LSTM-Based Time-Series Forecasting for Air Quality Prediction

1. Short Summary

Air pollution poses a significant health and environmental risks, which making an accurate air quality forecasting crucial for mitigating its effects. Where this project aims to develop a deep learning model which is based on Long Short-Term Memory (LSTM) networks that which is to predict air quality by using the historical sensor data. LSTMs, a specialized type of Recurrent Neural Network (RNN), which were well-suited for time-series forecasting due to their ability which to capture the long-term dependencies. By leveraging an advanced deep learning techniques, this project aims to improve air quality prediction accuracy and to provide insights into pollutant trends.

2. Objectives

- Developing a LSTM-based model to forecast the air quality levels.
- Optimizing the model performance through hyperparameter tuning and data preprocessing.
- Compare the LSTM model with other baseline models such as ARIMA and GRU.
- Conducting at least four experiments to evaluate different forecasting models.
- Deploying the trained model locally for real-world usability.

3. Methodology

- **Dataset:** This project will utilize the UCI Air Quality dataset, that which includes the time-series data on pollutant concentrations (e.g., CO, NO_2, PM10) which are collected from sensor networks.

- **Data Preprocessing:**
  - Handling the missing values through the interpolation.
  - Normalization of sensor readings to enhance the model training.
  - Feature engineering, which is including lag variables and moving averages.

- **Experiments:**
  - **Baseline Model (ARIMA):** Implementing and evaluating a traditional statistical model.
  - **Deep Learning Model (LSTM):** Training a deep LSTM model with the multiple layers, optimize hyperparameters, and apply dropout regularization.
  - **GRU-Based Model:** Implementing a GRU-based time-series forecasting model and comparing it with against to LSTM.

- **Hybrid or Transformer-Based Model:** Experimenting with the ConvLSTM (CNN + LSTM) or Transformer-based forecasting to test the performance improvements.
- **Training Strategy:**
    - Hyperparameter tuning (batch size, learning rate, number of layers).
    - Early stopping to prevent overfitting.
    - Data augmentation techniques to improve generalization.

4. Expected Results & Evaluation

- **Quantitative Metrics:**
    - Root Mean Squared Error (RMSE): Target < 0.07.
    - Mean Absolute Error (MAE): Target < 0.05.
    - R^2 Score: Target > 0.80.
- **Visualizations:**
    - Time-series plots comparing predicted vs. actual air quality values.
    - Correlation heatmaps of different pollutant levels.
    - Model loss curves to evaluate convergence.
- **Baseline Model Comparison:**
    - Performance comparison with the traditional models (ARIMA, GRU, and Transformer-based approaches).

5. Dataset Details

- **Source:** UCI Machine Learning Repository.
- **Features:** Air pollutant levels, temperature, humidity, pressure, timestamp.
- **Time Span:** Multiple years of historical data.
- **Data Availability:** Open source dataset which is accessible online.

6. References

- Hochreiter, S., & Schmidhuber, J. (1997). "Long Short-Term Memory." - Original LSTM paper.
- Deep Learning for Air Quality Prediction - Research paper on air quality forecasting using deep learning.
- UCI Machine Learning Repository - Air Quality Dataset - Dataset source.

- [TensorFlow LSTM Time Series Forecasting] - TensorFlow tutorial for LSTM-based time series forecasting.
- [PyTorch Time-Series Forecasting] - PyTorch documentation on time-series forecasting models.