

BAB I

PENDAHULUAN

A. Latar Belakang

Kualitas udara merupakan indikator yang sangat berpengaruh pada penilaian tingkat kesehatan lingkungan dan kesejahteraan masyarakat di suatu daerah [1]. Salah satu parameter dalam pengukuran kualitas udara adalah Kadar polutan *Particulate Matter 2.5*. PM_{2.5} yaitu partikel yang sangat halus berukuran kurang dari atau sama dengan 2,5 mikrometer yang dapat menembus sistem pernapasan manusia hingga ke alveoli paru-paru dan berdampak negatif terhadap kesehatan manusia ketika dihirup secara terus-menerus, akibat yang akan terjadi ialah gangguan pernapasan seperti ISPA, penyakit kardiovaskular, hingga peningkatan risiko kematian dini [2]. Menurut Organisasi Kesehatan Dunia (WHO), ambang batas PM_{2.5} yang direkomendasikan adalah rata-rata tahunan 5 $\mu\text{g}/\text{m}^3$ dan rata-rata 24 jam sebesar 15 $\mu\text{g}/\text{m}^3$ [3]. Menurut BMKG, batasan PM_{2.5} dikategorikan berdasarkan konsentrasi dalam $\mu\text{g}/\text{m}^3$ (mikrogram per meter kubik), dengan skala Baik (0-15.5), Sedang (15.6-55.4), Tidak Sehat (55.5-150.4), Sangat Tidak Sehat (150.5-250.4), hingga Berbahaya (>250.5) [4]. Namun, berbagai wilayah yang ada di Indonesia masih memiliki kadar PM_{2.5} yang melebihi batas tersebut [3]. Kondisi ini menandakan bahwa pencemaran udara akibat PM_{2.5} masih menjadi ancaman serius bagi kesehatan masyarakat dan memerlukan pengawasan serta prediksi yang akurat.

Salah satu kota besar di Indonesia ialah Kota Tangerang Selatan, yang merupakan wilayah perkotaan dengan pertumbuhan ekonomi, urbanisasi, dan aktivitas transportasi yang meningkat pesat dalam beberapa tahun terakhir [5]. Aktivitas tersebut berkontribusi terhadap peningkatan emisi polutan dari berbagai sumber seperti kendaraan bermotor, kegiatan industri, serta aktivitas domestik masyarakat. Selain itu, kondisi geografis Tangerang Selatan yang berada di wilayah beriklim tropis menjadikan distribusi dan konsentrasi polutan di udara menjadi fluktuatif, dinamis, dan sulit diprediksi [6].

Data pemantauan menunjukkan bahwa wilayah Serpong, Tangerang Selatan, beberapa kali mengalami kondisi kualitas udara tidak sehat hingga sangat tidak sehat, dengan konsentrasi PM_{2.5} pada periode Juni 2023 mencapai rentang

150-200 $\mu\text{g}/\text{m}^3$ atau 42 kali di atas batas paparan tahunan WHO, jauh melampaui ambang batas aman yang ditetapkan WHO yaitu rata-rata tahunan 5 $\mu\text{g}/\text{m}^3$ dan rata-rata 24 jam sebesar 15 $\mu\text{g}/\text{m}^3$ [7]. Kondisi pencemaran udara yang tinggi tersebut berimplikasi langsung pada kesehatan masyarakat, tercermin dari tingginya jumlah kasus Infeksi Saluran Pernapasan Akut (ISPA) di Kota Tangerang Selatan, yang mencapai 29.699 kasus selama periode Januari hingga Agustus 2023, Jumlah ini tergolong tinggi jika dibandingkan dengan beberapa kota besar lainnya, seperti Kota Jambi yang mencatat sekitar 7.717 kasus ISPA, Kota Serang dengan 2.226 kasus, serta Kota Semarang dengan sekitar 20.987 kasus pada periode dan penduduk yang relatif sama [8]. Perbandingan ini menunjukkan bahwa beban gangguan pernapasan di Tangerang Selatan relatif lebih besar dibandingkan sejumlah kota lain, yang mengindikasikan adanya kontribusi kuat dari faktor lingkungan, khususnya pencemaran udara. Penelitian terdahulu menyebutkan bahwa paparan PM_{2.5} merupakan salah satu penyebab utama meningkatnya kasus gangguan pernapasan karena partikel halus ini mampu menembus saluran pernapasan hingga alveoli paru-paru dan memicu peradangan ketika terhirup secara terus-menerus [9].

Oleh karena itu, wilayah ini dipilih sebagai lokasi penelitian karena memiliki karakteristik perkotaan tropis dengan tingkat polusi udara yang relatif tinggi dan relevan untuk menguji kinerja model prediksi konsentrasi PM_{2.5} berbasis data deret waktu. Data penelitian diperoleh dari sistem pemantauan kualitas udara (*Air Quality Monitoring System/AQMS*) yang berlokasi di depan Eka Hospital BSD, Serpong, dengan interval pencatatan setiap 30 menit selama periode satu tahun (1 Januari–31 Desember 2023), dengan beberapa data kosong telah diimputasi pada penelitian terdahulu [10].

Prediksi ialah proses memperkirakan nilai atau kejadian di masa mendatang berdasarkan pola data historis. Dalam konteks kualitas udara, prediksi memiliki peran penting untuk membantu pemerintah dan masyarakat dalam mengambil langkah antisipatif terhadap potensi peningkatan polusi [11]. Namun, prediksi kadar polusi konsentrasi PM_{2.5} bukanlah hal sederhana karena dipengaruhi oleh banyak faktor, baik dari sisi meteorologi, polutan, parameter kualitas udara lain, serta historis PM_{2.5} itu sendiri [11].

Data kualitas udara umumnya bersifat multivariat, yaitu terdiri dari banyak variabel yang saling berkaitan, seperti suhu udara, kelembapan, tekanan, kecepatan angin, curah hujan, dan konsentrasi berbagai jenis polutan. Kompleksitas hubungan antarvariabel ini dapat memengaruhi kemampuan model dalam mengenali pola yang relevan terhadap perubahan kadar polutan PM2.5. Banyaknya variabel yang terlibat dapat menimbulkan permasalahan *curse of dimensionality*, di mana model kesulitan mengenali pola signifikan akibat tingginya kompleksitas data, dan terjadinya *multikolinearitas*. Untuk mengatasi permasalahan tersebut, penelitian ini memanfaatkan *Principal Component Analysis* (PCA) sebagai dasar dalam proses seleksi fitur. Nilai *loading* yang dihasilkan PCA digunakan untuk mengidentifikasi variabel-variabel yang memiliki kontribusi paling besar terhadap pembentukan komponen utama. Variabel dengan kontribusi tinggi dipertahankan sebagai masukan model LSTM, sedangkan variabel dengan kontribusi rendah dieliminasi sehingga jumlah fitur yang digunakan menjadi lebih sedikit tanpa menghilangkan informasi penting. Pendekatan ini diharapkan dapat mengurangi kompleksitas data, meminimalkan pengaruh *multikolinearitas*, serta meningkatkan efisiensi dan stabilitas model prediksi [12].

Sebelum penerapan *Principal Component Analysis* (PCA), dilakukan analisis korelasi eksploratif menggunakan Korelasi Pearson (hubungan linear) untuk mengidentifikasi hubungan antarvariabel *input* serta keterkaitannya dengan konsentrasi PM2.5. Analisis ini bertujuan untuk mengevaluasi tingkat *multikolinearitas* antarvariabel dan memastikan bahwa PCA layak digunakan sebagai metode seleksi fitur pada data multivariat. Data yang digunakan dalam penelitian ini detailnya dapat dilihat pada jurnal [13]. Dan merupakan data kualitas udara yang telah melalui proses penanganan *missing value* pada penelitian sebelumnya [10]. *Dataset* tersebut telah bebas dari nilai hilang dan selanjutnya digunakan untuk tahap analisis korelasi, reduksi dimensi PCA, serta pemodelan LSTM.

Dalam konteks analisis data deret waktu, algoritma *Long Short-Term Memory* (LSTM) yang merupakan pengembangan dari *Recurrent Neural Network* (RNN) memiliki kemampuan untuk mengenali pola temporal jangka panjang. LSTM dirancang untuk mengatasi permasalahan *vanishing gradient* yang sering

terjadi pada RNN tradisional, sehingga mampu menangkap hubungan *non-linear* dan kompleks antar variabel yang berubah terhadap waktu [14]. Karakteristik ini menjadikan LSTM sangat sesuai digunakan dalam kasus prediksi kualitas udara, seperti perubahan konsentrasi PM2.5 yang bergantung pada kondisi meteorologi dan aktivitas manusia. Meskipun demikian, LSTM juga memiliki beberapa keterbatasan, antara lain kebutuhan komputasi yang tinggi, sensitivitas terhadap parameter pelatihan seperti jumlah *epoch*, *batch size*, dan *learning rate* [15]. Oleh karena itu, metode optimasi berupa pencarian parameter terbaik menggunakan *grid search* diperlukan dan kombinasi dengan metode lain seperti *Principal Component Analysis* (PCA) juga diperlukan untuk seleksi fitur berbasis PCA melalui analisis nilai *loading*.

Penelitian terdahulu seperti penelitian yang menggunakan kombinasi PCA-LSTM dalam prediksi curah hujan harian dengan menggunakan data meteorologi dari tanggal 1 Januari 2017 sampai 30 September 2022 untuk memprediksi curah hujan di Kabupaten Luwu Utara. Hasil PCA menunjukkan empat variabel paling berpengaruh, yaitu evaporasi, suhu maksimal, kelembaban rata-rata, dan lama penyinaran matahari. Variabel-variabel ini kemudian digunakan sebagai *input* pada model *Long Short-Term Memory* (LSTM) untuk memprediksi curah hujan. Parameter model yang diuji meliputi jumlah *hidden layer*, *batch size*, dan *learning rate drop period*. Prediksi terbaik yang diperoleh mencapai nilai *Mean Absolute Percentage Error* (MAPE) sebesar 0,0018, yang menunjukkan akurasi tinggi [16].

Penelitian terkait prediksi PM2.5 juga membandingkan metode ARCH/GARCH dengan *Long Short-Term Memory* (LSTM) dalam meramalkan konsentrasi PM2.5 di Jakarta Pusat. Model LSTM yang digunakan pada penelitian tersebut terdiri dari tiga lapisan dengan dua lapisan *bidirectional* dan dilatih menggunakan data kualitas udara per jam. Hasil evaluasi menunjukkan bahwa model LSTM memiliki performa terbaik dibandingkan ARCH/GARCH berdasarkan nilai MSE, RMSE, MAE, dan MAPE. Temuan ini menegaskan bahwa LSTM mampu menangani karakteristik deret waktu PM2.5 yang dinamis dan volatil sehingga lebih unggul dalam menghasilkan akurasi prediksi [17]. Selain itu, penelitian oleh Ayu dkk. [18] juga menggunakan pendekatan multivariat dengan enam variabel polutan yaitu PM10, SO₂, CO, O₃, NO₂, dan PM2.5, serta

menunjukkan bahwa model LSTM lebih unggul dibandingkan ANN dalam prediksi kualitas udara. Temuan ini menunjukkan bahwa penggunaan banyak variabel *input* pada prediksi PM2.5 telah umum digunakan dan LSTM mampu menangani karakteristik deret waktu PM2.5 yang dinamis sehingga menghasilkan akurasi prediksi yang baik.

Berdasarkan penelitian sebelumnya, penggunaan model *Deep Learning* seperti LSTM menunjukkan kinerja yang baik dalam memprediksi kualitas udara. Namun, hingga saat ini belum banyak studi yang mengkombinasikan PCA sebagai metode seleksi fitur dengan LSTM dalam konteks prediksi konsentrasi PM2.5 pada wilayah beriklim tropis seperti Serpong, Tangerang Selatan, dengan data multivariat yang dicatat secara frekuensi tinggi. Oleh karena itu, penelitian ini berjudul “Analisis Prediksi Polutan PM2.5 dengan Reduksi Dimensi PCA dan *Deep Learning* LSTM pada Data Deret Waktu Multivariat” yang bertujuan untuk menganalisis pengaruh PCA sebagai seleksi fitur melalui perbandingan akurasi model LSTM Murni dan akurasi model PCA-LSTM dalam memprediksi kualitas udara khususnya polutan PM2.5 berdasarkan data polutan dan variabel meteorologi yang direkam setiap 30 menit selama satu tahun dan model akan dievaluasi menggunakan metrik evaluasi RMSE, MAE dan *R-squared* (R^2).

B. Perumusan Masalah

Rumusan masalah dalam penelitian ini adalah Apakah penerapan PCA sebagai teknik seleksi fitur dapat meningkatkan akurasi prediksi konsentrasi PM2.5 pada model LSTM berdasarkan perbandingan nilai RMSE, MAE, dan R^2 antara model LSTM Murni dan model PCA-LSTM?

C. Batasan Penelitian

Untuk menjaga fokus penelitian, ada beberapa batasan yang diperlukan, antara lain :

1. Data yang digunakan dalam penelitian ini terbatas pada data deret waktu multivariat yang berasal dari sistem pemantauan kualitas udara (*Air Quality Monitoring System/AQMS*) yang berlokasi di Tangerang Selatan, Indonesia, selama periode 1 Januari hingga 31 Desember 2023 dengan

interval waktu setiap 30 menit, terdiri dari 14 variabel *input* termasuk *history* PM2.5 dan 1 variabel *output*. Pembatasan periode data pada satu tahun ini dilakukan karena data yang diperoleh dalam penelitian ini hanya mencakup rentang waktu tersebut. Data dibagi secara berurutan (*sequential*) menjadi data *training* (1 Januari–13 September 2023), data validasi (13 September–7 November 2023), dan data *testing* (7 November–31 Desember 2023), sehingga model hanya mempelajari pola hingga bulan September dan pola musiman akhir tahun (November–Desember) tidak tercakup dalam proses *training*, sementara evaluasi pada data *testing* juga hanya mewakili dua bulan terakhir tersebut. Karena seluruh data berasal dari rentang satu tahun yang sama, penelitian ini memiliki keterbatasan dalam menangkap variasi pola antar tahun (*inter-annual variability*), sehingga model belum tentu dapat digeneralisasi untuk tahun-tahun yang berbeda.

2. Model digunakan untuk memprediksi konsentrasi PM2.5 pada interval waktu berikutnya (30 menit ke depan) dari data masukan terakhir, diluar dari ini tidak dipertimbangkan.
3. Model prediksi yang digunakan terbatas pada kombinasi *Principal Component Analysis* (PCA) untuk seleksi fitur dan *Deep Learning LSTM* sebagai metode prediksi, serta model *Deep Learning LSTM* tanpa PCA sebagai pembanding.
4. Pembagian *dataset* dilakukan dengan perbandingan 70% untuk data *training*, 15% data validasi, dan 15% untuk data *testing*.
5. Parameter yang digunakan pada kedua model adalah sama, sehingga perbedaan performa yang dihasilkan semata-mata mencerminkan pengaruh seleksi fitur PCA.
6. Evaluasi performa model dilakukan dengan tiga metrik evaluasi, yaitu RMSE, MAE dan *R-squared* (R^2).

D. Tujuan Penelitian

Tujuan penelitian ini adalah untuk menganalisis pengaruh penerapan PCA sebagai teknik seleksi fitur terhadap akurasi model LSTM dalam memprediksi

konsentrasi PM2.5 dengan membandingkan nilai RMSE, MAE, dan R^2 antara model LSTM Murni dan model PCA-LSTM pada data deret waktu multivariat.

E. Manfaat Penelitian

Penelitian ini diharapkan dapat memberi manfaat sebagai berikut :

1. Menambah khasanah ilmu pengetahuan di bidang pemodelan prediksi kualitas udara, khususnya penerapan kombinasi *Principal component analysis* (PCA) dan *Long Short-Term Memory* (LSTM) pada data deret waktu multivariat.
2. Memberikan pengalaman dan wawasan bagi penulis dalam mengembangkan metode prediksi kualitas udara menggunakan pendekatan berbasis model *machine learning*.
3. Memberikan gambaran mengenai kelebihan dan keterbatasan model PCA-LSTM dalam memprediksi kualitas udara PM2.5, sehingga dapat menjadi bahan perbaikan bagi pengembangan model prediksi yang lebih optimal.

F. Keaslian Penelitian

Penelitian ini berfokus pada prediksi kadar polutan PM2.5 menggunakan *Principal Component Analysis* (PCA) sebagai metode seleksi fitur dan *Deep Learning Long Short-Term Memory* (LSTM) berdasarkan data historis polutan udara dan variabel meteorologi selama satu tahun, yaitu tahun 2023. Dalam penelitian ini, data historis yang bersifat multivariat dianalisis menggunakan PCA untuk memperoleh nilai *loading* yang digunakan sebagai dasar dalam mengidentifikasi variabel-variabel yang memiliki kontribusi terbesar terhadap pembentukan komponen utama. Variabel dengan kontribusi tinggi dipilih sebagai masukan model LSTM, sedangkan variabel dengan kontribusi rendah dieliminasi. Dengan demikian, model LSTM menggunakan variabel asli hasil seleksi fitur untuk mempelajari pola temporal dalam memprediksi konsentrasi PM2.5.

Pendekatan ini dapat dibandingkan dengan penelitian yang dilakukan oleh Kristiani et al [19], yang menerapkan LSTM untuk memprediksi konsentrasi PM2.5 jangka pendek menggunakan data polutan dan meteorologi dari Taiwan periode 2017–2019. Penelitian tersebut membandingkan berbagai arsitektur *Deep Learning*

yaitu CNN, RNN, LSTM, GRU, Bi-LSTM, dan Bi-GRU, serta menggunakan analisis korelasi Pearson dan Spearman. Hasil terbaik diperoleh oleh LSTM dengan RMSE sebesar 1,9. Namun ia tidak menganalisis pengaruh reduksi dimensi terhadap performa model prediksi. Perbandingan penelitian terdahulu dengan penelitian ini dapat dilihat pada Tabel 1.

Tabel 1. Keaslian Penelitian

Aspek	Penelitian Sebelumnya	Penelitian yang diusulkan
Fokus	Prediksi konsentrasi PM2.5 jangka pendek dengan membandingkan berbagai model <i>Deep Learning</i> (CNN, RNN, LSTM, GRU, Bi-LSTM, Bi-GRU) tanpa reduksi dimensi.	Prediksi konsentrasi PM2.5 menggunakan kombinasi PCA dan LSTM dengan fokus seberapa berguna PCA sebagai metode seleksi fitur melalui akurasi model.
Variabel yang Digunakan	Menggunakan variabel CO, NO, NO ₂ , NO _x , O ₃ , PM10, PM2.5, SO ₂ , THC, NMHC, CH ₄ dan faktor meteorologi berupa suhu, kelembapan, kecepatan angin, arah angin, dan curah hujan.	Menggunakan beberapa variabel polutan dan meteorologi yaitu SO ₂ , CO, O ₃ , NO ₂ , hidrokarbon, kecepatan angin, arah angin, kelembapan udara, suhu udara, tekanan udara, radiasi matahari, curah hujan PM10 dan PM2.5.
Tujuan Penelitian	Membandingkan berbagai arsitektur model <i>Deep Learning</i> untuk prediksi PM2.5 dan mengidentifikasi model terbaik.	Menganalisis pengaruh PCA terhadap model prediksi konsentrasi PM2.5 dengan membandingkan akurasi model LSTM Murni dan PCA-LSTM.
Hasil Penelitian	LSTM menghasilkan performa terbaik dengan RMSE 1,9 tanpa penerapan reduksi dimensi sehingga seluruh variabel digunakan langsung sebagai <i>input</i> model.	Diharapkan kombinasi PCA dan LSTM mampu memberikan performa prediksi lebih optimal pada data dengan kompleksitas multivariat.