UNIVERSITY OF TECHNOLOGY, SYDNEY

Personal Project

Artificial Intelligence

Urban Air Quality Predictor

Personal Project

Sophie Lam

December 27, 2025

Contents

List of Figures

List of Tables

Chapter 1

Overview

1.1 Background

Air quality monitoring and forecasting are critical for public health, urban planning, and environmental policy. PM2.5 (particulate matter with diameter ≤ 2.5 micrometers) is a key indicator of air pollution, capable of penetrating deep into the respiratory system and causing serious health effects.

1.2 Project Description

This project develops an LSTM-based neural network for hourly PM2.5 (fine particulate matter) prediction in the Sydney urban area. The system ingests six months of air quality measurements and meteorological data, processes and validates the data pipeline, and trains a deep learning model for short-term air quality forecasting.

1.3 Objectives

The main objectives of this project include:

- **Data Pipeline Development**: Create an automated system to collect air quality and weather data from multiple APIs
- **Data Quality Assurance**: Implement validation and preprocessing pipelines for reliable model training
- **Predictive Modeling**: Develop an LSTM neural network capable of forecasting hourly PM2.5 concentrations

- **Operational Forecasting**: Generate 24-hour ahead predictions for practical applications

1.4 Report Structure

The structure of this report is organized as follows:

- **Chapter 1:** Overview - Project background, objectives, and scope
- **Chapter 2:** Data Collection - Data sources, collection pipeline, and quality assessment
- **Chapter 3:** Model Architecture - LSTM network design and training configuration
- **Chapter 4:** Results and Conclusion - Performance evaluation and future work

Chapter 2

Data Collection

2.1 Data Sources

2.1.1 OpenAQ API v3 - Air Quality Data

OpenAQ provides open-access air quality data from government-operated monitoring stations worldwide. Table 2.1 shows the collected parameters.

Table 2.1: Collected Air Quality Parameters from OpenAQ

Parameter	**Units**	**Purpose**	**Sensor ID**
PM2.5	$\mu g/m^3$	Target variable	25196
PM10	$\mu g/m^3$	Strong predictor	25195
NO_2	ppm	Traffic pollution	25192
SO_2	ppm	Industrial source	25194
CO	ppm	Combustion indicator	23019
O_3	ppm	Atmospheric chemistry	25193

2.1.2 Open-Meteo API - Weather Data

Open-Meteo provides free historical and forecast weather data without API key requirements. Table 2.2 shows the collected weather variables.

Table 2.2: Collected Weather Variables from Open-Meteo

Category	**Variables**
Temperature	temperature_2m, dew_point_2m, apparent_temperature
Humidity	relative_humidity_2m
Precipitation	precipitation, rain
Pressure	pressure_msl, surface_pressure
Wind	wind_speed_10m, wind_direction_10m, wind_gusts_10m
Other	cloud_cover, is_day, sunshine_duration

2.2 Data Quality Summary

2.2.1 Pollutant Data

Table 2.3: Pollutant Data Quality Metrics

Metric	**Value**
Total Records	4,938
Columns	7
Duplicate Timestamps	611
Hourly Gaps	638

Missing value rates: PM2.5 (2.88%), PM10 (0.77%), NO_2 (2.47%), SO_2 (2.55%), CO (3.24%), O_3 (1.96%).

2.2.2 Weather Data

The weather data contained 4,416 records with 15 columns, zero duplicates, zero hourly gaps, and no missing values.

2.2.3 Merged Dataset

After merging pollutant and weather data:

- **Final Records**: 4,931
- **Total Features**: 21
- **Date Range**: 2025-06-01 07:00 UTC to 2025-12-01 23:00 UTC

2.2.4 Missing Data Handling

1. **Linear Interpolation**: Applied on time index for small gaps (< 3.3%)
2. **Forward/Backward Fill**: Used for edge cases
3. **Result**: Zero null values in processed dataset

2.3 PM2.5 Time Series Visualization

Figure 2.1 displays the hourly PM2.5 values over the last month of the study period. The chart reveals daily cyclic patterns typical of urban air pollution—concentrations tend to rise during morning and evening rush hours due to increased traffic and fall overnight. The visible day-to-day variability reflects weather influences (wind dispersing pollutants, rain washing particles from the air) and weekly patterns (lower pollution on weekends).

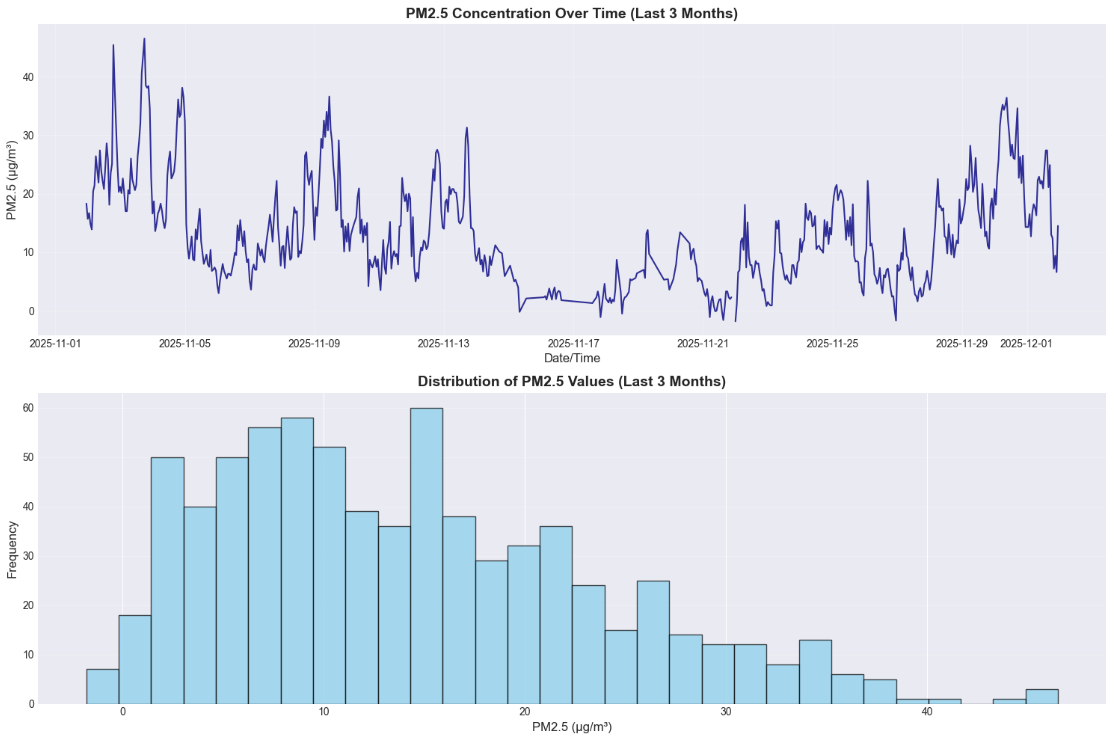

Figure 2.1: PM2.5 concentration over the last month of the study period.

2.4 Feature Correlations

Figure 2.2 shows the correlation matrix for all 21 features. Key observations:

- PM10 has the strongest correlation with PM2.5 (0.87), which is expected since both measure particulate matter from similar sources
- CO, NO_2, and SO_2 show moderate positive correlations with PM2.5, indicating shared emission sources (vehicles, industry)
- Wind speed shows negative correlation—higher winds disperse pollutants, reducing PM2.5 concentrations

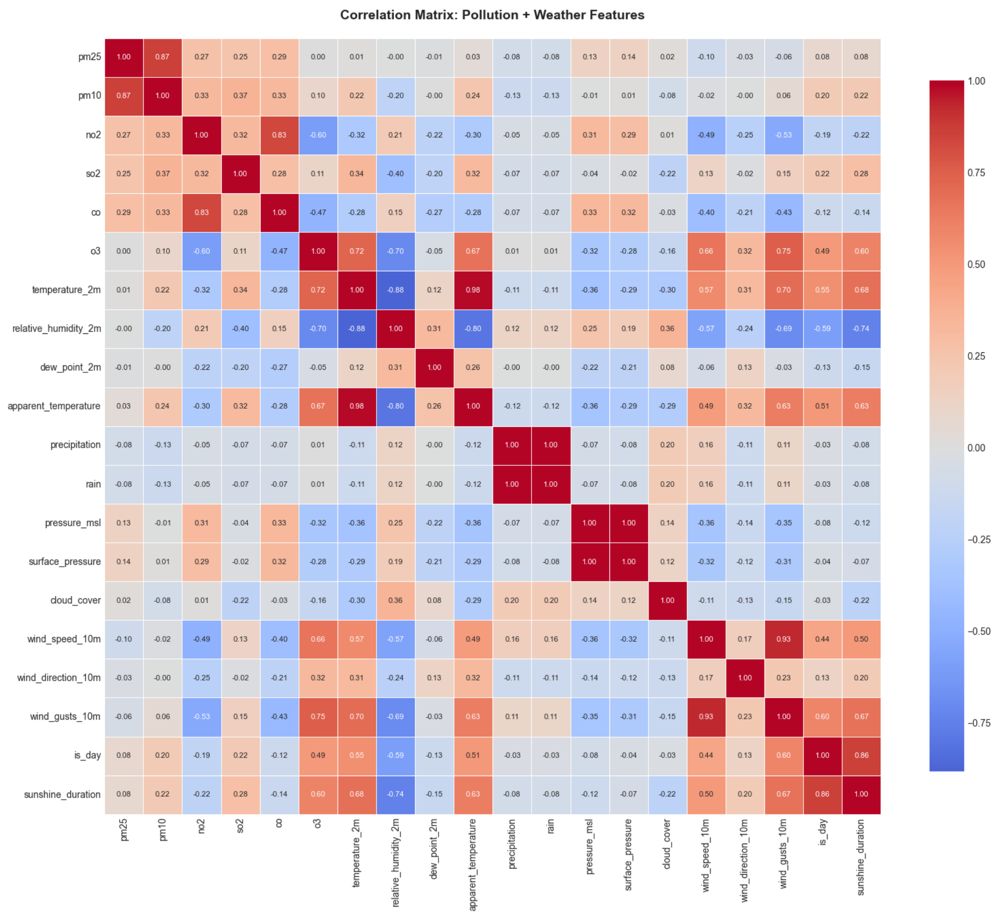

Figure 2.2: Correlation matrix showing relationships between all 21 features.

Chapter 3

Model Architecture

3.1 Why LSTM?

Long Short-Term Memory (LSTM) networks are a type of recurrent neural network (RNN) specifically designed for sequential data. They excel at air quality prediction because:

1. **Temporal dependencies**: Air pollution at any hour depends on conditions from previous hours—LSTM's memory cells capture these temporal patterns
2. **Long-range patterns**: Unlike standard RNNs, LSTMs can learn relationships spanning many time steps (e.g., pollution buildup over 24 hours)
3. **Non-linear relationships**: LSTMs model complex interactions between weather, traffic patterns, and pollution that linear models cannot capture

3.2 Network Architecture

The LSTM model architecture consists of the following layers:

- **Input**: 24-hour lookback window $\times$ 24 features (the model "sees" the last 24 hours to predict the next hour)
- **LSTM(64)**: First LSTM layer with 64 hidden units learns complex temporal patterns
- **Dropout(0.2)**: Randomly drops 20% of connections during training to prevent overfitting

- **LSTM(32)**: Second LSTM layer with 32 units further refines temporal representations
- **Dropout(0.2)**: Additional regularization layer
- **Dense(32, ReLU)**: Fully connected layer transforms LSTM output for prediction
- **Dense(1)**: Output layer produces the single PM2.5 prediction

Total Parameters: ∼36,000 trainable weights

3.3 Training Configuration

Table 3.1: Training Configuration Parameters

Parameter	**Value**
Optimizer	Adam
Learning Rate	1×10^{-3}
Loss Function	Mean Squared Error (MSE)
Metrics	MAE, MSE
Batch Size	32
Max Epochs	100
Validation Split	20% of training data

3.4 Regularization Strategies

- **Dropout**: 0.2 rate after each LSTM and Dense layer
- **EarlyStopping**: Patience of 15 epochs, restores best weights
- **ReduceLROnPlateau**: Factor of 0.5, patience of 5 epochs

3.5 Sequence Configuration

Table 3.2: Sequence Configuration

Parameter	**Value**
Lookback Window	24 hours
Prediction Horizon	1 hour ahead
Total Sequences	4,907
Train/Test Split	70% / 30%
Training Samples	3,434
Test Samples	1,473

3.6 Training Progress

Figure 3.1 shows the training and validation loss over epochs. The decreasing curves demonstrate that the model is learning effectively. The gap between training and validation loss indicates some overfitting—the model fits training data better than held-out validation data. Early stopping (patience=15) prevented excessive overfitting by stopping training when validation loss stopped improving.

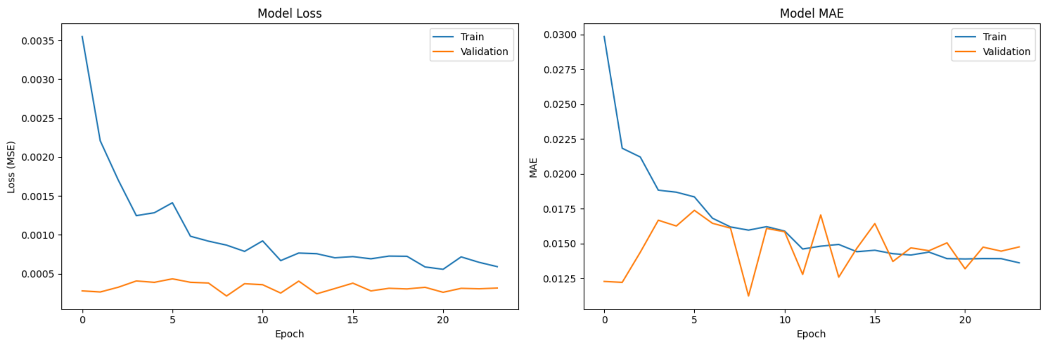

Figure 3.1: Training and validation loss over epochs.

Chapter 4

Results and Conclusion

4.1 Model Performance

Table 4.1 shows the model's performance on training and test sets.

Table 4.1: Model Performance Metrics

Metric	Training	Test	Gap
MAE	3.41 μg/m^3	5.93 μg/m^3	+2.52
RMSE	5.88 μg/m^3	8.09 μg/m^3	+2.21
R^2	0.74	0.17	-0.57

4.1.1 Metric Explanations

- **MAE (Mean Absolute Error)**: Average prediction error in μg/m^3. Training MAE of 3.41 means predictions are off by ~3.4 units on average.
- **RMSE (Root Mean Squared Error)**: Penalizes larger errors more heavily. Higher RMSE indicates some predictions have significant errors.
- **R^2 (Coefficient of Determination)**: Measures how much variance the model explains. Training R^2=0.74 means 74% of PM2.5 variation is captured; positive test R^2 (0.17) indicates the model has some predictive power on unseen data.

4.2 Actual vs Predicted

Figure 4.1 compares actual PM2.5 values (blue) with model predictions (orange) for both training and test sets. The training plot shows predictions closely tracking actual values, demonstrating the model learned the underlying patterns. However, the test plot reveals a generalization gap—the model struggles to capture the full variability of unseen data, often predicting values closer to the mean.

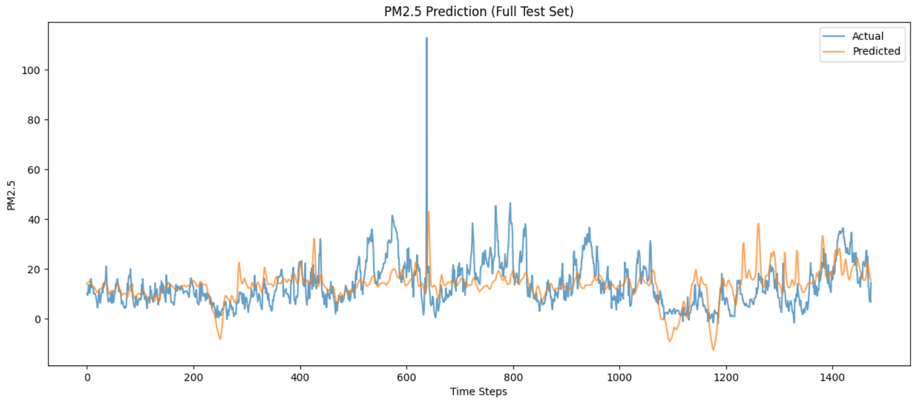

Figure 4.1: Actual vs predicted PM2.5 values for training and test sets.

4.3 24-Hour Forecast

Figure 4.2 shows the 24-hour ahead forecast using the most recent data. The chart overlays the predicted next 24 hours (orange) on top of the last 72 hours of actual data (blue). The forecast predicts stable PM2.5 values around 6.6–7.2 $\mu g/m^3$, which falls in the "Good" air quality category according to EPA standards.

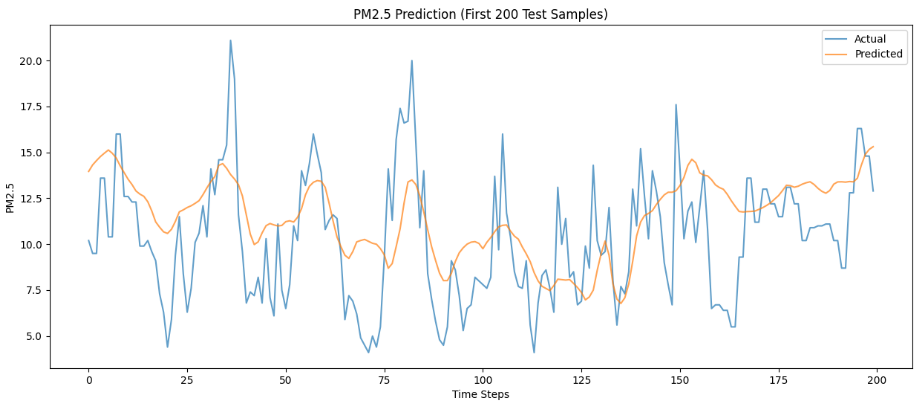

Figure 4.2: 24-hour ahead PM2.5 forecast overlaid on recent historical data.

4.4 Conclusion

The LSTM model successfully learns temporal patterns in the training data ($R^2 = 0.74$), demonstrating that air quality prediction using deep learning is feasible. However, the gap on test data ($R^2 = 0.17$) indicates challenges with generalization, possibly due to:

- Seasonal or event-based pollution patterns not seen in training
- Limited six-month training window
- Need for additional features (traffic data, fire events, etc.)

4.5 Future Work

Future improvements could include:

- **Extended Data**: Collect longer historical period (1–2 years) to capture seasonal patterns
- **Additional Features**: Include traffic data, wildfire events, and industrial activity
- **Alternative Architectures**: Explore Transformer models or Temporal Convolutional Networks
- **Ensemble Methods**: Combine LSTM with gradient boosting (XGBoost, LightGBM)
- **Multi-station Learning**: Train on data from multiple monitoring stations

Bibliography

[1] OpenAQ. Open air quality data platform, 2025.

[2] Open-Meteo. Free weather api, 2025.

[3] TensorFlow Team. Tensorflow: An end-to-end open source machine learning platform, 2025.

[4] U.S. Environmental Protection Agency. Air quality index (aqi) basics, 2025.

[5] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.