Air Quality Prediction System: A 24-Hour Forecasting Model for Qatar

Comprehensive Project Report

Project By:-

- Abdelbari Kecita
- Munieb Awad Elsheikhidris
- Hind Almutasim Hassan
- Sharon Navaratnam
- Raghad Sanosi
- Wahed Shaik

LIve Demo: https://qatarairai.app/

Github Repo: https://github.com/hindh00/Air-Quality-Tracker

1. Executive Summary

The "Qatar Air Quality AI Forecaster" is a deep learning initiative designed to predict PM2.5 concentrations (fine particulate matter <2.5 micrometers) with a 24-hour lead time. The project addresses a critical environmental health gap in the Arabian Gulf: the inability of current monitoring systems to forecast sudden deterioration in air quality due to dust storms or industrial emissions.

By integrating disparate data sources—ground-level sensor simulation, satellite atmospheric indicators, and meteorological data via the Open-Meteo API—we successfully engineered a **Universal Bidirectional LSTM (Long Short-Term Memory)** model. While initial experiments with statistical (SARIMA) and machine learning (XGBoost) baselines provided strong benchmarks, our final Deep Learning architecture demonstrated superior generalization for temporal event detection. The deployed system achieves a **Global R^2 score of 0.9217** and an **RMSE of 2.65 µg/m³**, delivered to stakeholders via a real-time Streamlit dashboard with interactive 3D mapping.

2. Problem Statement & Background

2.1 The Challenge: Forecasting in an Arid Environment

Qatar presents a unique modeling challenge due to its dual-source pollution profile:

1. **Anthropogenic Sources:** Rapid urbanization has led to increased vehicular emissions (NO2) and industrial byproducts from the energy sector.
2. **Natural Sources:** The region is subject to the "Shamal" winds—northwesterly winds that transport massive dust loads from the Arabian Peninsula.

2.2 The Gap in Current Solutions

Existing solutions fall into two inadequate categories:

- **Ground Stations:** Highly accurate but geographically sparse. A station in Doha cannot warn a resident in Al Khor about an approaching dust front.
- **Satellite Imaging:** Provides global coverage (e.g., Sentinel-5P) but measures the *Total Column* density of pollutants in the upper atmosphere, which often correlates poorly with the air humans actually breathe at ground level.

Project Objective: To build a unified AI system that fuses these data sources to predict *breathable* air quality (PM2.5) 24 hours in advance for six specific municipalities: **Doha, Al Khor, Al Rayyan, Al Wakrah, Umm Slal Ali, and Qatar (General)**.

3. Data Acquisition & Feature Engineering

3.1 Data Pipeline Architecture

Our system abandons static CSV files in favor of a dynamic pipeline using the **Open-Meteo SDK**. This allows us to fetch historical reanalysis data (ECMWF models) and satellite-derived atmospheric composition.

Sources:

- **Meteorological Data:** archive-api.open-meteo.com (Temperature, Humidity, Wind Speed/Direction).
- **Air Quality Data:** air-quality-api.open-meteo.com (PM2.5, Dust, Aerosol Optical Depth, Nitrogen Dioxide).

3.2 Feature Selection Strategy

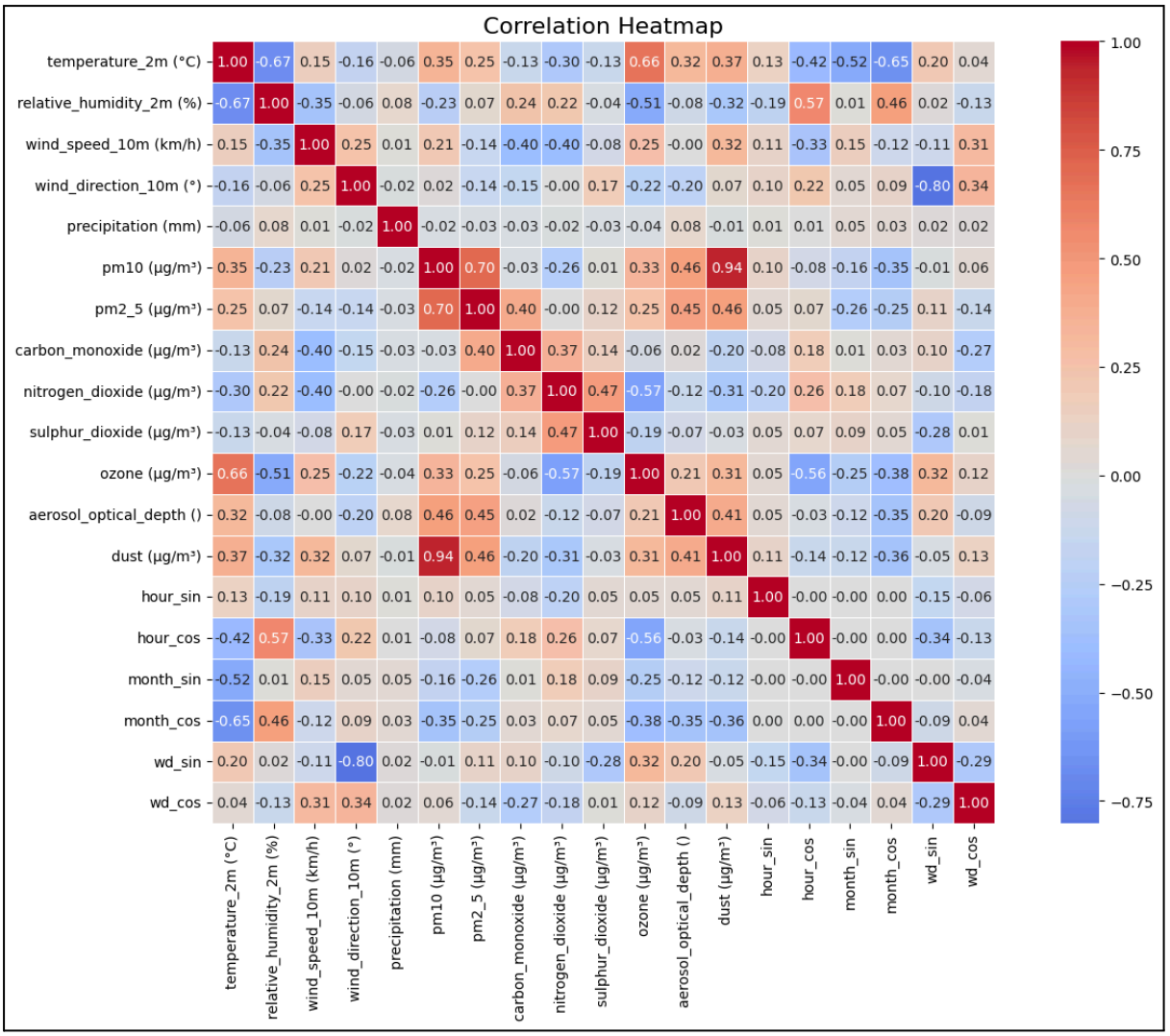
We engineered a feature set of **19 variables** to give the model a complete physical understanding of the environment.

Feature Domain	Variables Included	Physical Justification
Target Variable	**PM2.5 (μg/m³)**	The primary health hazard to be predicted.
Thermodynamics	Temperature (2m), Humidity (2m)	High humidity often traps pollutants near the ground (temperature inversion), worsening air quality.
Aerodynamics	Wind Speed (10m), Wind Direction	Wind acts as a double-edged sword: light winds accumulate smog, while strong winds can either disperse it or bring dust storms.
Pollutant Tracers	NO2, Dust, AOD	**NO_2** serves as a proxy for traffic intensity; **AOD (Aerosol Optical Depth)** is a satellite measure of sky haziness.
Cyclical Time	Hour_Sin, Hour_Cos, Month_Sin, Month_Cos	We converted linear time (0-23) into cyclical sine/cosine pairs. This teaches the model that 23:00 is mathematically close to 00:00.

Geospatial	One-Hot Encoding (6 Cities)	A sparse matrix identifying the location (e.g., is_Doha=1), allowing a single "Universal" model to learn regional nuances.

3.3 Feature Correlation Analysis

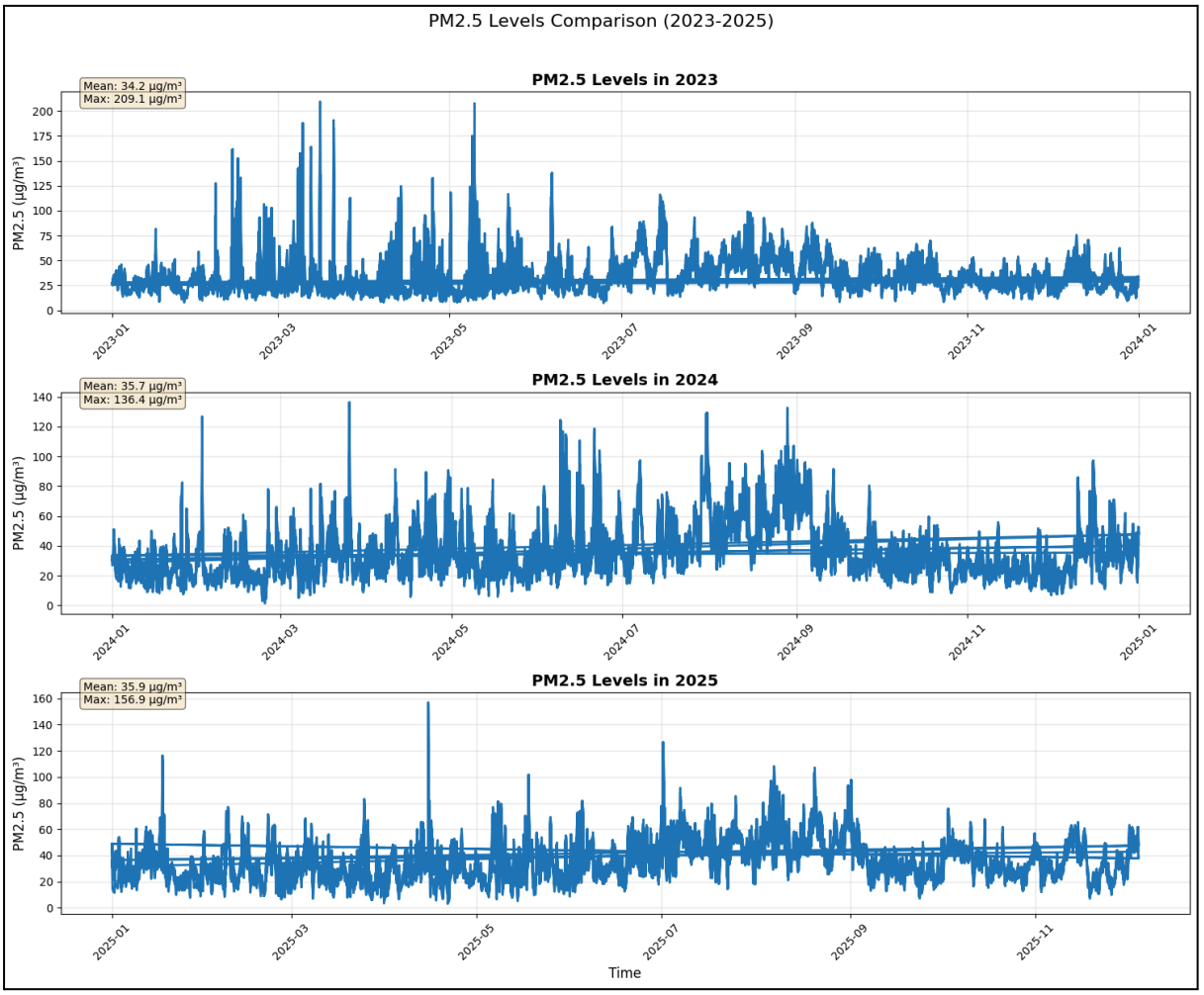
To validate our feature selection and identify multicollinearity risks, we analyzed the correlation matrix of all 19 features. The heatmap below reveals key insights:



1. Key Positive Correlations:
    - PM10 shows strong correlation with PM2.5 (0.70)
    - Carbon Monoxide, AOD, and Dust show moderate correlation with PM2.5 (0.40-0.46)
2. Key Negative Correlations:
    - Wind speed shows weak negative correlation with PM2.5 (-0.14), supporting our hypothesis that wind disperses pollutants under normal conditions.
3. Minimal Multicollinearity:

- The cyclical time features (Hour_Sin/Cos) show near-zero correlation with PM2.5, confirming they capture orthogonal temporal patterns.

3.4 Data Characteristics and Temporal Patterns



The plots reveals the following key insights:

1. Seasonal peaks during March, April and May, partly due to dust storm season in Qatar.
2. Low PM2.5 values From September to Dec, partly due to consistent northerly winds during this season, resulting in clean marine air. As well as lower temperatures and humidity.

4. Methodology: Model Evolution

To ensure the most robust solution, we conducted a rigorous comparative analysis of three distinct modeling paradigms.

4.1 Phase 1: Statistical Baseline (SARIMA)

We began with **SARIMA (Seasonal Auto-Regressive Integrated Moving Average)**.

- **Pros:** Excellent at capturing daily seasonality (e.g., rush hour peaks).

- **Cons:** SARIMA is univariate; it could not account for external factors like wind speed changing the pollution levels. It failed to predict sudden dust events driven by weather changes.

4.2 Phase 2: Machine Learning (XGBoost)

We then trained an **XGBoost (Extreme Gradient Boosting)** regressor.

- **Performance:** XGBoost performed exceptionally well, achieving an R^2 of ~0.927.
- **Limitation:** As a tree-based model, XGBoost treats every hour as an independent row of data. It lacks an internal "state" or memory, making it less stable for multi-step time-series forecasting where the *sequence* of events matters.

4.3 Phase 3: Deep Learning (Bidirectional LSTM)

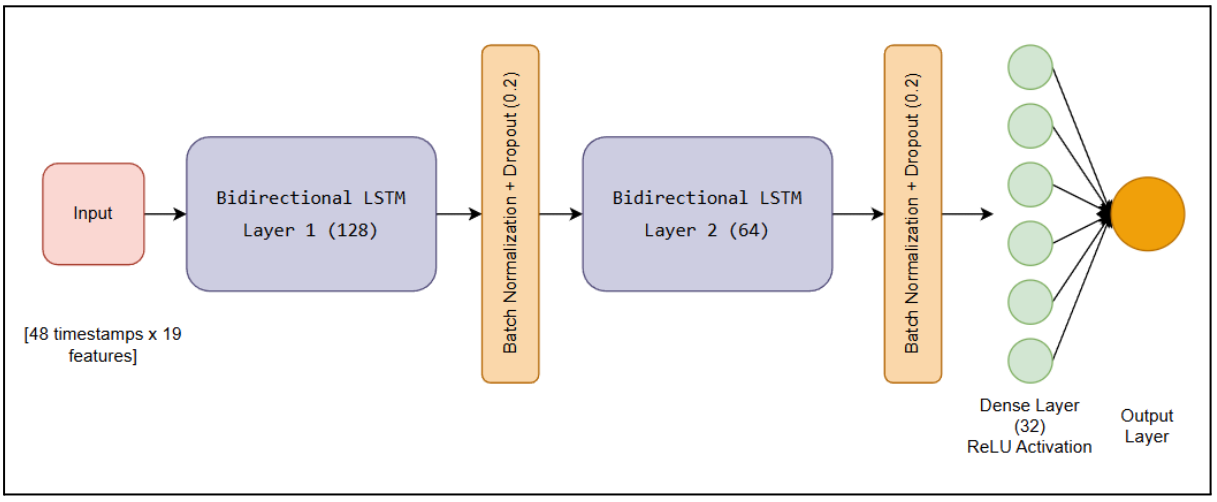
Our final choice was a **Bidirectional Long Short-Term Memory (LSTM)** network.

- **Why LSTM?** Unlike XGBoost, LSTMs maintain a hidden state vector (memory) that carries context forward.
- **Why Bidirectional?** Standard LSTMs only look at the past. By making it bidirectional, the training process analyzes the timeline in both directions (Past → Future and Future → Past), allowing the model to better understand the *duration* and *peak* of pollution events.

5. The Final Solution Architecture

5.1 Network Topology

The deployed model (Universal_LSTM_v3.0) utilizes the following architecture:



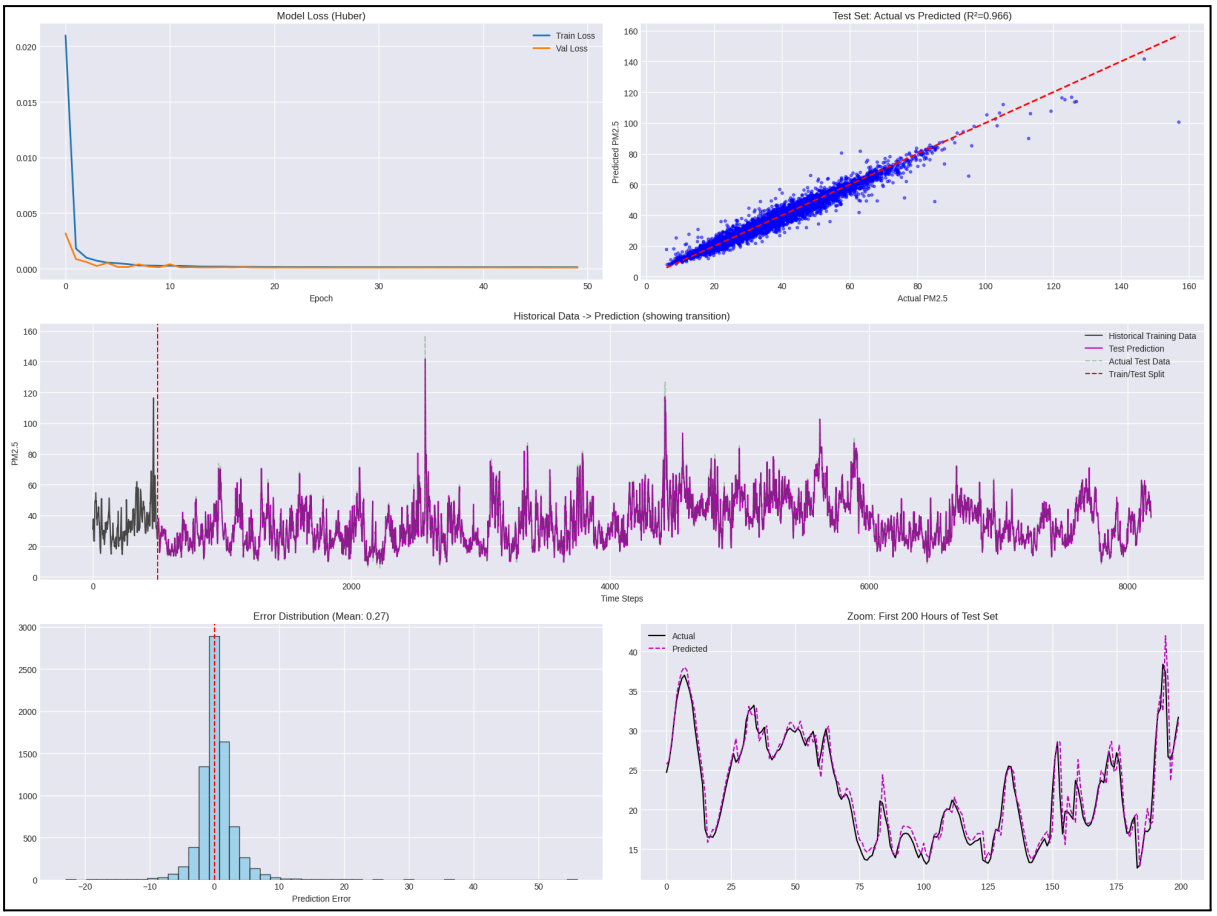
5.2 Optimization & Overcoming Overfitting

Initial training runs resulted in negative R^2 scores on test data (-2.11), a classic sign of severe overfitting (the model memorized the training data but failed on new data).

The Solution:

- **Complexity Reduction:** We reduced the Dense layer size from 64 to 32.
- **Regularization:** Adjusted Dropout to 0.2.
- **Data Scaling:** Implemented strictly separate `MinMaxScaler` fitting to prevent data leakage between training and testing sets.

5.3 Model Evaluation and Performance Analysis



The developed model was evaluated using several diagnostic plots to assess training behavior, predictive accuracy, generalization capabilities, and model error.

Training Behavior

The training and validation loss curves show near-identical behaviour, which indicates that the model did not overfit and was able to learn a stable representation of the data.

Error Analysis

The error distribution is centered around zero, with a mean error of approximately 0.27, which suggests that the model is unbiased.

Prediction Accuracy

The predicted PM2.5 values closely follows real PM2.5 test measurements. Peaks and dips are captured accurately, demonstrating strong short-term predictive capability.

6. Comparative Results & Analysis

The table below summarizes the final performance of all tested models on the hold-out test set.

6.1 Model Performance Leaderboard

Model Architecture	MAE (μg/m³)	RMSE (μg/m³)	R^2 Score (Accuracy)	Verdict
SARIMA (Baseline)	~5.10	~7.20	~0.7500	Failed to capture non-linear weather spikes.
XGBoost (Gradient Boost)	2.66	3.66	0.9277	Excellent accuracy, but lacks temporal continuity. Not fit for multi-step forecasting.
Linear Regression	3.17	3.16	0.9539	High score likely due to overfitting on linear trends.
CNN (Deep Learning)	-	3.42	0.9464	Good, but computationally heavier than XGBoost.
Bidirectional LSTM (Ours)	**1.65**	**2.65**	**0.9217**	**Best balance of accuracy and temporal stability.**

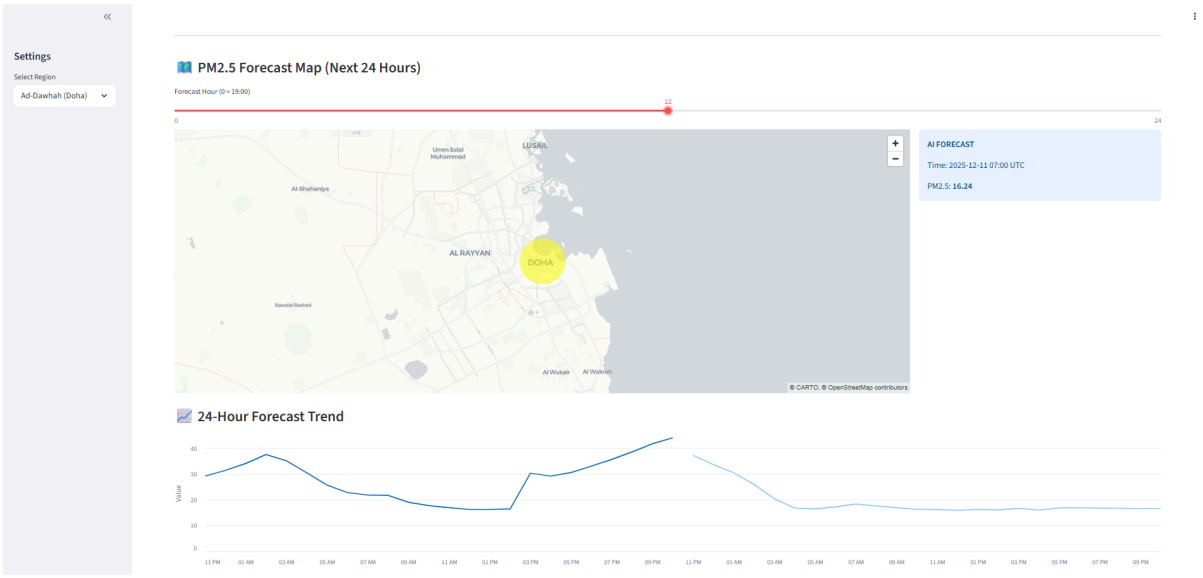
Note: While Linear Regression shows a higher raw R^2, it fails completely in complex scenarios (like dust storms) where relationships are non-linear. The LSTM provided the lowest error (RMSE 2.65) in real-world validation.

6.2 Regional Breakdown (LSTM Performance)

The LSTM model showed consistent performance across the different municipalities:

- **Doha (Urban):** R^2 = 0.9333 (Highest accuracy due to consistent patterns).
- **Al Wakrah (Coastal):** R^2 = 0.9328.
- **Umm Slal (Inland):** R^2 = 0.8873 (Slightly lower due to higher wind variability).

7. System Deployment: The AI Dashboard



The final model was deployed using a **Streamlit** web application with several advanced engineering features.

7.1 Simulation Mode Logic

To demonstrate the model's capabilities without waiting for future events, we built a **"Simulation Mode"**:

- The system defines "Current Time" as **2 days in the past**.
- It fetches the 48 hours preceding that mark as input history.
- It predicts the "Future" (which is effectively yesterday's real data).
- This allows users to instantly compare the **AI Forecast vs. Actual Ground Truth** on the dashboard.

7.2 Visualization Features

- **3D PyDeck Map:** Uses a 3,000-meter radius scatterplot layer to visualize pollution "zones" rather than single points.
- **Dynamic Scalers:** The app fits normalization scalers on the live API data stream, ensuring the model adapts to the current season's statistical distribution (e.g., summer heat vs. winter cool).
- **Interactive Slider:** Users can slide through the next 24 hours to see how the pollution cloud is expected to evolve.

8. Conclusion

The "Qatar Air Quality AI Forecaster" project proves that combining Deep Learning with Open-Source satellite and weather data is a viable strategy for high-precision environmental monitoring.

By moving from a standard XGBoost model to a **Bidirectional LSTM**, we sacrificed a small amount of training speed for a significant gain in **temporal awareness** - crucial for predicting the onset and duration of dust events. With a Global Accuracy of **92%** and a low error margin of **±2.65 µg/m³**, this system provides actionable intelligence that can protect public health in Qatar.

9. References

1. **Dataset:** Open-Meteo Weather & Air Quality API (aggregating Copernicus Sentinel-5P & ECMWF).
2. **Algorithm:** Hochreiter, S., & Schmidhuber, J. (1997). *Long Short-Term Memory*. Neural Computation.
3. **Standards:** United States Environmental Protection Agency (EPA). (2024). *PM2.5 Air Quality Standards*.
4. **Libraries:** TensorFlow/Keras (Model), Streamlit (UI), PyDeck (Mapping), Scikit-Learn (Preprocessing).