1. Introduction and Methodology

The project aims to develop a **time series regression model** for forecasting the **Air Quality Index (AQI)**. The model utilizes a Recurrent Neural Network (RNN) architecture, specifically a **Stacked Gated Recurrent Unit (GRU)**, to predict the next hour's AQI based on the preceding 24 hours of multivariate data.

Data Source and Preprocessing

Component	Detail
Dataset File	/content/final_cleaned_interpolated.csv
Data Type	Hourly Time Series Data
Time Period	2020-11-25 01:00:00 to 2025-10-24 06:00:00
Total Samples	43,062 entries
Target Variable (y)	**AQI** (Air Quality Index)
Scaling Method	**MinMaxScaler** was applied separately to the input features X and the target variable y.

Input Feature Engineering (X)

The model was trained using **18 input features** (len(feature_cols) = 18), which include a combination of air pollutant concentrations, meteorological parameters, and engineered temporal features.

- **Air Pollutants & Meteorological Variables (12 features):**
  - pm2_5_ugm3, pm10_ugm3, co_ppm, no2_ppb, o3_ppm, so2_ppb
  - temperature_2m, relative_humidity_2m, surface_pressure, precipitation, cloudcover, windspeed_10m
- **Temporal Features (6 features):**
  - hour, month
  - **Cyclic Encoding:** hour_sin, hour_cos, month_sin, month_cos (used to capture the cyclical nature of time variables)

Data Splitting and Sequence Creation

The time series data was split chronologically to maintain the temporal order.

Set	Proportion	Role
Training Set	80%	Used for model training
Validation Set	10%	Used for hyperparameter tuning and early stopping
Test Set	10%	Used for final, unbiased performance evaluation

Sequence Creation:

- **Input Sequence Length (SEQ_LEN): 24 hours** (The model uses the past 24 hourly data points to predict the AQI at the next time step).
- **Training Sequence Shape:** (34425, 24, 18) (Samples, Time Steps, Features).

---

2. Model Architecture

The forecasting model is a **Sequential Stacked GRU Network**.

Layer Type	Units/Filters	Output Shape	Activation	Dropout Rate	Parameters
GRU (1)	64	(None, 24, 64)	Tanh (Default)	-	16,320
Dropout (1)	-	(None, 24, 64)	-	**0.2**	-
GRU (2)	32	(None, 32)	Tanh (Default)	-	9,408
Dropout (2)	-	(None, 32)	-	**0.2**	-
Dense (Output)	1	(None, 1)	Linear (Default)	-	33
Total Parameters	-	-	-	-	**25,569** (all trainable)

Key Architectural Details:

- The first **GRU** layer uses return_sequences=True to pass the full sequence output to the subsequent layer, which is a standard practice for stacked RNNs.
- The second **GRU** layer implicitly uses return_sequences=False (default), returning only the output for the last time step in the sequence (the predicted value).
- The output layer is a **Dense** layer with a single unit, appropriate for a single-value regression prediction.

---

3. Training Configuration

Parameter	Value
Optimization Algorithm	**Adam** (optimizer='adam')
Loss Function	**Mean Squared Error (MSE)** (loss='mse')
Primary Metric	Mean Absolute Error (MAE) (metrics=['mae'])
Maximum Epochs	50
Batch Size	32

Parameter	Value
Hardware	GPU accelerated (Colab GPU Type: T4)
Early Stopping Callback	Monitored: val_loss
	Patience: **10 epochs**
	Restoration: restore_best_weights=True (Ensures the model state from the epoch with the lowest validation loss is used)
Best Epoch (Observed)	The model achieved its best validation loss in **Epoch 21**.

---

4. Evaluation Metrics and Results

The model performance was evaluated using three key regression metrics, calculated on the predictions that were inverse-transformed back to the original AQI scale.

Evaluation Metrics

1. **Mean Absolute Error (MAE):** Measures the average magnitude of the errors in a set of predictions, without considering their direction.
2. **Root Mean Squared Error (RMSE):** Represents the square root of the average squared errors, giving higher weight to larger errors.
3. **Coefficient of Determination (R^2 Score):** Indicates the proportion of the variance in the dependent variable that is predictable from the independent variables (a value close to 1 indicates an excellent fit).

Performance Summary

```
📌 Validation Evaluation
MAE  = 9.0525
RMSE = 14.7338
R²   = 0.9543
--------------------------------------
📌 Test Evaluation
MAE  = 7.4107
RMSE = 11.1473
R²   = 0.9487
--------------------------------------
```