaitu $PM_{2.5}$, PM_{10}, SO_2, CO, NO_2, dan O_3 pada tahun 2024 di Jakarta. Penelitian ini disusun berdasarkan metode yang digunakan hingga evaluasi model yang diperoleh serta kesimpulan. Hasil yang diharapkan dari penelitian ini yaitu mampu memprediksi kualitas udara dengan lebih akurat sehingga dapat membantu pemerintah dan masyarakat dalam mengambil kebijakan dan langkah yang tepat untuk menjaga kualitas udara di Jakarta.

2. METODE

2.1. Pengumpulan Data

Data penelitian menggunakan data geospasial yang diambil melalui proses data mining dengan menggunakan *Earth Engine Code Editor* (https://code.earthengine.google.com/?hl=id, diakses pada 27 April 2025). Data yang dikumpulkan mencakup indikator pembangun indeks kualitas udara, yaitu konsentrasi polutan $PM_{2.5}$, PM_{10}, SO_2, CO, NO_2, dan O_3 periode 1 Januari 2024 hingga 30 Maret 2025 di wilayah Jakarta, Indonesia. Data kualitas udara ini utamanya diambil dari koleksi citra satelit Sentinel-5P TROPOMI. Pemilihan koleksi citra satelit ini berdasarkan hasil penelitian [12] yang menyatakan bahwa konsentrasi polutan yang diukur oleh instrumen TROPOMI pada Sentinel-5P dapat diandalkan untuk memantau polusi. Dataset mencakup pengukuran harian polutan udara dengan resolusi spasial.

2.2. Preprocessing

Preprocessing data merupakan langkah awal dalam membangun model pembelajaran statistik [6]. Pada penelitian ini, preprocessing atau persiapan data dilakukan dengan menggunakan software R. *Preprocessing data* terbagi menjadi tiga tahap utama, yaitu integrasi data *(data integration)*, pembersihan data *(data cleaning)*, dan transformasi data *(data transformation)*. Langkah-langkah ini dilakukan untuk memastikan konsistensi, akurasi, dan kegunaan data harian kualitas udara di Jakarta untuk periode 1 Januari 2024 hingga 30 Maret 2025.

2.2.1. Data Integration

Tahap pertama yang dilakukan dalam preprocessing adalah integrasi data. Dataset yang digunakan dalam penelitian ini awalnya merupakan data masing-masing variabel yang terpisah saat diambil dari *Google Earth Engine*.

Integrasi data dilakukan untuk menggabungkan data dari berbagai sumber dalam katalog *Google Earth Engine*, seperti koleksi Sentinel-5P TROPOMI untuk NO_2, SO_2, CO, dan O_3, serta dataset tambahan seperti CAMS *(Copernicus Atmosphere Monitoring Service)* untuk $PM_{2.5}$ dan PM_{10}. Data-data tersebut diambil secara terpisah sehingga perlu dilakukan penggabungan dengan memastikan kesesuaian tanggal dan waktu pada setiap pengukuran. Proses ini menghasilkan dataset terpadu yang mencakup semua variabel polutan dalam format konsisten, dengan kolom waktu sebagai kunci utama untuk sinkronisasi. Pengintegrasian dilakukan melalui software R kemudian di *export* ke dalam tipe file csv. Berikut tampilan dataset setelah dilakukan pengintegrasian.

Tabel 1. Dataset Penelitian

date	NO2	$PM_{2.5}$	PM_{10}	CO	O3	SO2
1/1/24	4,7	11,5	14,3	0,8	0,02	-48
2/1/24	8,8	13,0	13,0	0,9	0,02	-29
3/1/24	11,2	9,3	11,2	0,9	0,02	-31
4/1/24	16,1	10,0	12,2	1,3	0,02	-42
5/1/24	7,15	9,75	11,80	0,72	0,02	29,50
6/1/24	7,91	9,95	12,10	0,90	0,02	25,00
7/1/24	14,84	10,11	12,33	0,79	0,02	26,94
8/1/24	5,72	10,45	12,81	0,75	0,02	13,46
9/1/24	5,83	9,83	11,94	0,84	0,02	-77,82
10/1/24	7,43	8,98	10,72	1,00	0,02	51,15

2.2.2. Data Cleaning

Setelah data diintegrasikan, selanjutnya dilakukan pembersihan data untuk mengatasi inkonsistensi dan anomali dalam dataset. Aspek-aspek yang diperiksa meliputi format waktu, mengatasi *missing value* jika ada, dan pendeteksian *outlier* (pencilan).

https://doi.org/10.25077/TEKNOSI.v11i2.2025.161-169

Dataset diperiksa untuk mengidentifikasi nilai kosong (NaN), null, atau nilai nol yang tidak valid. Adapun satuan dari setiap variabel diperiksa untuk memastikan kesesuaian data dengan standar yang dibutuhkan.

2.2.3. *Data Transformation*

Apabila data sudah dalam kondisi *clean*, untuk meningkatkan kegunaan dataset dalam analisis dilakukan transformasi data. Salah satu langkah utama dalam penelitian ini adalah konstruksi atribut baru, yaitu AQI. AQI dihitung berdasarkan standar internasional seperti yang ditetapkan oleh USEPA *(United States Environmental Protection Agency)* dengan menggunakan enam polutan pembangun $PM_{2.5}$, PM_{10}, SO_2, CO, NO_2, dan O_3. Indeks individu untuk setiap polutan dihitung berdasarkan konsentrasinya, dan nilai AQI ditentukan sebagai nilai maksimum dari indeks-indeks tersebut. Perhitungan ini dilakukan menggunakan formula seperti berikut [13].

$$AQI = \frac{AQI_u - AQI_l}{X_u - X_l}(X - X_l) + AQI_l \quad (1)$$

keterangan :

AQI_u : Batas atas AQI
AQI_l : Batas bawah AQI
X : Data aktual polutan
X_u : Batas atas konsentrasi polutan
X_l : Batas bawah konsentrasi polutan

2.2.4. *Splitting Data*

Setelah proses persiapan data selesai, dataset dibagi menjadi dua bagian, yaitu 80% sebagai data latih *(training)* dan 20% sebagai data uji *(testing)*. Data yang sudah dibagi menjadi dua diubah dalam format matriks. Adapun proses pelatihan model XGboost dilakukan selama 200 iterasi dengan evaluasi kinerja menggunakan RMSE *(Root Mean Square Error)* pada data latih dan data uji.

2.3. *Metode* XGBoost

Extreme Gradient Boosting (XGBoost) adalah model prediksi yang terintegrasi dengan performa lebih unggul dan stabilitas yang baik karena dapat meminimalkan kesalahan prediksi dan meningkatkan akurasi model [14]. Algoritma dalam pemodelan ini menggunakan gradien negatif dari fungsi kerugian untuk memecahkan nilai minimum.

XGBoost menjadi metode yang sangat cocok untuk prediksi AQI karena mampu menangani hubungan non-linear antara variabel input seperti konsentrasi polutan dengan nilai AQI. Dalam [15] dinyatakan bahwa prediksi AQI mendatang menggunakan metode ini memiliki tingkat keandalan dan akurasi yang tinggi. Algoritma yang digunakan oleh model ini adalah dengan menggabungkan beberapa pohon keputusan *(decision trees)* yang lemah secara berurutan, di mana setiap pohon baru berusaha memperbaiki kesalahan dari pohon sebelumnya untuk meningkatkan akurasi prediksi. Berikut ini algoritma lengkap model XGBoost [6].

1. Input :
   Kita mulai dengan dataset berupa pasangan data input dan output (xt,yt), serta tentukan berapa kali kita ingin melakukan *bootstrap* (misalnya B kali).
2. Output :
   Hasilnya nanti adalah B kali hasil prediksi dari proses *bootstrap* tersebut
3. Bangun model y(t) = f(xt) + εt pada data yang sudah di bersihkan
4. Prediksi dari model data yang sudah dibersihkan
   $\hat{y}_t = \hat{f}(x_t)$ (2)
5. Perhitungan dari residual :
   $\varepsilon t = y_t - \hat{y}_t$ (3)
6. Langkah-langkah *bootstrap* :
   - Atur berapa kali bootstrap yang diinginkan dengan b : 1,2,...,B
   - Sampel baru $\varepsilon_t^{(b)}$ diperoleh dari *resampling* sebelumnya secara acak dengan pengembalian
   - Mendapatkan sampel baru diperoleh dengan $\hat{y}_t^{(b)} = \hat{y}_t + \varepsilon_t$
   - Hasil prediksi *bootstrap* $\hat{y}_t^{(b)}$ diperoleh dengan melatih model XGBoost menggunakan data baru $(x_t, \hat{y}_t^{(b)})$ sehingga menghasilkan model $\hat{y}_t^{(b)} = \hat{f}_b(x_t)$
7. Berdasarkan serangkaian hasil prediksi *bootstrap* $\hat{y}_t^{(b)}$ dihitung deviasi standar prediksi dan interval prediksi 95%.

2.4. *Evaluasi Model*

Model XGBoost merupakan model supervised learning yang menggabungkan beberapa klasifikasi dengan akurasi rendah untuk menghasilkan model dengan akurasi yang lebih tinggi. Model ini bertujuan untuk mengurangi kesalahan prediksi [11]. Algoritma XGBoost memanfaatkan kombinasi beberapa pohon keputusan CART *(Classification and Regression Trees)* dan memiliki kemampuan generalisasi yang kuat. XGBoost bekerja dengan cara memperbaiki kesalahan prediksi secara bertahap dan menjaga model agar tidak terlalu kompleks dengan menambahkan aturan khusus [6].

Dalam penelitian ini, untuk mengevaluasi performa model XGBoost dalam memprediksi AQI di Jakarta, digunakan empat metrik evaluasi, yaitu *Root Mean Square Error* (RMSE), *Mean Absolute Error* (MAE), Koefisien Determinasi (R^2), dan *Relative Squared Error* (RSE). RMSE, MAE, dan RSE mengukur tingkat kesalahan prediksi sedangkan R^2 menunjukkan seberapa baik model dalam menjelaskan variasi data [11]. Persamaan matematis metrik-metrik tersebut sebagai berikut.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \quad (4)$$

$$MAE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|} \quad (5)$$

Di mana $\hat{y}_i$ merupakan nilai AQI yang diprediksi, y_i nilai aktual dari AQI, dan N merupakan jumlah data.

Adapun R^2 memiliki persamaan sebagai berikut.

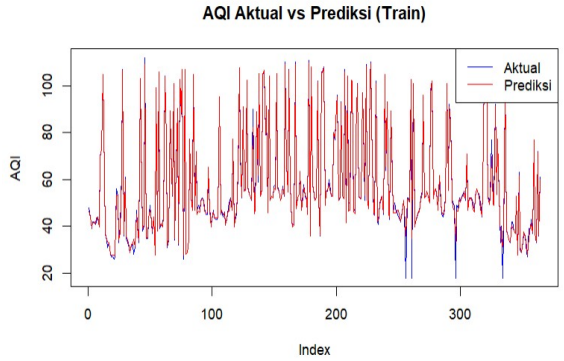
$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \quad (6)$$

Di mana $\bar{y}$ adalah nilai rata-rata dari data AQI aktual dengan persamaan RSE sebagai berikut.

$$RSE = \frac{\sum_{i=1}^{N}(y_i - \hat{y})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \quad (7)$$

3. HASIL

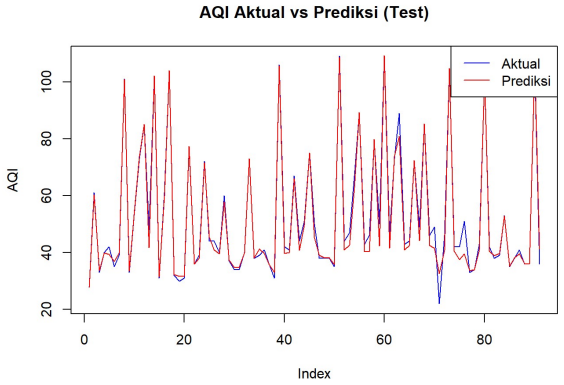
Prediksi Nilai AQI menggunakan algoritma XGBoost dilakukan baik untuk data *training* ataupun data *testing*. Untuk membandingkan hasil prediksi dengan nilai aktual, diagram yang digunakan adalah diagram garis atau *line chart*. Berikut perbandingan AQI aktual dan hasil prediksi.



Gambar 1. Perbandingan AQI Aktual dan Hasil Prediksi pada *training dataset*

Gambar 1 menunjukkan perbandingan antara nilai AQI aktual dengan hasil prediksi model XGBoost pada data pelatihan dalam penelitian ini. Secara umum, model XGBoost mampu mengikuti tren dan pola perubahan nilai AQI aktual dengan cukup baik pada data train. Model ini dapat menangkap pergerakan naik-turun serta fluktuasi yang terjadi pada data AQI, sehingga garis prediksi (merah) tampak sejalan dengan garis aktual (biru) pada sebagian besar rentang waktu.

Model ini juga dapat mengidentifikasi puncak dan lembah utama pada data AQI, meskipun pada beberapa periode terdapat perbedaan antara nilai prediksi dan aktual, terutama pada lonjakan ekstrem. Hal ini menandakan bahwa XGBoost sudah cukup efektif dalam memahami interaksi kompleks antara berbagai faktor yang memengaruhi kualitas udara di Jakarta.



Gambar 2. Perbandingan AQI Aktual dan Hasil Prediksi pada *testing dataset*

Gambar 2 menampilkan perbandingan antara nilai AQI aktual dan hasil prediksi model XGBoost pada dataset pengujian *(testing)*. Dari grafik tersebut, terlihat bahwa model XGBoost mampu mengikuti tren utama pergerakan AQI aktual, yang ditunjukkan oleh garis prediksi (merah) yang cenderung bergerak searah dengan garis aktual (biru).

Model ini cukup baik dalam menangkap pola fluktuasi harian dan perubahan nilai AQI yang terjadi pada data uji, dengan deviasi yang relatif kecil di sebagian besar titik pengamatan. Hal ini mengindikasikan bahwa model XGBoost dapat mempelajari pola historis data AQI dan menerapkannya pada data baru yang sebelumnya belum pernah dilihat, sehingga prediksi yang dihasilkan tetap relevan dan informatif.

Namun, jika diamati lebih detail, terdapat beberapa perbedaan pada titik-titik tertentu, terutama saat terjadi lonjakan ekstrem *(spike)* pada data aktual. Pada beberapa kasus, garis prediksi tampak tidak selalu mengikuti lonjakan tajam yang terjadi pada nilai AQI aktual. Hal ini menunjukkan bahwa model XGBoost cenderung memberikan prediksi yang lebih moderat pada data yang belum pernah dilihat, sehingga tidak seluruh lonjakan ekstrem dapat direpresentasikan dengan akurat.

Meski demikian, secara keseluruhan, model tetap mampu memberikan gambaran tren dan fluktuasi utama yang terjadi pada kualitas udara di Jakarta. Hasil ini memperlihatkan potensi XGBoost sebagai alat bantu prediksi AQI yang andal untuk mendukung pemantauan kualitas udara dan pengambilan keputusan berbasis data di lingkungan perkotaan.

.

Tabel 2. Hasil prediksi nilai AQI pada dataset *testing*

date	Aktual	Prediksi	Standar Deviasi	Prediksi Interval
30/12/24	28	27,80	2,78	[25,55 , 35,41]
31/12/24	61	60,36	2,63	[58,10 , 67,96]
1/1/25	33	33,66	2,95	[28,40 , 41,27]
2/1/25	40	39,80	2,77	[37,47 , 47,40]
3/1/25	42	39,45	2,56	[37,12 , 46,88]
4/1/25	35	36,89	2,79	[34,64 , 44,50]
5/1/25	39	39,86	2,82	[37,60 , 47,46]
6/1/25	101	100,9	2,71	[98,56 , 108,3]
7/1/25	33	33,53	3,05	[28,26 , 41,62]
8/1/25	53	53,33	2,64	[51,08 , 61,42]

Tabel 2 menyajikan hasil prediksi nilai AQI pada dataset *testing* untuk periode 30 Desember 2024 hingga 8 Januari 2025, dengan membandingkan nilai aktual, prediksi, standar deviasi, dan interval prediksi pada masing-masing tanggal. Secara umum, nilai prediksi yang dihasilkan oleh model XGBoost berada sangat dekat dengan nilai AQI aktual, yang terlihat dari selisih yang kecil antara kolom aktual dan prediksi. Selain itu, interval prediksi dengan tingkat kepercayaan 95% pada setiap tanggal juga relatif sempit, misalnya pada 30/12/24 interval prediksi berada pada rentang [25,55 , 31,45], sementara nilai aktual dan prediksi keduanya berada di dalam rentang tersebut. Interval yang sempit ini menunjukkan tingkat kepastian model yang tinggi dalam melakukan prediksi, serta menandakan bahwa model mampu memberikan estimasi yang konsisten dan dapat diandalkan untuk memproyeksikan nilai AQI di masa mendatang.

Selain itu, nilai standar deviasi yang tercantum pada tabel juga menunjukkan konsistensi prediksi model. Hampir seluruh nilai

https://doi.org/10.25077/TEKNOSI.v11i2.2025.161-169

standar deviasi berada di bawah angka 3, yang menandakan bahwa variasi atau penyimpangan hasil prediksi terhadap rata-rata prediksi sangat kecil. Hal ini memperkuat keyakinan bahwa model XGBoost mampu memberikan prediksi yang stabil pada data uji, terutama dalam jangka pendek. Namun, seperti yang dijelaskan dalam tabel, untuk prediksi jangka panjang, model tetap memerlukan pertimbangan faktor-faktor eksternal lain yang dapat memengaruhi kualitas udara, seperti perubahan cuaca, aktivitas industri, dan pola lalu lintas. Dengan demikian, hasil ini menunjukkan bahwa model XGBoost sangat efektif untuk prediksi AQI harian di Jakarta, tetapi untuk proyeksi jangka panjang, integrasi data tambahan akan meningkatkan akurasi dan relevansi prediksi.

Setelah dilakukan pemodelan, langkah selanjutnya adalah mengevaluasi model yang didapatkan dengan beberapa ukuran seperti RMSE, MAE, R-Square, dan RSE. Berikut ini hasil evaluasi model.

Tabel 3. Evaluasi Model

Data	RMSE	MAE	R Square	RSE
Training	2,1108	0,9658	0,9917	2,1108
Testing	2,9449	1,7831	0,9831	2,9949

Tabel 3 menampilkan hasil evaluasi performa model prediksi yang telah dikembangkan untuk data AQI di Jakarta. Penelitian ini menggunakan beberapa metrik statistik yang umum dipakai untuk mengukur akurasi dan keandalan model regresi, yaitu RMSE, MAE, R-Square, dan RSE. Metrik-metrik tersebut memberikan gambaran komprehensif tentang seberapa efektif model dalam menangkap pola hubungan antar variabel serta tingkat ketepatan model dalam memprediksi nilai AQI.

Pada data pelatihan, hasil evaluasi menunjukkan bahwa model memiliki kinerja yang sangat baik. Nilai RMSE sebesar 2,1108 mengindikasikan bahwa rata-rata kesalahan prediksi model terhadap nilai AQI aktual di Jakarta adalah sekitar 2,11 satuan. Nilai MAE sebesar 0,9658 mendukung temuan ini, dengan rata-rata selisih absolut antara prediksi dan nilai aktual tidak lebih dari 1 satuan. Selain itu, nilai R-Square yang mencapai 0,9917 atau 99,17% menunjukkan bahwa hampir seluruh variasi pada data pelatihan dapat dijelaskan oleh model. Nilai RSE sebesar 2,1108 juga menandakan bahwa kesalahan model relatif kecil jika dibandingkan dengan variasi alami data, sehingga dapat disimpulkan bahwa model berhasil mempelajari pola data dengan baik.

Saat diterapkan pada data pengujian, model tetap menunjukkan performa yang memuaskan. Meskipun terdapat sedikit peningkatan nilai kesalahan, secara keseluruhan akurasi model masih tergolong tinggi. Pada data uji, nilai RMSE tercatat sebesar 2,9449, yang berarti rata-rata kesalahan prediksi terhadap nilai AQI aktual sekitar 2,94 satuan. Nilai MAE sebesar 1,7831 menunjukkan bahwa rata-rata selisih absolut antara prediksi dan nilai aktual sedikit lebih besar dibandingkan pada data pelatihan. Namun demikian, nilai R-Square tetap sangat tinggi, yakni 0,9831 atau 98,31%, yang menunjukkan bahwa model masih mampu menjelaskan sebagian besar variasi dalam data pengujian. Nilai RSE pada data uji juga sedikit meningkat menjadi 2,9949, namun masih dalam batas yang dapat diterima untuk prediksi AQI.

Dari hasil tersebut, dapat disimpulkan bahwa model prediksi yang dikembangkan memiliki kemampuan yang sangat baik dalam memahami pola hubungan antar variabel dan menghasilkan prediksi yang akurat, baik pada data pelatihan maupun data pengujian. Tingginya nilai R-Square serta rendahnya nilai RMSE, MAE, dan RSE pada kedua dataset menjadi bukti bahwa model mampu merepresentasikan data dengan baik. Model ini juga terhindar dari *overfitting* karena *overfitting* terjadi jika performa model data pada data uji jauh lebih buruk daripada data latih [16]

4. PEMBAHASAN

4.1. Interpretasi Hasil Evaluasi Model

Model XGBoost yang dikembangkan menunjukkan performa yang baik dalam memprediksi kualitas udara di Jakarta, dengan nilai RMSE, MAE, R-Square, dan RSE seperti yang tertera pada Tabel 3. Nilai RMSE dan MAE yang rendah mengindikasikan bahwa prediksi model cukup akurat dan kesalahan prediksi relatif kecil [17]. Nilai R-Square yang mendekati 1 menunjukkan bahwa model mampu menjelaskan variabilitas data dengan baik [18]. Perbandingan hasil pada data training dan testing juga menunjukkan bahwa model tidak mengalami overfitting yang signifikan, sehingga dapat diandalkan untuk prediksi pada data baru.

Temuan ini sejalan dengan penelitian [19] yang menunjukkan bahwa model ini mampu meningkatkan akurasi prediksi $PM_{2.5}$ secara signifikan dengan koefisien determinasi (R-Square) yang tinggi dan kesalahan prediksi yang rendah. Model berhasil mengatasi tantangan prediksi nilai ekstrem $PM_{2.5}$ dengan stabilitas yang lebih baik dibandingkan metode konvensional. Selain itu, model ini juga menangkap efek musiman dan perbedaan spasial polusi udara di berbagai distrik, dengan konsentrasi PM2.5 yang lebih tinggi di pusat kota dan daerah industri. Secara keseluruhan, penelitian ini menegaskan bahwa penggunaan XGBoost dapat menjadi metode yang sangat efektif untuk prediksi kualitas udara di wilayah perkotaan besar dan sebagai alat yang kuat untuk manajemen polusi dan peringatan dini.

Sebagai tambahan, penelitian [20] menunjukkan bahwa metode XGBoost memiliki performa prediksi yang sangat baik dalam memodelkan konsentrasi $PM_{2.5}$. Hasil penelitian ini mengungkapkan bahwa XGBoost mampu meningkatkan akurasi prediksi secara signifikan dibandingkan dengan model prediksi tradisional, terutama dalam menangani fluktuasi nilai $PM_{2.5}$ yang ekstrem. Model ini berhasil mengurangi kesalahan prediksi dan memberikan hasil yang stabil untuk prakiraan kualitas udara harian di Shanghai. Selain itu, model ini juga efektif dalam menangkap pola musiman dan spasial polusi udara, sehingga sangat berguna untuk peringatan dini dan pengelolaan kualitas udara di kota besar.

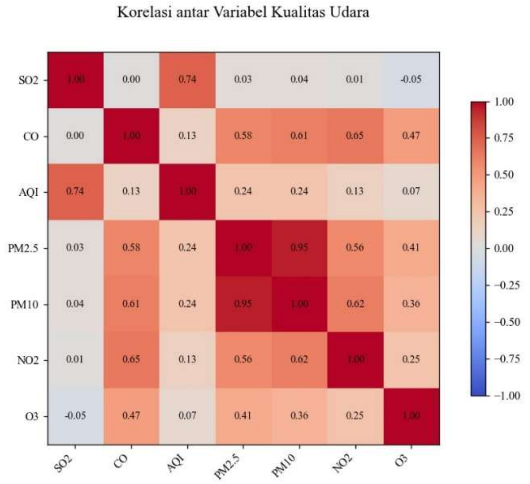
Selain itu, penelitian [6] juga menyatakan bahwa *Bootstrap*-XGBoost dapat digunakan untuk menghasilkan prediksi AQI

beserta interval prediksi 95% yang sempit, yang menandakan tingkat kepastian model yang tinggi.

Ketiga penelitian ini menegaskan bahwa XGBoost adalah metode *machine learning* yang sangat efektif untuk prediksi kualitas udara, dengan kemampuan menangani data yang kompleks dan tidak seimbang, serta memberikan hasil prediksi yang akurat dan stabil secara spasial dan temporal.

4.2. Analisis Feature Importance

Untuk mengetahui polutan mana yang paling memengaruhi nilai AQI, analisis dapat dilakukan dengan melihat korelasi antar variabel atau dengan menganalisis *feature importance* pada model XGBoost yang didapatkan. Gambar 3 ini *heatmap* yang menunjukkan korelasi antar variabel.



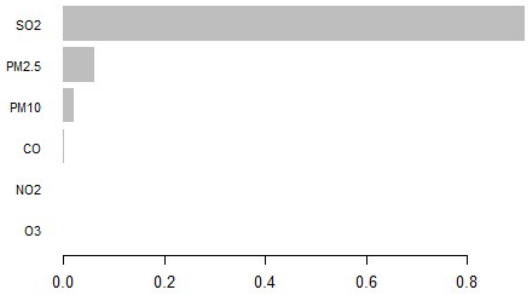
Gambar 3. *Heatmap* Korelasi antar Variabel

Gambar 3 menunjukkan hasil visualisasi heatmap yang menggambarkan korelasi antar variabel kualitas udara di Jakarta. Heatmap ini menggunakan skala warna untuk menunjukkan seberapa kuat hubungan linear antara pasangan variabel, di mana warna yang semakin merah gelap menunjukkan korelasi yang semakin kuat, baik positif maupun negatif, sedangkan warna biru atau merah muda menunjukkan korelasi yang lemah atau tidak signifikan.

Berdasarkan hasil tersebut, terlihat bahwa variabel $PM_{2.5}$ dan PM_{10} memiliki korelasi yang sangat kuat, dengan nilai sebesar 0,95. Angka ini menunjukkan bahwa saat konsentrasi $PM_{2.5}$ meningkat, konsentrasi PM_{10} juga cenderung meningkat, atau sebaliknya. Kuatnya hubungan ini dapat dijelaskan secara ilmiah karena kedua variabel tersebut sama-sama merupakan partikel materi di udara *(particulate matter)* dengan ukuran berbeda, sehingga pergerakan atau peningkatan salah satunya biasanya diikuti oleh yang lain, terutama di wilayah perkotaan dengan aktivitas transportasi dan industri yang tinggi.

Selain itu, variabel SO_2 juga menunjukkan korelasi yang cukup kuat dengan nilai 0,74 terhadap indeks kualitas udara (AQI). Korelasi ini menunjukkan bahwa konsentrasi SO_2 memiliki kontribusi yang cukup besar terhadap penurunan atau peningkatan kualitas udara di Jakarta. Hal ini dapat dikaitkan dengan fakta bahwa gas SO_2 berasal dari pembakaran bahan bakar fosil, seperti batu bara dan minyak, yang banyak terjadi di kawasan industri maupun kendaraan bermotor, sehingga peningkatan konsentrasi SO_2 akan berdampak signifikan terhadap penurunan kualitas udara.

Sementara itu, variabel O_3 menunjukkan korelasi yang lemah terhadap AQI, dengan nilai korelasi sebesar 0,07. Hal ini menunjukkan bahwa kadar O_3 tidak memiliki pengaruh yang dominan terhadap fluktuasi nilai AQI di Jakarta dalam periode pengamatan ini. Meski demikian, bukan berarti konsentrasi O_3 dapat diabaikan sepenuhnya, sebab gas ini tetap berperan dalam dinamika kualitas udara, namun dalam penelitian ini kontribusinya terhadap AQI relatif kecil.



Gambar 4. *Feature Importance* Model XGBoost

 https://doi.org/10.25077/TEKNOSI.v11i2.2025.161-169

=

Dari hasil *heatmap* ini dapat disimpulkan bahwa variabel $PM_{2.5}$, PM_{10}, dan SO_2 merupakan variabel-variabel utama yang memiliki pengaruh signifikan terhadap nilai AQI di Jakarta. Artinya, peningkatan konsentrasi ketiga parameter polutan tersebut berpotensi besar mendorong kenaikan nilai AQI, yang mencerminkan memburuknya kualitas udara.

Dengan memahami pola hubungan antar variabel ini, dapat dilakukan langkah-langkah strategis untuk mengendalikan polusi udara. Upaya pengurangan emisi partikel halus ($PM_{2.5}$ dan PM_{10}) serta gas SO_2 melalui kebijakan pengendalian emisi industri, transportasi, dan penggunaan bahan bakar fosil menjadi langkah yang sangat relevan untuk meningkatkan kualitas udara di Jakarta.

Gambar 4 menampilkan visualisasi *feature importance* dari model XGBoost yang digunakan untuk memprediksi kualitas udara. Hasil analisis menunjukkan bahwa variabel SO_2 (sulfur dioksida) merupakan fitur yang memiliki kontribusi paling besar dalam menentukan prediksi model. Hal ini berarti bahwa keberadaan SO_2 di udara menjadi indikator yang sangat signifikan dalam menentukan buruk atau tidaknya kualitas udara. Setelah SO_2, fitur berikutnya yang memiliki pengaruh penting adalah $PM_{2.5}$ dan PM_{10}. Hal ini sejalan dengan penelitian [21] yang menemukan bahwa pada analisis feature importance untuk model XGBoost, polutan $PM_{2.5}$ paling berpengaruh dalam prediksi kualitas udara di wilayah perkotaan.

Model XGBoost sendiri menggunakan algoritma boosting yang kuat dalam menangani data non-linear dan kompleksitas hubungan antar variabel. Ketika model menunjukkan bahwa SO_2, $PM_{2.5}$, dan PM_{10} adalah fitur yang paling berpengaruh, hal ini mencerminkan pentingnya ketiga indikator ini sebagai komponen utama dalam pengawasan dan manajemen kualitas udara. Dengan demikian, hasil ini dapat menjadi landasan ilmiah yang kuat untuk perumusan kebijakan lingkungan, khususnya dalam hal prioritas pengendalian polusi udara.

Misalnya, di kota besar seperti Jakarta yang mengalami polusi udara cukup parah, informasi ini sangat berguna untuk menentukan fokus intervensi. Pemerintah dan pemangku kepentingan dapat memprioritaskan upaya pengurangan emisi dari sektor-sektor yang diketahui menghasilkan konsentrasi tinggi dari SO_2 dan partikel halus, seperti sektor transportasi berbahan bakar diesel, pembangkit listrik berbasis batubara, dan aktivitas industri berat. Selain itu, pemantauan berkelanjutan terhadap variabel-variabel dominan ini dapat membantu dalam peringatan dini *(early warning system)* terhadap potensi bahaya kesehatan masyarakat akibat polusi udara.

Dengan kata lain, feature importance tidak hanya memberikan wawasan tentang kinerja teknis model prediksi, tetapi juga dapat menjadi alat bantu pengambilan keputusan berbasis data dalam konteks kebijakan publik yang lebih luas, khususnya dalam pengendalian dan peningkatan kualitas udara perkotaan.

4.3. Implikasi Praktis bagi Kebijakan Publik

Prediksi kualitas udara yang akurat memiliki peran strategis dalam mendukung proses pengambilan keputusan, khususnya dalam upaya mitigasi dan pengendalian polusi udara di wilayah perkotaan seperti Provinsi DKI Jakarta. Model XGBoost yang dikembangkan dalam studi ini dapat dimanfaatkan sebagai alat bantu oleh pemerintah daerah untuk memantau tingkat polusi udara secara lebih presisi dan *real-time*. Model ini mampu memprediksi konsentrasi polutan dengan mempertimbangkan berbagai parameter penting yang berkontribusi terhadap penurunan kualitas udara, seperti SO_2, $PM_{2.5}$, dan PM_{10}.

Pemerintah dapat memberikan peringatan dini kepada masyarakat, terutama saat terjadi peningkatan signifikan terhadap konsentrasi polutan berbahaya di udara. Hal ini penting untuk melindungi kelompok rentan seperti anak-anak, lansia, dan individu dengan gangguan pernapasan. Selain itu, sistem prediksi ini juga dapat membantu mengoptimalkan langkah-langkah tanggap darurat, seperti pembatasan aktivitas luar ruangan, penutupan sekolah sementara, atau peningkatan pelayanan kesehatan di wilayah terdampak.

Dari sisi kebijakan, informasi mengenai *feature importance* dari model XGBoost memungkinkan pemerintah untuk lebih fokus pada pengendalian sumber pencemar yang paling signifikan. Misalnya, jika SO_2 terbukti sebagai kontributor utama terhadap penurunan kualitas udara, maka kebijakan bisa diarahkan pada pengurangan emisi dari sumber-sumber yang menghasilkan gas ini, seperti pembangkit listrik tenaga batubara, kendaraan diesel, dan sektor industri. Begitu pula dengan partikel halus $PM_{2.5}$ dan PM_{10} yang sering berasal dari aktivitas konstruksi, pembakaran terbuka, dan polusi kendaraan bermotor [22].

Lebih jauh lagi, hasil dari model ini dapat digunakan untuk menyusun perencanaan pembangunan yang lebih berwawasan lingkungan *(environmentally sound planning)*, termasuk dalam hal tata ruang wilayah, pengembangan sistem transportasi publik rendah emisi, serta regulasi terhadap aktivitas industri yang berdampak terhadap udara. Dengan integrasi model prediksi ini ke dalam sistem pengambilan kebijakan, pemerintah dapat mengadopsi pendekatan berbasis data *(evidence-based policy making)* yang lebih responsif dan efektif dalam jangka panjang.

Selain itu, keterbukaan data hasil prediksi ini kepada masyarakat juga berpotensi meningkatkan kesadaran publik terhadap pentingnya menjaga kualitas udara. Akses terhadap informasi yang akurat dapat mendorong perubahan perilaku masyarakat, seperti beralih ke transportasi umum, mengurangi pembakaran sampah, atau menggunakan energi bersih dalam aktivitas sehari-hari.

Dengan demikian, *insight* utama yang didapat dari penelitian ini adalah bahwa penerapan model XGBoost tidak hanya meningkatkan akurasi prediksi kualitas udara, tetapi juga memperkuat landasan ilmiah bagi kebijakan publik yang lebih terarah, adaptif, dan berorientasi pada perlindungan kesehatan masyarakat serta keberlanjutan lingkungan perkotaan seperti Jakarta.

5. KESIMPULAN

Penelitian ini berhasil mengembangkan model prediksi kualitas udara di Jakarta menggunakan XGBoost yang menunjukkan performa sangat baik berdasarkan metrik evaluasi yang digunakan. Model XGBoost mampu menangkap pola kompleks dalam data polutan dan variabel lingkungan dengan akurasi yang tinggi. Analisis *feature importance* menegaskan bahwa SO_2, $PM_{2.5}$ dan PM_{10} adalah faktor utama yang mempengaruhi kualitas udara di wilayah tersebut. Hasil penelitian ini dapat dimanfaatkan oleh pemerintah untuk meningkatkan sistem monitoring dan peringatan dini polusi udara, serta mendukung kebijakan pengendalian polusi yang lebih efektif. Untuk pengembangan selanjutnya, disarankan untuk mengintegrasikan data meteorologi yang lebih lengkap dan memperluas cakupan data agar model dapat lebih akurat dan aplikatif.

Selain itu, terdapat beberapa saran yang diberikan untuk penelitian selanjutnya. Penelitian ini masih terbatas hanya pada polutan pembagun indeks kualitas udara. Untuk meningkatkan akurasi model, penelitian selanjutnya disarankan menggunakan data yang lebih komprehensif, termasuk parameter meteorologis seperti suhu, kelembapan, kecepatan angin, curah hujan, dan juga data aktivitas lalu lintas, serta faktor topografi yang juga dapat mempengaruhi sebaran polutan. Penelitian ke depan juga dapat menggabungkan model prediksi kualitas udara dengan pemetaan spasial menggunakan SIG. Hal ini memungkinkan visualisasi sebaran polusi udara secara geografis dan dinamis, sehingga mendukung pengambilan keputusan berbasis lokasi.

DAFTAR PUSTAKA

[1] N. Sunusi, A. A. Sikib dan S. Pasari, "A novel hybrid CLARA and fuzzy time series Markov chain model for predicting air pollution in Jakarta," *MethodsX,* vol. 14, 2025.

[2] Z. Guo, X. Jing, Y. Ling, Y. Yang dan N. Jing, "Optimized air quality management based on air quality index prediction and air pollutants identification in representative cities in China," *Scientific Reports,* vol. 14, 2024.

[3] "Kota paling berpolusi di dunia," IQAir, 2024. [Online]. Available: https://www.iqair.com/id/world-most-polluted-cities?continent=59af92b13e70001c1bd78e53. [Diakses 24 Maret 2025].

[4] "Kebijakan Komprehensif Atasi Polusi Udara Jakarta," Pemprov DKI Jakarta, 6 November 2023. [Online]. Available: https://www.jakarta.go.id/page/kebijakan-komprehensif-atasi-polusi-udara-jakarta. [Diakses 24 Maret 2025].

[5] W. Gao, T. Xiao, L. Zou, H. Li dan S. Gu, "Analysis and Prediction of Atmospheric Environmental Quality Based on the Autoregressive Integrated Moving Average Model (ARIMA Model) in Hunan Province, China," *Sustainability (Switzerland),* vol. 16, 2024.

[6] J. Yang, Y. Tian dan C. Wu, "Air Quality Prediction and Ranking Assessment Based on Bootstrap-XGBoost Algorithm and Ordinal Classification Models," *Atmosphere,* vol. 15, 2024.

[7] S. Qian, T. Peng, Z. Tao, X. Li dan M. , "An evolutionary deep learning model based on XGBoost feature selection and Gaussian data augmentation for AQI prediction," *Process Safety and Environmental Protection,* vol. 191, pp. 836-851, 2024.

[8] A. Vatresia, R. Nafila, W. Agwil dan F. P. Utama, "Forecasting air quality index data with autoregressive integrated moving average models," *EQA,* 2025.

[9] T. Liu dan S. You, "Analysis and Forecast of Beijing's Air Quality Index Based on ARIMA Model and Neural Network Model," *Atmosphere,* vol. 13, 2022.

[10] Z. Alfasanah, M. Z. H. Niam dan S. Wardia, "Monitoring air quality index with EWMA and individual charts using XGBoost and SVR residuals," *MethodsX,* vol. 14, 2025.

[11] T. Chen dan C. Guestrin, "Xgboost: A scalable tree boosting system," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* pp. 785-794, 2016.

[12] X. Guo, Z. Zhang, Z. Cai, L. Wang, Z. Gu, Y. Xu dan J. Zhao, "Analysis of the Spatial–Temporal Distribution Characteristics of NO2 and Their Influencing Factors in the Yangtze River Delta Based on Sentinel-5P Satellite Data," *Atmosphere,* vol. 13, 2022.

[13] B. Pardamean, R. Rahutomo, T. W. Cenggoro, A. Budiarto dan A. S. Perbangsa, "The Impact of Large-Scale Social Restriction Phases on the Air Quality Index in Jakarta," *Atmosphere,* vol. 12, 2021.

[14] Almaliki, A. H, A. Derdour dan E, "Air Quality Index (AQI) Prediction in Holy Makkah Based on Machine Learning Methods," *Sustainability,* vol. 15, no. 17, 2023.

[15] D. Q. Duong, Q. M. Le, T.-L. Nguyen-Tai, D. Bo, D. Nguyen, M.-S. Dao dan B. T. Nguyen, "Multi-source machine learning for AQI estimation," *IEEE International Conference on Big Data (Big Data),* 2020.

[16] J. M. Kuhn dan K. , Applied Predictive Modeling, New York: Springer, 2013.

[17] A. Jierula, S. Wang, T.-M. OH dan P. Wang, "Study on Accuracy Metrics for Evaluating the Predictions of Damage Locations in Deep Piles Using Artificial Neural Networks with Acoustic Emission Data," *Applied Sciences,* vol. 11, no. 5, 2021.

[18] K. Li, "Informativeness of Performance Measures: Coefficients or R-Squareds?," *Risk and Financial Management,* vol. 17, no. 11, 2024.

[19] Z. Wang, X. Wu dan Y. Wu, "A spatiotemporal XGBoost model for PM2.5 concentration prediction and its application in Shanghai.," *Heliyon,* vol. 9, 2023.

[20] J. Ma, Z. Yu, Y. Qu, J. Xu dan Y. Cao, "Application of the XGBoost Machine Learning Method in PM2.5 Prediction: A Case Study of Shanghai," *Aerosol Air Qual,* vol. 20, pp. 128-138, 2020.

[21] M. Z. Joharestani, C. Cao, X. Ni, B. Bashir dan S. Talebiesfandarani, "PM2.5 Prediction Based on Random Forest, XGBoost,," *Atmosphere,* vol. 10, no. 3, 2019.

[22] S. A. Meo, M. A. Salih, J. M. Alkhalifah, A. H. Alsomali dan A. A. Almushawah, "Environmental pollutants particulate matter (PM2.5, PM10), Carbon Monoxide (CO), Nitrogen dioxide (NO2), Sulfur dioxide (SO2), and Ozone (O3) impact on lung functions," *Journal of King Saud University - Science,* vol. 36, no. 7, 2024.