Peningkatan Akurasi Prediksi PM10 Jakarta Menggunakan Model Hybrid Random Forest-ARIMA Berbasis Feature Engineering

Proposal Tugas Akhir

Oleh

Muhammad Rafi Dhiyaulhaq
18222069

PROGRAM STUDI SISTEM DAN TEKNOLOGI INFORMASI
SEKOLAH TEKNIK ELEKTRO DAN INFORMATIKA
INSTITUT TEKNOLOGI BANDUNG

Desember 2025

LEMBAR PENGESAHAN

Peningkatan Akurasi Prediksi PM10 Jakarta Menggunakan Model Hybrid Random Forest-ARIMA Berbasis Feature Engineering

Proposal Tugas Akhir

Oleh

Muhammad Rafi Dhiyaulhaq
18222069

Program Studi Sistem dan Teknologi Informasi
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung

Proposal Tugas Akhir ini telah disetujui dan disahkan
di Bandung, pada tanggal 2 Desember 2025

Pembimbing

Dr. Fetty Fitriyanti Lubis, S.T., M.T.

NIP. 118110071

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB I

PENDAHULUAN

I.1 Latar Belakang

DKI Jakarta sebagai ibu kota Indonesia dengan populasi lebih dari 10 juta jiwa menghadapi permasalahan serius terkait polusi udara. Analisis WRI Indonesia terhadap data tahun 2019 hingga 2021 menunjukkan bahwa konsentrasi rata-rata bulanan $PM_{2,5}$ mengikuti pola musiman dengan puncak polusi mencapai 40–80 µg/m³ pada musim kemarau (Firdaus dkk. September 2023). Kualitas udara Jakarta pada tahun 2019 tercatat lima kali lebih buruk dibandingkan pedoman WHO, dengan peningkatan rata-rata *Air Quality Index* (AQI) sebesar 69% antara Juni 2017 dan Juni 2020 (Aulia 2024). Pusat Krisis Kesehatan Kementerian Kesehatan Republik Indonesia melaporkan bahwa nilai ISPU di Jakarta Timur, Selatan, dan Barat berada pada kategori tidak sehat, dengan dampak langsung berupa gangguan pernapasan dan iritasi mata (Pusat Krisis Kesehatan Kementerian Kesehatan Republik Indonesia July 2019).

Partikulat PM_{10} memiliki dampak signifikan terhadap kesehatan. Zhang dkk. (June 2021) menunjukkan bahwa setiap peningkatan 10 µg/m³ PM_{10} berhubungan dengan peningkatan kunjungan unit gawat darurat sebesar 0,14% untuk penyakit kardiovaskular dan 0,56% untuk aritmia. Kyung dan Jeong (March 2020) melaporkan peningkatan 2,7% pada hospitalisasi COPD dan 1,1% pada mortalitas COPD untuk setiap kenaikan 10 µg/m³ PM_{10}. Pemerintah Indonesia telah mengatur pemantauan kualitas udara melalui Indeks Standar Pencemar Udara (ISPU) yang berfungsi sebagai sistem peringatan dini bagi masyarakat dan bahan pertimbangan dalam upaya pengendalian pencemaran udara (Kementerian Lingkungan Hidup dan Kehutanan Republik Indonesia July 2020; Chaniago, Zahara, dan Ramadhani September 2020).

Dalam konteks prediksi kualitas udara menggunakan *machine learning*, model *hybrid* yang menggabungkan *Random Forest* dan ARIMA telah menunjukkan performa

menjanjikan. Yenkikar dkk. (July 2025) mengembangkan model *hybrid Random Forest Regressor* (RFR) dengan ARIMA yang mencapai $R^2 = 0{,}94$ untuk prediksi AQI di India. Namun, penelitian tersebut secara eksplisit menyatakan keterbatasan bahwa model mengecualikan faktor eksternal dan hanya menggunakan nilai konsentrasi polutan mentah tanpa mengeksplorasi pola temporal, fitur statistik, dan interaksi antar polutan yang dapat meningkatkan akurasi prediksi. Penelitian terkini menunjukkan pentingnya *feature engineering*, dengan Naz dkk. (2024) melaporkan peningkatan performa model sebesar 5–86% melalui pendekatan *feature engineering* dua tahap yang menghasilkan 22 fitur mencakup kategori temporal, statistik, dan polutan. Chen dkk. (January 2025) menunjukkan bahwa integrasi fitur tambahan dapat mengurangi RMSE sebesar 4,216–8,458 untuk prediksi polutan udara.

Penelitian prediksi kualitas udara Jakarta menggunakan *machine learning* masih terbatas. Arsy dan Yasir (April 2025) mengembangkan model prediksi $PM_{2,5}$ Jakarta menggunakan *Random Forest* yang mencapai $R^2 = 0{,}61$, sementara Radjabaycolle, Wattimena, dan Pattiradjawane (2025) menerapkan LSTM dengan analisis SHAP untuk Jakarta, menunjukkan pentingnya *explainability* dalam mendukung intervensi kebijakan berbasis bukti.

Berdasarkan analisis tersebut, penelitian ini mengembangkan model prediksi PM_{10} Jakarta dengan meningkatkan arsitektur *hybrid Random Forest*-ARIMA melalui strategi *feature engineering* sistematis yang mengekstraksi informasi temporal, statistikal, dan interaksi polutan. Model diterapkan pada dataset Jakarta 15 tahun (2010–2025) dengan fokus PM_{10} untuk meningkatkan kepraktisan dan skalabilitas. Analisis *explainability* menggunakan SHAP (Molnar 2025) diintegrasikan untuk mengidentifikasi faktor-faktor kunci yang mendorong polusi PM_{10} Jakarta, memberikan wawasan untuk pengambilan keputusan kebijakan berbasis bukti.

I.2 Rumusan Masalah

Berdasarkan latar belakang pada subbab I.1, maka ditetapkan rumusan masalah untuk proposal tugas akhir ini adalah: “Bagaimana meningkatkan akurasi prediksi PM_{10} Jakarta melalui *feature engineering* pada model *hybrid* Random Forest-ARIMA, dan bagaimana mengidentifikasi faktor-faktor utama penyebab polusi menggunakan SHAP?”

Model prediksi AQI yang ada hanya menggunakan nilai konsentrasi polutan mentah tanpa mengeksplorasi pola temporal dan statistik (Yenkikar dkk. July 2025), padahal *feature engineering* dapat meningkatkan performa hingga 86% (Naz dkk.

2024). Kesenjangan ini menciptakan peluang untuk mengembangkan model dengan 15 fitur yang menangkap pola temporal, trend, dan variasi musiman Jakarta.

Urgensi penyelesaian masalah ini tinggi karena peningkatan akurasi prediksi PM_{10} akan memungkinkan Dinas Lingkungan Hidup Jakarta untuk *early warning system* yang lebih efektif dan mendukung keputusan kebijakan untuk melindungi 10 juta penduduk dari dampak kesehatan polusi.

Solusi yang diusulkan adalah mengembangkan model *hybrid Random Forest*-ARIMA dengan 15 fitur *engineering* (lag, *rolling statistics*, temporal, interaksi polutan) yang diharapkan mencapai peningkatan akurasi 15–20%, serta menggunakan SHAP untuk mengidentifikasi faktor-faktor kunci yang mendorong polusi PM_{10}.

I.3 Tujuan

Beberapa tujuan dari penelitian tugas akhir ini sebagai berikut:

1. Meningkatkan akurasi prediksi PM_{10} Jakarta melalui pengembangan model *hybrid* Random Forest-ARIMA dengan *feature engineering* sistematis yang menghasilkan peningkatan akurasi minimal 15% dibandingkan baseline dalam hal metrik RMSE.
2. Mengidentifikasi fitur-fitur terpenting yang berkontribusi terhadap prediksi PM_{10} Jakarta menggunakan analisis SHAP untuk memberikan wawasan tentang faktor-faktor utama penyebab polusi.
3. Mendemonstrasikan efektivitas pendekatan *feature engineering* pada dataset Jakarta dengan data historis 15 tahun (2010–2025) untuk membangun model prediksi yang dapat diterapkan dalam sistem peringatan dini kualitas udara.

I.4 Batasan Masalah

Pada bagian ini, dituliskan batasan-batasan masalah yang digunakan secara rinci untuk memperjelas cakupan penelitian yang dilakukan. Batasan-batasan tersebut, antara lain:

1. Prediksi PM_{10} dalam penelitian ini spesifik untuk wilayah DKI Jakarta menggunakan data dari lima stasiun pemantauan kualitas udara.
2. Dataset yang digunakan adalah data historis harian ISPU Jakarta mulai dari Januari 2010 hingga Februari 2025 yang bersumber dari Kaggle.
3. Model prediksi menggunakan lima polutan utama (PM_{10}, SO_2, CO, O_3, NO_2) tanpa mengintegrasikan data eksternal seperti data cuaca, lalu lintas, atau sa-

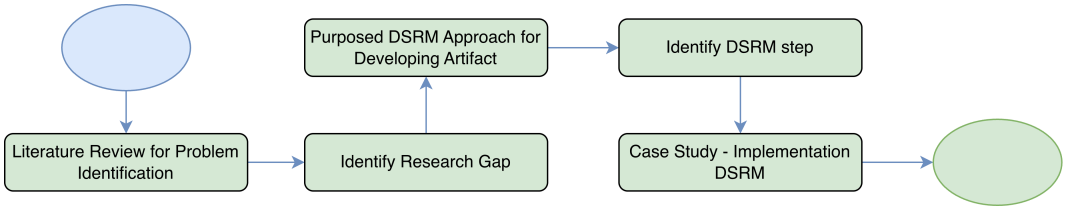
telit.

4. Pengembangan model prediksi tidak mencakup tahap *deployment* ke dalam sistem operasional Dinas Lingkungan Hidup DKI Jakarta, tetapi fokus pada pengembangan dan evaluasi model.
5. Prediksi bersifat prediksi harian (daily forecast) dan tidak mencakup prediksi dengan granularitas lebih tinggi seperti per jam atau per lokasi spesifik.

I.5 Metodologi

Metodologi yang digunakan dalam pelaksanaan penelitian tugas akhir ini adalah *Design Science Research Methodology* (DSRM) seperti pada Gambar I.1. DSRM merupakan kerangka kerja terstruktur untuk mengembangkan dan mengevaluasi artefak berbasis teknologi informasi (Haryanti dkk. 2022). Penelitian ini mengembangkan model *hybrid* Random Forest-ARIMA dari Yenkikar dkk. (July 2025) dengan tiga pengembangan utama:

1. Penerapan pada konteks Jakarta dengan data historis 15 tahun (2010–2025)
2. Penambahan 10 fitur hasil *feature engineering* dari 5 fitur baseline menjadi 15 fitur, mencakup fitur lag, *rolling statistics*, temporal, dan interaksi polutan
3. Fokus pada PM_{10} karena merupakan polutan dominan di Jakarta yang berasal dari sumber mekanis seperti debu jalan, konstruksi, dan keausan kendaraan (Firdaus dkk. September 2023), memiliki dampak signifikan pada sistem respirasi dan kardiovaskular (Zhang dkk. June 2021; Kyung dan Jeong March 2020), serta merupakan salah satu parameter utama dalam regulasi ISPU Indonesia (Kementerian Lingkungan Hidup dan Kehutanan Republik Indonesia July 2020)



Gambar I.1 Metodologi *Design Science Research Methodology* (DSRM) (Haryanti dkk. 2022)

Terdapat lima tahap utama dalam prosesnya yang diadaptasi untuk penelitian ini.

1. *Literature Review for Problem Identification*
   Pada tahap ini, dilakukan kajian literatur untuk mengidentifikasi permasalahan terkait polusi udara Jakarta dan keterbatasan model prediksi yang ada.

Pencarian literatur dilakukan menggunakan basis data Google Scholar, IEEE Xplore, dan PubMed dengan kata kunci "air quality prediction", "PM10 forecasting", "Random Forest ARIMA hybrid", dan "Jakarta air pollution". Literatur difilter berdasarkan relevansi, tahun publikasi (2019–2025), dan kualitas jurnal. Paper Yenkikar dkk. (July 2025) diidentifikasi sebagai *primary methodological foundation* yang akan dikembangkan.

2. *Identify Research Gap*

   Tahap ini bertujuan untuk mengidentifikasi kesenjangan penelitian dari literatur yang telah dikaji. Yenkikar dkk. (July 2025) secara eksplisit menyatakan keterbatasan bahwa model mengecualikan faktor eksternal dan hanya menggunakan 6 fitur polutan mentah tanpa mengeksplorasi pola temporal, fitur statistik, dan interaksi antar polutan. Kesenjangan ini menjadi peluang untuk mengembangkan model dengan pendekatan *feature engineering* sistematis yang dapat meningkatkan akurasi prediksi.

3. *Proposed DSRM Approach for Developing Artifact*

   Pada tahap ini, diusulkan pendekatan pengembangan artefak berupa model prediksi PM_{10} Jakarta yang meningkatkan model baseline (Yenkikar dkk. July 2025). Pengembangan yang dilakukan meliputi ekspansi fitur dari 5 fitur baseline menjadi 15 fitur melalui *feature engineering*, integrasi analisis SHAP untuk *explainability* dan identifikasi faktor kunci polusi Jakarta, serta penerapan pada konteks geografis Jakarta dengan data 15 tahun untuk meningkatkan kepraktisan dan skalabilitas model.

4. *Identify DSRM Step*

   Tahap ini mendefinisikan langkah-langkah detail dalam pengembangan artefak sebagai berikut:

   (a) *Data preprocessing* untuk menangani data hilang dan agregasi multistasiun
   (b) *Feature engineering* untuk menghasilkan 15 fitur
   (c) Pelatihan model *hybrid* Random Forest-ARIMA dengan pembagian data 80:20
   (d) *Hyperparameter tuning* untuk memaksimalkan performa model
   (e) Evaluasi menggunakan metrik RMSE, MAE, dan R^2 dengan target peningkatan akurasi minimal 15% dibandingkan baseline

5. *Case Study – Implementation DSRM*

   Tahap terakhir adalah implementasi melalui studi kasus prediksi PM_{10} Jakarta menggunakan dataset ISPU dari Kaggle (Pohan 2025) dengan rentang waktu Januari 2010 hingga Februari 2025. Perbandingan performa dilakukan antara

model baseline (5 fitur) dan model yang dikembangkan (15 fitur). Analisis SHAP dilakukan untuk mengidentifikasi kontribusi setiap fitur, memberikan wawasan tentang faktor-faktor utama yang mendorong polusi PM_{10} di Jakarta untuk mendukung kebijakan berbasis bukti.

Seluruh source code LaTeX proposal tugas akhir ini tersedia di repository GitHub berikut: https://github.com/rafidhiyaulh/Proposal-Tugas-Akhir

BAB II

Studi Literatur

II.1 Kualitas Udara dan PM_{10}

Kualitas udara merupakan indikator penting kesehatan lingkungan yang dipengaruhi oleh berbagai polutan. Salah satu polutan utama adalah *Particulate Matter* (PM), yaitu partikel padat atau cair yang tersuspensi di udara. PM dikategorikan berdasarkan diameter aerodinamisnya, dengan PM_{10} merujuk pada partikel dengan diameter kurang dari 10 mikrometer dan $PM_{2,5}$ merujuk pada partikel dengan diameter kurang dari 2,5 mikrometer (California Air Resources Board 2025).

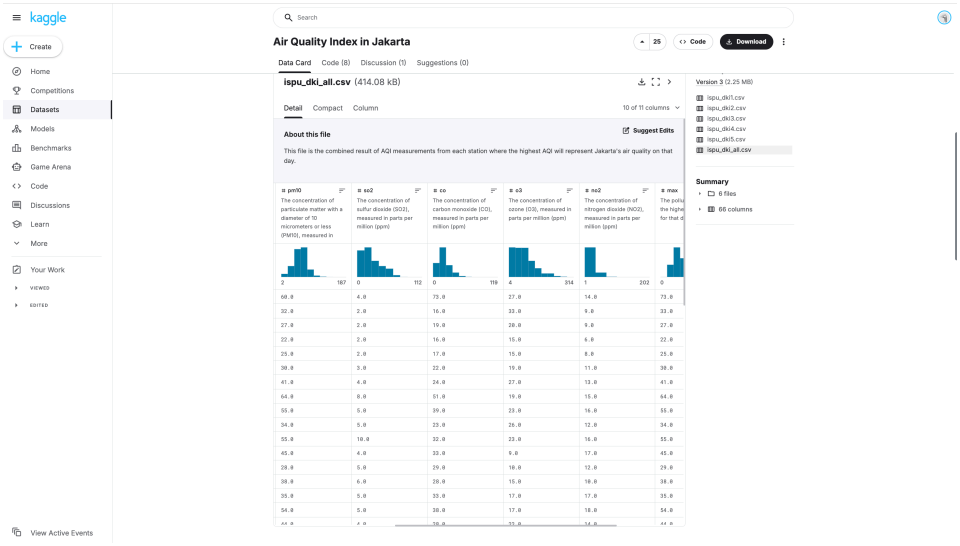
PM_{10} dan $PM_{2,5}$ memiliki karakteristik dan dampak kesehatan yang berbeda seperti ditunjukkan pada Tabel II.1. PM_{10} umumnya berasal dari sumber mekanis seperti debu jalan, aktivitas konstruksi, dan keausan ban kendaraan, sedangkan $PM_{2,5}$ lebih banyak berasal dari proses pembakaran seperti emisi kendaraan dan industri (Firdaus dkk. September 2023).

Tabel II.1 Perbandingan karakteristik PM_{10} dan $PM_{2,5}$

Aspek	**PM_{10}**	**$PM_{2,5}$**
Ukuran	$<$ 10 mikrometer	$<$ 2,5 mikrometer
Sumber utama	Debu jalan, konstruksi, keausan ban	Pembakaran, emisi kendaraan, industri
Penetrasi	Saluran pernapasan atas	Saluran pernapasan dalam, alveoli
Dampak kesehatan	Iritasi, asma, COPD	Kardiovaskular, stroke, kanker paru

Dampak kesehatan PM_{10} telah terdokumentasi dengan baik dalam literatur. Zhang dkk. (June 2021) melaporkan bahwa setiap peningkatan 10 $\mu g/m^3$ konsentrasi PM_{10} berhubungan dengan peningkatan kunjungan unit gawat darurat sebesar 0,14% untuk penyakit kardiovaskular dan 0,56% untuk aritmia. Pada sistem pernapasan,

Kyung dan Jeong (March 2020) menemukan bahwa peningkatan 10 µg/m³ PM_{10} berasosiasi dengan peningkatan 2,7% pada hospitalisasi *Chronic Obstructive Pulmonary Disease* (COPD) dan 1,1% pada mortalitas COPD.



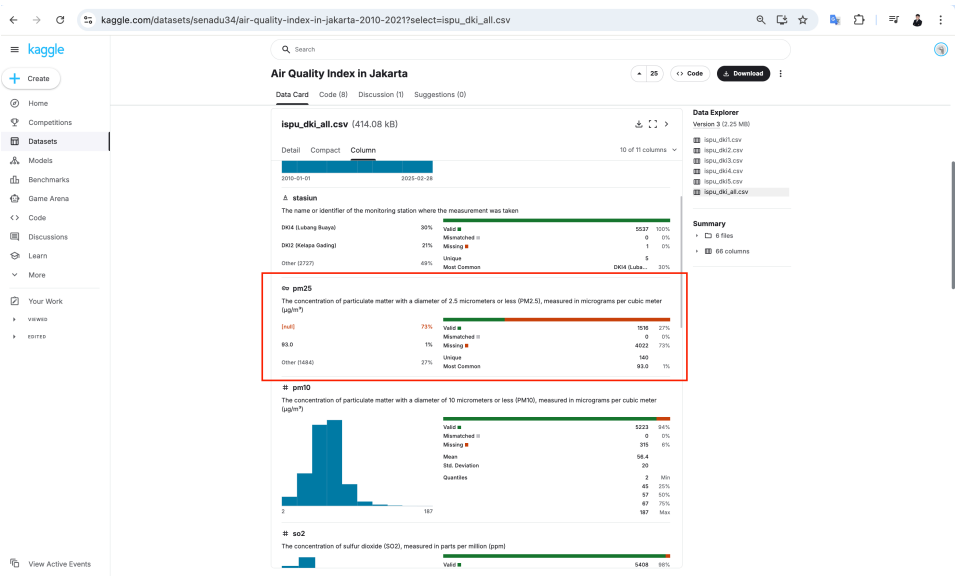
Gambar II.1 Dataset ISPU Jakarta 2010–2025 dari Kaggle menunjukkan distribusi 6 polutan utama

Dataset ISPU Jakarta yang digunakan dalam penelitian ini mencakup periode 2010–2025 yang terdiri dari data temporal (tanggal, stasiun), konsentrasi lima polutan utama (PM_{10}, SO_2, CO, O_3, NO_2), dan nilai ISPU agregat dari lima stasiun pemantauan (DKI1–DKI5). Gambar II.1 menunjukkan distribusi konsentrasi keenam polutan utama dengan PM_{10} menunjukkan variasi yang signifikan (rentang 2–187 µg/m³), mengindikasikan pentingnya prediksi akurat untuk sistem peringatan dini.

Di Jakarta, PM_{10} merupakan polutan dominan yang berasal dari kombinasi lalu lintas padat dengan lebih dari 20 juta kendaraan bermotor, aktivitas konstruksi yang intensif, dan debu jalan akibat kondisi infrastruktur yang bervariasi (Firdaus dkk. September 2023). Kondisi ini menjadikan prediksi PM_{10} sangat relevan untuk sistem peringatan dini kualitas udara Jakarta.

Selain pertimbangan karakteristik dan dampak kesehatan, pemilihan PM_{10} sebagai fokus prediksi juga didasarkan pada ketersediaan data. Seperti ditunjukkan pada Gambar II.2, kolom $PM_{2,5}$ memiliki persentase *missing values* yang sangat tinggi mencapai 73% (hanya 1.516 data valid dari total dataset), sedangkan PM_{10} memiliki kelengkapan data yang jauh lebih baik (94% valid). Keterbatasan data ini menjadikan pemodelan $PM_{2,5}$ tidak memungkinkan untuk dilakukan secara akurat menggunakan metode *data-driven* pada rentang waktu 2010–2025, sehingga penelitian ini

memfokuskan prediksi pada PM_{10} yang memiliki kontinuitas data yang memadai untuk analisis deret waktu jangka panjang.



Gambar II.2 Statistik dataset ISPU Jakarta menunjukkan persentase *missing values* untuk kolom pm25 mencapai 73%

II.2 Indeks Standar Pencemar Udara (ISPU)

Indeks Standar Pencemar Udara (ISPU) adalah angka tanpa satuan yang menggambarkan kondisi mutu udara ambien di lokasi dan waktu tertentu berdasarkan dampak terhadap kesehatan manusia, nilai estetika, dan makhluk hidup lainnya (Kementerian Lingkungan Hidup dan Kehutanan Republik Indonesia July 2020). ISPU diatur dalam Peraturan Menteri Lingkungan Hidup dan Kehutanan Nomor P.14/MENLHK/SETJEN/KUM.1/7/2020 yang mendefinisikan tujuh parameter polutan: PM_{10}, $PM_{2,5}$, SO_2, CO, O_3, NO_2, dan HC.

ISPU berfungsi sebagai bahan informasi kepada masyarakat tentang kualitas udara di lokasi dan waktu tertentu, bahan pertimbangan pemerintah dalam melakukan upaya pengendalian pencemaran udara, dan sebagai sistem peringatan dini (*early warning system*) bagi masyarakat (Chaniago, Zahara, dan Ramadhani September 2020). Kategori ISPU dan dampaknya terhadap kesehatan ditunjukkan pada Tabel II.2.

Tabel II.2 Kategori ISPU dan dampak kesehatan (Septiani June 2024)

Rentang	**Kategori**	**Dampak Kesehatan**
0–50	Baik	Tidak memberikan efek bagi kesehatan
51–100	Sedang	Tidak berpengaruh pada kesehatan manusia atau hewan, namun berpengaruh pada tumbuhan sensitif
101–200	Tidak Sehat	Merugikan manusia atau kelompok yang sensitif
201–300	Sangat Tidak Sehat	Dapat merugikan kesehatan pada sejumlah segmen populasi yang terpapar
$>$300	Berbahaya	Dapat menimbulkan efek kesehatan serius pada seluruh populasi

II.3 *Machine Learning* untuk Prediksi Kualitas Udara

Pendekatan *machine learning* telah banyak diterapkan untuk prediksi kualitas udara karena kemampuannya menangkap pola kompleks dan non-linier dalam data lingkungan. Penelitian ini menggunakan model *hybrid* yang menggabungkan *Random Forest* dan ARIMA untuk memanfaatkan kelebihan kedua pendekatan tersebut.

II.3.1 *Random Forest*

Random Forest adalah algoritma *ensemble learning* yang mengkombinasikan prediksi dari banyak pohon keputusan (*decision trees*) untuk menghasilkan prediksi yang lebih akurat dan *robust* (Arsy dan Yasir April 2025). Setiap pohon dalam *Random Forest* dilatih menggunakan subset acak dari data pelatihan (*bootstrap sampling*) dan subset acak dari fitur pada setiap *split node*.

Untuk tugas regresi, prediksi *Random Forest* dihitung sebagai rata-rata prediksi dari semua pohon dalam ensemble seperti pada Persamaan II.1.

$$\hat{y}_{RF} = \frac{1}{B} \sum_{b=1}^{B} T_b(x) \tag{II.1}$$

Keterangan:

- $\hat{y}_{RF}$ adalah prediksi *Random Forest*
- B adalah jumlah pohon dalam *forest*
- $T_b(x)$ adalah prediksi dari pohon ke-b untuk input x

Keunggulan *Random Forest* untuk prediksi kualitas udara meliputi kemampuan menangani hubungan non-linier antar variabel, *robust* terhadap *outlier* dan *noise*, dapat menangani data dengan banyak fitur tanpa memerlukan seleksi fitur eksplisit, serta memberikan estimasi *feature importance* yang berguna untuk interpretasi (Abdallah dan Elameen June 2025).

II.3.2 ARIMA (*AutoRegressive Integrated Moving Average*)

ARIMA adalah model statistik klasik untuk analisis dan peramalan data *time series*. Model ARIMA cocok untuk data yang menunjukkan pola temporal seperti *trend* dan musiman. ARIMA terdiri dari tiga komponen utama: *AutoRegressive* (AR), *Integrated* (I), dan *Moving Average* (MA) (Yunis, Andri, dan Djoni June 2024).

Model ARIMA(p, d, q) didefinisikan pada Persamaan II.2.

$$\phi(L)(1 - L)^d y_t = \theta(L)\epsilon_t \tag{II.2}$$

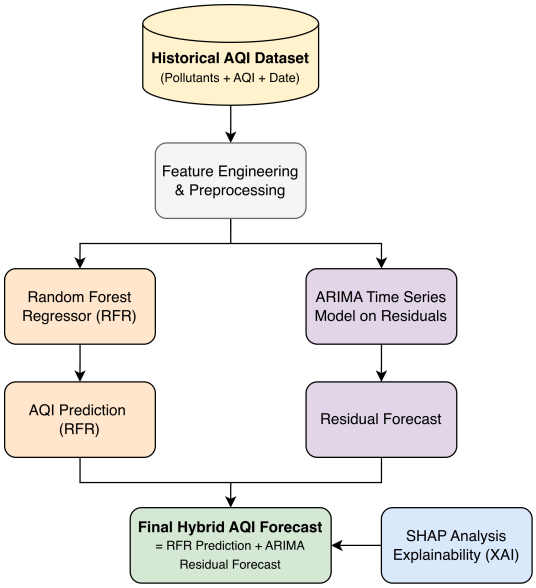
Keterangan:

- y_t adalah nilai *time series* pada waktu t
- L adalah *lag operator* ($Ly_t = y_{t-1}$)
- $\phi(L) = 1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p$ adalah polinomial AR
- $\theta(L) = 1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q$ adalah polinomial MA
- d adalah orde diferensiasi untuk mencapai stasioneritas
- ϵ_t adalah *white noise error*

Sebelum menerapkan ARIMA, data harus diuji stasioneritas menggunakan uji *Augmented Dickey-Fuller* (ADF). Jika data tidak stasioner, diferensiasi dilakukan hingga data menjadi stasioner (Vasconcelos July 2025).

II.3.3 Model *Hybrid Random Forest*-ARIMA

Model *hybrid* menggabungkan *Random Forest* dan ARIMA untuk memanfaatkan kelebihan kedua pendekatan. *Random Forest* efektif menangkap hubungan non-linier antar fitur, sedangkan ARIMA efektif memodelkan pola temporal pada residual (Yenkikar dkk. July 2025). Arsitektur model *hybrid* ditunjukkan pada Gambar II.3.



Gambar II.3 Arsitektur model *hybrid Random Forest*-ARIMA (Yenkikar dkk. July 2025)

Yenkikar dkk. (July 2025) melaporkan bahwa model *hybrid* RF-ARIMA mencapai $R^2 = 0{,}94$ untuk prediksi AQI di India, menunjukkan efektivitas pendekatan ini dalam prediksi kualitas udara.

Proses prediksi *hybrid* terdiri dari empat tahap. Pertama, *Random Forest* dilatih menggunakan fitur polutan dan fitur hasil *engineering* untuk memprediksi nilai PM_{10}. Kedua, residual dihitung sebagai selisih antara nilai aktual dan prediksi *Random Forest*. Ketiga, model ARIMA diterapkan pada residual untuk menangkap pola temporal yang tidak tertangkap *Random Forest*. Keempat, prediksi akhir dihitung sebagai penjumlahan prediksi *Random Forest* dan prediksi residual ARIMA seperti pada Persamaan II.3.

$$\hat{y}_{hybrid} = \hat{y}_{RF} + \hat{r}_{ARIMA} \tag{II.3}$$

Keterangan:

- $\hat{y}_{hybrid}$ adalah prediksi akhir model *hybrid*
- $\hat{y}_{RF}$ adalah prediksi dari *Random Forest*
- $\hat{r}_{ARIMA}$ adalah prediksi residual dari ARIMA

II.4 *Feature Engineering* untuk *Time Series*

Feature engineering adalah proses membuat fitur baru dari data yang ada untuk meningkatkan performa model *machine learning*. Dalam konteks prediksi kualitas udara, *feature engineering* dapat mengekstraksi informasi temporal, statistikal, dan interaksi yang tidak tersedia dalam data mentah (Naz dkk. 2024).

Naz dkk. (2024) menunjukkan bahwa pendekatan *feature engineering* dua tahap dapat meningkatkan performa prediksi polutan udara sebesar 5–86% tergantung pada jenis polutan. Jiménez-Navarro dkk. (December 2024) juga melaporkan bahwa teknik seleksi fitur temporal secara signifikan meningkatkan akurasi prediksi untuk pola *time series* yang kompleks.

Penelitian ini menggunakan 15 fitur yang terdiri dari 5 fitur baseline dan 10 fitur hasil *engineering* seperti ditunjukkan pada Tabel II.3. Fitur lag menangkap persistensi polusi dari hari-hari sebelumnya. Fitur *rolling statistics* menangkap tren jangka pendek dan volatilitas. Fitur temporal menangkap pola musiman Jakarta (musim kemarau April–Oktober cenderung memiliki polusi lebih tinggi). Fitur interaksi CO $\times$ O_3 menangkap aktivitas fotokimia yang berkontribusi pada pembentukan PM_{10} sekunder.

Tabel II.3 Daftar 15 fitur yang digunakan dalam penelitian

Kategori	**Fitur**	**Deskripsi**
Baseline	pm10	Konsentrasi PM_{10} hari ini
Baseline	so2	Konsentrasi SO_2
Baseline	co	Konsentrasi CO
Baseline	o3	Konsentrasi O_3
Baseline	no2	Konsentrasi NO_2
Lag	pm10_lag1	PM_{10} 1 hari sebelumnya
Lag	pm10_lag2	PM_{10} 2 hari sebelumnya
Lag	pm10_lag7	PM_{10} 7 hari sebelumnya
Rolling	pm10_ma3	*Moving average* 3 hari
Rolling	pm10_ma7	*Moving average* 7 hari
Rolling	pm10_std7	Standar deviasi 7 hari
Temporal	month	Bulan (1–12)
Temporal	is_weekend	Akhir pekan (0/1)
Temporal	season	Musim (Hujan/Kemarau)
Interaksi	co_times_o3	Interaksi CO $\times$ O_3

II.5 *Explainable AI* dengan SHAP

Explainable AI (XAI) adalah pendekatan untuk membuat model *machine learning* lebih transparan dan dapat diinterpretasi. SHAP (*SHapley Additive exPlanations*) adalah metode XAI yang didasarkan pada teori permainan kooperatif untuk menjelaskan prediksi individual (Molnar 2025).

SHAP menghitung kontribusi setiap fitur terhadap prediksi menggunakan nilai Shapley seperti pada Persamaan II.4.

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)] \tag{II.4}$$

Keterangan:

- ϕ_i adalah nilai SHAP untuk fitur i
- N adalah himpunan semua fitur
- S adalah subset fitur yang tidak mengandung i
- $f(S)$ adalah prediksi model menggunakan fitur dalam S

Nilai SHAP positif menunjukkan fitur meningkatkan prediksi, sedangkan nilai negatif menunjukkan fitur menurunkan prediksi. Jumlah nilai SHAP semua fitur sama dengan selisih antara prediksi model dan nilai baseline (Awan June 2023).

Dalam konteks prediksi kualitas udara, SHAP memungkinkan identifikasi faktor-faktor utama yang mendorong polusi. Radjabaycolle, Wattimena, dan Pattiradjawane (2025) menunjukkan bahwa integrasi SHAP dengan model prediksi kualitas udara Jakarta memungkinkan identifikasi faktor lingkungan dan antropogenik kunci, mendukung intervensi kebijakan berbasis bukti. Antonini dkk. (September 2024) juga melaporkan bahwa SHAP efektif untuk interpretasi model *machine learning* dalam tugas klasifikasi dan regresi kompleks.

II.6 Penelitian Terdahulu

Beberapa penelitian terdahulu telah mengembangkan model prediksi kualitas udara menggunakan berbagai pendekatan *machine learning*.

Yenkikar dkk. (July 2025) mengembangkan model *hybrid* RF-ARIMA dengan SHAP untuk prediksi AQI di India yang mencapai $R^2 = 0{,}94$. Namun, penelitian tersebut secara eksplisit menyatakan keterbatasan bahwa model mengecualikan faktor eks-

ternal dan hanya menggunakan 6 fitur polutan mentah tanpa mengeksplorasi *feature engineering*.

Naz dkk. (2024) menunjukkan pentingnya *feature engineering* dengan pendekatan dua tahap yang menghasilkan 22 fitur dan meningkatkan performa model LSTM sebesar 5–86%. Chen dkk. (January 2025) mengembangkan model KSC-ConvLSTM untuk prediksi $PM_{2,5}$ di Beijing yang memanfaatkan data spasial dan temporal, namun memerlukan infrastruktur data yang kompleks.

Untuk konteks Jakarta, Arsy dan Yasir (April 2025) mengembangkan model *Random Forest* untuk prediksi $PM_{2,5}$ yang mencapai $R^2 = 0{,}61$, sedangkan Radjabaycolle, Wattimena, dan Pattiradjawane (2025) menerapkan LSTM dengan SHAP untuk prediksi kualitas udara Jakarta. Kedua penelitian Jakarta tersebut belum mengeksplorasi pendekatan *hybrid* RF-ARIMA dengan *feature engineering* sistematis.

II.7 Kesenjangan Penelitian

Berdasarkan kajian literatur, teridentifikasi beberapa kesenjangan penelitian yang menjadi peluang pengembangan.

Pertama, model *hybrid* RF-ARIMA dari Yenkikar dkk. (July 2025) hanya menggunakan 6 fitur polutan mentah tanpa mengeksplorasi *feature engineering*. Penelitian menunjukkan bahwa *feature engineering* dapat meningkatkan performa prediksi secara signifikan (Naz dkk. 2024), namun belum diterapkan pada arsitektur *hybrid* RF-ARIMA.

Kedua, penelitian prediksi kualitas udara Jakarta yang ada belum mengeksplorasi pendekatan *hybrid* RF-ARIMA. Arsy dan Yasir (April 2025) hanya menggunakan *Random Forest* tunggal, sedangkan Radjabaycolle, Wattimena, dan Pattiradjawane (2025) menggunakan LSTM tanpa komponen ARIMA untuk residual.

Ketiga, belum ada penelitian yang menerapkan model *hybrid* RF-ARIMA dengan *feature engineering* sistematis pada dataset Jakarta jangka panjang (lebih dari 10 tahun) untuk prediksi PM_{10}.

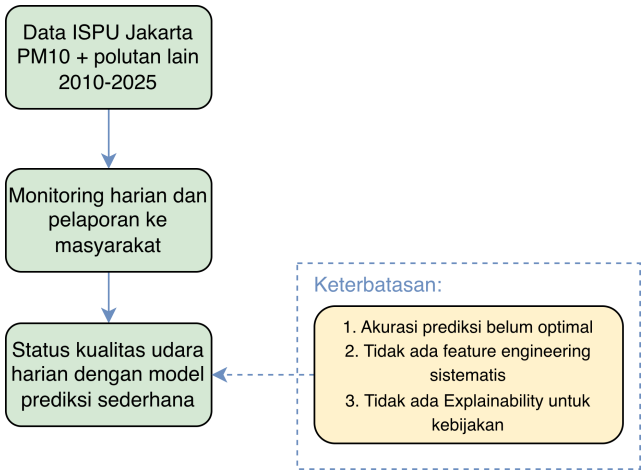
Penelitian ini mengisi kesenjangan tersebut dengan mengembangkan model *hybrid* RF-ARIMA dari Yenkikar dkk. (July 2025) melalui penambahan 10 fitur hasil *feature engineering*, penerapan pada konteks Jakarta dengan data 15 tahun, fokus pada PM_{10} yang relevan dengan sumber polusi Jakarta, dan integrasi SHAP untuk mengidentifikasi faktor kunci polusi yang mendukung kebijakan berbasis bukti.

BAB III

ANALISIS MASALAH

III.1 Analisis Kondisi Saat Ini

Sistem pemantauan kualitas udara Jakarta saat ini dikelola oleh Dinas Lingkungan Hidup DKI Jakarta melalui lima stasiun pemantauan (DKI1–DKI5) yang mengukur konsentrasi polutan secara harian. Data yang dikumpulkan meliputi PM_{10}, SO_2, CO, O_3, dan NO_2 yang kemudian dihitung menjadi nilai Indeks Standar Pencemar Udara (ISPU). Model konseptual sistem pemantauan kualitas udara Jakarta saat ini ditunjukkan pada Gambar III.1.



Gambar III.1 Model konseptual sistem pemantauan kualitas udara Jakarta saat ini

Berdasarkan Gambar III.1, sistem saat ini memiliki alur sebagai berikut: data ISPU Jakarta yang mencakup PM_{10} dan polutan lain dari periode 2010–2025 dikumpulkan dari lima stasiun pemantauan, kemudian dilakukan monitoring harian dan pelaporan ke masyarakat melalui website dan aplikasi KLHK, dan menghasilkan status kualitas udara harian dengan model prediksi sederhana.

Meskipun sudah terdapat beberapa penelitian prediksi kualitas udara Jakarta, sistem yang ada masih memiliki keterbatasan. Pertama, akurasi prediksi belum optimal karena penelitian sebelumnya seperti Arsy dan Yasir (April 2025) hanya mencapai R^2 = 0,61 untuk prediksi $PM_{2,5}$ Jakarta menggunakan *Random Forest* tunggal. Kedua, tidak ada *feature engineering* sistematis karena model yang ada hanya menggunakan fitur polutan mentah tanpa mengeksplorasi pola temporal, statistik, dan interaksi antar polutan (Yenkikar dkk. July 2025). Ketiga, tidak ada *explainability* untuk kebijakan karena model prediksi yang ada belum mengintegrasikan analisis SHAP untuk mengidentifikasi faktor-faktor kunci yang mendorong polusi dan mendukung pengambilan keputusan berbasis bukti.

III.2 Analisis Kebutuhan

III.2.1 Identifikasi Masalah Pengguna

Pengguna utama sistem prediksi PM_{10} Jakarta adalah Dinas Lingkungan Hidup DKI Jakarta dan Dinas Kesehatan DKI Jakarta. Berdasarkan analisis kondisi saat ini, teridentifikasi beberapa masalah yang dihadapi pengguna.

Pertama, Dinas Lingkungan Hidup membutuhkan prediksi PM_{10} yang lebih akurat untuk sistem peringatan dini (*early warning system*) kepada masyarakat. Prediksi yang akurat memungkinkan pemberian peringatan sebelum tingkat polusi mencapai kategori tidak sehat, memberikan waktu bagi masyarakat untuk mengambil tindakan pencegahan.

Kedua, Dinas Kesehatan membutuhkan informasi tentang faktor-faktor utama yang mendorong polusi PM_{10} untuk merumuskan kebijakan kesehatan yang tepat sasaran. Tanpa *explainability*, sulit untuk mengidentifikasi sumber polusi yang paling berpengaruh dan merancang intervensi yang efektif.

Ketiga, kedua dinas membutuhkan model yang dapat memanfaatkan data historis jangka panjang (15 tahun) untuk menangkap pola musiman dan tren polusi Jakarta yang khas, termasuk perbedaan antara musim hujan dan kemarau.

III.2.2 Kebutuhan Fungsional

Berdasarkan identifikasi masalah pengguna, kebutuhan fungsional sistem prediksi PM_{10} Jakarta ditunjukkan pada Tabel III.1.

Tabel III.1 Kebutuhan fungsional sistem prediksi PM_{10}

Kode	Deskripsi Kebutuhan
F1	Sistem harus dapat memprediksi nilai PM_{10} harian untuk DKI Jakarta
F2	Sistem harus menggunakan model *hybrid* Random Forest-ARIMA
F3	Sistem harus memanfaatkan minimal 15 fitur (5 baseline + 10 *feature engineering*)
F4	Sistem harus dapat menghasilkan metrik performa (RMSE, MAE, R^2) pada data uji
F5	Sistem harus memberikan keluaran kontribusi fitur menggunakan SHAP
F6	Sistem harus mampu memproses data historis rentang 2010–2025

III.2.3 Kebutuhan Nonfungsional

Selain kebutuhan fungsional, sistem juga harus memenuhi kebutuhan nonfungsional seperti ditunjukkan pada Tabel III.2.

Tabel III.2 Kebutuhan nonfungsional sistem prediksi PM_{10}

Kode	Deskripsi Kebutuhan
NF1	Akurasi: RMSE model harus minimal 15% lebih rendah dibandingkan baseline
NF2	*Robustness*: Model tidak boleh *overfitting* (divalidasi dengan *train-test split* 80:20)
NF3	Interpretabilitas: Model harus dapat dijelaskan menggunakan SHAP (*global* dan *local*)
NF4	*Reproducibility*: Eksperimen harus dapat direplikasi dengan skrip Python yang terdokumentasi

III.3 Analisis Pemilihan Solusi

III.3.1 Alternatif Solusi

Dalam proposal tugas akhir ini, disusun tiga alternatif model yang dapat menjadi pilihan untuk mengembangkan model prediksi PM_{10} Jakarta. Alternatif model prediktif pertama merupakan model analisis statistik untuk data deret waktu, yaitu ARIMA. Alternatif model prediktif kedua merupakan model pembelajaran mesin berbasis pohon (*tree-based model*), yaitu *Random Forest*. Alternatif model prediktif ketiga adalah model *hybrid* yang menggabungkan *Random Forest* dan ARIMA dengan *feature engineering*.

Model analisis statistik untuk data deret waktu dapat menjadi alternatif yang relevan dan efektif, terutama pada data yang memiliki pola musiman yang konsisten. Dengan memanfaatkan tren dan pola musiman dalam data, model ini mampu memberikan proyeksi yang akurat dan dapat dijelaskan secara transparan (Yunis, Andri, dan Djoni June 2024). Transparansi ini menjadi keunggulan tersendiri dibandingkan dengan beberapa model *machine learning* yang bersifat *black-box*. Dalam kasus prediksi PM_{10} Jakarta, data ISPU menunjukkan pola musiman yang jelas antara musim hujan dan kemarau sehingga model statistik berpotensi memberikan hasil prediksi yang akurat.

Model pembelajaran mesin berbasis pohon juga terbukti dapat menjadi pilihan solusi yang sangat andal untuk melakukan prediksi, termasuk memprediksi data deret waktu. Dalam konteks prediksi PM_{10} Jakarta, model berbasis pohon berpotensi menghasilkan prediksi yang lebih akurat karena terdapat beberapa variabel polutan lain yang mungkin mempengaruhi konsentrasi PM_{10}. Dengan memanfaatkan model berbasis pohon, hubungan non-linier antar variabel dapat ditangkap dengan baik sehingga model dapat melakukan analisis yang lebih menyeluruh (Abdallah dan Elameen June 2025).

Sebagai pilihan yang menggabungkan kedua pendekatan, model *hybrid* Random Forest-ARIMA dengan *feature engineering* dapat menjadi alternatif solusi yang optimal. Yenkikar dkk. (July 2025) menunjukkan bahwa model *hybrid* RF-ARIMA mencapai $R^2 = 0{,}94$ untuk prediksi AQI. Namun, penelitian tersebut hanya menggunakan 6 fitur polutan mentah. Naz dkk. (2024) menunjukkan bahwa *feature engineering* dapat meningkatkan performa model hingga 86%. Dengan demikian, kombinasi *hybrid* RF-ARIMA dengan *feature engineering* berpotensi menghasilkan akurasi yang lebih tinggi.

III.3.2 Analisis Penentuan Solusi

Dalam proses penentuan alternatif solusi yang telah dirumuskan, terdapat lima kriteria yang menjadi bahan pertimbangan melalui metode *Analytical Hierarchy Process* (AHP) pada Tabel III.3, dengan rentang nilai 1–5 (Sangat Buruk – Sangat Baik).

1. Interpretabilitas: Hasil prediksi harus berdasar dan dapat dijelaskan dengan baik untuk mendukung kebijakan.
2. Akurasi: Kesalahan hasil prediksi minimal, diukur dengan RMSE dan R^2.
3. Kemampuan menangkap pola data kompleks: Model harus mampu menangkap hubungan non-linier dan pola temporal.
4. Sumber daya komputasi: Kebutuhan komputasi yang efisien.

5. Kemudahan implementasi: Model dapat diimplementasikan dengan *library* Python yang tersedia.

Tabel III.3 AHP pemilihan alternatif solusi

Kriteria (Bobot)	**ARIMA**	**RF**	***Hybrid* + FE**
Interpretabilitas (0,35)	5	4	4
Akurasi (0,30)	3	4	5
Pola kompleks (0,15)	3	4	5
Komputasi (0,10)	5	4	3
Implementasi (0,10)	5	5	4
Total	**3,95**	**4,10**	**4,40**

Kriteria interpretabilitas menjadi kriteria yang paling diutamakan dalam memprediksi PM_{10} Jakarta. Interpretabilitas informasi yang mempengaruhi hasil prediksi dianggap lebih penting bagi pembuat kebijakan untuk membuat keputusan. Oleh karena itu, model ARIMA mendapatkan nilai 5 karena model statistik deret waktu tersebut biasanya memiliki transparansi yang tinggi. Model *Random Forest* dan *hybrid* mendapatkan nilai 4 karena model-model tersebut dapat diintegrasikan dengan SHAP untuk meningkatkan transparansi (Radjabaycolle, Wattimena, dan Pattiradjawane 2025).

Akurasi menjadi faktor yang tidak terpisahkan dalam hal memprediksi. Model *hybrid* dengan *feature engineering* mendapatkan nilai 5 karena Yenkikar dkk. (July 2025) menunjukkan $R^2 = 0{,}94$ untuk model *hybrid*, dan Naz dkk. (2024) menunjukkan *feature engineering* dapat meningkatkan akurasi hingga 86%. Model *Random Forest* mendapatkan nilai 4 karena memiliki tingkat akurasi yang cukup tinggi (Abdallah dan Elameen June 2025). Di sisi lain, model ARIMA mendapatkan nilai 3 karena tingkat akurasinya cenderung lebih rendah dibandingkan model pembelajaran mesin.

Dalam hal kemampuan menangkap pola data yang kompleks, model ARIMA memperoleh nilai 3 karena model ini cenderung kurang andal ketika harus menghadapi data yang dipengaruhi oleh banyak faktor eksternal. *Random Forest* mendapatkan nilai 4 karena memiliki kemampuan yang lebih baik dalam memahami hubungan non-linier antar variabel serta menangani data dengan banyak fitur secara efisien. Model *hybrid* dengan *feature engineering* mendapatkan nilai 5 karena menggabungkan kelebihan kedua pendekatan dan diperkaya dengan fitur tambahan (Yenkikar dkk. July 2025).

Dari sisi efisiensi sumber daya komputasi, model ARIMA menjadi yang paling ringan karena tidak memerlukan komputasi yang besar, berbeda dengan *Random Fo-*

rest dan terutama model *hybrid* yang membutuhkan lebih banyak daya komputasi. Selain itu, model ARIMA dan *Random Forest* juga unggul dari segi kemudahan implementasi karena dapat langsung digunakan dengan memanfaatkan *library* yang tersedia seperti *statsmodels* dan *scikit-learn*. Model *hybrid* memperoleh nilai 4 karena membutuhkan usaha tambahan dalam menyusun arsitektur gabungan.

Berdasarkan analisis AHP pada Tabel III.3, alternatif model *hybrid* Random Forest-ARIMA dengan *feature engineering* memperoleh skor tertinggi sebesar 4,40 dan terpilih sebagai solusi terbaik. Pemilihan ini didasarkan pada beberapa alasan. Pertama, model *hybrid* memiliki potensi akurasi tertinggi karena menggabungkan kelebihan *Random Forest* dalam menangkap hubungan non-linier dengan ARIMA dalam menangkap pola temporal pada residual, dan diperkaya dengan *feature engineering* untuk mengekstraksi informasi tambahan dari data (Yenkikar dkk. July 2025). Kedua, model *hybrid* tetap memiliki interpretabilitas yang baik karena dapat diintegrasikan dengan SHAP untuk mengidentifikasi kontribusi setiap fitur terhadap prediksi (Radjabaycolle, Wattimena, dan Pattiradjawane 2025). Ketiga, meskipun kompleksitas implementasi lebih tinggi, kompleksitas tersebut masih dapat dikelola dengan ketersediaan *library* Python.

Dengan demikian, penelitian ini akan mengembangkan model *hybrid* Random Forest-ARIMA dengan *feature engineering* untuk prediksi PM_{10} Jakarta, dengan integrasi SHAP untuk *explainability*.

BAB IV

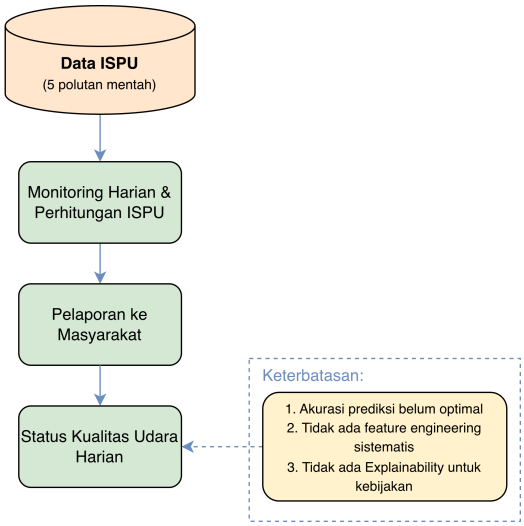
DESAIN KONSEP SOLUSI

IV.1 Diagram Konseptual Sistem

Bagian ini menjelaskan perbandingan sistem prediksi kualitas udara Jakarta sebelum dan sesudah pengembangan model *hybrid* Random Forest-ARIMA dengan *feature engineering*.

IV.1.1 Sistem Sebelum (*Before*)

Sistem pemantauan kualitas udara Jakarta saat ini memiliki alur seperti ditunjukkan pada Gambar IV.1. Data ISPU yang terdiri dari lima polutan mentah (PM_{10}, SO_2, CO, O_3, NO_2) dikumpulkan dari stasiun pemantauan, kemudian dilakukan monitoring harian dan perhitungan nilai ISPU, lalu dilaporkan ke masyarakat sebagai status kualitas udara harian.

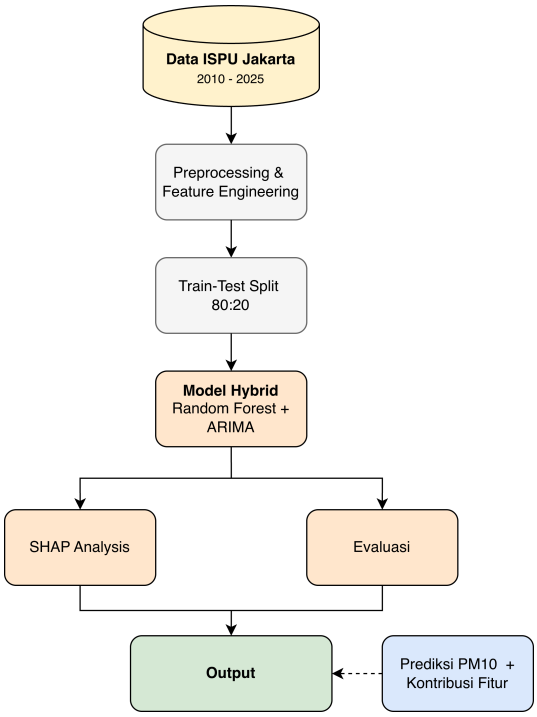


Gambar IV.1 Diagram sistem pemantauan kualitas udara Jakarta saat ini

Berdasarkan Gambar IV.1, sistem saat ini memiliki beberapa keterbatasan. Pertama, akurasi prediksi belum optimal karena tidak ada model *machine learning* yang dioptimasi untuk prediksi PM_{10}. Kedua, tidak ada *feature engineering* sistematis karena hanya menggunakan lima polutan mentah tanpa mengeksplorasi pola temporal dan interaksi antar polutan. Ketiga, tidak ada *explainability* untuk kebijakan karena sistem tidak memberikan informasi tentang faktor-faktor yang mendorong polusi.

IV.1.2 Sistem Sesudah (*After*)

Sistem prediksi PM_{10} Jakarta yang diusulkan memiliki alur seperti ditunjukkan pada Gambar IV.2. Sistem ini mengembangkan model *hybrid* Random Forest-ARIMA dengan *feature engineering* dan integrasi SHAP untuk *explainability*.



Gambar IV.2 Diagram sistem prediksi PM_{10} Jakarta yang diusulkan

Berdasarkan Gambar IV.2, sistem yang diusulkan memiliki enam tahap utama. Pertama, data ISPU Jakarta periode 2010–2025 dikumpulkan sebagai input. Kedua, dilakukan *preprocessing* dan *feature engineering* untuk menghasilkan 15 fitur dari 5 polutan mentah. Ketiga, data dibagi menjadi 80% untuk pelatihan dan 20% untuk pengujian (*train-test split*). Keempat, model *hybrid* Random Forest dan ARIMA dilatih menggunakan data pelatihan. Kelima, dilakukan analisis SHAP untuk mengidentifikasi kontribusi setiap fitur terhadap prediksi. Keenam, model dievaluasi menggunakan metrik RMSE, MAE, dan R^2. Output akhir berupa prediksi PM_{10} harian beserta kontribusi fitur yang dapat digunakan untuk mendukung kebijakan.

IV.1.3 Perbandingan Sistem *Before* dan *After*

Perbandingan antara sistem sebelum dan sesudah pengembangan ditunjukkan pada Tabel IV.1.

Tabel IV.1 Perbandingan sistem sebelum dan sesudah pengembangan

Aspek	**Sistem Sebelum**	**Sistem Sesudah**
Data input	5 polutan mentah	15 fitur (5 baseline + 10 *engineered*)
Model	Tidak ada model ML khusus	*Hybrid* Random Forest-ARIMA
Akurasi target	Belum optimal	Target $R^2 \geq 0{,}85$
Explainability	Tidak ada	Analisis SHAP (*global* dan *local*)
Pola temporal	Tidak dieksploitasi	Fitur lag, *rolling*, dan temporal
Dukungan kebijakan	Tidak ada	Identifikasi faktor kunci polusi

IV.2 Penjelasan Desain Solusi

Bagian ini menjelaskan desain setiap komponen sistem prediksi PM_{10} Jakarta yang diusulkan secara ringkas dan relevan dengan masalah serta kebutuhan yang telah diidentifikasi pada Bab III.

IV.2.1 Desain *Data Preprocessing*

Tahap *data preprocessing* bertujuan untuk menyiapkan data mentah agar dapat digunakan untuk pelatihan model. Desain *preprocessing* mencakup dua proses utama.

Pertama, penanganan data hilang (*missing value handling*) menggunakan metode *rolling mean* dengan *window* 5 hari. Metode ini dipilih karena menjaga kontinuitas pola temporal dan tidak menimbulkan kebocoran data dari masa depan (*data leakage*). Berdasarkan analisis data, terdapat sekitar 5–6% data hilang yang perlu ditangani.

Kedua, agregasi multi-stasiun dilakukan dengan menghitung rata-rata nilai polutan dari lima stasiun pemantauan (DKI1–DKI5) untuk setiap hari. Agregasi ini menghasilkan satu nilai representatif untuk Jakarta secara keseluruhan, menyederhanakan model tanpa kehilangan informasi penting.

Desain ini menjawab kebutuhan F6 (memproses data historis 2010–2025) dan NF2 (*robustness*).

IV.2.2 Desain *Feature Engineering*

Tahap *feature engineering* bertujuan untuk mengekstraksi informasi tambahan dari data mentah yang dapat meningkatkan akurasi prediksi. Desain *feature engineering* menghasilkan 10 fitur tambahan dari 5 fitur baseline.

Fitur lag (pm10_lag1, pm10_lag2, pm10_lag7) menangkap persistensi polusi dari hari-hari sebelumnya. Fitur *rolling statistics* (pm10_ma3, pm10_ma7, pm10_std7) menangkap tren jangka pendek dan volatilitas konsentrasi PM_{10}. Fitur temporal (month, is_weekend, season) menangkap pola musiman Jakarta, dengan musim kemarau (April–Oktober) cenderung memiliki polusi lebih tinggi dibandingkan musim hujan. Fitur interaksi (co_times_o3) menangkap aktivitas fotokimia yang berkontribusi pada pembentukan PM_{10} sekunder.

Desain ini menjawab kebutuhan F3 (memanfaatkan minimal 15 fitur) dan didasarkan pada temuan Naz dkk. (2024) bahwa *feature engineering* dapat meningkatkan performa model hingga 86%.

IV.2.3 Desain Model *Hybrid* Random Forest-ARIMA

Tahap pemodelan menggunakan arsitektur *hybrid* yang menggabungkan *Random Forest* dan ARIMA secara sekuensial. Desain model terdiri dari empat langkah.

Pertama, data dibagi menjadi 80% untuk pelatihan dan 20% untuk pengujian menggunakan pembagian temporal (data terlama untuk pelatihan, data terbaru untuk pengujian). Kedua, *Random Forest* dilatih menggunakan 15 fitur untuk memprediksi nilai PM_{10}. Ketiga, residual (selisih antara nilai aktual dan prediksi *Random Forest*) dihitung dan dimodelkan menggunakan ARIMA untuk menangkap pola temporal yang tidak tertangkap *Random Forest*. Keempat, prediksi akhir dihitung sebagai penjumlahan prediksi *Random Forest* dan prediksi residual ARIMA.

Hyperparameter tuning akan dilakukan menggunakan *Grid Search* dengan *cross-validation* untuk memaksimalkan performa model. Parameter yang akan di-*tuning* meliputi jumlah pohon (*n_estimators*), kedalaman maksimum (*max_depth*), dan *order* ARIMA (p, d, q).

Desain ini menjawab kebutuhan F1 (prediksi PM_{10} harian), F2 (model *hybrid* RF-

ARIMA), F4 (metrik performa), NF1 (akurasi 15% lebih baik), dan NF2 (*train-test split* 80:20).

IV.2.4 Desain *Explainability* dengan SHAP

Tahap *explainability* menggunakan SHAP (*SHapley Additive exPlanations*) untuk menjelaskan kontribusi setiap fitur terhadap prediksi model. Desain SHAP mencakup dua jenis interpretasi.

Interpretasi *global* mengidentifikasi fitur-fitur yang secara umum paling berpengaruh terhadap prediksi PM_{10} Jakarta. Visualisasi yang dihasilkan berupa *summary plot* yang menunjukkan *ranking* kontribusi fitur. Interpretasi *local* menjelaskan faktor-faktor yang mendorong prediksi spesifik pada hari tertentu. Visualisasi yang dihasilkan berupa *force plot* yang menunjukkan kontribusi positif dan negatif setiap fitur.

Informasi kontribusi fitur ini dapat digunakan oleh Dinas Lingkungan Hidup dan Dinas Kesehatan DKI Jakarta untuk mengidentifikasi sumber polusi utama dan merancang intervensi kebijakan yang tepat sasaran.

Desain ini menjawab kebutuhan F5 (kontribusi fitur menggunakan SHAP) dan NF3 (interpretabilitas *global* dan *local*).

BAB V

RENCANA SELANJUTNYA

V.1 Rencana Implementasi

V.1.1 Langkah-langkah Implementasi

Implementasi penelitian prediksi PM_{10} Jakarta dilakukan melalui enam tahap utama dengan timeline total 11 minggu. Tahapan-tahapan tersebut dirancang secara sekuensial untuk memastikan kualitas *research* yang optimal dan *manageability* dari segi waktu dan *resources*.

Tabel V.1 Timeline implementasi penelitian

No.	Tahap	Kegiatan	Durasi
1	Persiapan Data	Pengumpulan dataset ISPU Jakarta dari Kaggle dan KLHK, *exploratory data analysis*	2 minggu
2	Data Preprocessing	Penanganan *missing values*, agregasi multi-stasiun, persiapan data untuk *feature engineering*	1 minggu
3	Feature Engineering	Pembuatan 15 fitur (5 baseline + 10 *engineered*) meliputi lag, *rolling statistics*, temporal, dan interaksi	1.5 minggu
4	Pengembangan Model	*Training* dan *hyperparameter tuning* untuk Random Forest, ARIMA, dan model *hybrid*	3 minggu
5	Integrasi SHAP	Implementasi SHAP untuk *global* dan *local explainability*, visualisasi kontribusi fitur	1.5 minggu
6	Evaluasi & Dokumentasi	Pengujian model, analisis hasil, penulisan laporan akhir	2 minggu
		TOTAL	**11 minggu**

Tahap-tahap ini disusun dengan mempertimbangkan ketergantungan antar tahap dan keseimbangan antara *quality* dan *feasibility* dalam konteks jadwal semester di ITB.

V.1.2 Alat yang Dibutuhkan

Penelitian ini menggunakan perangkat keras, perangkat lunak, dan *libraries* yang telah teruji untuk penelitian *machine learning* dalam konteks *time series forecasting*. Berikut adalah daftar alat yang dibutuhkan:

Tabel V.2 Alat dan bahan penelitian

No.	Kategori	Alat/Bahan	Keterangan
1	Perangkat Keras	MacBook Air M3 2024	Laptop dengan spesifikasi: CPU Apple M3, RAM 16 GB, Storage 512 GB. Cukup untuk *testing* dan dokumentasi lokal.
2	Bahasa Pemrograman	Python 3.10+	Bahasa pemrograman untuk implementasi model, dipilih karena ekosistem *library machine learning* yang lengkap.
3	Perangkat Lunak - Data	pandas, NumPy	*Library* untuk manipulasi data dan komputasi numerik. Pandas untuk *structured data handling*, NumPy untuk operasi *array*.
4	Perangkat Lunak - Visualisasi	Matplotlib, Seaborn	*Library* untuk visualisasi data dan hasil analisis. Matplotlib untuk *basic plotting*, Seaborn untuk *statistical visualizations*.
5	Perangkat Lunak - Time Series	statsmodels	*Library* untuk analisis deret waktu dan pemodelan ARIMA, menyediakan API lengkap dan *diagnostic tools*.
6	Perangkat Lunak - ML	scikit-learn, XGBoost	scikit-learn untuk implementasi Random Forest. XGBoost digunakan sebagai alternatif *baseline model*.

Bersambung ke halaman berikutnya

Tabel V.2 Alat dan bahan penelitian (lanjutan)

No.	Kategori	Alat/Bahan	Keterangan
7	Perangkat Lunak - XAI	SHAP	*Library* untuk interpretasi model menggunakan Shapley values, menyediakan *global* dan *local explainability*.
8	IDE	Jupyter Notebook	Lingkungan interaktif untuk pengembangan, pengujian, dan dokumentasi. Dijalankan via Google Colab Pro.
9	Version Control	Git, GitHub	Digunakan untuk *version control* dan kolaborasi. Seluruh kode akan disimpan di *repository* GitHub.
10	Cloud Computing	Google Colab Pro	Lingkungan komputasi awan berbayar dengan GPU yang lebih cepat dan *reliable*, 100 unit komputasi per bulan. Digunakan untuk *training* model dengan estimasi waktu 20–30 menit per iterasi.
11	Dataset	ISPU Jakarta 2010–2025	Dataset dari Kaggle (2010–2021) yang dilengkapi dengan data 2021–2025 dari KLHK API sehingga mencakup periode 2010–2025.

Pemilihan *tools* tersebut didasarkan pada beberapa pertimbangan utama:

1. Semua *software* adalah *open-source* dan gratis, tidak ada biaya lisensi yang membebani penelitian (kecuali Google Colab Pro untuk *training acceleration*).
2. Ekosistem yang *mature* dan *well-documented*, dengan komunitas pengguna yang besar sehingga memudahkan *troubleshooting* dan *knowledge sharing*.
3. *Support* untuk *Mac architecture* (Apple Silicon M3), memastikan kompatibilitas dengan *hardware* yang digunakan tanpa perlu *dual-boot* atau *virtual*

machine.

4. *Best practices* dalam industri untuk *research* serupa, *tools* ini sudah digunakan di berbagai publikasi penelitian kualitas udara dan *time series forecasting*.

Pemanfaatan Google Colab Pro memungkinkan *training* model dengan GPU yang *reliable* dan *efficient*, tanpa membebani *resources* MacBook. Internet yang digunakan sudah termasuk dalam biaya kosan dan akses ITB, sehingga tidak ada biaya tambahan untuk konektivitas. Dengan demikian, seluruh *pipeline research* dapat berjalan dengan performa optimal.

V.1.3 Analisis Biaya Implementasi

Biaya pengembangan model mencakup perangkat keras (MacBook Air M3 sudah dimiliki), perangkat lunak (semua *open-source* gratis), dan Google Colab Pro untuk *accelerated computing*. Internet yang digunakan sudah termasuk dalam biaya kosan dan akses ITB, sehingga tidak ada biaya internet tambahan.

Google Colab Pro dipilih untuk memastikan *training* model yang *reliable* dan *efficient*, dengan GPU yang lebih cepat, 100 unit komputasi per bulan, dan waktu *idle timeout* yang lebih panjang. Dengan *resources* ini, estimasi waktu *training* berkurang menjadi 20–30 menit per iterasi, memastikan timeline 11 minggu tetap *feasible*.

Model dirancang sebagai *open-source* dan dapat diimplementasikan secara gratis di server lokal Dinas Lingkungan Hidup DKI Jakarta, tanpa biaya lisensi atau *subscription*. Dengan pendekatan ini, tidak ada *barrier* finansial untuk adopsi model oleh pemerintah, sejalan dengan tujuan *social impact* dari penelitian ini.

V.2 Rencana Evaluasi

V.2.1 Metode Pengujian

Model prediksi PM_{10} akan diuji menggunakan metodologi yang *rigorous* untuk memastikan validitas hasil dan generalisasi ke data baru. Metodologi pengujian mencakup beberapa komponen utama.

Pertama, ***train-test split* 80:20 temporal**. Data dibagi menjadi 80% untuk *training* (periode 2010–2023, $\sim$4.430 hari) dan 20% untuk *testing* (periode 2024–2025, $\sim$1.108 hari). Pembagian dilakukan secara temporal (tidak *random*) untuk menghindari *data leakage* dan mencerminkan *real-world scenario* di mana model diprediksi untuk masa depan.

Tabel V.3 Estimasi biaya implementasi

No.	Kategori	Item	Biaya	Catatan
1	Pengembangan	Hardware (MacBook M3)	Rp 0	Sudah dimiliki
2	Pengembangan	Software tools (Python, libraries)	Rp 0	*Open-source*, gratis selamanya
3	Pengembangan	Google Colab Pro	Rp 170K/bulan	*Reliable* GPU, 100 unit komputasi, *training* lebih cepat
4	Pengembangan	Internet	Rp 0	Sudah termasuk biaya kosan dan akses ITB
5	Implementasi	Biaya *deployment* pengguna	Rp 0	*Open-source, local installation*
		TOTAL	**Rp 470K**	**Untuk 11 minggu *development***

Kedua, **perbandingan dengan *baseline***. Model yang diusulkan (*Hybrid* RF-ARIMA + *Feature Engineering*) akan dibandingkan dengan tiga *baseline*:

1. Baseline 1: ARIMA tunggal (tanpa Random Forest)
2. Baseline 2: Random Forest tunggal (tanpa ARIMA, dengan 5 fitur *baseline* saja)
3. Baseline 3: *Hybrid* RF-ARIMA tanpa *feature engineering*, sesuai (Yenkikar dkk. July 2025)

Ketiga, **metrik evaluasi**. Performa model dievaluasi menggunakan tiga metrik:

1. **RMSE** (*Root Mean Square Error*): Mengukur *magnitude* dari *error*, *sensitive* terhadap *outliers*.
2. **MAE** (*Mean Absolute Error*): Mengukur *average absolute error*, lebih *robust* terhadap *outliers*.
3. **R^2** (*Coefficient of Determination*): Mengukur proporsi varians yang dijelaskan oleh model (target: $\geq 0{,}85$).

Keempat, **analisis residual**. Residual model dianalisis untuk:

1. Uji normalitas menggunakan uji *Shapiro-Wilk* untuk memverifikasi bahwa residual terdistribusi normal.
2. Uji autokorelasi menggunakan plot ACF/PACF dan uji *Ljung-Box* untuk mendeteksi korelasi temporal yang tersisa dalam residual.
3. Deteksi *overfitting* dengan membandingkan metrik (RMSE, MAE, R^2) antara

data *training* dan *test* untuk memastikan model tidak hanya menghafal data *training*.

V.2.2 Kriteria Keberhasilan

Kesuksesan penelitian dievaluasi berdasarkan lima kriteria yang terukur dan objektif, yang semuanya terkait dengan kebutuhan fungsional (F1-F6) dan nonfungsional (NF1-NF4) yang telah didefinisikan pada Bab III.

Tabel V.4 Kriteria keberhasilan penelitian

No.	Kriteria	Target	Keterangan
1	RMSE model *hybrid* lebih rendah dari *baseline* terbaik	$\geq$ 15% *improvement*	Menjawab kebutuhan NF1 (akurasi)
2	R^2 model *hybrid*	$\geq$ 0,85	Menjawab kebutuhan F1 dan F4 (prediksi akurat)
3	Model tidak *overfitting*	Selisih R^2 *train-test* $<$ 0,10	Menjawab kebutuhan NF2 (*robustness*)
4	SHAP mengidentifikasi *top* 5 fitur penting	Ya, teridentifikasi dengan jelas	Menjawab kebutuhan F5 dan NF3 (*explainability*)
5	Eksperimen dapat direplikasi	Skrip Python terdokumentasi di GitHub	Menjawab kebutuhan NF4 (*reproducibility*)

Kelima kriteria ini dirancang agar *objective*, *measurable*, dan *achievable* dalam timeline 11 minggu dengan *resources* yang tersedia. Jika semua kriteria terpenuhi, penelitian dianggap berhasil mengembangkan model prediksi PM_{10} Jakarta yang akurat, *robust*, dan *interpretable*.

V.3 Analisis Risiko

Bagian ini mengidentifikasi risiko potensial yang mungkin menghambat penelitian beserta strategi mitigasi yang realistis dan *actionable*.

Tabel V.5 Analisis risiko dan strategi mitigasi

No.	Risiko	Penyebab	Dampak	Mitigasi
1	Keterbatasan *hardware* untuk *training*	Proses *training* model kompleks memerlukan komputasi besar; MacBook M3 hanya mengandalkan CPU	*Training* berlangsung lama, berpotensi melewati jadwal yang direncanakan	Menggunakan Google Colab Pro dengan dukungan GPU yang *reliable*. Colab Pro menyediakan *resources* yang cukup untuk model ini dengan estimasi *training* 20–30 menit per iterasi.
2	*Data quality issues* (*missing values* terlalu banyak)	Data ISPU dari berbagai sumber mungkin tidak lengkap atau tidak konsisten	Akurasi model menurun dan hasil prediksi menjadi kurang dapat dipercaya	Menggunakan beberapa metode imputasi (*rolling mean*, interpolasi). Jika persentase data hilang $>10\%$, menambahkan data dari BMKG atau stasiun cuaca alternatif sebagai pendukung.

Bersambung ke halaman berikutnya

Tabel V.5 Analisis risiko dan strategi mitigasi (lanjutan)

No.	Risiko	Penyebab	Dampak	Mitigasi
3	Model tidak mencapai target akurasi $R^2 \geq 0{,}85$	*Feature selection* kurang optimal atau *hyperparameter tuning* tidak konvergen	Kriteria keberhasilan tidak terpenuhi, sehingga perlu revisi desain atau timeline	Melakukan *intensive hyperparameter tuning* menggunakan *Grid Search*. Mencoba kombinasi *hyperparameter* yang berbeda (misalnya jumlah pohon RF, *order* ARIMA p-d-q). Mencoba metode *ensemble* lain seperti *Stacking* atau *Voting* menggunakan kombinasi RF, ARIMA, dan XGBoost. Jika tetap tidak tercapai, dokumentasikan sebagai *limitation* dan investigasi *root cause* di hasil penelitian.
4	Komunikasi dengan pembimbing terhambat	Jadwal pembimbing padat, waktu respon lambat, atau terjadi miskomunikasi terkait arah penelitian	*Progress* penelitian melambat, muncul banyak *rework*, dan *deadline* berpotensi terlewat	Menjadwalkan pertemuan rutin (misalnya dua minggu sekali). Mengirimkan *progress report* tertulis melalui GitHub atau email untuk memungkinkan umpan balik asinkron. Menjaga dokumentasi yang rapi agar diskusi lebih efektif.

Bersambung ke halaman berikutnya

Tabel V.5 Analisis risiko dan strategi mitigasi (lanjutan)

No.	Risiko	Penyebab	Dampak	Mitigasi
5	Timeline tidak realistis	Adanya *unexpected issues* (*bug*, masalah data), atau *underestimation effort*	*Deadline* terlewat dan kualitas hasil penelitian menurun	Menyediakan *buffer* waktu sekitar 20% di setiap tahap. Memprioritaskan *core deliverables*, sedangkan fitur tambahan dapat ditunda. Menerapkan pendekatan *Agile* dengan *weekly sprint review* untuk memantau progres dan menyesuaikan rencana.

Semua risiko yang teridentifikasi memiliki strategi mitigasi yang *concrete* dan *feasible* dengan *resources* yang ada. Strategi tersebut dirancang agar *proactive* (mencegah risiko terjadi) dan *reactive* (*handle* jika risiko terjadi). *Monitoring* berkelanjutan terhadap risiko akan dilakukan setiap minggu untuk memastikan *early warning* dan *quick corrective actions* jika diperlukan.

DAFTAR PUSTAKA

Abdallah, Asma Salah Aldeen Mohammed, dan Islam Hamad Eljack Elameen. June 2025. "Predicting PM2.5 Levels in Seoul Using Random Forest Regression". *International Journal of Engineering and Science Invention* 14, no. 6 (): 45–48. ISSN: 2319-6734. https://doi.org/10.35629/6734-14064548. https://ijesi.org/papers/Vol(14)i6/14064548.pdf.

Antonini, Antonella S., Juan Tanzola, Lucía Asiain, Gabriela R. Ferracutti, Silvia M. Castro, Ernesto A. Bjerg, dan María Luján Ganuza. September 2024. "Machine Learning model interpretability using SHAP values: Application to Igneous Rock Classification task". *Applied Computing and Geosciences* 23 (): 100178. https://doi.org/10.1016/j.acags.2024.100178. https://www.sciencedirect.com/science/article/pii/S2590197424000259.

Arsy, Muhammad Naufal Afif Al, dan Ahmad Meijlan Yasir. April 2025. "PM2.5 Concentration Prediction Model in Jakarta Area Using Random Forest Algorithm". *Journal of Computation Physics and Earth Science* 5, no. 1 (): 31–39. https://journal.physan.org/index.php/jocpes/article/view/52.

Aulia, Salza Afifa. 2024. "Air Pollution Threatens Health and Climate Change in Jakarta". *Human Error and Safety* 1 (1): 36–46. https://doi.org/10.61511/hes.v1i1.2024.648. https://journal-iasssf.com/index.php/HES/article/download/648/562/4393.

Awan, Abid Ali. June 2023. "An Introduction to SHAP Values and Machine Learning Interpretability". DataCamp. Accessed November 21, 2025. https://www.datacamp.com/tutorial/introduction-to-shap-values-machine-learning-interpretability.

California Air Resources Board. 2025. “Inhalable Particulate Matter and Health ($PM_{2.5}$ and PM_{10})”. California Air Resources Board. Accessed November 21, 2025. https://ww2.arb.ca.gov/resources/inhalable-particulate-matter-and-health.

Chaniago, Dasrul, Annisa Zahara, dan Indah Suci Ramadhani. September 2020. “Indeks Standar Pencemar Udara (ISPU) sebagai Informasi Mutu Udara Ambien di Indonesia”. Direktorat Pengendalian Pencemaran Udara KLHK. Accessed November 21, 2025. https://ditppu.menlhk.go.id/portal/read/indeks-standar-pencemar-udara-ispu-sebagai-informasi-mutu-udara-ambien-di-indonesia.

Chen, Gang, Shen Chen, Dong Li, dan Cai Chen. January 2025. “A hybrid deep learning air pollution prediction approach based on neighborhood selection and spatio-temporal attention”. *Scientific Reports* 15, no. 3685 (). https://doi.org/10.1038/s41598-025-88086-1. https://www.nature.com/articles/s41598-025-88086-1.

Firdaus, Fadhil Muhammad, Beth Elliott, Jorie Malsch, dan Paulista Surjadi. September 2023. “7 Things to Know About Jakarta’s Air Pollution Crisis”. WRI Indonesia. Accessed November 20, 2025. https://wri-indonesia.org/en/insights/7-things-know-about-jakartas-air-pollution-crisis.

Haryanti, Tining, Nur Aini Rakhmawati, Apol Pribadi Subriadi, dan Aris Tjahyanto. 2022. “The Design Science Research Methodology (DSRM) for Self-Assessing Digital Transformation Maturity Index in Indonesia”. Dalam *Proceedings of the IEEE,* 1–6. IEEE. https://repository.um-surabaya.ac.id/id/eprint/10034/2/4.%20IEEE%20Prociding%20(DSRM).pdf.

Jiménez-Navarro, Manuel J., María Martínez-Ballesteros, Mario Lovrić, Simonas Kecorius, dan Emmanuel Karlo Nyarko. December 2024. “Explainable Deep Learning on Multi-Target Time Series Forecasting: An Air Pollution Use Case”. *Results in Engineering* 24 (): 103290. https://doi.org/10.1016/j.rineng.2024.103290. https://www.sciencedirect.com/science/article/pii/S2590123024015445.

Kementerian Lingkungan Hidup dan Kehutanan Republik Indonesia. July 2020. *Peraturan Menteri Lingkungan Hidup dan Kehutanan Republik Indonesia Nomor P.14/MENLHK/SETJEN/KUM.1/7/2020 tentang Indeks Standar Pencemar Udara.* Berita Negara Republik Indonesia Tahun 2020 Nomor 774. Diakses 21 November 2025. https://jdih.menlhk.go.id/new2/uploads/files/P_14_2020_ISPU_menlhk_07302020074834.pdf.

Kyung, Sun Young, dan Sung Hwan Jeong. March 2020. "Particulate-Matter Related Respiratory Diseases". *Tuberculosis and Respiratory Diseases* 83, no. 2 (): 116–121. https://doi.org/10.4046/trd.2019.0025. https://pmc.ncbi.nlm.nih.gov/articles/PMC7105434/.

Molnar, Christoph. 2025. "Interpretable Machine Learning: A Guide for Making Black Box Models Explainable". Chapter SHAP (SHapley Additive exPlanations), 2nd. Accessed November 21, 2025. Leanpub. https://christophm.github.io/interpretable-ml-book/shap.html.

Naz, Fareena, Muhammad Fahim, Adnan Ahmad Cheema, Nguyen Trung Viet, Tuan-Vu Cao, dan Ruth Hunter. 2024. "Two-Stage Feature Engineering to Predict Air Pollutants in Urban Areas". *IEEE Access* 12:114073–114085. https://doi.org/10.1109/ACCESS.2024.3443810. https://ieeexplore.ieee.org/document/10636142.

Pohan, Taufiq. 2025. *Air Quality Index in Jakarta.* Kaggle. Accessed November 21, 2025. Dataset contains daily AQI/ISPU values from January 2010 to February 2025. https://www.kaggle.com/datasets/senadu34/air-quality-index-in-jakarta-2010-2021.

Pusat Krisis Kesehatan Kementerian Kesehatan Republik Indonesia. July 2019. "Beware of Poor Air Quality Escalation". Kementerian Kesehatan Republik Indonesia. Accessed November 21, 2025. https://pusatkrisis.kemkes.go.id/beware-of-poor-air-quality-escalation.

Radjabaycolle, Jefri E. T., Emanuella M. C. Wattimena, dan Victor Eric Pattiradjawane. 2025. "Improving Air Quality Forecasts with LSTM and SHAP Explainability: A Case Study in Jakarta". *Journal of Embedded System Security and Intelligent Systems* 6 (3): 337–347. ISSN: 2722-273X. https://journal.unm.ac.id/index.php/JESSI/article/view/9512.

Septiani, Faradila. June 2024. “Combating Air Pollution in Indonesia: What You Need to Know About ISPU Regulation”. MUSA Green. Accessed November 21, 2025. https://www.musagreen.com/2024/06/20/combating-air-pollution-in-indonesia-what-you-need-to-know-about-ispu-regulation/.

Vasconcelos, Izairton. July 2025. “Stationarity in Time Series: The Power of ADF and KPSS Tests for ARIMA Models”. Developer Service Blog. Accessed November 21, 2025. https://developer-service.blog/stationarity-in-time-series-the-power-of-adf-and-kpss-tests-for-arima-models/.

Yenkikar, Anuradha, Ved Prakash Mishra, Manish Bali, dan Tabassum Ara. July 2025. “Explainable Forecasting of Air Quality Index Using a Hybrid Random Forest and ARIMA Model”. *MethodsX* 15 (): 103517. https://doi.org/10.1016/j.mex.2025.103517. https://pmc.ncbi.nlm.nih.gov/articles/PMC12329590/.

Yunis, Roni, Andri, dan Djoni. June 2024. “Hybridization Model for Air Pollution Prediction Using Time Series Data”. *COGITO Smart Journal* 10, no. 1 (): 1–14. https://cogito.unklab.ac.id/index.php/cogito/article/view/619.

Zhang, Yifan, Yuxia Ma, Fengliu Feng, Bowen Cheng, Hang Wang, Jiahui Shen, dan Haoran Jiao. June 2021. “Association between PM_{10} and specific circulatory system diseases in China”. *Scientific Reports* 11, no. 12129 (). https://doi.org/10.1038/s41598-021-91637-x. https://www.nature.com/articles/s41598-021-91637-x.