Laporan Akhir Tugas MK. Data Mining (KOM1338), Semester Genap 2024/2025
Program Studi Sarjana Ilmu Komputer
SSMI IPB

Klasifikasi Kondisi Udara Jakarta menggunakan LSTM untuk Mendukung Peningkatan Kualitas Udara

Syifa Izzatul Rahmah (G641221016), Khansa Fitri Zhafirah (G6401221017), Dwiamalina Qurratuain Najla (G641221040), Syira Rijannati Rosadi (G6401221094) [1*]
Kelompok: 4, Kelas Paralel: 3

Abstrak

Pencemaran udara di Jakarta merupakan isu serius yang berdampak langsung terhadap kesehatan masyarakat. Diperlukan sistem prediksi kualitas udara yang andal guna mendukung pengambilan kebijakan dan meningkatkan kesadaran publik. Penelitian ini bertujuan untuk mengklasifikasikan kondisi udara di DKI Jakarta menggunakan model Long Short-Term Memory (LSTM). Teknik data mining diterapkan untuk mengolah data Indeks Standar Pencemar Udara (ISPU) dari Stasiun Bundaran HI selama periode 2010 hingga 2021. Proses *preprocessing* melibatkan imputasi nilai hilang, normalisasi, encoding, dan windowing data sebelum pelatihan model. Arsitektur model terbaik terdiri dari dua lapisan LSTM, dilengkapi mekanisme *dropout* dan *early stopping* untuk menghindari *overfitting*. Hasil evaluasi menunjukkan bahwa model mampu mengklasifikasikan kualitas udara dengan akurasi 83,45% dan F1-score maupun nilai AUC yang cukup tinggi untuk setiap kelas, meskipun masih perlu perbaikan dalam menangani kelas minoritas. Temuan ini menunjukkan bahwa pendekatan LSTM dapat dimanfaatkan secara efektif untuk mendukung sistem peringatan dini dan strategi mitigasi pencemaran udara di wilayah perkotaan.

Kata Kunci: Jakarta, Klasifikasi, Kualitas Udara, LSTM, Pencemaran Udara

PENDAHULUAN

Latar Belakang

Seiring berkembangnya kawasan perkotaan dan meningkatnya aktivitas industri, transportasi, serta pembakaran lahan, tingkat pencemaran udara di kota-kota besar, termasuk Jakarta, mengalami peningkatan yang signifikan. Proses transformasi masyarakat dari sektor agraris menuju sektor industri telah memicu berbagai bentuk polusi lingkungan, khususnya di wilayah urban dan sub-urban (Kannajmi dan Saputra 2025). Pencemaran udara sendiri merupakan bentuk pencemaran lingkungan fisik akibat kehadiran zat cair, padat, atau gas di atmosfer, termasuk gas berbahaya seperti karbon monoksida (CO), nitrogen oksida (NOx), ozon (O_3), sulfur dioksida (SO_2), serta partikel seperti PM1, PM2.5, dan PM10 (Dyana *et al*. 2025).

[1]Program Studi Sarjana Ilmu Komputer, Sekolah Sains Data, Matematika dan Informatika (SSMI), Institut Pertanian Bogor, Bogor 16680
*Mahasiswa Program Studi Sarjana Ilmu Komputer, SSMI IPB; Surel: izzarahmah@apps.ipb.ac.id, khansa26zhafirah@apps.ipb.ac.id, najladwiamalina@apps.ipb.ac.id, syirarosadi@apps.ipb.ac.id

Dampak dari polusi udara terhadap kesehatan sangat serius. Beberapa di antaranya meliputi gangguan pernapasan, peningkatan risiko keguguran, autisme, penyakit kardiovaskular, kanker, dan bahkan kematian (Kannajmi dan Saputra 2025). Studi terbaru menunjukkan adanya hubungan sebab-akibat antara paparan polusi udara dan berbagai penyakit seperti infeksi saluran pernapasan bawah akut, asma, kanker paru-paru, dan penyakit paru obstruktif kronik (COPD) (Maio *et al*. 2023). Polusi udara juga terbukti mempengaruhi sistem imun tubuh, terutama pada saluran pernapasan dan pencernaan, serta dapat menyebabkan disregulasi respons imun adaptif seperti Th2 dan Th17 (Glencross *et al*. 2020).

DKI Jakarta sebagai ibu kota negara menjadi salah satu wilayah dengan tingkat pencemaran udara tertinggi di dunia. Berdasarkan data Air Quality Life Index (AQLI) pada April 2021, Jakarta menempati posisi ke-6 kota dengan kualitas udara terburuk di dunia (Amalia *et al*. 2022). Hal ini tidak terlepas dari tingginya pertumbuhan penduduk dan aktivitas industri, seperti industri tekstil, kelistrikan, logam, dan kendaraan bermotor yang turut menyumbang emisi pencemar. Selain itu, jumlah kendaraan pribadi di Jakarta yang mencapai lebih dari 18 juta unit pada tahun 2017 juga berkontribusi besar terhadap polusi udara (Agista *et al*. 2020). Untuk memantau dan mengevaluasi kualitas udara, digunakan sistem pengukuran Indeks Standar Pencemar Udara (ISPU) atau Air Quality Index (AQI). Perhitungannya mencakup parameter utama seperti PM10, PM2.5, SO_2, NO_2, CO, dan O_3 (Insani dan Darlianti 2019).

Pengolahan dan analisis data kualitas udara memerlukan metode yang mampu menangkap pola dan tren dalam data historis. Salah satu metode yang banyak digunakan dalam bidang kecerdasan buatan adalah machine learning, khususnya model Long Short-Term Memory (LSTM). LSTM merupakan jenis Recurrent Neural Network (RNN) yang efektif dalam mempelajari data deret waktu (*time-series*) dan digunakan untuk membuat prediksi akurat berdasarkan data historis (Syabani 2022). Dalam konteks peramalan, metode LSTM bertujuan untuk meminimalkan tingkat kesalahan prediksi sehingga hasil klasifikasi kualitas udara dapat diandalkan (Yanti *et al*. 2016).

Penelitian ini dilakukan untuk mengklasifikasikan kondisi udara di wilayah DKI Jakarta menggunakan model LSTM. Dataset yang digunakan berisi data pengukuran ISPU dari satu stasiun pemantauan kualitas udara (SPKU), yaitu Bundaran HI di Jakarta selama periode 2010 hingga 2021. Melalui pendekatan ini, diharapkan dapat mendukung upaya peningkatan kualitas udara dengan memberikan informasi prediktif yang akurat terkait kondisi polutan utama seperti NO_2, SO_2, O_3, CO, dan PM10.

Tujuan

Tujuan dari tugas akhir ini adalah untuk membangun model prediksi kualitas udara di wilayah DKI Jakarta menggunakan metode Long Short-Term Memory (LSTM) berdasarkan data historis Indeks Standar Pencemar Udara (ISPU) dari satu stasiun pemantauan selama periode 2010 hingga 2021. Model ini diharapkan mampu memprediksi kondisi kategori kualitas udara di masa mendatang secara lebih akurat.

Ruang Lingkup

Ruang lingkup kegiatan dalam tugas akhir ini meliputi pengumpulan dan pemrosesan data ISPU yang terdiri dari parameter polutan seperti PM10, PM2.5, SO_2, CO, O_3, dan NO_2 yang dikumpulkan dari satu stasiun pemantauan di DKI Jakarta. Teknik yang digunakan mencakup pembersihan data (data *cleaning*), pembuatan model LSTM, dan evaluasi model menggunakan Python dengan pustaka deep learning seperti TensorFlow.

Manfaat

Hasil dari tugas akhir ini diharapkan dapat memberikan kontribusi dalam bentuk sistem prediksi kualitas udara yang bermanfaat bagi masyarakat dan pihak terkait dalam pengambilan keputusan. Dengan prediksi yang lebih akurat, masyarakat dapat lebih waspada terhadap potensi paparan polusi udara berbahaya, sementara pemerintah atau lembaga lingkungan dapat menggunakan hasil prediksi ini sebagai dasar untuk menyusun strategi mitigasi dan kebijakan lingkungan.

TINJAUAN PUSTAKA

Pencemaran Udara

Pencemaran udara merupakan kondisi di mana udara tercemar oleh zat, energi, atau komponen lain yang mengganggu kualitas udara bersih. Berdasarkan Peraturan Pemerintah No. 22 Tahun 2021, pencemaran udara diartikan sebagai masuk atau dimasukkannya zat, energi, dan/atau komponen lain ke udara ambien oleh kegiatan manusia sehingga melampaui baku mutu udara ambien. Menurut Akhmad (2000 dalam Dwangga M), pencemaran udara adalah adanya bahan-bahan atau zat-zat asing di dalam udara yang menyebabkan perubahan komposisi udara dari keadaan normal. Sumber pencemaran udara dapat berasal dari proses alami maupun aktivitas manusia (antropogenik). Proses alami antara lain letusan gunung berapi, dekomposisi biotik, debu, spora tumbuhan, dan penguapan garam dari laut. Namun, pencemaran dari aktivitas antropogenik cenderung lebih dominan dan berbahaya, seperti emisi dari transportasi, pembakaran sampah, kegiatan industri, serta aktivitas rumah tangga (Ardhaningtyas dan Mahmudah 2019).

Zat pencemar udara dapat berupa partikel maupun gas. Partikel pencemar terdiri atas partikel tersuspensi total (Total Suspended Particulate/TSP) yang berdiameter hingga 100 μm, partikel kurang dari 10 μm (PM10), dan partikel sangat halus kurang dari 2.5 μm (PM2.5). Selain itu, terdapat gas pencemar seperti sulfur dioksida (SO_2), nitrogen dioksida (NO_2), karbon monoksida (CO), dan ozon permukaan (O_3) (Rita *et al*. 2016). Particulate Matter (PM) merupakan salah satu jenis polutan udara yang sangat berbahaya bagi kesehatan. PM2.5 atau fine particles berukuran sangat kecil sehingga dapat masuk ke dalam sistem pernapasan hingga ke alveoli dan menyebabkan berbagai gangguan kesehatan seperti infeksi saluran pernapasan akut, bronkitis, hingga penyakit paru obstruktif menahun (Arba 2019). Konsentrasi polutan seperti PM2.5, PM10, NO_2, dan O_3 terbukti lebih tinggi di area dengan lalu lintas padat dan berkorelasi dengan peningkatan kadar *Fractional Exhaled Nitric Oxide* (FeNO), yaitu indikator peradangan saluran napas (Meo *et al*. 2024).

Dampak dari pencemaran udara sangat merugikan baik dalam jangka pendek maupun jangka panjang. Paparan jangka pendek dapat menyebabkan iritasi mata, kelelahan, dan gangguan pernapasan, sementara paparan jangka panjang berisiko menyebabkan penyakit jantung, kanker paru-paru, serta gangguan perkembangan anak (Sivarethinamohan *et al*. 2021). Selain itu, perubahan kualitas udara yang signifikan juga berkontribusi terhadap perubahan iklim global. Ketika konsentrasi zat pencemar melebihi ambang batas, akan terjadi fenomena efek rumah kaca, di mana panas matahari yang seharusnya dipantulkan ke luar angkasa kembali ke permukaan bumi (Clayton 2019). Hal ini menyebabkan peningkatan suhu global yang dikenal sebagai pemanasan global (*global warming*). Pemanasan global berdampak besar terhadap iklim dunia, yang semula stabil menjadi ekstrem. Beberapa dampaknya antara lain mencairnya es di kutub, meningkatnya intensitas dan frekuensi kebakaran hutan, serta seringnya terjadi cuaca ekstrem seperti badai hebat dan anomali iklim lainnya (Jorquera *et al*. 2019).

Kualitas Udara dan Indeks Standar Pencemaran Udara (ISPU)

ISPU adalah angka tanpa satuan yang mencerminkan mutu udara ambien berdasarkan dampaknya terhadap kesehatan manusia, estetika, dan makhluk hidup lainnya (Kurniawan 2018). Meskipun Indeks Standar Pencemaran Udara (ISPU) dirancang lebih sesuai untuk wilayah perkotaan, secara umum indeks ini tetap dapat diterapkan di berbagai jenis wilayah. Rincian parameter yang digunakan dalam perhitungan ISPU dijelaskan secara lebih lengkap dalam Lampiran Keputusan Kepala BAPEDAL No. 107 Tahun 1997 mengenai Tata Cara Perhitungan, Pelaporan, dan Informasi ISPU.

Tabel 1 Parameter dasar untuk pengukuran ISPU dan periode waktu pengukurannya sesuai dengan lampiran Keputusan Kepala Bapedal No. 107 Tahun 1997

Parameter	**Waktu Pengukuran (rata-rata)**
Partikulat (PM_{10})	24 jam
Sulfurdioksida (SO_2)	24 jam
Karbonmonoksida (CO)	8 jam
Ozon (O_3)	1 jam
Nitrogendioksida (NO_2)	1 jam

Nilai ISPU yang telah dihitung dimanfaatkan untuk mengelompokkan tingkat kualitas udara di suatu lokasi. Kategori ini ditentukan berdasarkan nilai ISPU dari parameter pencemar dominan. Rangkuman klasifikasi kondisi kualitas udara tersebut dapat dilihat pada Tabel 2 berikut.

Tabel 2 Kategori kualitas udara berdasarkan nilai ISPU sesuai dengan lampiran Keputusan Kepala Bapedal No. 107 Tahun 1997

Nilai ISPU	**Kategori**
0-50	Baik
50-100	Sedang
101-199	Tidak Sehat
200-299	Sangat Tidak Sehat
>300	Berbahaya

Long Short Term Memory Network

Pengolahan data berurutan seperti teks maupun deret waktu (*time series*) memerlukan arsitektur jaringan saraf yang mampu memahami hubungan antar elemen dalam urutan tersebut. Jaringan saraf tradisional seperti *feed-forward neural network* umumnya tidak efektif dalam menangani jenis data ini karena tidak memiliki memori internal yang dapat menyimpan informasi dari langkah-langkah sebelumnya. Recurrent Neural Network (RNN) hadir sebagai solusi khusus untuk data sekuensial karena mampu menyimpan informasi historis melalui koneksi antar unit tersembunyi (*hidden units*) yang disertai dengan penundaan waktu (Yu *et al*. 2019).

RNN pertama kali diperkenalkan pada tahun 1980-an sebagai model pemrosesan data urut yang memungkinkan pendeteksian korelasi temporal antara peristiwa yang terjadi dalam interval waktu berbeda (Elman 1990). Meskipun secara teoritis mampu mempelajari ketergantungan jangka panjang, model ini sering mengalami kesulitan dalam

mempertahankan informasi yang relevan akibat permasalahan *vanishing gradient* (Bengio *et al*. 1994).

Permasalahan tersebut mendorong pengembangan Long Short Term Memory (LSTM), sebuah arsitektur yang diperkenalkan oleh Hochreiter dan Schmidhuber pada tahun 1997. Hochreiter dan Schmidhuber (1997) dalam Wiranda (2019) mengatakan LSTM dirancang untuk mengatasi keterbatasan RNN konvensional, khususnya permasalahan *vanishing* dan *exploding gradient* yang menyebabkan hilangnya informasi penting seiring meningkatnya panjang urutan data.

Struktur LSTM terdiri atas tiga gerbang utama: *input gate*, *forget gate*, dan *output gate*. (Vinayakumar *et al*. 2017 dalam Wiranda 2019) *Input gate* berfungsi mengontrol seberapa banyak informasi baru yang disimpan dalam sel memori. *Forget gate* menentukan bagian informasi lama yang perlu dipertahankan atau dilupakan. *Output gate* mengatur seberapa banyak informasi dari sel memori yang digunakan untuk menghasilkan output (Mathisen 2017). Secara matematis, fungsi dari masing-masing gerbang dijelaskan sebagai berikut:

1. *Input Gate* (i_t):
   $i_t = \sigma(W_i S_{t-1} + W_i X_t)$
   Dimana:
   W_i merupakan bobot dari *input gate*
   S_{t-1} adalah *state* sebelumnya
   X_t adalah input saat ini
   σ adalah fungsi aktivasi sigmoid.
2. *Forget Gate* (f_t):
   $f_t = \sigma(WfS_{t-1} + WfX_t)$
   Dimana:
   Wf adalah bobot dari *forget gate*.
3. *Output Gate* (o_t) dan *Output State* (h_t):
   $o_t = \sigma(W_o S_{t-1} + W_o X_t)$
   $h_t = o_t \times tanh(c_t)$

Sel memori dalam LSTM menyimpan nilai atau *state* untuk periode waktu pendek maupun panjang, menjadikannya sangat efektif dalam berbagai aplikasi yang melibatkan data sekuensial seperti pengenalan suara, terjemahan mesin, hingga prediksi kerusakan sistem.

Penerapan LSTM dalam konteks dunia nyata ditunjukkan dalam studi yang dilakukan di Kaohsiung, Taiwan. Penelitian tersebut menggunakan LSTM untuk memprediksi konsentrasi PM2.5 berdasarkan data kualitas udara tahun 2017 hingga 2018. Hasilnya menunjukkan bahwa model LSTM berhasil mencapai nilai *Root Mean Square Error* (RMSE) sebesar 2.759, yang mencerminkan tingkat akurasi yang cukup tinggi dalam memprediksi kadar PM2.5 harian (Yang dan Guo 2021).

METODE

Data

Dataset yang digunakan dalam penelitian ini diambil dari platform Kaggle dan bersumber dari Dinas Lingkungan Hidup Provinsi DKI Jakarta serta Satu Data Jakarta. Dataset ini memuat hasil pengukuran Indeks Standar Pencemar Udara (ISPU) atau *Air Quality Index* (AQI) yang dikumpulkan dari satu Stasiun Pemantauan Kualitas Udara (SPKU) di wilayah DKI Jakarta, yaitu Stasiun DKI1 (Bundaran HI). Data berisi 4383 baris

dan 11 atribut mencakup periode waktu dari tahun 2010 hingga 2021, dengan parameter utama pencemar udara yang digunakan dalam penilaian kualitas udara.

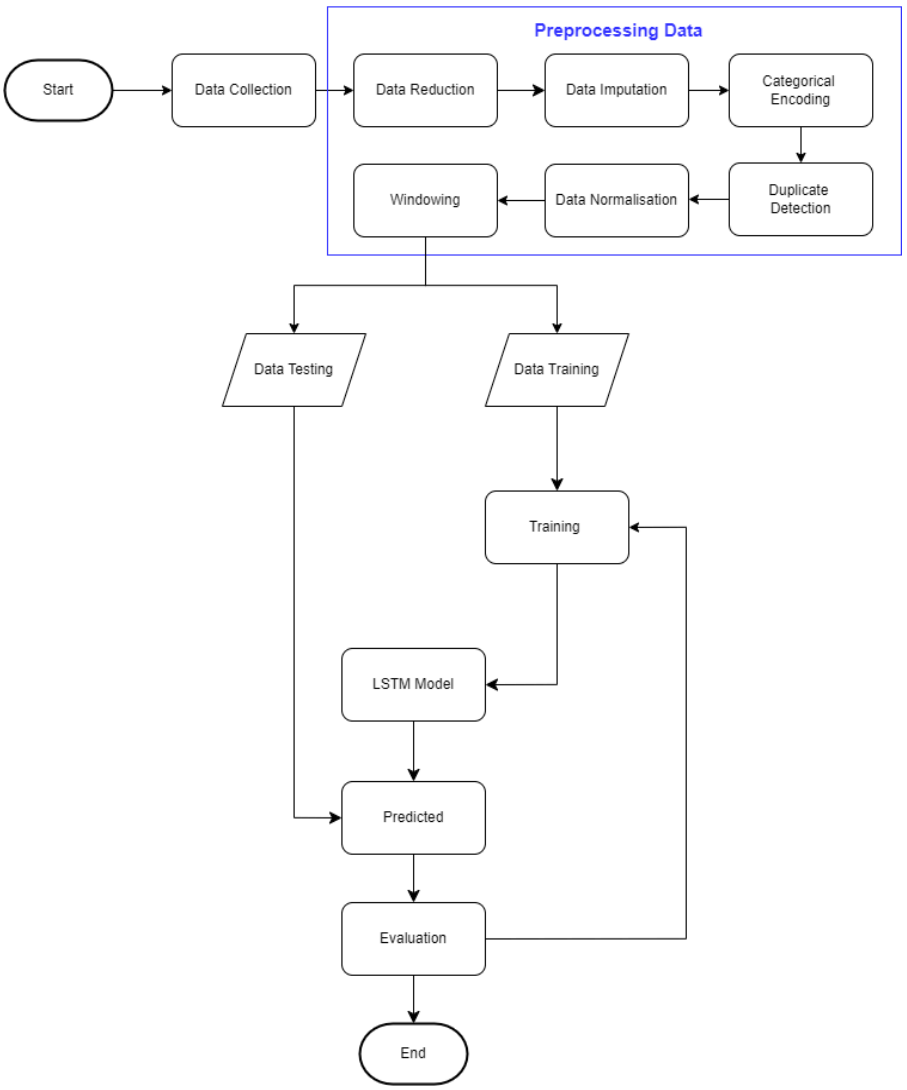
Tabel 3 Atribut data kualitas udara

Nama Atribut	Tipe Data	Deskripsi
tanggal	(YYYY-MM-DD)	Tanggal pencatatan pengukuran kualitas udara
stasiun	String	Nama atau kode identifikasi stasiun pemantauan kualitas udara tempat pengukuran dilakukan
pm 10	Float	Konsentrasi partikel udara berukuran ≤10 mikrometer (PM10), dalam satuan mikrogram per meter kubik (µg/m³)
pm25	Float	Konsentrasi partikel udara berukuran ≤2.5 mikrometer (PM2.5), dalam satuan mikrogram per meter kubik (µg/m³)
so2	Float	Konsentrasi gas sulfur dioksida (SO_2), dalam satuan parts per million (ppm)
co	Float	Konsentrasi gas karbon monoksida (CO), dalam satuan parts per million (ppm)
o3	Float	Konsentrasi gas ozon (O_3), dalam satuan parts per million (ppm)
no2	Float	Konsentrasi gas nitrogen dioksida (NO_2), dalam satuan parts per million (ppm)
max	Float	Nilai tertinggi (maksimum) dari lima polutan utama (PM10, SO2, CO, O3, NO2) pada tanggal dan stasiun tersebut. Mewakili tingkat polusi terburuk

critical	String	Nama polutan yang memiliki konsentrasi tertinggi pada hari dan stasiun tersebut
category	String	Kategori kualitas udara berdasarkan nilai max. Menunjukkan tingkat kesehatan udara, mengikuti standar baku mutu ISPU (Baik, Sedang, Tidak Sehat, Sangat Tidak Sehat, Berbahaya)

Atribut pada dataset yang digunakan masih memerlukan tahapan *preprocessing* sebelum melakukan teknik penambangan data terutama untuk teknik LSTM.

Tahapan Kegiatan



Gambar 1 Tahapan kegiatan

Pada tahap awal dilakukan pencarian dan pengumpulan data dari internet hingga kemudian memilih dataset ISPU dari Dinas Lingkungan Hidup Provinsi DKI Jakarta. Data kemudian dieksplorasi untuk memahami struktur, pola, dan sifat guna mendapatkan pemahaman yang lebih baik terkait karakteristik data sehingga kedepannya dapat memilih teknik penambangan data yang sesuai.

Setelah itu, tahapan selanjutnya adalah melakukan *preprocessing* data yang terdiri dari beberapa proses berurutan:

1. Data Reduction
   Pada tahap ini, data yang tidak diperlukan dihilangkan untuk menyederhanakan dataset dan menghindari beban komputasi dan menghindari hal yang dapat menyebabkan *overfitting* atau tidak stabilnya model.
2. Data Imputation
   Pada tahap ini, data yang memiliki nilai kosong *(missing value)* akan diisi dengan menggunakan teknik *forward fill (LOCF)* sebagai salah satu metode imputasi data untuk teknik LSTM. Dengan menggunakan metode ini, kita dapat mengisi nilai yang hilang pada deret waktu menggunakan nilai sebelumnya yang paling baru. Tahapan ini dilakukan untuk menjaga integritas data dan memastikan tidak ada informasi hilang yang dapat mempengaruhi model. Hasil imputasi kemudian dievaluasi sebelum melakukan tahapan selanjutnya.
3. Categorical encoding
   Tahapan ini dilakukan untuk mengubah data kategorik menjadi representasi numerik agar dapat diproses oleh model pembelajaran mesin. Pada dataset yang digunakan, terdapat kolom label kategori dengan tiga nilai yaitu *Baik, Sedang,* dan *Tidak Sehat.*
4. Duplicated Detection
   Pada tahap ini dilakukan pengecekan terhadap baris data yang sama persis (duplikat) untuk memastikan bahwa tidak ada data yang berulang yang dapat menyebabkan bias dalam pelatihan model.
5. Normalisasi data
   Proses ini dilakukan untuk menyamakan skala data numerik sehingga semua fitur memiliki kontribusi yang setara dalam *training* model. Teknik normalisasi yang digunakan pada projek ini adalah *StandardScaler*. Dengan menggunakan rumus z-score, skala data numerik diubah hingga setiap fiturnya memiliki rata-rata 0 dan standar deviasi sebesar 1.
6. Windowing
   Pada tahap ini, data diolah menjadi bentuk *window* untuk menyesuaikan format input yang dibutuhkan oleh model LSTM. Model LSTM bekerja optimal pada data berurutan dan mempertimbangkan informasi historis, maka dilakukan proses *windowing* dengan cara membagi data ke dalam segmen-segmen waktu.

Selain tahapan diatas, sebelum maupun setelah tahap *preprocessing* juga dilakukan eksplorasi data, pengecekan *noise* atau *outlier* dengan boxplot, serta melakukan pengecekan *trend* dan *seasonality* diakhir tahapan *preprocessing* .

Setelah dilakukan *preprocessing,* langkah selanjutnya adalah tahap pembuatan model menggunakan algoritma Long Short-Term Memory (LSTM) yang merupakan

bagian dari Recurrent Neural Network (RNN). Model dibangun dengan menggunakan bahasa pemrograman *Python.* Tahapan awal pembuatan model adalah dengan membagi data menjadi data *training* (untuk melatih model) sebanyak 80% dan data *testing* (untuk menguji performa model) sebanyak 20%. Pada proses pelatihan model menggunakan LSTM, data training dari hasil split sebelumnya kemudian dibagi kembali menjadi dua bagian, yaitu data training dan data validasi dengan proporsi 80:20. Artinya, 20% dari data training awal digunakan sebagai data validasi guna memantau kinerja model selama proses pelatihan dan menghindari overfitting,

Selain itu, dilakukan juga evaluasi terhadap nilai *learning rate* dengan mencoba beberapa model percobaan untuk menentukan *learning rate* yang paling optimal. Proses ini bertujuan untuk memastikan bahwa model dapat belajar secara efektif tanpa mengalami *underfitting* maupun *overfitting* akibat laju pembelajaran yang terlalu lambat atau terlalu cepat.

Tahapan terakhir adalah evaluasi model. Pada tahap ini, model yang didapat kemudian diprediksi terhadap data *testing* untuk melihat performa model. Hasil prediksi selanjutnya dibandingkan dengan nilai aktual dan diukur sejauh mana performa model dalam melakukan klasifikasi. Implementasi evaluasi dilakukan dengan memanfaatkan pustaka *sklearn.metrics* diantaranya Akurasi*, Recall, Precision,* dan *F1-Score* yang memberikan gambaran kuantitatif terkait performa model terhadap data uji.

Tahapan ini diakhiri dengan penarikan kesimpulan terhadap performa model, yaitu dengan mengevaluasi apakah model sudah mampu mengklasifikasikan data dengan baik atau masih memerlukan perbaikan. Jika hasil evaluasi menunjukkan performa yang belum optimal, maka perlu dilakukan pelatihan ulang (*retraining*) dengan menyesuaikan parameter dan arsitektur.

Lingkungan Pengembangan

Pengembangan penelitian dilakukan dengan menggunakan Google Colab. Pemrosesan data hingga evaluasi dilakukan dalam satu notebook untuk memudahkan dokumentasi, replikasi proses, serta visualisasi hasil secara terstruktur. Lingkungan pengembangan ini dapat diakses dengan mudah sehingga memudahkan kolaborasi.

HASIL DAN PEMBAHASAN

Hasil eksplorasi data di awal menghasilkan informasi dataset yang memiliki 4383 baris data dengan jumlah 11 kolom atribut. Selain itu, juga diperoleh bahwa 7 dari 11 atribut data memiliki *missing value*. Hal ini menandakan bahwa dataset memerlukan tindakan *preprocessing* sebelum nantinya dimasukkan menjadi input model LSTM. Hasil dari reduksi adalah berjumlah 7 kolom, dimana pada kondisi ini kolom `stasiun` dan `critical` dilakukan penghapusan dikarenakan setiap baris pada kolom `stasiun` memiliki data yang sama, begitu pun pada kolom `critical` sehingga dapat dikatakan kolom tersebut tidak dibutuhkan dalam pembuatan model. Selain itu, pada kolom `max` dan `pm25` memiliki korelasi tinggi (lebih dari 0.75) terhadap atribut lain yaitu `pm10` dan `o3`. Tingginya korelasi antar atribut dapat dianggap redundan dan bisa mengganggu proses pembelajaran model yang bisa menyebabkan *overfitting* atau tidak stabil. Sehingga dalam hal ini kedua kolom tersebut akan dihilangkan. Selanjutnya, tahapan imputasi data menggunakan teknik *forward fill (LOCF)* dikatakan berhasil dan tidak tidak merusak struktur data asli karena hasil statistik dan korelasi tidak berubah drastis dengan data

sebelum dilakukan imputasi. Tahapan *encoding* juga dilakukan pada dataset yang digunakan khususnya pada kolom label `kategori` yang menyatakan kualitas udara, yaitu *Baik*, *Sedang*, atau *Tidak Sehat*. *Encoding* dilakukan menggunakan skema "Baik" menjadi 0, "Sedang" menjadi 1, dan "Tidak Sehat" menjadi 2. Selain itu, hasil dari pengecekan duplikasi data menyatakan bahwa dalam dataset tidak terdapat data yang duplikat sehingga tidak diperlukan tindakan penghapusan baris data. Hal ini juga didukung oleh sifat dataset yang dimiliki yaitu *time-series* dimana tindakan penghapusan data dapat mengganggu kontinuitas dan urutan kronologis data yang tentunya sangat penting dalam analisis dan pembuatan model dari data *time-series*.

Noise yang didapat pada kondisi dataset ini tidak dilakukan penghapusan atau pengisian karena sifat data *time-series* yang mengandalkan urutan dan kontinuitas waktu, sehingga tindakan seperti penghapusan atau pengisian dapat mengganggu pola temporal yang penting bagi model. Selain itu, nilai noise yang ditemukan masih berada dalam batas wajar dan tidak menunjukkan adanya outlier ekstrem, sehingga tidak memberikan dampak signifikan terhadap kualitas data maupun hasil analisis. Teknik normalisasi juga tidak lupa dilakukan menggunakan *StandardScaler*. Berikut visualisasi singkat dataset setelah dilakukan berbagai *preprocessing* tahapan diatas.

```
RangeIndex: 4383 entries, 0 to 4382
Data columns (total 7 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   tanggal  4383 non-null   datetime64[ns]
 1   pm10     4383 non-null   float64
 2   so2      4383 non-null   float64
 3   co       4383 non-null   float64
 4   o3       4383 non-null   float64
 5   no2      4383 non-null   float64
 6   label    4383 non-null   int64
dtypes: datetime64[ns](1), float64(5), int64(1)
memory usage: 239.8 KB
```

Gambar 2 Informasi struktur dan tipe data setelah *preprocessing*

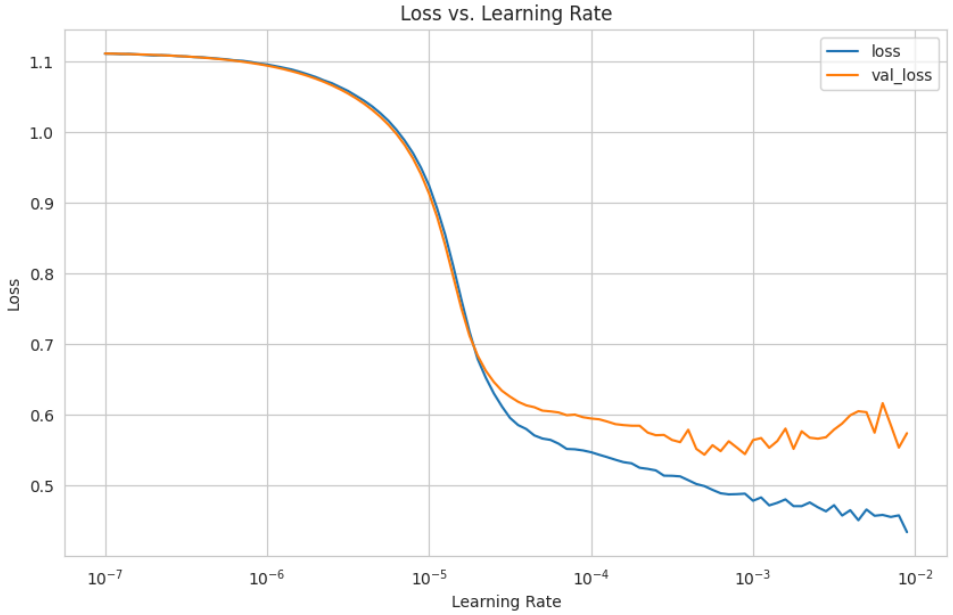
	tanggal	pm10	so2	co	o3	no2	label
0	2010-01-01	0.527261	-1.294768	4.306084	-0.868359	0.002510	1
1	2010-01-02	-1.376910	-1.480237	-0.793864	-0.634612	-0.558836	0
2	2010-01-03	-1.716940	-1.480237	-0.525446	-1.141063	-0.558836	0
3	2010-01-04	-2.056971	-1.480237	-0.793864	-1.335851	-0.895644	0
4	2010-01-05	-1.852953	-1.480237	-0.704391	-1.335851	-0.671106	0
5	2010-01-06	-1.512922	-1.387503	-0.257028	-1.180020	-0.334298	0
6	2010-01-07	-0.764855	-1.294768	-0.078082	-0.868359	-0.109759	0
7	2010-01-08	0.799285	-0.923831	2.337683	-1.180020	0.114780	1
8	2010-01-09	0.187230	-1.202034	1.264010	-1.024189	0.227049	1
9	2010-01-10	-1.240898	-1.202034	-0.167555	-0.907316	-0.222028	0

Gambar 3 Cuplikan awal data setelah *preprocessing*

Tahapan terakhir dari *preprocessing* yaitu *windowing* yang dilakukan untuk menyesuaikan format input yang dibutuhkan oleh model LSTM. Dalam proyek ini,

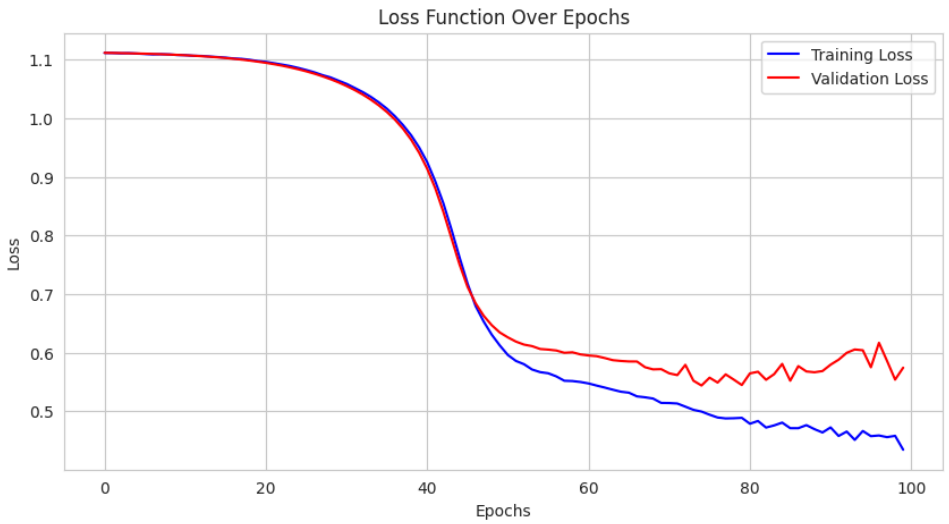
digunakan pendekatan jendela berjalan (*sliding window*) dengan panjang 7 hari sebagai input yang digunakan untuk memprediksi kondisi pada hari ke-8. Dengan kata lain, setiap 7 baris data sebelumnya akan digunakan sebagai input fitur untuk mempelajari pola deret waktu, sedangkan label target-nya adalah kategori kondisi kualitas udara pada hari ke-8. Proses ini dilakukan secara berulang hingga seluruh data terbagi sesuai kebutuhan model.

Setelah dataset sukses dilakukan pembagian antara data *training* dan juga data *test* langkah selanjutnya adalah pemodelan dimana pada tahapan awal akan dicari terlebih dahulu parameter *Learning Rate* dan *epoch* yang optimal untuk diterapkan pada model. Model yang digunakan pada pencarian evaluasi parameter maupun model seungguhnya menggunakan *Stacked Multivariate LSTM for Multiclass Classification*. Model ini dilatih menggunakan optimizer Adam yang adaptif dan efisien untuk data *time-series*, dengan fungsi loss *sparse_categorical_crossentropy* yang sesuai untuk klasifikasi multi kelas dengan label dalam bentuk bilangan bulat. Evaluasi performa dilakukan menggunakan metrik akurasi, dan pelatihan model dilakukan dengan ukuran *batch* sebesar 32.



Gambar 4 Evaluasi loss terhadap variasi *learning rate*

Berdasarkan grafik evaluasi terhadap berbagai nilai *learning rate* pada gambar diatas, terlihat bahwa penurunan nilai loss yang signifikan mulai terjadi di sekitar *learning rate* 10^{-5} dan mencapai titik paling optimal sebelum kembali meningkat di atas nilai 10^{-3} atau 0.001. Nilai *loss* dan *validation loss* terlihat menurun secara konsisten hingga sekitar *learning rate* 0.001 tanpa adanya fluktuasi tajam yang menandakan ketidakstabilan pelatihan. Oleh karena itu, *learning rate* sebesar 0.001 dapat dianggap sebagai titik optimal untuk digunakan pada model selanjutnya.

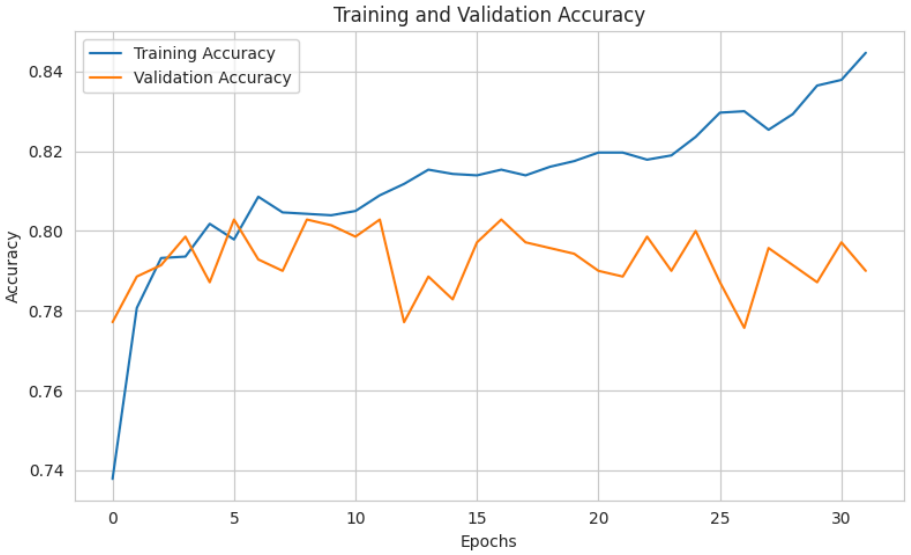


Gambar 5 Evaluasi loss terhadap *epoch*

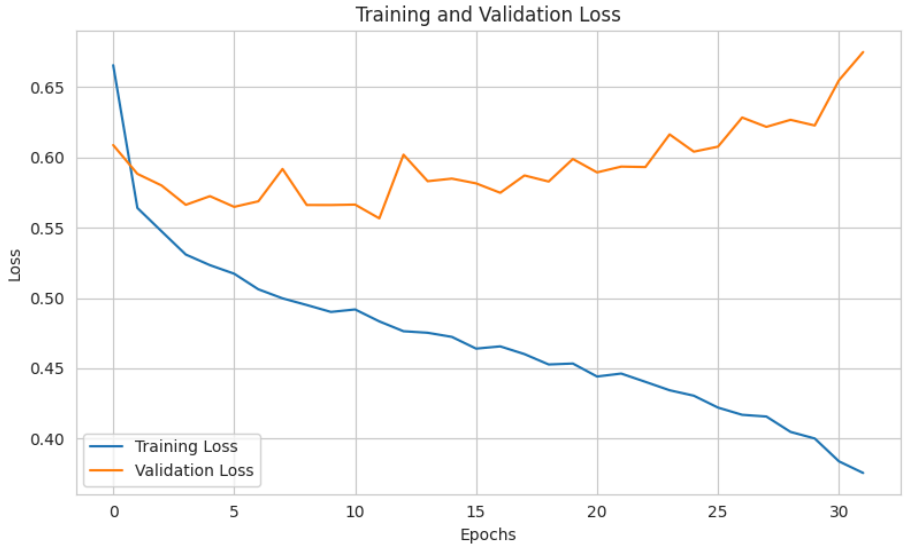
Berdasarkan hasil pelatihan model, penggunaan 100 *epoch* menunjukkan performa yang masih membaik pada *training loss* tanpa menyebabkan *overfitting* yang signifikan pada *validation loss*. Meskipun *validation loss* mulai berfluktuasi setelah *epoch* ke-70, tidak terjadi lonjakan tajam yang menandakan *overfitting* berat. Hal ini menunjukkan bahwa model masih belajar secara efektif hingga akhir. Oleh karena itu, penggunaan 100 epoch tetap layak dan direkomendasikan untuk memperoleh model selanjutnya yang lebih optimal.

Setelah dilakukan penyesuaian terhadap parameter jumlah epoch dan *learning rate*, proses dilanjutkan dengan membuat dan menguji model sesungguhnya. Pada penelitian kali ini terdapat empat arsitektur model berbeda yang dibangkitkan. Dari keseluruhan model yang dikembangkan, terdapat satu model yang memiliki performa terbaik dibanding lainnya berdasarkan evaluasi pada data pelatihan, validasi, dan pengujian. Adapun arsitektur model terbaik terdiri dari dua lapisan Long Short-Term Memory (LSTM), yang dirancang khusus untuk menangani data sekuensial. Lapisan pertama merupakan LSTM dengan 128 unit dan konfigurasi `return_sequences=True` untuk mempertahankan bentuk sekuensial data menuju lapisan berikutnya. Lapisan kedua berupa LSTM dengan 64 unit. Setelah itu, ditambahkan lapisan `Dropout` sebesar 0.2 sebagai bentuk regularisasi tambahan. Pada tahap akhir, digunakan lapisan `Dense` dengan tiga neuron dan aktivasi softmax untuk menghasilkan output klasifikasi multi-kelas. Model dikompilasi menggunakan algoritma optimasi Adam dengan *learning rate* sebesar 0.001, serta menggunakan fungsi *loss* `sparse_categorical_crossentropy` yang sesuai untuk label dalam bentuk integer. Selain itu, proses pelatihan dilengkapi dengan mekanisme *early stopping* dengan *patience* sebanyak 20 *epoch* dan parameter `restore_best_weights=True`. Mekanisme ini bertujuan untuk mencegah *overfitting* serta memastikan bahwa bobot terbaik berdasarkan performa pada data validasi disimpan dan digunakan sebagai model akhir. Berdasarkan hasil pelatihan, proses *training* secara otomatis berhenti pada *epoch* ke-32, ketika tidak lagi terjadi peningkatan signifikan pada performa validasi, sehingga efisiensi pelatihan tetap terjaga tanpa mengorbankan akurasi model.

Model terbaik ini menunjukkan kinerja yang unggul dengan akurasi sebesar 0.8402 dan nilai *loss* sebesar 0.3780 pada data pelatihan. Sementara itu, pada data validasi, model mencatatkan akurasi validasi sebesar 0.7900 dan *loss* validasi sebesar 0.6748, yang menunjukkan kemampuan yang cukup akurat tanpa indikasi *overfitting* yang signifikan.



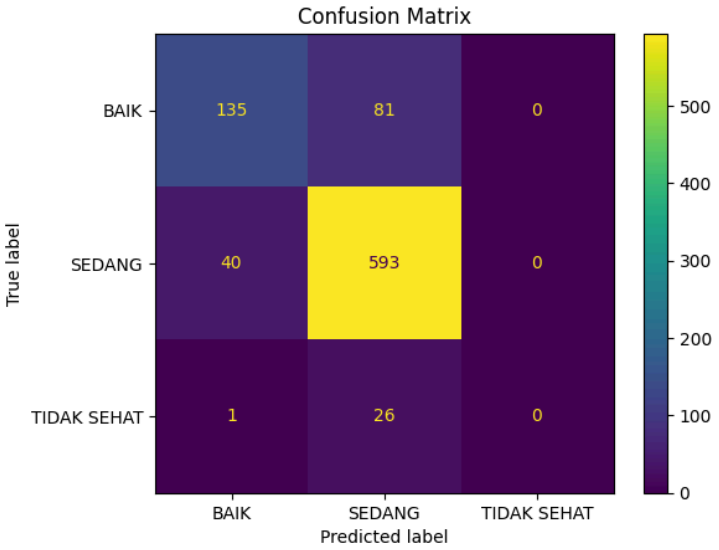
Gambar 6 Perbandingan akurasi proses pelatihan



Gambar 7 Perbandingan loss proses pelatihan

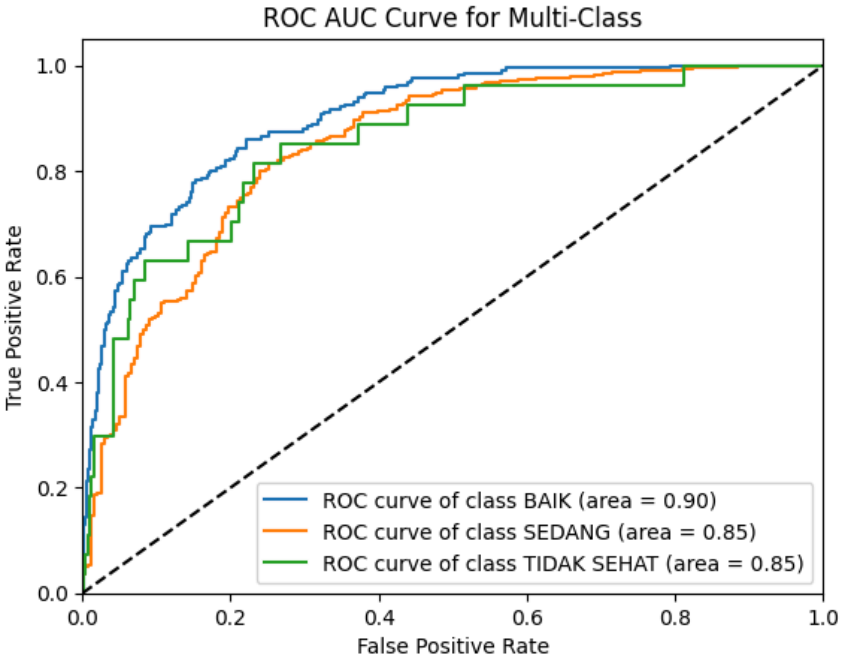
Berdasarkan grafik akurasi, terlihat bahwa nilai akurasi pelatihan menunjukkan peningkatan yang konsisten hingga mencapai nilai akhir 0.8402. Sementara itu, akurasi validasi cenderung fluktuatif namun tetap berada dalam rentang yang relatif stabil dengan nilai tertinggi mendekati 0.80. Meskipun terdapat perbedaan antara akurasi pelatihan dan validasi, tidak terdapat indikasi *overfitting* yang ekstrem karena selisih nilai masih dalam batas wajar. Sementara itu, grafik *loss* memperlihatkan penurunan *loss* yang signifikan pada data pelatihan seiring bertambahnya jumlah *epoch*. Menunjukkan bahwa model mampu mempelajari pola dari data pelatihan dengan baik. Namun, nilai *loss* pada data validasi tampak mengalami fluktuasi dan cenderung meningkat setelah melewati *epoch* ke-20. Meskipun demikian, fluktuasi tersebut tidak menunjukkan lonjakan tajam yang mengindikasikan *overfitting* berat.

Evaluasi lebih lanjut dilakukan pada data *test* untuk menilai performa model terhadap data yang tidak pernah dilihat sebelumnya. Hasil evaluasi menunjukkan bahwa model memiliki akurasi sebesar 0.8345 dan *loss* sebesar 0.4531.



Gambar 8 Confusion matrix

Selain itu, evaluasi model juga dilakukan menggunakan *confusion matrix* untuk tiga kelas yang ada. Dari hasil dari *confusion matrix* didapatkan *sensitivity/recall* sebesar 0.8310, *precision* sebesar 0.8013, dan *F1-score* sebesar 0.8128. Nilai *F1-score* yang tinggi mencerminkan keseimbangan yang baik antara presisi dan sensitivitas yang penting dalam konteks klasifikasi multi-kelas. Namun demikian, model menunjukkan kelemahan yang signifikan dalam mengklasifikasikan kelas *Tidak Sehat*, di mana tidak ada satu pun sampel yang berhasil diprediksi dengan benar, dan seluruh sampel dari kelas tersebut justru diklasifikasikan sebagai *Baik* atau *Sedang*. Hal ini mengindikasikan bahwa model belum mampu membedakan fitur-fitur khas dari kelas *Tidak Sehat* yang kemungkinan besar disebabkan oleh ketidakseimbangan distribusi data atau representasi fitur yang kurang informatif untuk kelas minoritas tersebut.



Gambar 9 Kurva ROC AUC

Selain evaluasi metrik umum, analisis lebih lanjut dilakukan menggunakan kurva ROC AUC (*Receiver Operating Characteristic - Area Under Curve*) untuk menilai

kemampuan klasifikasi model pada setiap kelas. Berdasarkan grafik ROC AUC yang dihasilkan, model menunjukkan bahwa model memiliki performa diskriminatif yang cukup baik terhadap ketiga kelas, dengan nilai AUC sebesar 0.90 untuk kelas *Baik*, 0.85 untuk kelas *Sedang*, dan 0.85 untuk kelas *Tidak Sehat*. Nilai AUC yang tinggi ini menunjukkan bahwa meskipun model tidak dapat memprediksi label kelas *Tidak Sehat* dengan tepat, model tetap mampu memberikan nilai probabilitas yang relatif akurat dalam membedakan kelas tersebut dari kelas lainnya.

KESIMPULAN DAN SARAN

Penelitian ini berhasil mengklasifikasikan kualitas udara di DKI Jakarta menggunakan model Long Short-Term Memory (LSTM) berbasis *multivariate time-series*. Model yang dibangun dapat menginternalisasi pola temporal kualitas udara dengan cukup baik, dibuktikan dari hasil akurasi *training* sebesar 84,02%, akurasi validasi sebesar 79%, dan akurasi *testing* sebesar 83,45%. F1-Score rerata adalah 0.8128 dan ROC AUC lebih dari 0,85 untuk semua kelas menunjukkan klasifikasi yang dihasilkan model relatif baik, terutama dalam membedakan antara kelas kualitas udara dengan distribusi tidak seimbang. Walaupun pada hasilnya model tidak begitu dapat memprediksi label kelas *Tidak Sehat* dengan tepat. Secara keseluruhan, model LSTM menunjukkan kinerja yang menjanjikan namun masih perlu perbaikan khususnya dalam menangani kelas minoritas.

Dalam penelitian selanjutnya, disarankan dilakukan *hyperparameter tuning* secara menyeluruh guna memperoleh konfigurasi model LSTM yang lebih optimal. Penyesuaian terhadap parameter seperti jumlah unit LSTM, jumlah lapisan tersembunyi, ukuran *batch*, tingkat *learning rate*, serta teknik regularisasi seperti dropout dapat secara signifikan meningkatkan performa model. Selain itu, strategi yang dapat dipertimbangkan adalah melakukan penyeimbangan ulang dataset menggunakan penggunaan *class weights* atau *focal loss* dalam pelatihan, serta penyesuaian ambang batas klasifikasi (*threshold*) untuk meningkatkan sensitivitas terhadap kelas *Tidak Sehat*. Langkah-langkah tersebut diharapkan dapat meningkatkan akurasi prediksi dan memastikan bahwa model memiliki performa yang seimbang untuk seluruh kelas.

DAFTAR PUSTAKA

Agista PI, Gusdini N, Maharani MDD. 2020. Analisis kualitas udara dengan indeks standar pencemaran udara (ISPU) dan sebaran kadar polutannya di Provinsi DKI Jakarta. *SEOI*. 2(2):39–57. doi: https://doi.org/10.36441/seoi.v2i2.491.

Alex Sherstinsky. 2020. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network, physica d: nonlinear phenomena. 404. doi: https://doi.org/10.1016/j.physd.2019.132306.

Al-Selwi SM, Hassan MF, Abdulkadir SJ, Muneer A, Sumiea EH, Alqushaibi A, Ragab MG. 2024. RNN-LSTM: From applications to modeling techniques and beyond—Systematic review. *Journal of King Saud University-Computer and Information Sciences*. doi: https://doi.org/10.1016/j.jksuci.2024.102068.

Amalia A, Zaidiah A, Isnainiyah IN. 2022. Prediksi kualitas udara menggunakan algoritma K-Nearest Neighbor. *JIPI*. 7(2):496–507. doi: https://doi.org/10.29100/jipi.v7i2.2843.

Arba S. 2019. Kosentrasi respirable debu particulate matter (PM2, 5) dan gangguan kesehatan pada masyarakat di pemukiman sekitar PLTU. *Promotif: Jurnal Kesehatan Masyarakat*, *9*(2):178-184. doi: https://doi.org/10.56338/pjkm.v9i2.963.

Ardhaningtyas RU, Mahmudah L. 2019. Verifikasi metode pengujian NO_2 dan SO_2 dalam udara ambient (*Verification of Method for Testing NO_2 and SO_2 in Ambient Air*). *Jurnal Teknologi Proses dan Inovasi Industri*, 4(1): 9–15.

Bengio Y, Simard P, Frasconi P. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*. *5*(2):157-166. doi: https://doi.org/10.1109/72.279181.

Clayton S. 2019. Psychology and climate change. *Current Biology*, *29*(19), R992-R995. doi: https://doi.org/10.1016/j.cub.2019.07.017.

Dwangga M. 2018. Intensitas polusi udara untuk penunjang penataan ruang Kota Pelaihari Kabupaten Tanah Laut. *Metode Jurnal Teknik Industri*. 4(2): 69-77. doi: https://doi.org/10.33506/mt.v4i2.1461.

Dyana JS, Amelia RN, Davita SAM, Arafah YA, Sede AI, Hidayati AR. 2025. Dampak bahaya pencemaran udara terhadap kesehatan masyarakat di perkotaan. *Jurnal Ilmiah Wahana Pendidikan*. 11(1A):132–140. url: https://jurnal.peneliti.net/index.php/JIWP/article/view/9583.

Elman JL. 1990. Finding structure in time. *Cognitive science*. 14(2):179-211. doi: https://doi.org/10.1207/s15516709cog1402_1.

Glencross DA, Ho TR, Camina N, Hawrylowicz CM, Pfeffer PE. 2020. Air pollution and its effects on the immune system. *Free Radical Biology and Medicine*. 151:56–68. doi: https://doi.org/10.1016/j.freeradbiomed.2020.01.179.

Insani F, Darlianti SI. 2019. Pembentukan model regresi linier menggunakan algoritma genetika untuk prediksi parameter Indeks Standar Pencemar Udara (ISPU). *Jurnal CoreIT: Jurnal Hasil Penelitian Ilmu Komputer dan Teknologi Informasi*. 5(2):110–117. doi: http://dx.doi.org/10.24014/coreit.v5i2.9157.

Jorquera H, Montoya LD, Rojas NY. 2019. Urban air pollution. In *Urban Climates in Latin America* (pp. 137-165). Cham: Springer International Publishing. doi : https://doi.org/10.1007/978-3-319-97013-4_7.

Kannajmi AR, Saputra D. 2025. Penentuan model algoritma klasifikasi terbaik untuk klasifikasi kualitas udara di Jakarta 2023. *Jurnal Informatika dan Teknik Elektro Terapan*. 13(1). doi: http://dx.doi.org/10.23960/jitet.v13i1.5664.

Keputusan Kepala Bapedal No. 107 Tahun 1997 Tentang : Perhitungan Dan Pelaporan Serta Informasi Indeks Standar Pencemar Udara, no. 107. 1997.

Kurniawan A. 2018. Pengukuran parameter kualitas udara (CO, NO_2, SO_2, O_3 dan PM_{10}) di Bukit Kototabang berbasis ISPU. *Jurnal Teknosains*. 7(1):1–13. doi: https://doi.org/10.22146/teknosains.34658.

Rita R, Lestiani DD, Panjaitan EH, Santoso M, Yulinawati H. 2016. Kualitas udara (Pm10 dan Pm2. 5) untuk melengkapi kajian indeks kualitas lingkungan hidup. *Ecolab*. 10(1):1-7. doi: http://dx.doi.org/10.20886/jklh.2016.10.1.1-7

Mathisen G. 2018. *Forecasting multivariate time series data using neural networks*. [Tesis]. Norwegian University of Science and Technology, Trondheim. url: http://hdl.handle.net/11250/2559922.

Maio S, Sarno G, Tagliaferro S, Pirona F, Stanisci I, Baldacci S, Viegi G. 2023. Outdoor air pollution and respiratory health. *The International Journal of Tuberculosis and Lung Disease*. 27(1):7–12. doi: https://doi.org/10.5588/ijtld.22.0249.

Meo SA, Salih MA, Alkhalifah JM, Alsomali AH, Almushawah AA. 2024. Environmental pollutants particulate matter (PM2. 5, PM10), Carbon Monoxide (CO), Nitrogen dioxide (NO2), Sulfur dioxide (SO2), and Ozone (O3) impact on lung functions. *Journal of King Saud University-Science*, *36*(7), 103280. doi: https://doi.org/10.1016/j.jksus.2024.103280.

Sivarethinamohan R, Sujatha S, Priya S, Gafoor A, Rahman Z. 2021. Impact of air pollution in health and socio-economic aspects: review on future approach. *Materials Today: Proceedings*, *37*, 2725-2729. doi: https://doi.org/10.1016/j.matpr.2020.08.540.

Syabani DR. 2022. Klasifikasi buah segar dan busuk menggunakan algoritma Convolutional Neural Network dengan TFLite sebagai media penerapan model machine learning. [Disertasi]. *Institut Sains dan Teknologi AKPRIND Yogyakarta*. url: http://eprints.akprind.ac.id/id/eprint/3050.

Wiranda L dan Sadikin M. 2020. Penerapan long short term memory pada data time series untuk memprediksi penjualan produk PT. Metiska Farma. *Jurnal Nasional Pendidikan Teknik Informatika: JANAPATI*. 8(3): 184–196. doi: https://doi.org/10.23887/janapati.v8i3.19139.

Yang SW, Guo HR. 2021. To predict PM2.5 by a deep learning method of long-short term memory network – A case study of Kaohsiung City. *ISEE 2021: 33rd Annual Conference of the International Society of Environmental Epidemiology*. ISEE Conference Abstracts. doi: https://doi.org/10.1289/isee.2021.P-526.

Yanti NPLP, Tuningrat IM, Wiranatha AAPA. 2016. Analisis peramalan penjualan produk kecap pada perusahaan kecap Manalagi Denpasar Bali. *Jurnal Rekayasa Dan Manajemen Agroindustri*. 4:72–81. url: https://ojs.unud.ac.id/index.php/jtip/article/view/19588.

Yu Y, Si X, Hu C, Zhang J. 2019. A review of recurrent neural networks: LSTM cells and network architectures. *Neural computation*. *31*(7):1235-1270. doi: https://doi.org/10.1162/neco_a_01199.