# Name: Aisun Zarghami

# Pollution Prediction Using LSTM and RNN Models

## Introduction

This is an example of the implementation of a machine learning approach to predict pollution levels using Long Short-Term Memory (LSTM) and Simple Recurrent Neural Network (RNN) models. The dataset utilized for this analysis is derived from a CSV file named `LSTM-Multivariate_pollution.csv`, which contains multiple variables related to pollution and meteorological conditions.

## Data Loading and Preprocessing

The data was loaded using the Pandas library, and the following preprocessing steps were conducted:

1. **Label Encoding**: The categorical feature `wnd_dir` (wind direction) was transformed into numerical format using `LabelEncoder` to make it suitable for model training.

2. **Feature Selection**: The relevant features selected for the model included:

    - `temp` (temperature)
    - `press` (pressure)
    - `wnd_dir` (wind direction)
    - `wnd_spd` (wind speed)
    - `pollution` (target variable)
3. **Handling Missing Values**: Any missing values in the dataset were filled using forward fill (`ffill`) to maintain continuity in the time series data.
4. **Data Scaling**: The features were scaled to a range between 0 and 1 using `MinMaxScaler`, which is crucial for improving the performance of neural networks.

5. **Sequence Creation**: A function was implemented to create sequences of the time series data. Each sequence consisted of 24 time steps (representing 24 hours of data), which allowed the models to learn temporal dependencies.

## Data Splitting

The dataset was split into training and testing sets, with 80% of the data allocated for training and 20% for testing. The shapes of the training and testing datasets were confirmed to ensure proper partitioning.

## Model Building

Two types of models were built using the Keras Sequential API:

1. **LSTM Model**: An LSTM model was created with one LSTM layer containing 50 units, followed by a dense output layer designed to predict pollution levels.
2. **RNN Model**: A Simple RNN model was similarly constructed with one RNN layer containing 50 units and a dense output layer.

Both models were compiled using the Adam optimizer and mean squared error as the loss function.

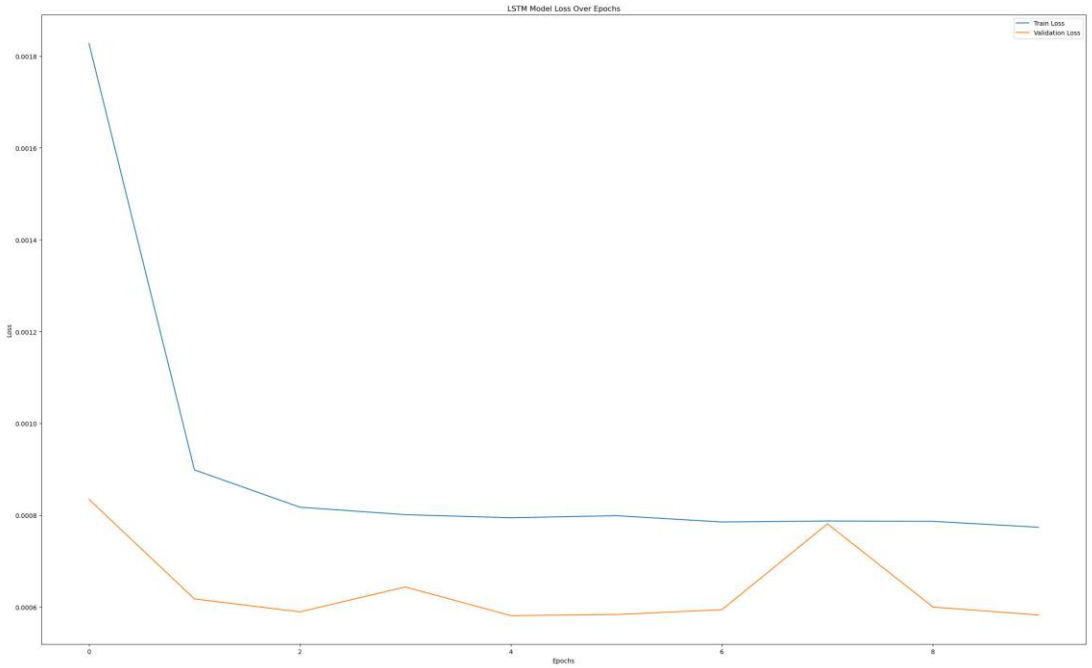
## Model Training

The models were trained for 10 epochs with a batch size of 64. Training history was captured to monitor the loss during training and validation phases.

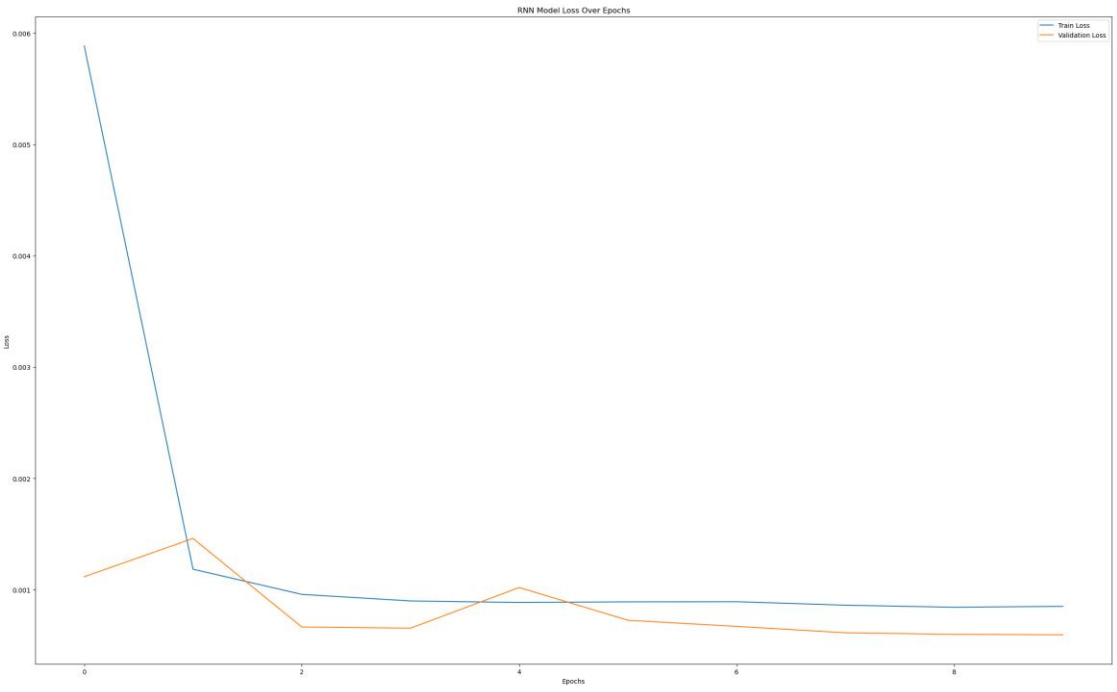
# Model Evaluation

After training, both models were evaluated on the test dataset. The loss values for each model were reported:

- **LSTM Loss**: 0.0005825438420288265




- **RNN Loss**: 0.0005936746019870043



These loss values provide insight into the models' performance, with lower values indicating better predictive accuracy.

# Training History Visualization

The training history was visualized through loss curves for both models. This helped in understanding how well each model learned over the epochs and whether they experienced overfitting.

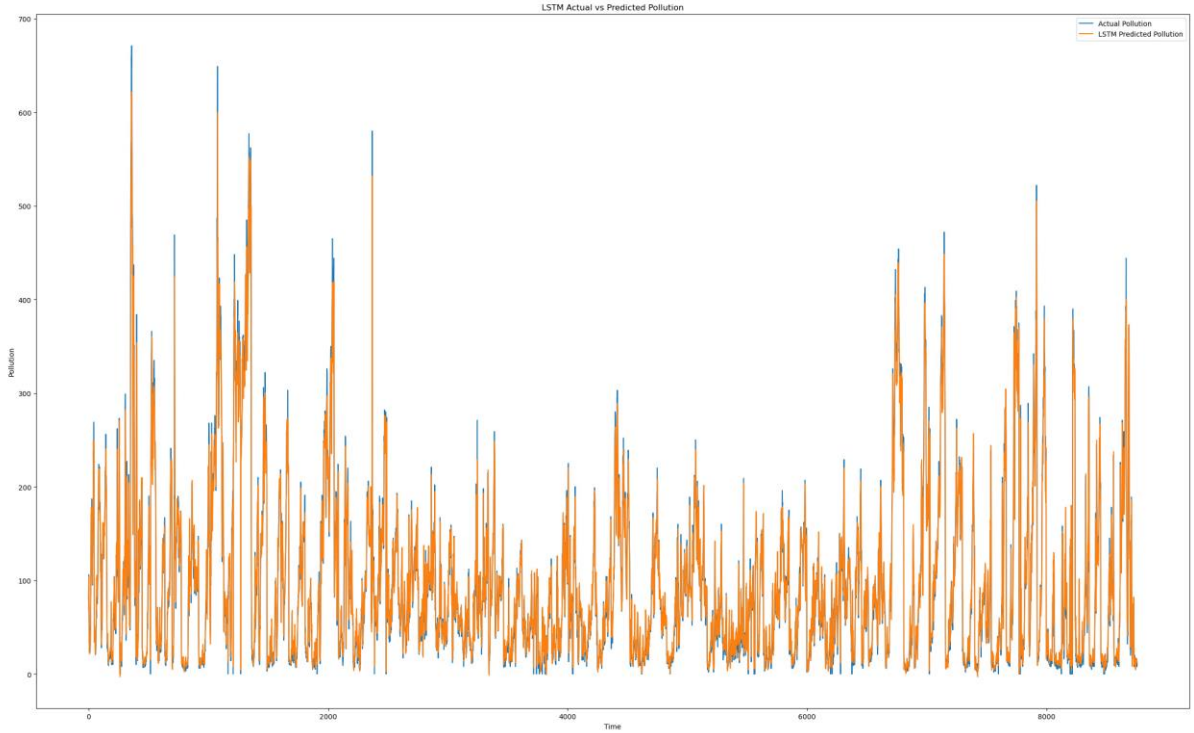
# Predictions and Results

Predictions were made on the test set using both models. The predicted pollution levels were rescaled back to their original scale for comparison with actual values.

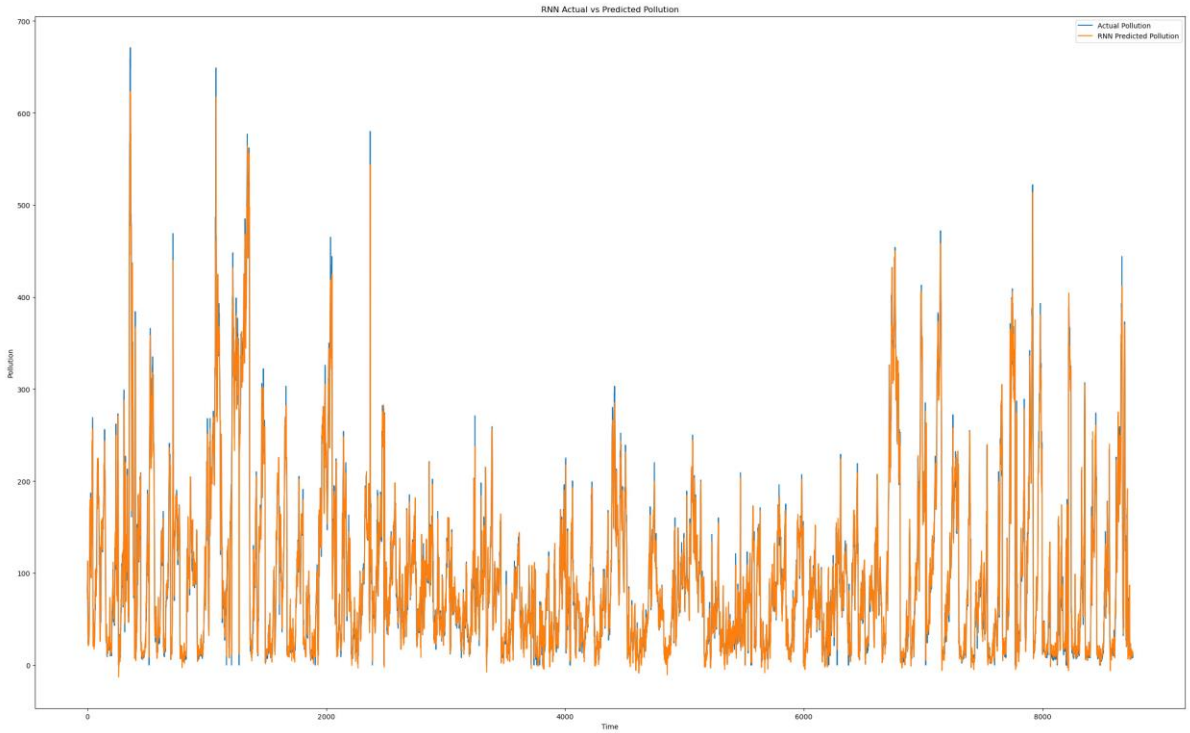
## Actual vs Predicted Pollution

The actual pollution levels were plotted against the predicted values from both models:

- **LSTM Model Predictions**:



- **RNN Model Predictions**:




These plots visually demonstrate the models' capabilities to predict pollution levels over time.

# Conclusion

The analysis successfully demonstrated the application of LSTM and RNN models in predicting pollution levels based on multivariate time series data. The LSTM model, in particular, is expected to perform better due to its ability to capture long-term dependencies in the data.