

BAB I

PENDAHULUAN

A. Latar Belakang

Polusi udara telah menjadi salah satu ancaman lingkungan dan kesehatan masyarakat paling mendesak di tingkat global. Menurut *World Health Organization* (WHO), polusi udara menyebabkan sekitar 7 juta kematian dini setiap tahun, dengan beban terbesar dialami oleh negara berkembang di Asia dan Afrika. Di antara berbagai polutan atmosfer, *Particulate Matter* dengan diameter aerodinamis kurang dari 2,5 mikrometer (PM_{2.5}) dianggap paling berbahaya karena ukurannya yang sangat kecil memungkinkannya menembus hingga alveoli paru-paru dan masuk ke aliran darah. Komponen PM_{2.5} dalam polusi udara bertanggung jawab atas diperkirakan 4,2 juta kematian prematur secara global pada tahun 2015, terutama akibat penyakit kardiovaskular, gangguan pernapasan kronis, dan kanker paru-paru [1]. WHO menetapkan ambang batas panduan harian PM_{2.5} sebesar 15 µg/m³ sebagai upaya melindungi kesehatan publik. Namun, realitas di wilayah perkotaan Indonesia jauh melampaui standar tersebut. Studi perbandingan kualitas udara di kota-kota Indonesia menunjukkan bahwa Tangerang Selatan lokasi pengambilan data penelitian ini masuk dalam kategori “tidak sehat”, dengan konsentrasi PM_{2.5} rata-rata mencapai 97,2 µg/m³ pada jam puncak pagi (07.00–08.00), atau lebih dari 6 kali lipat batas harian WHO [2]. Penelitian lapangan di Jakarta Barat mengungkap bahwa konsentrasi PM_{2.5} mencapai 72,71 µg/m³ di permukiman sekitar jalan raya padat, dengan *Risk Quotient* (RQ) tertinggi sebesar 1,74, yang mengindikasikan adanya risiko kesehatan nyata bagi masyarakat terpapar [3]. Kesenjangan tajam antara standar kesehatan global dan paparan riil di wilayah tropis perkotaan ini menegaskan urgensi pengembangan sistem prediksi yang tidak hanya akurat, tetapi juga sensitif terhadap lonjakan episodik PM_{2.5} yang kritis bagi kesehatan masyarakat.

Serpong, Tangerang Selatan, dipilih sebagai lokasi studi kasus karena merepresentasikan tipikal kawasan tropis perkotaan yang mengalami dinamika polusi udara kompleks. Di satu sisi, kualitas udara latar relatif baik dalam kondisi normal, namun di sisi lain, sangat rentan terhadap lonjakan episodik PM_{2.5} akibat transportasi lintas batas, khususnya selama musim kemarau ketika kabut asap dari

kebakaran hutan dan lahan di Sumatra dan Kalimantan menyebar ke wilayah Jabodetabek [4]. Karakteristik ini menghasilkan data $PM_{2.5}$ yang sangat tidak seimbang (*imbalanced*), di mana hari-hari dengan konsentrasi "normal" (di bawah ambang batas) mendominasi sepanjang tahun, sementara peristiwa polusi ekstrem meski jarang dapat menyebabkan konsentrasi $PM_{2.5}$ melonjak hingga $75,6 \mu g/m^3$, jauh melampaui baku mutu harian nasional sebesar $65 \mu g/m^3$ [4].

Analisis risiko kesehatan di kawasan KST BJ Habibie Serpong mengungkap bahwa meskipun nilai RQ rata-rata masih <1 , paparan jangka pendek terhadap $PM_{2.5}$ episodik tetap berpotensi memicu gangguan pernapasan akut, terutama pada populasi sensitif seperti petugas keamanan yang menghabiskan >12 jam/hari di luar ruangan [4]. Temuan penelitian menunjukkan bahwa nilai RQ $PM_{2.5}$ untuk seluruh responden berada di bawah 1 ($RQ \leq 1$) dengan rata-rata sebesar 0,13, namun responden dengan durasi paparan lebih lama, misalnya 13 tahun, mencatatkan RQ tertinggi sebesar 0,28, yang meskipun masih di bawah 1, mengindikasikan adanya akumulasi risiko seiring waktu [4]. Fenomena "jarang tapi kritis" ini konsisten dengan pola yang teramati di wilayah Jabodetabek secara umum, sebagaimana ditunjukkan oleh sebuah studi di Jakarta Pusat selama periode Maret 2020 yang melaporkan bahwa konsentrasi $PM_{2.5}$ berada pada rentang $15-55 \mu g/m^3$ dan masih di bawah baku mutu udara ambien nasional untuk 24 jam sebesar $65 \mu g/m^3$, namun rentang tersebut menunjukkan bahwa konsentrasi latar tetap berada pada tingkat yang cukup dekat dengan ambang batas dan tidak mencerminkan potensi lonjakan episodik yang dapat terjadi di luar periode pengamatan [5]. Kondisi ini menegaskan perlunya pendekatan yang tidak hanya akurat secara umum, tetapi juga sensitif terhadap kelas minoritas suatu kemampuan yang sangat dibutuhkan untuk sistem peringatan dini di kawasan rawan asap seperti Serpong.

Menghadapi kompleksitas data $PM_{2.5}$ yang bersifat episodik dan tidak seimbang di wilayah tropis, pendekatan berbasis *machine learning* telah terbukti mampu menangkap hubungan non-linear antara polutan, meteorologi, dan konsentrasi $PM_{2.5}$. Di antara berbagai algoritma, *XGBoost* menonjol karena kemampuannya dalam menangani data berdimensi tinggi, toleransi terhadap nilai hilang, serta akurasi prediktif yang konsisten pada skenario perkotaan [6]. Kinerja optimal *XGBoost* sangat bergantung pada proses tuning hiperparameter dan seleksi

fitur yang sistematis, langkah yang sering kali diabaikan dalam konfigurasi default dan berisiko menyebabkan *underestimation* pada hari-hari polusi tinggi, terutama dalam konteks data *imbalanced*, di mana jumlah sampel dengan konsentrasi ekstrem sangat terbatas sehingga menghambat kapasitas pembelajaran model [7]. Untuk mengatasi keterbatasan tersebut, rekayasa fitur berbasis lag menjadi strategi krusial. Dengan menambahkan nilai historis PM_{2.5} (misalnya 30 menit, 1,5 jam, dan 12 jam sebelumnya) sebagai fitur *input*, model memperoleh “memori temporal” yang memungkinkannya menangkap autokorelasi jangka pendek serta pola akumulasi partikulat selama periode prakondisi *haze* suatu dinamika temporal yang khas di wilayah tropis [8]. Tanpa representasi temporal semacam ini, model cenderung mengandalkan hanya pada variabel meteorologi sesaat, yang sering kali tidak cukup untuk mendeteksi prekursor episodik. Kuatnya peran *fitur lag* ini juga dikonfirmasi secara empiris oleh Chimla et al. (2025), yang menemukan bahwa variabel PM_{2.5_lag1} secara konsisten menjadi prediktor paling berpengaruh berdasarkan analisis *Random Forest feature importance* dan SHAP dalam memprediksi PM_{2.5} di wilayah dengan karakteristik polusi episodik akibat pembakaran biomassa [9].

Untuk mengoptimalkan seleksi fitur dan penyetelan hiperparameter secara simultan dan adaptif, *Differential Evolution* (DE) sebagai algoritma metaheuristik berbasis populasi menawarkan solusi global yang tangguh. DE mampu mengeksplorasi ruang parameter multidimensi secara efektif dan menghindari jebakan optimum lokal, menjadikannya ideal untuk mengoptimalkan model *boosting* yang kompleks [10]. Pendekatan ini tidak hanya meningkatkan stabilitas pelatihan, tetapi juga memastikan bahwa kombinasi fitur terbaik termasuk *fitur lag* kritis dipilih secara otomatis, sehingga menghasilkan model yang lebih sensitif terhadap peristiwa episodik yang jarang namun berisiko tinggi.

XGBoost telah banyak diterapkan dalam prediksi kualitas udara, namun sebagian besar studi masih mengandalkan konfigurasi default atau metode optimasi lokal yang rentan terjebak pada solusi suboptimal, terutama ketika menghadapi data yang tidak seimbang dan bersifat episodik [11]. Beberapa pendekatan *hibrida*, seperti integrasi *XGBoost* dengan teknik *oversampling* misalnya *Synthetic Minority Over-sampling Technique* (SMOTE), memang mampu meningkatkan sensitivitas

terhadap kelas minoritas, namun tidak secara langsung mengoptimalkan struktur model atau seleksi fitur sehingga masih berisiko menghasilkan *false alarm* atau *missed detection* pada lonjakan $PM_{2.5}$ yang kritis [12]. Di sisi lain, pemanfaatan DE sebagai optimizer untuk algoritma *boosting* memang telah menunjukkan keberhasilan dalam bidang lain, seperti klasifikasi data kebugaran fisik mahasiswa, namun penerapannya dalam konteks prediksi polusi udara khususnya untuk $PM_{2.5}$ di wilayah tropis masih sangat terbatas. Selain itu, banyak model prediktif saat ini mengandalkan representasi spasial seperti *Aerosol Optical Depth* (AOD) dari satelit atau arsitektur *deep learning* yang kompleks, yang memerlukan data eksternal tambahan atau komputasi intensif, sehingga kurang praktis untuk implementasi *real-time* di stasiun pemantauan lokal [13].

Penelitian ini mengisi celah dalam literatur terkini yakni belum adanya model prediksi $PM_{2.5}$ yang secara terpadu menggabungkan optimasi global, rekayasa fitur temporal, dan penanganan data *imbalanced* dalam konteks episodik *haze* di wilayah tropis dengan mengusulkan model hibrid DE-*XGBoost* yang dioptimalkan secara *end-to-end* menggunakan data deret waktu lokal dari Serpong tanpa ketergantungan pada data satelit atau model numerik eksternal. Kontribusi utamanya terletak pada tiga aspek: (1) penerapan DE untuk optimasi simultan terhadap *hyperparameter XGBoost* dan seleksi fitur, (2) pemanfaatan rekayasa *fitur lag* historis sebagai representasi temporal yang ringkas namun efektif dalam menangkap pola prekursor *haze*, dan (3) fokus eksplisit pada prediksi akurat pada hari-hari polusi tinggi dalam konteks data *imbalanced* khas wilayah tropis. Dengan demikian, penelitian ini tidak hanya meningkatkan akurasi prediksi secara umum, tetapi juga memperkuat kemampuan deteksi dini terhadap peristiwa episodik $PM_{2.5}$ yang berisiko tinggi bagi kesehatan masyarakat di kawasan perkotaan tropis seperti Serpong.

B. Perumusan Masalah

Rumusan masalah dalam penelitian ini adalah bagaimana perbedaan tingkat akurasi antara model *XGBoost* standar dan model *XGBoost hibrida* yang dioptimalkan menggunakan DE serta diperkaya dengan *fitur lag* historis dalam memprediksi konsentrasi $PM_{2.5}$ di wilayah Serpong, Tangerang Selatan?

C. Batasan Penelitian

Agar penelitian ini dapat berjalan secara terarah dan fokus pada tujuan yang telah ditetapkan, maka penelitian ini dibatasi pada beberapa hal berikut:

1. Data yang digunakan dalam penelitian ini adalah data historis berjenis deret waktu (*time series*) yang mencakup parameter polutan dan meteorologi dari satu sumber data yaitu sistem pemantauan kualitas udara (*Air Quality Monitoring System / AQMS*) yang berada di Serpong, Indonesia.
2. Model prediksi yang dikembangkan tidak hanya mengandalkan data historis PM_{2.5} itu sendiri, tetapi juga mempertimbangkan 12 fitur meteorologi dan polutan lain dari dataset AQMS sebagai variabel pendukung (*supporting features*) dalam proses prediksi, meliputi parameter polutan (SO₂, CO, O₃, NO₂, HC) serta parameter meteorologi (kecepatan angin, arah angin, kelembapan relatif, suhu udara, tekanan udara, radiasi matahari, curah hujan).
3. Fokus utama dari model prediksi yang dibangun dalam penelitian ini adalah pada polutan PM_{2.5}. Model tidak dirancang untuk memprediksi konsentrasi polutan lain seperti SO₂, CO, O₃, maupun NO₂.
4. Model prediksi yang dikembangkan dirancang untuk melakukan prediksi dalam jangka waktu pendek (*short-term forecasting*), yaitu untuk 30 menit ke depan, sesuai dengan resolusi data yang digunakan, dan tidak ditujukan untuk prediksi jangka panjang seperti harian, mingguan, atau bulanan.
5. Algoritma yang dianalisis dalam penelitian ini terbatas pada implementasi DE untuk tahap optimasi fitur dan *XGBoost* sebagai algoritma prediksi utama, sehingga tidak melakukan studi komparatif dengan berbagai algoritma *machine learning* atau optimasi lainnya.
6. Penelitian ini membatasi penggunaan fitur hanya pada data yang tersedia dalam *dataset* utama. Atribut eksternal lain yang berpotensi memengaruhi, seperti data titik api (*hotspot*) dari citra satelit, tidak diintegrasikan ke dalam model.
7. Implementasi sistem dalam bentuk aplikasi *web* interaktif yang dihasilkan merupakan sebuah purwarupa (*prototype*) yang bertujuan untuk

demonstrasi konsep dan fungsionalitas model, bukan merupakan aplikasi skala produksi yang siap untuk penerapan publik.

D. Tujuan Penelitian

Penelitian ini bertujuan untuk membandingkan akurasi prediksi konsentrasi PM_{2.5} antara model *XGBoost* standar dan model *hibrida XGBoost* yang dioptimalkan dengan DE serta diperkaya *fitur lag* historis, menggunakan data deret waktu dari Serpong, Tangerang Selatan.

E. Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat, sebagai berikut:

1. Manfaat Teoretis

Penelitian ini diharapkan dapat memperkaya literatur ilmiah mengenai penerapan metode hybrid yang mengintegrasikan *feature engineering* berbasis waktu, optimasi fitur dengan DE, dan prediksi menggunakan *XGBoost*. Hasil penelitian menyediakan referensi empiris tentang efektivitas optimasi fitur menggunakan DE dalam meningkatkan performa model *XGBoost* untuk prediksi kualitas udara. Selain itu, penelitian ini memberikan *baseline* performa untuk model *hybrid* DE-*XGBoost* yang dapat dijadikan acuan komparatif untuk penelitian sejenis dengan metode atau lokasi berbeda.

2. Manfaat Praktis

Secara praktis, penelitian ini menyediakan model prediksi PM_{2.5} dengan akurasi tinggi, yang dapat diimplementasikan sebagai inti dari sistem peringatan dini (*early warning system*) kualitas udara di kawasan daerah tropis yang rentan asap lintas batas. Model yang dikembangkan mengoptimalkan pemanfaatan data historis dengan mengidentifikasi dan fokus pada atribut prediktor yang paling berpengaruh, termasuk data lag historis, untuk menghasilkan informasi prediktif yang akurat. Lebih lanjut, penelitian ini memberikan dasar untuk pengembangan aplikasi yang mampu menyajikan informasi prediksi kualitas udara yang akurat kepada masyarakat, guna mendukung perencanaan aktivitas harian dan tindakan preventif untuk melindungi kesehatan.

F. Keaslian Penelitian

Penelitian ini berfokus pada prediksi konsentrasi polutan $PM_{2.5}$ menggunakan model *hibrida XGBoost* yang dioptimalkan dengan DE dan diperkaya *fitur lag* historis, berdasarkan data observasi lokal dari sistem pemantauan kualitas udara (*Air Quality Monitoring System / AQMS*) di Serpong, Tangerang Selatan, periode 1 Januari hingga 31 Desember 2023, dengan interval pencatatan setiap 30 menit dan mencakup 12 variabel prediktor serta 1 variabel target ($PM_{2.5}$). Berbeda dengan penelitian sebelumnya yang memprediksi $PM_{2.5}$ di DKI Jakarta menggunakan *XGBoost* standar hanya dengan data historis $PM_{2.5}$ tanpa variabel meteorologi dan tanpa algoritma optimasi metaheuristik, penelitian ini mengintegrasikan 12 variabel sebagai fitur prediktor, menerapkan rekayasa *fitur lag* untuk menangkap dinamika temporal episodik, serta menggunakan DE untuk optimasi simultan terhadap *hyperparameter* dan seleksi fitur suatu pendekatan yang belum diterapkan dalam penelitian prediksi $PM_{2.5}$ berbasis data lokal di wilayah perkotaan tropis Indonesia. Perbandingan penelitian terdahulu dengan penelitian ini dapat dilihat pada Tabel 1.

Tabel 1. Keaslian Penelitian

Aspek	Penelitian Sebelumnya	Penelitian yang Diusulkan
Lokasi Studi	DKI Jakarta	Serpong, Tangerang Selatan
Fokus	Prediksi konsentrasi $PM_{2.5}$ di DKI Jakarta menggunakan model <i>Extreme Gradient Boosting</i> (<i>XGBoost</i>) standar berbasis data historis $PM_{2.5}$.	Prediksi konsentrasi $PM_{2.5}$ di Serpong, Tangerang Selatan menggunakan model <i>hibrida</i> DE- <i>XGBoost</i> , yaitu <i>XGBoost</i> yang dioptimalkan dengan algoritma <i>Differential Evolution</i> (DE), serta diperkaya <i>fitur lag</i> historis.
Variabel yang digunakan	Menggunakan data historis konsentrasi $PM_{2.5}$ sebagai variabel utama model.	Menggunakan 12 variabel prediktor polutan dan meteorologi, ditambah <i>fitur lag</i> historis $PM_{2.5}$, dengan DE menyeleksi 4 fitur terpilih dari 15 kandidat.
Tujuan Penelitian	Mengembangkan model prediktif $PM_{2.5}$ di Jakarta menggunakan <i>XGBoost</i> dan mengevaluasi kinerjanya dengan metrik RMSE,	Membandingkan akurasi prediksi $PM_{2.5}$ antara <i>XGBoost</i> standar (<i>Baseline</i>) dan model <i>hibrida</i> DE- <i>XGBoost</i> , serta menguji signifikansi perbedaan performa keduanya menggunakan uji

Aspek	Penelitian Sebelumnya	Penelitian yang Diusulkan
	MAE, dan R^2 .	<i>Wilcoxon Signed-Rank</i> .
Hasil Penelitian	Pada data uji, model menghasilkan RMSE sebesar 16,7755, MAE sebesar 11,3960, dan R^2 sebesar 0,6812.	Model DE- <i>XGBoost</i> menunjukkan peningkatan performa dibanding <i>Baseline</i> pada seluruh metrik: RMSE turun 4,55%, MAE turun 4,26%, dan R^2 meningkat dari 0,8602 menjadi 0,8726, dengan uji <i>Wilcoxon</i> ($p = 0,0086$) mengonfirmasi peningkatan ini signifikan secara statistik.

