E-ISSN: 2715-2731

Machine Learning Untuk Peramalan Kualitas Indeks Standar Pencemar Udara DKI Jakarta Dengan Metode *Hibrid ARIMAX-LSTM*

Diaz Perdana[1*], Ahmad Muklason[2]

[1]Departemen Magister Manajemen Teknologi, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
[2]Departemen Sistem Informasi, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

[1]perdanadiaz@gmail.com; [2]muklason@is.its.ac.id

**corresponding author*

INFO ARTIKEL

Sejarah Artikel
Diterima: 28 November 2023
Direvisi: 27 Desember 2023
Diterbitkan: 31 Desember 2023

Kata Kunci
Forecasting
Hibrid
ISPU
Machine Learning

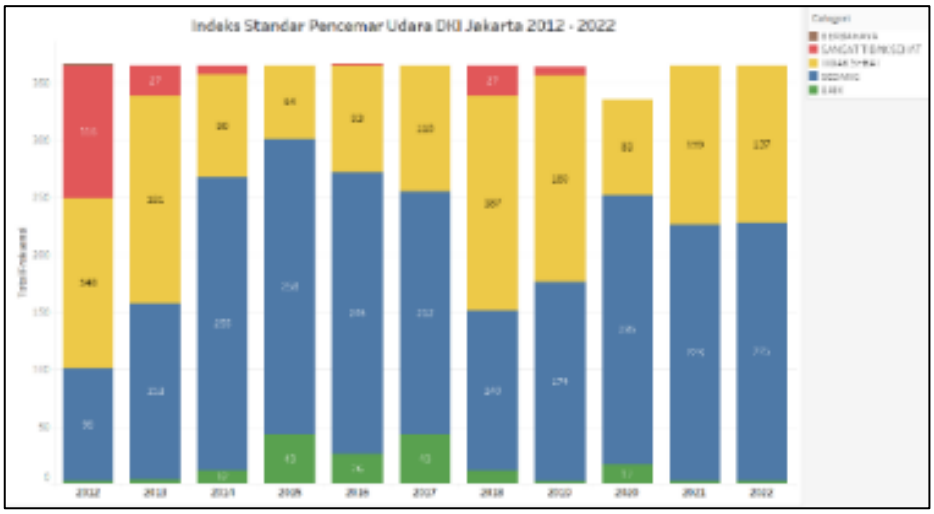
ABSTRAK

Polusi udara merupakan tantangan lingkungan global yang signifikan, menyebabkan dampak serius terhadap kesehatan masyarakat tidak terkecuali di Jakarta. Dengan Indeks Standar Pencemaran Udara (ISPU) sebagai parameter utama untuk memantau kualitas udara. Meskipun ISPU memberikan informasi saat ini, informasi yang diberikan untuk kualitas udara di masa mendatang masih terbatas. Oleh karena itu, diperlukan pendekatan yang lebih canggih, dan salah satu metode yang menjanjikan adalah menggunakan teknik *machine learning* (ML). Metode ML telah terbukti efektif dalam pemantauan dan peramalan kualitas udara. Namun, untuk meningkatkan akurasi peramalan, pendekatan hibridisasi, seperti menggabungkan dua model telah diusulkan. Pendekatan ini dapat memberikan deteksi pola yang lebih komprehensif dan meningkatkan akurasi hasil peramalan. Penelitian ini bertujuan untuk memprediksi ISPU di DKI Jakarta menggunakan model hibrid ARIMAX-LSTM. Data polusi udara dari tahun 2012 hingga 2022 akan digunakan, bersama dengan variabel eksternal seperti volume kendaraan, temperatur, kelembapan, dan kecepatan angin. Model ini mampu menghasilkan model prediksi dengan RMSE 13.00; 20.51; dan 17.10 untuk masing-masing polutan PM_{10}, $PM_{2.5}$, dan O_3. Sedangkan metriks MAPE yang dihasilkan dari model hibrid adalah 0.1916; 0.1917; dan 0.2869 untuk masing-masing polutan PM_{10}, $PM_{2.5}$, dan O_3. Model hibrid mampu menghasilkan model prediksi yang lebih baik dari model ARIMAX itu sendiri.

PENDAHULUAN

Polusi udara merupakan masalah lingkungan utama di seluruh dunia. Menurut laporan tahun 2021 yang diterbitkan oleh Organisasi Kesehatan Dunia (WHO), 13 orang di seluruh dunia meninggal setiap menit akibat polusi udara dan penyakit serius seperti penyakit kardiovaskular, stroke, dan kanker paru-paru [1]. Polusi udara adalah masalah yang berkembang di Jakarta, mempengaruhi kesehatan masyarakat dan lingkungan. Indeks Standar Pencemaran Udara (ISPU) digunakan untuk memantau kualitas udara dan menunjukkan polutan utama yang mempengaruhi kualitas udara di Jakarta. Namun, ISPU hanya memberikan informasi terkini dan kualitas udara di masa mendatang tidak dapat diprediksi. Indeks Standar Pencemar Udara (ISPU) adalah parameter yang mengukur kualitas udara dengan menunjukkan tingkat pencemaran udara yang disebabkan oleh bahan kimia dan partikel yang dikandungnya. Parameter yang digunakan untuk mengukur suatu kualitas udara adalah $PM_{2.5}$, PM_{10}, SO_2, CO, O_3, dan NO_2. Terdapat lima titik Stasiun Pemantau Kualitas

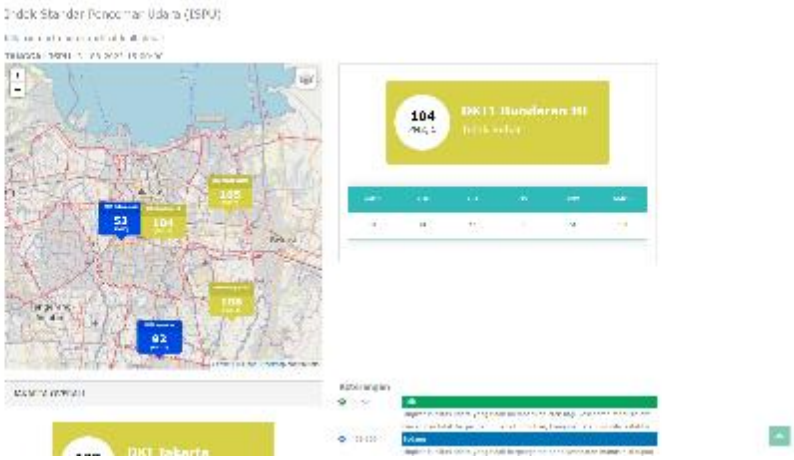
Udara (SPKU) di DKI Jakarta yaitu di Bundaran Hotel Indonesia, Jakarta Pusat (DKI 1), Kelapa Gading, Jakarta utara (DKI 2), Jagakarsa, Jakarta Selatan (DKI 3), Lubang Buaya, Jakarta Timur (DKI 4), dan Kebon Jeruk, Jakarta Barat (DKI 5). Berikut kondisi ISPU pada periode tahun 2012 – 2022.



Gambar 1. Kualitas Udara Dki Jakarta Periode 2012-2022

Berdasarkan Gambar 1, dapat dilihat bahwa kondisi ISPU provinsi di DKI Jakarta mayoritas berada di kategori ‘Sedang’ tetapi semenjak tahun 2021 frekuensi kategori ‘Tidak Sehat’ mengalami peningkatan. Penting untuk mengukur dan memantau ISPU secara teratur dan Jakarta tidak terkecuali karena kualitas udara yang buruk dapat berdampak negatif pada kesehatan manusia dan lingkungan. Berdasarkan informasi halaman *https://www.iqair.com/*, Jakarta berada di peringkat 21 dari 101 di antara beberapa kota besar di dunia dengan skor indeks 121 data dari Februari 2023. Nilai tersebut termasuk dalam kategori kelompok sensitif yang tidak sehat [2]. Bahkan, IQAir juga menyatakan kualitas udara di Jakarta selama bulan September 2023 berada di kategori “Tidak Sehat Bagi Kelompok Sensitif” dan menjadi kota dengan kualitas udara terburuk ketiga di dunia. Untuk mengatasi permasalahan tersebut, pemerintah harus segera mengeluarkan kebijakan untuk memperbaiki kondisi tersebut.

Pemantauan ISPU dilakukan berdasarkan data meteorologi yang mempengaruhi konsentrasi udara ambien. Data yang dimaksud meliputi kecepatan dan arah angin, kelembapan, suhu udara, intensitas matahari, dan curah hujan. Data ini dapat digunakan untuk memprediksi parameter ISPU harian. Perkiraan harian parameter ISPU memungkinkan pengambil kebijakan mengembangkan kebijakan dengan lebih akurat dan cepat. Pada website SILIKA DKI Jakarta belum terdapat fungsi untuk menampilkan prediksi ISPU padahal fungsi ini sangat bermanfaat bagi masyarakat. Keuntungan dari peramalan ISPU adalah masyarakat dapat memperkirakan terlebih dahulu kondisi kualitas udara seperti apa yang akan terjadi, termasuk pencemaran udara.



Gambar 2. Laman Website SILIKA DKI

Berdasarkan studi yang telah dilakukan sebelumnya, ada beberapa algoritma forecasting yang dapat digunakan untuk mengetahui kondisi ISPU dalam beberapa waktu kedepan. Beberapa penelitian telah dilakukan untuk memprediksi data pemantauan udara untuk mengetahui tren konsentrasi polutan. Salah satunya adalah metode hybrid. Keuntungan utama model hybrid yaitu meningkatkan akurasi peramalan karena deteksi pola dan pemodelan yang komprehensif, mengurangi risiko penggunaan model yang tidak sesuai akibat kombinasi perkiraan, dan menyederhanakan prosedur pemilihan model karena penggunaan komponen yang berbeda. Kombinasi yang bisa dilakukan adalah menggabungkan metode statistic dan metode *machine learning*.

Peneliti menggunakan *Autoregressive Integrated Moving Average with Exogenous* (ARIMAX). Kelebihan metode ARIMAX adalah fleksibilitasnya yaitu mengikuti model data yang ada dan mempunyai akurasi peramalan yang cukup tinggi, cocok untuk meramalkan sejumlah variabel dengan cepat karena hanya memerlukan data historis untuk melakukan peramalan dan mampu menggunakan variable eksternal dalam peramalannya. [3]. Selain itu, model *machine learning* yang bisa digunakan adalah *Long Short Term Memory* (LSTM). Keuntungan utama LSTM sebagai metode analisis regresi adalah kemampuannya untuk mengaproksimasi fungsi kontinyu sambil meminimalkan kesalahan dalam menyesuaikan data pengamatan yang terbatas. Dengan menggunakan data polusi udara dari rentang tahun 2012 hingga 2022 dan menggunakan variabel eksternal seperti volume kendaraan, temperatur, kelembapan, dan kecepatan angin, hasil yang diharapkan dari penelitian ini mampu memprediksi kondisi kualitas udara dalam periode harian dan diharapkan dapat membantu pemerintah dan masyarakat untuk mengambil kebijakan dan langkah yang tepat untuk menjaga kualitas udara di Jakarta.

METODE

Gambar 3. Alur Penelitian

Metode penelitian dapat dilihat pada Gambar 3. Untuk setiap langkah-langkah dapat dijelaskan sebagai berikut:

Data preprocessing

Pada tahap ini terdapat beberapa proses sebagai berikut:

Input Data

Data yang digunakan dalam penelitian ini merupakan data primer yang diperoleh dari Dinas Lingkungan Hidup Provinsi DKI Jakarta. Data tersebut meliputi parameter pengukuran indeks standar pencemar udara mulai tanggal 01 Januari 2012 sampai 31 Desember 2022. Variabel penelitian terdiri dari variabel respon yaitu polutan yang digunakan (PM_{10}, $PM_{2.5}$, dan O_3) dikarenakan ketiga polutan tersebut merupakan polutan yang mendominasi dalam periode 2012 – 2022. Kemudian, juga ditambahkan variabel eksternal yang digunakan untuk mendukung hasil prediksi yang digunakan, diantaranya rata-rata temperatur, tingkat curah hujan dan kelembapan.

Normalisasi Data

Melakukan normalisasi data dengan menggunakan scaling data dengan *MinMaxScaler* untuk merubah angka menjadi lebih kecil dari rentang 0 hingga 1.

Uji Korelasi

Melakukan uji korelasi sebagai screening awal apakah ada keterkaitan yang erat antara variabel utama dengan variabel eksternal.

Data splitting

Data *splitting* merupakan sebuah teknik untuk membagi dataset menjadi 2 bagian yaitu data testing, dan data training. Data *training* adalah data yang digunakan untuk mengembangkan model hingga didapatkan parameter yang optimal. Kemudian, untuk menguji bagaimana prediksi ISPU digunakan data testing. Pada penelitian ini, proporsi yang digunakan untuk data splitting adalah 80% data *training* dan 20% data *testing.*

Pengembangan model *Hibrid Forecasting*

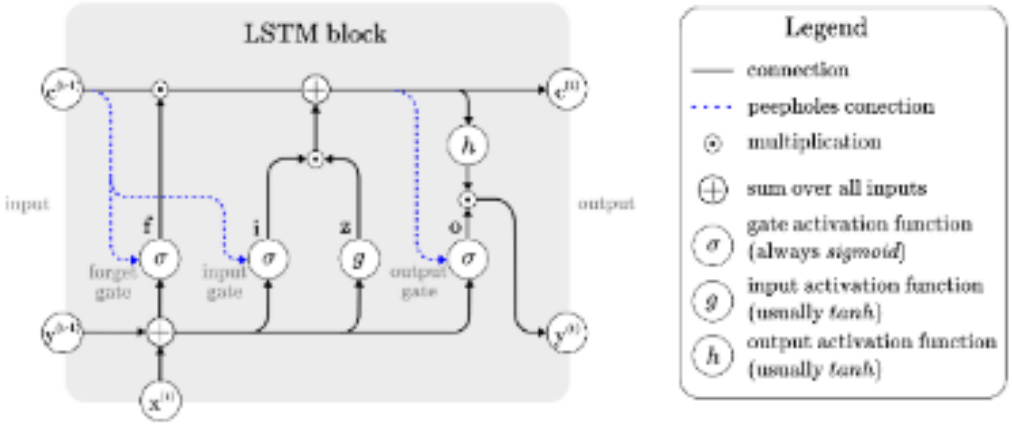
Pada tahap ini akan dilakukan pembangunan model Hibrid forecasting yaitu dengan model ARIMAX kemudian hasil prediksi dari model ARIMAX dijadikan input untuk dikembangkan kembali dengan model LSTM. ARIMAX merupakan versi lanjutan dari model ARIMA yang menggunakan peramalan deret waktu multivariat dengan menggunakan beberapa deret waktu yang disediakan sebagai variabel eksogen untuk meramalkan variabel terikat. Fungsi Auto ARIMA biasanya digunakan untuk memilih model terbaik dengan secara otomatis menghasilkan sekumpulan parameter optimal dengan menguji semua kemungkinan kombinasi (*p,d,q*) dan mengembalikan gambar model dengan *Akaike Information Criterion* (AIC) terendah dan *Bayesian Information Criterion* (nilai BIC) [4]. Secara umum bentuk model ARIMAX (p,d,q) dapat diberikan dengan Persamaan (1).

$$(1-B)^d \phi_p Z_t = \mu + \theta_q(B) a_t + \beta_1 X_{1,t} + \cdots + \beta_k X_{k,t} \qquad (1)$$

dengan Z_t adalah variabel dependen waktu ke-t, $X_{1,t}$, $X_{k,t}$. B adalah variabel eksogen pada waktu ke-t, $\phi_p(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$ dan $\theta_q(B) = 1 + \theta_1 B + \cdots + \theta_q B^p$ adalah komponen *autoregressive* dan *moving average* pada pola non musiman [5].

Long Short-Term Memory merupakan pengembangan dari jenis arsitektur jaringan saraf berulang (RNN) yang dirancang khusus untuk menangani masalah ketergantungan jangka panjang dalam data urutan, seperti data deret waktu. Kelemahan dari RNN adalah pengaruh *input* yang diberikan pada lapisan tersembunyi, dan oleh karena itu, pada *output* jaringan, sensitivitas akan berkurang atau meledak secara eksponensial saat ia melakukan siklusnya

di sekitar *recurrent networks conection*. Efek ini sering disebut dalam literatur sebagai masalah gradien hilang atau *vanished gradient* [6]. Dalam unit LSTM terdiri dari sel, *input gate*, *output gate*, dan *forget gate*. *Forget gate* ini yang memiliki fungsi untuk mengatur ulang kondisi yang terjadi. Sel mengingat nilai-nilai dalam interval waktu yang berubah-ubah dan ketiga *gates* mengatur aliran informasi yang terkait dengan sel. Hal tersebut dijelaskan pada Gambar 4.



Gambar 4. Arsitektur LSTM [7]

Evaluasi Model

Model yang telah dibangun kemudian ditampilkan visualisasinya dengan menggunakan line chart kemudian diukur dengan *Root Mean Square Error* (RMSE) dan *Mean Absolute Percentage Error* (MAPE). *Root Mean Square Error* adalah metrik yang umum digunakan untuk mengevaluasi keakuratan prediksi yang diperoleh suatu model. Dibutuhkan sisa antara nilai aktual dan prediksi dan membandingkan kesalahan prediksi model yang berbeda untuk data tertentu. Manfaat utama penggunaan RMSE adalah memberikan penalti terhadap kesalahan besar dan menskalakan skor dalam satuan yang sama dengan nilai perkiraan. Misalkan $\hat{y}_i$ adalah nilai prediksi dan y_i adalah nilai aktual dan untuk n observasi [8]. RMSE untuk n nilai tertentu dihitung menggunakan Persamaan (2).

$$RMSE\ (f,x) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \qquad (2)$$

dengan $\hat{y}_i$ = jumlah prediksi ke-i dan y_i= jumlah observasi ke-i

Kemudian *Mean Absolute Percentage Error* ditambahkan karena masalah ukuran *error* dapat dihindari ketika data aktual lebih besar dari data prediksi atau sebaliknya dengan menggunakan MAPE. *Mean Absolute Percentage Error* (MAPE) dihitung dengan mencari kesalahan absolut untuk setiap periode waktu, kemudian membaginya dengan nilai observasi untuk periode waktu tersebut, dan terakhir merata-ratakan persentase absolut tersebut sesuai Persamaan (3) [9].

$$MAPE = \frac{\sum_{i=1}^{n}\left|\frac{Y_i - \hat{Y}_i}{Y_i}\right|}{n} \qquad (3)$$

dengan Y_i = observasi ke-i dan $\hat{Y}_i$ = prediksi ke-i

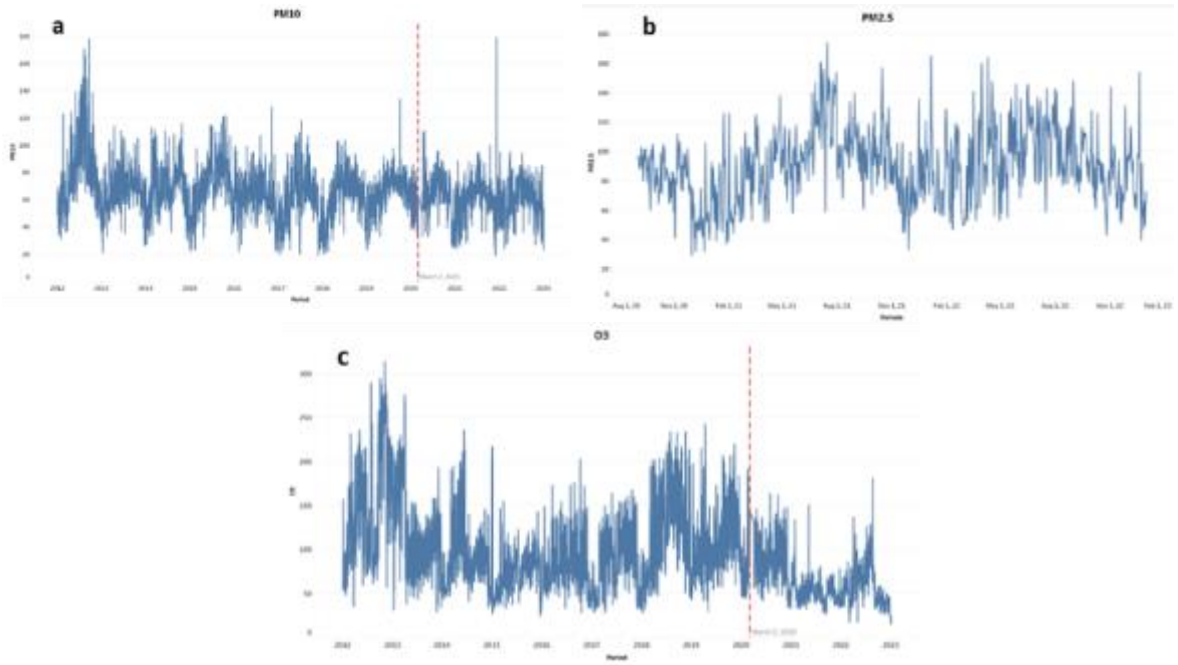
HASIL DAN PEMBAHASAN

Dilakukan pengecekan *missing value* untuk masing-masing variabel supaya hasil prediksi menjadi lebih akurat. Hasil pengecekan *missing value* dapat dilihat pada Tabel 1.

Tabel 1. Pengecekan *Missing Values* Data

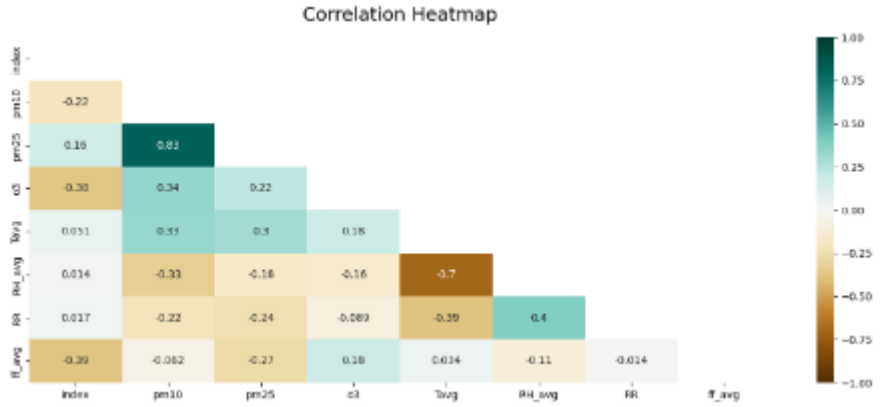
No	VARIABEL	N	N	% n
1	Tanggal	3986	0	0.00%
2	Polutan PM_{10}	3986	0	0.00%
3	Polutan $PM_{2.5}$	852	3134	78.63%
4	Polutan O_3	3986	0	0.00%
5	Temperatur	3960	26	0.65%
6	Kelembapan	3959	27	0.68%
7	Kecepatan Angin	3972	14	0.35%
8	Curah Hujan	3986	0	0.00%

Dengan N adalah jumlah observasi, n adalah jumlah *missing data*, dan %n adalah persentase *missing data* dari jumlah observasi. Untuk data yang memiliki *missing values rate* kurang dari 1% maka kita bisa melakukan imputasi dengan metode interpolasi. Metode seperti interpolasi dan *moving average* mempunyai sifat komputasi berdasarkan informasi sebelum dan sesudah. karena menunjukkan kesesuaian yang lebih baik, sesuai dengan semua kriteria kinerja [10].



Gambar 5. Grafik Harian dari Polutan (a) PM_{10}, (b) $PM_{2.5}$, dan (c) O_3

Untuk grafik harian dari masing-masing polutan dapat dilihat pada Gambar 5. Dari Gambar 4 dapat dilihat bahwa kondisi polutan PM_{10} dan O_3 dari periode Januari 2012 hingga Desember 2022. Sedangkan untuk polutan $PM_{2.5}$ dari periode September 2020 hingga Desember 2022. Secara keseluruhan kondisi udara di Jakarta sangat fluktuatif dan berada di kategori sedang (50-100). Sehingga data historis merupakan data yang akan dijadikan input untuk pengembangan model hibrid. Kemudian, untuk mengetahui apakah variabel cuaca memiliki hubungan terhadap polutan, dilakukan pengecekan dengan metode Pearson untuk mengetahui apakah ada korelasi dari variabel cuaca dengan polutan. Plot korelasi dapat dilihat pada Gambar 6. Berdasarkan Gambar V korelasi diatas, bahwa polutan PM_{10} dan $PM_{2.5}$ memiliki korelasi yang cukup signifikan (> 0.3) dengan variabel cuaca. Sedangkan polutan O_3 berkorelasi tidak terlalu signifikan dengan variabel cuaca. Hal ini menandakan bahwa variabel cuaca cukup berpengaruh terhadap kualitas udara di DKI Jakarta



Gambar 6. Korelasi antara Variabel Polutan dengan Cuaca

Prediksi dengan ARIMAX

Sebelum dilakukannya prediksi untuk polutan, dilakukan pengujian kestasioneran, penentuan lag optimum, dan Estimasi model ARIMAX untuk menglihat model terbaik berdasarkan nilai RMSE dan MAPE terkecil. Syarat pertama dalam pemodelan ARIMAX adalah data *series* harus stasioner pada seluruh variabel baik stasioner dalam varians maupun stasioner dalam rata-rata. Identifikasi kestasioneran dalam varians dapat dilakukan dengan melihat plot *Box-Cox*. Hasil pengujian kestasioneran data dapat dilihat pada Tabel 2.

Tabel 2. Nilai Box-Cox Plot

Variabel	**Hasil**			
	Estimate	***Lower CL***	***Upper CL***	***Rounded Value***
PM_{10}	0.48	0.39	0.57	0.50
$PM_{2.5}$	0.51	0.27	0.72	0.50
O_3	-0.15	-0.21	-0.09	-0.15

Pada Tabel 2, menunjukkan bahwa $PM_{2.5}$ dan PM_{10} memiliki nilai *rounded value* bernilai 0.5 dan O_3 bernilai -0.15 sehingga dapat disimpulkan bahwa data tersebut sudah memenuhi syarat yaitu stasioner dalam varians ketika dilakukan transformasi lebih lanjut sesuai nilai *rounded value* lambda nya.

Setelah semua variabel stasioner dalam varians, selanjutnya menguji apakah semua variabel stasioner dalam rata-rata sehingga dilakukan uji ADF (*Augmented Dickey Fuller).*

Data *time series* dikatakan stasioner dalam rata-rata apabila nilai *p-value* kurang dari α = 5% atau 0,05. Hasil pengujian stasioneritas disajikan dibawah ini: Hipotesis yang dipakai untuk uji ini adalah:

H_0: Data tidak stasioner

H_1: Data stasioner

Tabel 3. Nilai ADF Test

Variabel	*ADF-Test (p-value)*	Keputusan	Keterangan
PM_{10}	0.01	Tolak H_0	Data Stasioner
$PM_{2.5}$	0.01	Tolak H_0	Data Stasioner
O_3	0.01	Tolak H_0	Data Stasioner

Dengan menggunakan α = 5% dari Tabel 3 menunjukkan bahwa hasil pengujian yang diperoleh bahwa nilai ADF test untuk variable PM2,5 dan PM10, memiliki nilai *p-value* < 0.05 yang menunjukkan bahwa variable tersebut stasioner pada tingkat signifikansi 5%. Untuk menentukan order untuk masing-masing AR dan MA. Maka dilakukan *Autocorrelation Plot* untuk order MA dan *Partial Autocorrelation Plot* untuk order AR dengan rentang lag dari 1-30. Dari hasil yang didapatkan, nilai lag optimum dapat dilihat pada Tabel 4.

Tabel 4. Penentuan Order ARIMAX dengan Nilai AIC Minimum

Variabel	ARMA (ORDER P,Q)	
	Order	AIC minimum
PM_{10}	1,1	9333.16
$PM_{2.5}$	1,1	7443.4
O_3	1,2	37769.8

Prediksi dilakukan untuk masing-masing variabel menggunakan order p,q yang menghasilkan AIC minimum dan dikarenakan seluruh polutan berdasarkan transformasi *box-cox* sudah ditransformasi berdasarkan nilai lambda optimal.

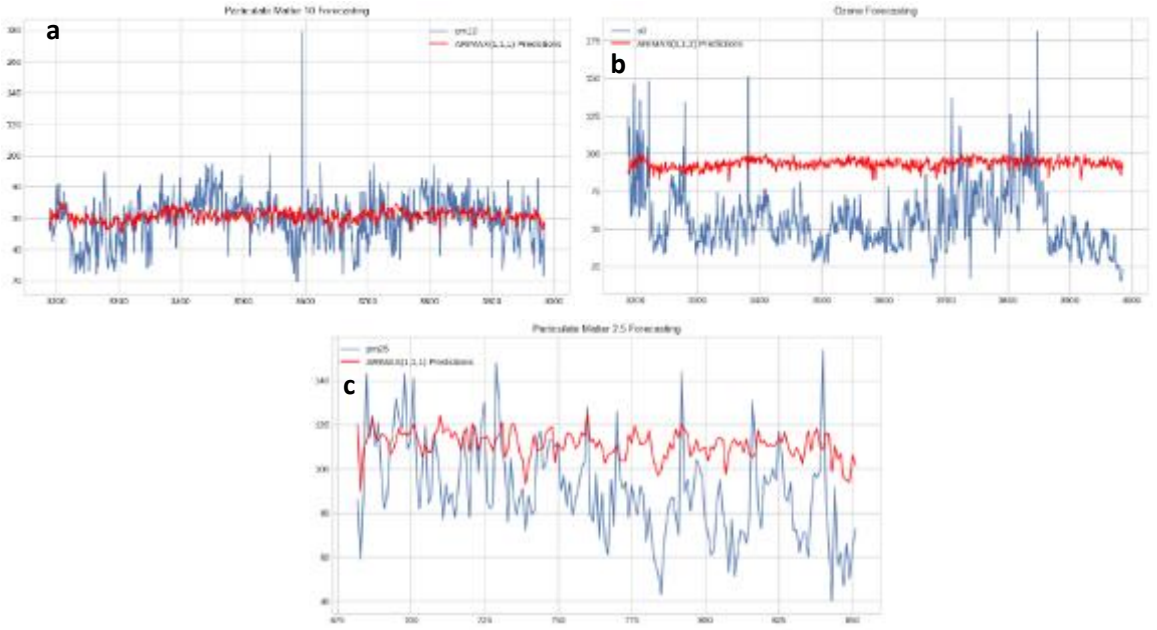
Tabel 5. Perbandingan Nilai Aktual dan Prediksi Berdasarkan Model ARIMAX

Variabel	Periode ke-n	Nilai Aktual	Nilai Prediksi
PM_{10}	3189	57	55.69
	3190	52	54.36
	3191	66	59.96
	3192	57	64.96
	3193	63	64.64
	…	...	...
	3981	36	53.96

Variabel	Periode ke-n	Nilai Aktual	Nilai Prediksi
	3982	46	52.40
	3983	23	53.31
	3984	40	57.48
	3985	54	55.41
O_3	3189	123	87.66
	3190	98	85.57
	3191	117	89.74
	3192	88	94.86
	3193	58	94.65
	...	...	...
	3981	20	88.11
	3982	15	84.39
	3983	16	88.85
	3984	17	92.89
	3985	23	89.08
$PM_{2.5}$	682	86.00	120.02
	683	59.00	89.72
	684	89.00	109.03
	685	143.00	112.62
	686	114.00	117.06
	...	...	...
	847	47.00	95.98
	848	66.00	94.78
	849	50.00	93.76
	850	64.00	106.35
	851	73.00	101.56

Tabel 5 menunjukkan hasil perbandingan antara data prediksi dengan data aktual berdasarkan periode waktu pada data testing. Terlihat bahwa nilai prediksi yang dihasilkan berdasarkan model ARIMAX yang sudah dikembangkan sesuai order p,d, dan q masing-masing. Kemudian dari hasil prediksi tersebut, dibentuk visualisasi dengan menggunakan

line chart untuk mengetahui bagaimana pola data prediksi yang sudah dihasilkan. Visualisasi yang dihasilkan dapat dilihat pada Gambar 7.



Gambar 7. Grafik Prediksi ARIMAX Polutan (a) PM_{10}, (b) $PM_{2.5}$, dan (c) O_3

Berdasarkan Gambar 7 dapat dilihat bahwa secara pola data yang dihasilkan dari metode ARIMAX untuk variabel PM_{10} dan O_3 masih linier sedangkan untuk variabel $PM_{2.5}$ terlihat bisa mengikuti pola dari data aktual. Hal ini bisa terjadi karena perbedaan data *training* dan *testing*, dimana data yang digunakan lebih besar untuk variabel PM_{10} dan O_3 sedangkan $PM_{2.5}$ lebih kecil. Hal ini tentu berpengaruh terhadap model prediksi yang dihasilkan karena ARIMA hanya memiliki kemampuan memprediksi data time series jangka pendek (*short memory*) [11] sehingga perlu adanya model yang mampu mengatasi permasalahan dalam memprediksi data jangka panjang.

Tabel 6. Evaluasi Model ARIMAX

Variabel	RMSE	MAPE
PM_{10}	13.91	0.2121
$PM_{2.5}$	28.41	0.3228
O_3	41.09	0.8631

Untuk evaluasi model digunakan metriks MAPE dan RMSE. Hasil RMSE dan MAPE masing-masing variabel untuk model prediksi ARIMAX dapat dilihat pada Tabel IV. Untuk variabel PM_{10}, RMSE yang dihasilkan sebesar 13.91 dan untuk MAPE yang dihasilkan 21.21%. Kemudian, untuk variabel $PM_{2.5}$, RMSE yang dihasilkan sebesar 28.41 dan untuk MAPE sebesar 32.28%. Untuk variabel O_3. RMSE yang dihasilkan 41.09 dan untuk MAPE sebesar 86.31%. MAPE yang dihasilkan terutama untuk variabel O_3 masih cukup besar dikarenakan komponen data non linier dengan mampu menurunkan keakuratan dalam memprediksi. [12]. Sehingga model akan dikembangkan kembali dengan LSTM yang memiliki kemampuan untuk mengatasi masalah non-linieritas data.

Prediksi dengan Model Hibrid

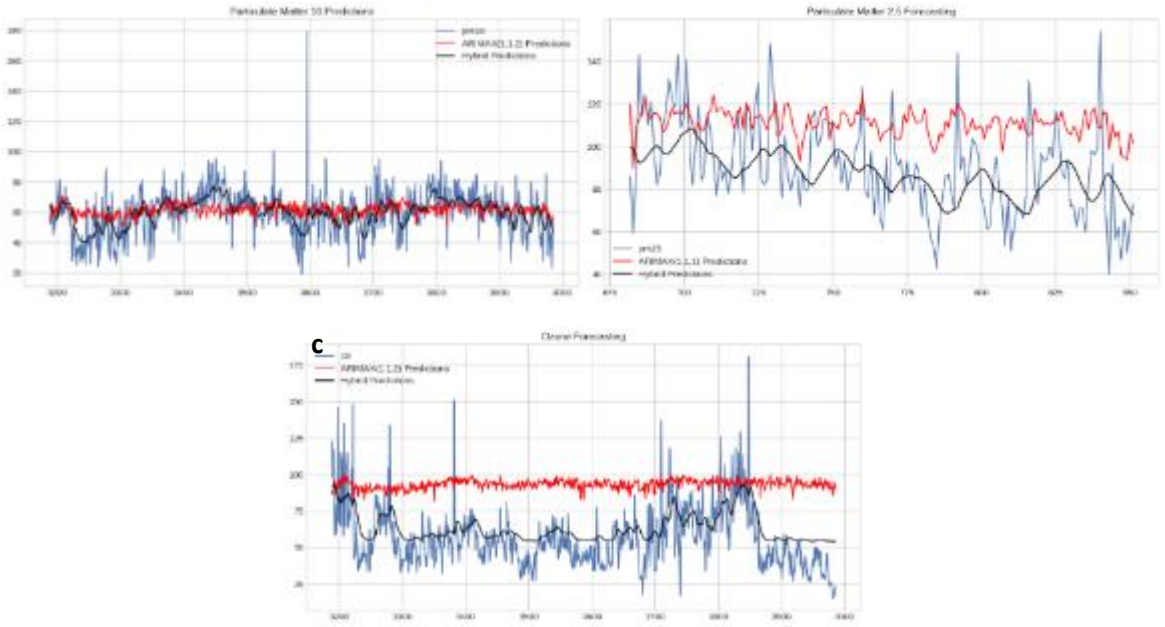
Setelah mendapatkan hasil prediksi dengan model ARIMAX untuk masing-masing variabel. Perlu dikembangkan kembali dengan LSTM supaya mendapatkan hasil prediksi yang lebih baik. Penelitian ini menggunakan hasil prediksi ARIMAX sebagai input variabel untuk model LSTM. Sebelum dilakukan prediksi menggunakan LSTM, dilakukan terlebih dahulu normalisasi data, data *splitting*, penentuan batch size dan epoch untuk model LSTM, dan evaluasi model menggunakan nilai RMSE dan MAPE. Penelitian ini menggunakan metode *MinMax Scaling* yang akan me-*rescale* data dari suatu range ke range baru lain dimana range yang digunakan adalah dari range 0 sampai 1. Kemudian, dilakukan data *splitting* dengan proporsi 80% data *training* (3198 *record* data untuk variabel PM_{10} dan O_3; 682 *records* data untuk variabel $PM_{2.5}$. 20% data *testing* (710 *records* data untuk variabel PM_{10} dan O_3; 170 *records* data untuk variabel $PM_{2.5}$). Jumlah hidden layer yang digunakan adalah 1, *batch size* sebanyak 90 dan dilakukan iterasi *epoch* sebanyak 30 untuk masing-masing variabel digunakan untuk pembangunan model kali ini. Kemudian model yang sudah dibangun dievaluasi dengan RMSE dan MAPE.

Tabel 7. Perbandingan Nilai Aktual dan Prediksi Model Hibrid

Variabel	Periode ke-n	Nilai Aktual	Nilai Prediksi
PM_{10}	3189	57	64.94
	3190	52	63.68
	3191	66	62.22
	3192	57	61.46
	3193	63	60.71
	…	...	...
	3981	36	54.45
	3982	46	52.46
	3983	23	50.70
	3984	40	48.52
	3985	54	46.61
O_3	3189	123	92.17
	3190	98	92.36
	3191	117	92.49
	3192	88	93.58
	3193	58	93.75
	...	...	...
	3981	20	54.13
	3982	15	54.11

	3983	16	54.09
	3984	17	54.07
	3985	23	54.05
$PM_{2.5}$	682	86.00	104.45
	683	59.00	103.38
	684	89.00	100.34
	685	143.00	97.72
	686	114.00	98.10
	...	...	...
	847	47.00	82.26
	848	66.00	79.05
	849	50.00	76.32
	850	64.00	73.81
	851	73.00	71.82

Tabel 7 menunjukkan hasil perbandingan antara data prediksi dengan data aktual berdasarkan periode waktu pada data testing. Terlihat bahwa nilai prediksi yang dihasilkan berdasarkan model hibrid. Kemudian dari hasil prediksi tersebut, dibentuk visualisasi dengan menggunakan *line chart* untuk mengetahui bagaimana pola data prediksi yang sudah dihasilkan. Visualisasi yang dihasilkan dapat dilihat pada Gambar 8



Gambar 8. Grafik Prediksi Model Hibrid Polutan (a) PM_{10}, (b) $PM_{2.5}$, dan (c) O_3

Berdasarkan Gambar 8 dapat dilihat bahwa secara pola data yang dihasilkan dari metode Hibrid untuk seluruh variabel PM_{10}, O_3, dan $PM_{2.5}$ terlihat lebih bisa mengikuti pola dari data aktual dibandingkan model ARIMAX. Hal ini bisa terjadi karena kelebihan dari model LSTM dimana mampu mempelajari pola kompleks dari data sesuai dengan jumlah *epochs* dan fungsi *dropout.*

Tabel 7. Perbandingan Evaluasi Model ARIMAX dan Hibrid

Variabel	ARIMAX		Hibrid	
	RMSE	MAPE	RMSE	MAPE
PM_{10}	13.91	0.2121	13.00	0.1916
$PM_{2.5}$	28.41	0.3228	20.51	0.1917
O_3	41.09	0.8631	17.10	0.2869

Untuk evaluasi model hibrid juga menggunakan metriks MAPE dan RMSE. Hasil RMSE dan MAPE masing-masing variabel untuk model hibrid dapat dilihat pada Tabel 7. Secara keseluruhan, model hibrid yang dihasilkan lebih baik dibandingkan model prediksi ARIMAX karena menghasilkan metriks evaluasi yang lebih baik. Untuk variabel PM_{10}, RMSE yang dihasilkan meningkat menjadi 13.00 dan untuk MAPE yang dihasilkan 19.16%. Kemudian, untuk variabel $PM_{2.5}$, RMSE yang dihasilkan sebesar 20.51 dan untuk MAPE sebesar 19.17%. Untuk variabel O_3. RMSE yang dihasilkan 17.10 dan untuk MAPE sebesar 28.69%.

KESIMPULAN

Berdasarkan hasil prediksi yang dilakukan pada masing-masing polutan dengan model hibrid menunjukkan bahwa model hibrid dapat memprediksi kualitas udara dengan metriks evaluasi error yang cukup baik. Model ini mampu menghasilkan model prediksi dengan RMSE 13.00; 20.51; dan 17.10 untuk masing-masing polutan PM_{10}, $PM_{2.5}$, dan O_3. Sedangkan metriks MAPE yang dihasilkan dari model hibrid adalah 0.1916; 0.1917; dan 0.2869 untuk masing-masing polutan PM_{10}, $PM_{2.5}$, dan O_3. Dari hasil prediksi yang dihasilkan menunjukkan bahwa polutan $PM_{2.5}$ yang membuat kondisi udara di Jakarta berada pada kategori "Tidak Sehat". Salah satu sumber polutan $PM_{2.5}$ adalah emisi asap kendaraan bermotor dan menyumbang sebanyak 42% dari total *Particulate Matter* yang ada di Jakarta berdasarkan data dari BPS Provinsi DKI Jakarta sehingga pemerintah perlu mengontrol bagaimana frekuensi kendaraan yang melewati ruas jalan di Jakarta.

REFERENSI

[1] World Health Organization, "Air pollution causes 13 deaths per minute worldwide," 2021. [Online]. Available: https://www.who.int/multi-media/details/air-pollution-climate-change. [Accessed 3 March 2023].

[2] IQAir, 2023. [Online]. Available: https://www.iqair.com/id/indonesia/jakarta. [Accessed 14 February 2023].

[3] H. WS, "Penggunaan Metode ARIMA (Autoregressive Integrated Moving Average) Untuk Prakiraan Jumlah Permintaan Gula Rafinasi," Universitas Islam Negeri Alaludin, Makassar, 2018.

[4] B. Dissanayake, N. Lakshitha, O. Hemachandra and D. Haputhanthri, "A Comparison of ARIMAX, VAR and LSTM on Multivariate Short-Term Traffic Volume Forecasting," *PROCEEDING OF THE 28TH CONFERENCE OF FRUCT ASSOCIATION*, pp. 564-570, 2021.

[5] A. R. Suryani, "Peramalan Curah Hujan Dengan Metode Autoregressive Integrated Moving Average With Exogenouse Input (ARIMAX)," Universitas Negeri Semarang, Semarang, 2016.

[6] A. Graves, "Long Short-Term Memory," in *Supervised Sequence Labelling with Recurrent Neural Networks*, Berlin, Springer, 2012, pp. 37-45.

[7] G. V. Houdt, C. Mosquera and G. Napoles, "A review on the long short-term memory model," *Artificial Intelligence Review*, vol. 53, pp. 5929-5955, 2020.

[8] T. O. Hodson, "Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not," *Geoscientific Model Development*, pp. 5481-5487, 2022.

[9] P. Jana, "Aplikasi Triple Exponential Smoothing Untuk Forecasting Jumlah Penduduk Miskin," *Jurnal Derivat*, vol. 3, pp. 75-81, 2016.

[10] P. Saeipourdizaj, P. Sarbakhsh and A. Gholampour, "Application of imputation methods for missing values of PM10 and O3 data: Interpolation, moving average and K-nearest neighbor methods," *Environmental Health Engineering and Management Journal*, vol. 8, no. 3, pp. 215-226, 2021.

[11] F. Sowell, "Maximum likelihood estimation of stationary univariate fractionally integrated time series models," *Journal of Econometrics*, vol. 1992, no. 3, pp. 165-188, 1992.

[12] C. V. M. Sihombing, S. Martha and N. M. Huda, "Analisis Metode Hybrid ARIMA–SVR Pada Indeks Harga Saham Gabungan," *Jurnal Ilmiah Math. Stat. dan Terapannya (Bimaster)*, vol. 11, no. 3, pp. 413-422, 2021.