মেট্রোপলিটন ইউনিভার্সিটি
Metropolitan University

Course Code: CSE - 422
Course Title: Artificial Intelligence Lab

Predicting Air Quality: A Machine Learning Approach using the UCI Air Quality Dataset

Presented by

Israt Jahan
ID: 222-115-173

Ahmed.
ID: 222-115-190
Department of Computer Science & Engineering
Metropolitan University

Presented to

Khulud Binte Harun
Lecturer
Department of Computer Science & Engineering
Metropolitan University

Presentation Outline

- Introduction
- Dataset Overview
- Data Preprocessing
- Exploratory Data Analysis (EDA)
- Feature Engineering
- Data Splitting – Train & Split
- Machine Learning Models Implemented
- Deep Learning Models Implemented
- Hyper parameter Tuning & Cross Validation
- Model Comparison
- Result

Introduction

Air pollution is a major global concern with significant impacts on public health and the environment. Accurate prediction of air quality is crucial for issuing timely warnings and informing policy decisions.

Objective

To leverage machine learning models to analyze the relationship between various chemical sensor responses and environmental factors to predict the concentration of key pollutants.

Methodology

We will walk through the entire machine learning pipeline, from data cleaning and exploration to training and evaluating advanced models.

Dataset Overview: Air Quality UCI Dataset

This project utilizes the Air Quality UCI Dataset from the UCI Machine Learning Repository. It is a multivariate time-series dataset comprising **9,471 hourly** instances. The data, collected from March 2004 to February 2005 in a polluted urban area in Italy, includes **15 raw features** from *five* metal oxide chemical sensors (PT08.S1(CO) to PT08.S5(O3)) and environmental variables such as Temperature (T), Relative Humidity (RH), and Absolute Humidity (AH). Ground truth pollutant concentrations include Carbon Monoxide (CO(GT)), NMHC(GT), Benzene (C6H6(GT)), NOx(GT), and NO2(GT). Initial inspection revealed issues like inconsistent data formats, use of commas as decimal separators, the presence of **-200** as a missing value indicator, and unnecessary columns, which required thorough cleaning.

Dataset Overview: Air Quality UCI Dataset

After preprocessing, the dataset was transformed into a structured format with **12** numerical features used as predictors. The target variable is CO(GT), representing Carbon Monoxide concentration in mg/m^3. All features were standardized for comparability and optimized for machine learning model training, including deep learning regression techniques.

3. Data Preprocessing

Initial Challenges: Discuss raw data issues such as semicolon delimiters, commas for decimals, the -200 missing value indicator, and incorrect data types.

Cleaning Steps:

- Handling missing values with forward-fill (ffill).
- Correcting data types.
- Converting Date and Time into a single date time index.

4. Exploratory Data Analysis (EDA)

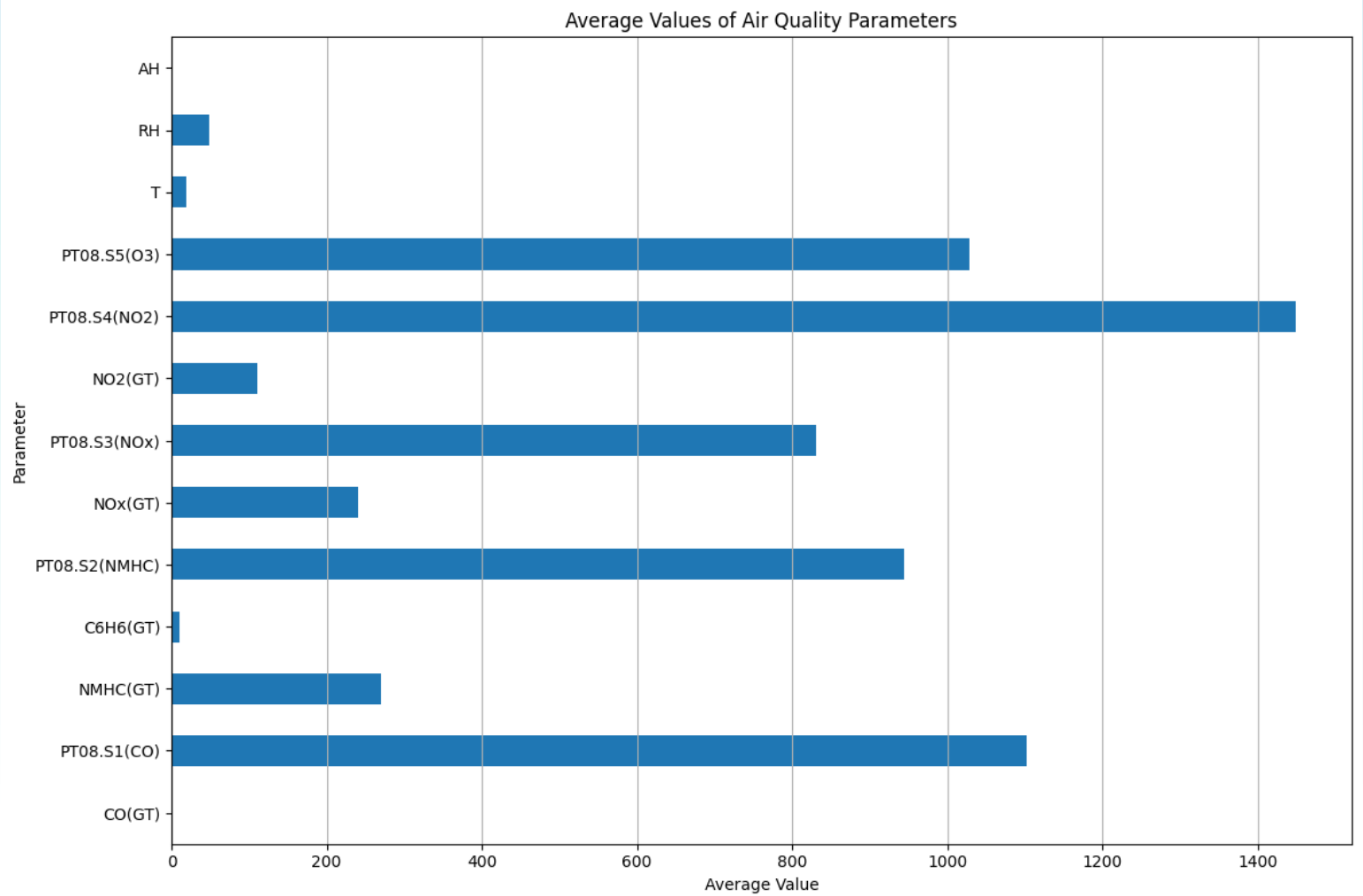


Fig: Average Values of Air Quality Parameters

4. Exploratory Data Analysis (EDA)

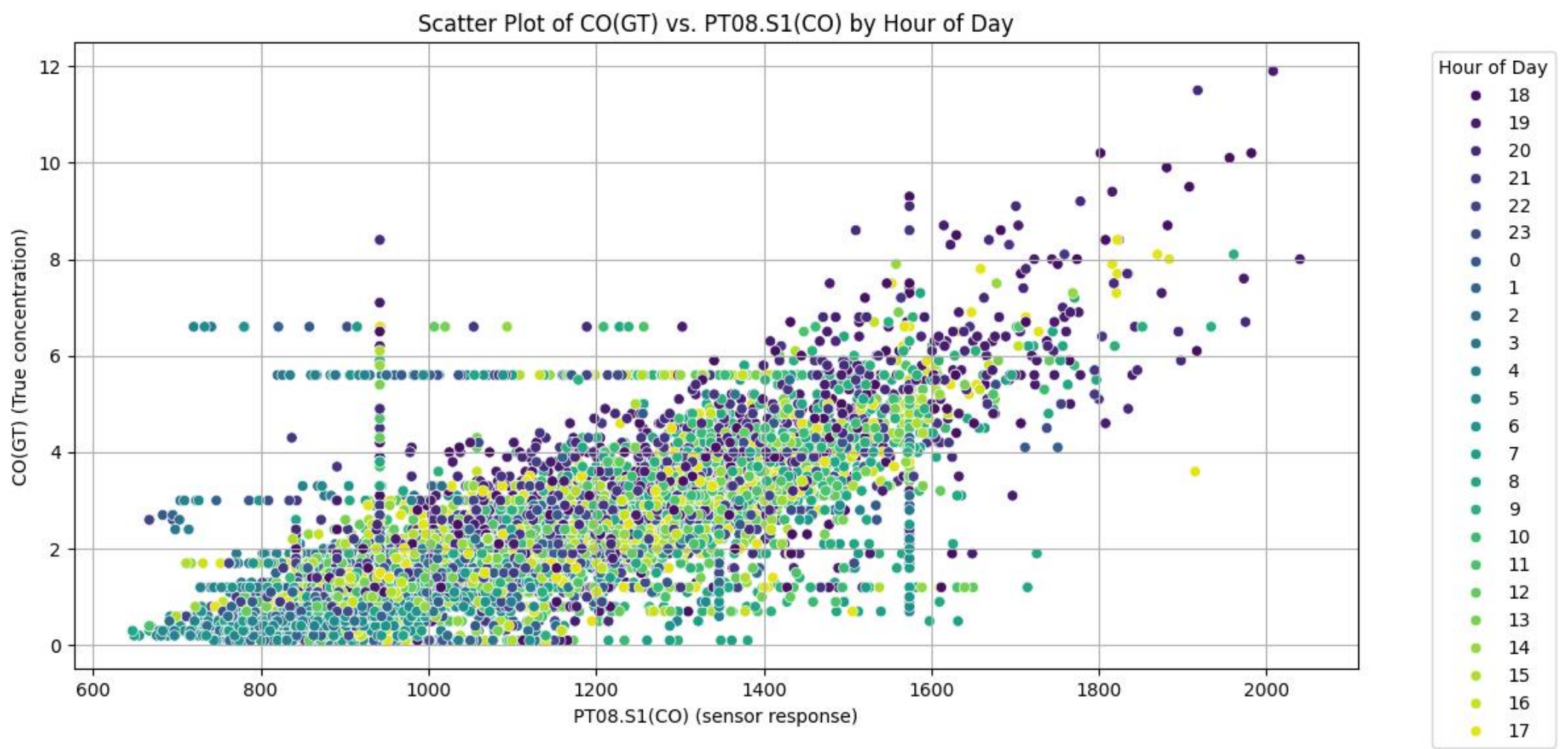


Fig: Scatter Plot of CO(GT) vs PT08.S1(CO) Colored by Hour of Day

4. Exploratory Data Analysis (EDA)

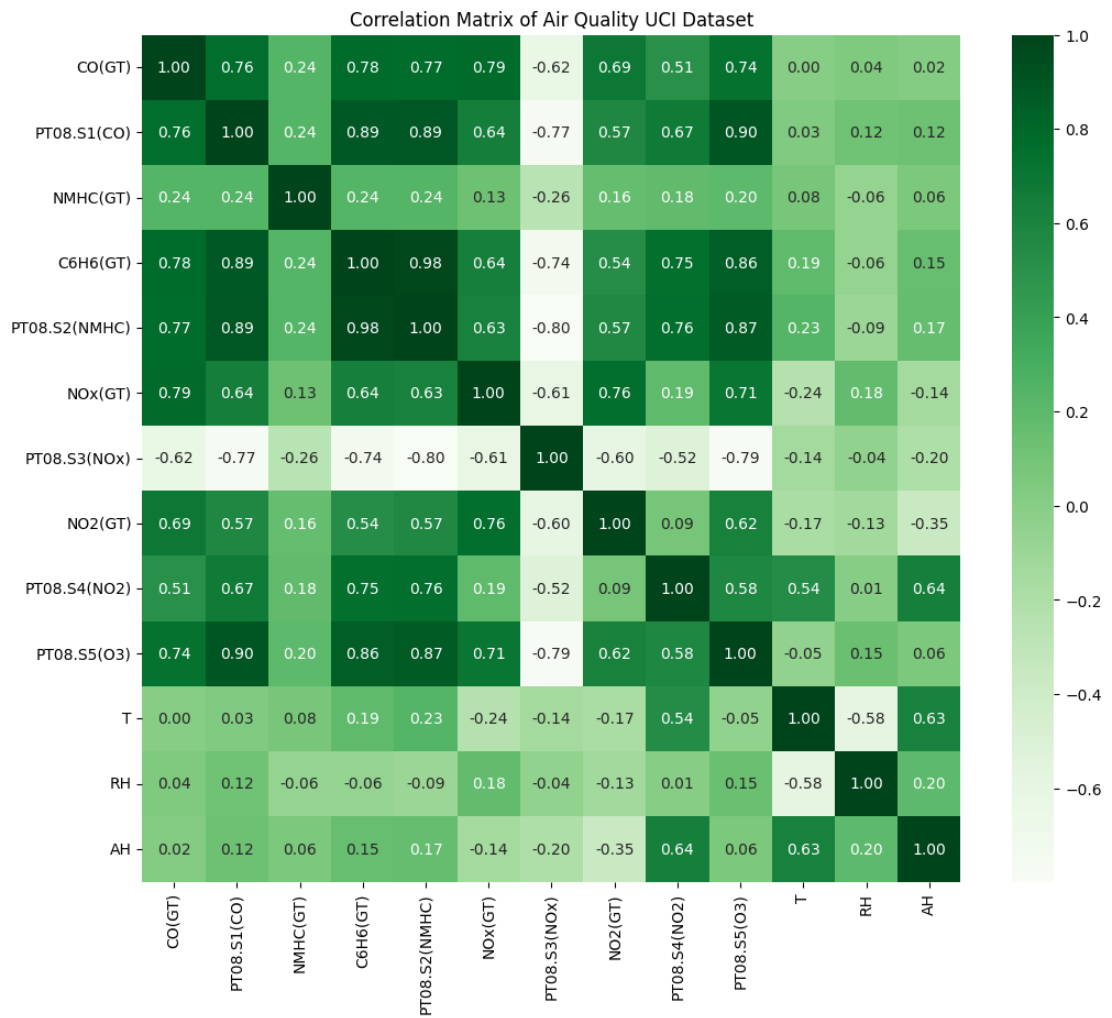


Fig: Correlation Matrix of Air Quality UCI Dataset

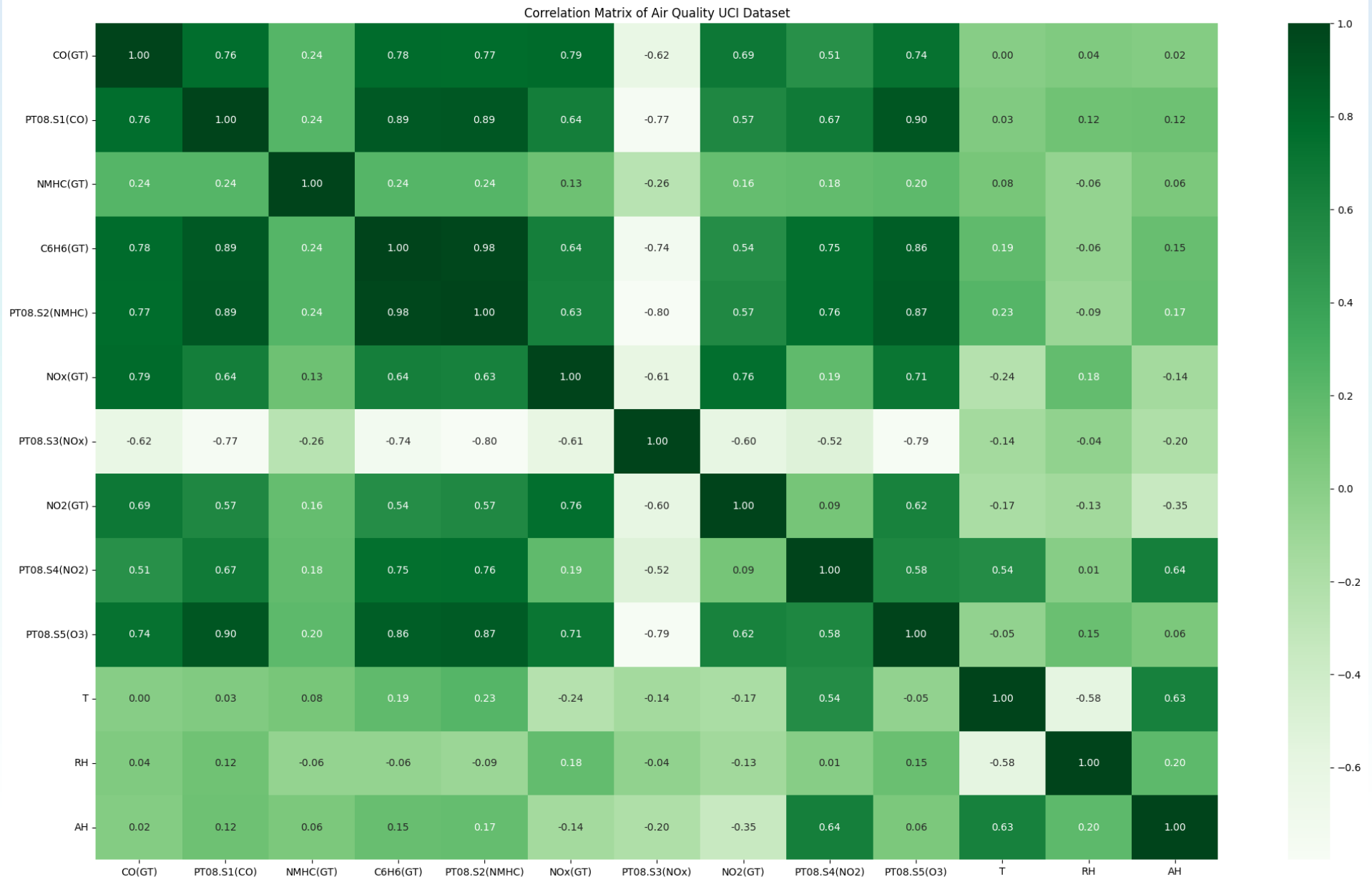


Fig: Correlation Matrix of Air Quality UCI Dataset

5. Feature Engineering

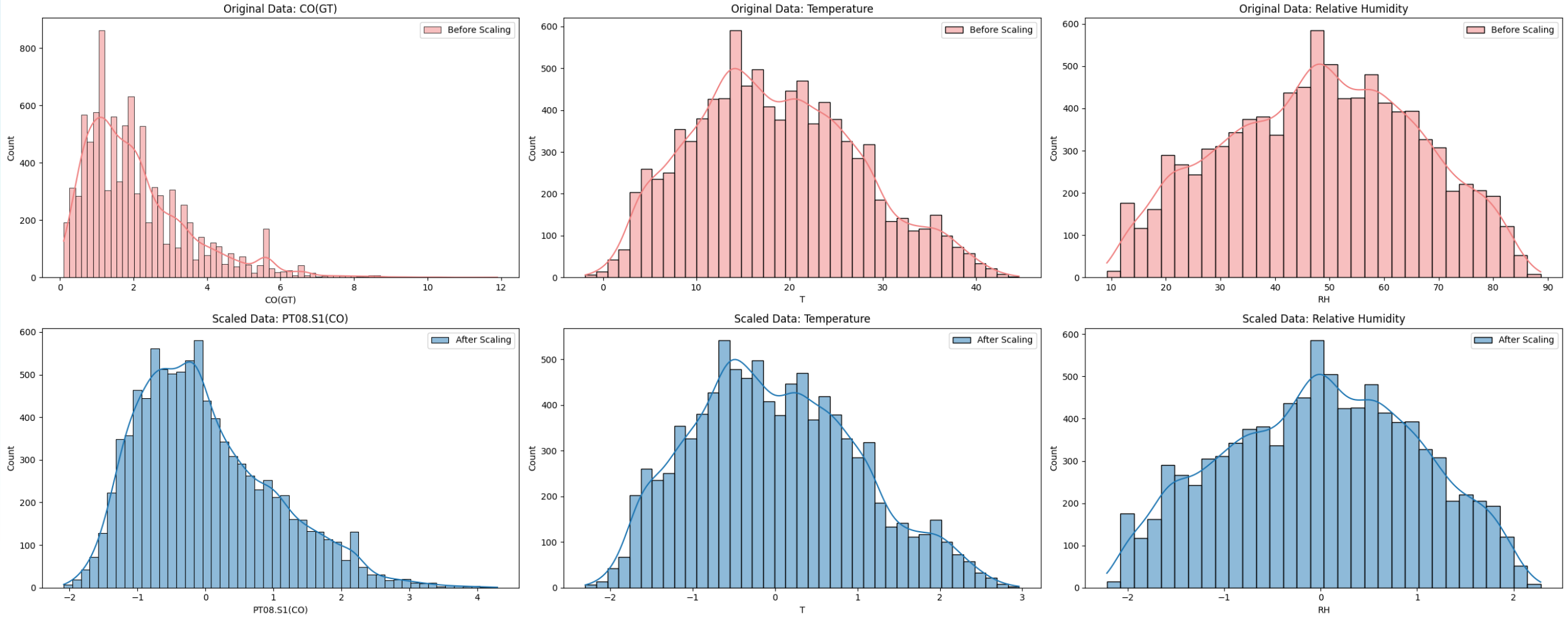


Fig: Visualize data distribution before and after scaling with histograms for key features

5. Feature Engineering

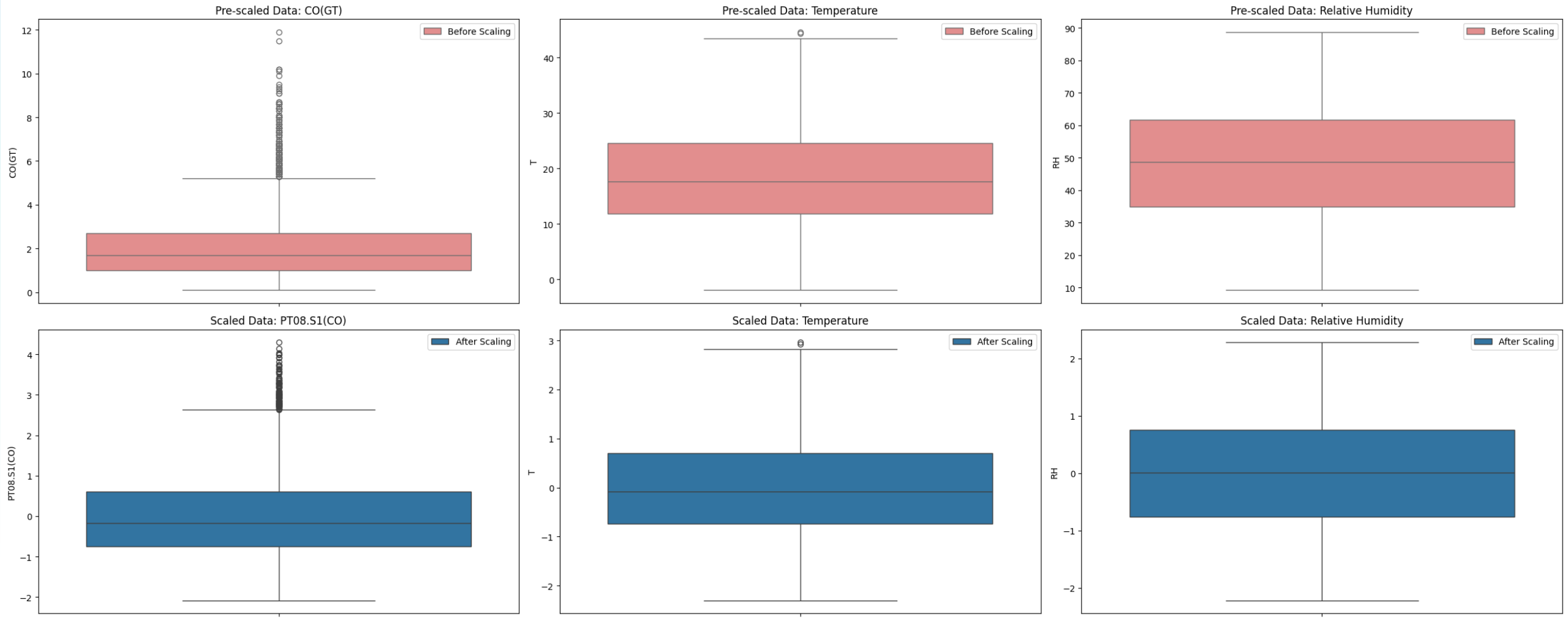


Fig: Visualize data distribution before and after scaling with box plots for key features

5. Feature Engineering

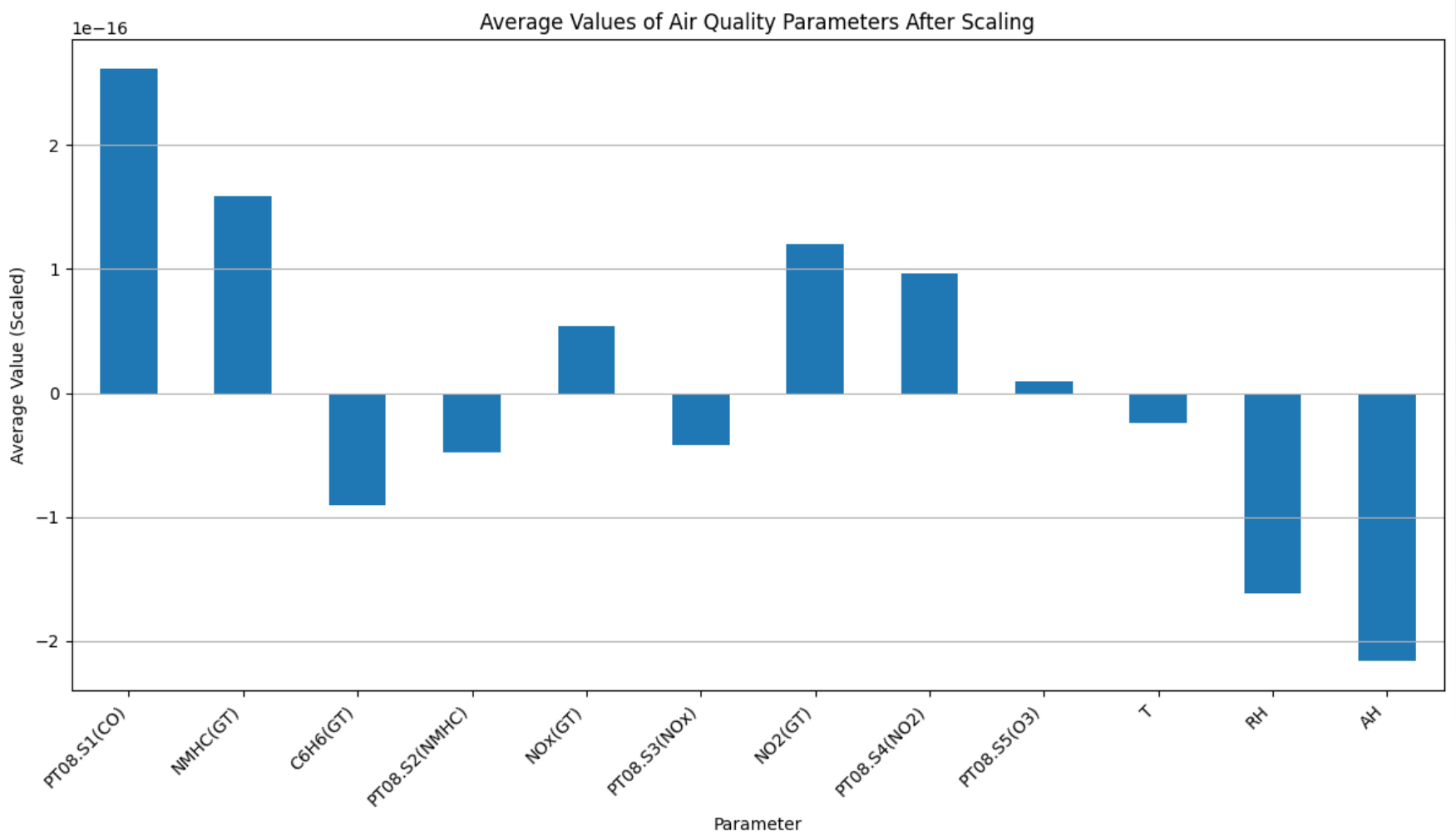


Fig: Average Values of Air Quality Parameters After Scaling

5. Feature Engineering

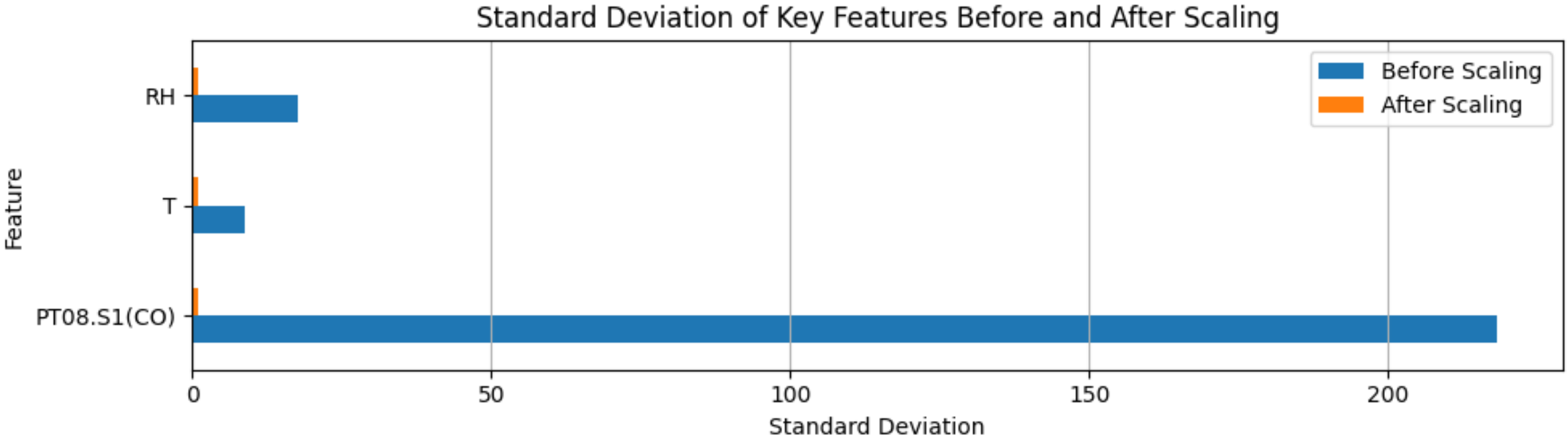


Fig: Standard Deviation of Key Features Before and After Scaling

5.Feature Engineering

Fig: Scatter Plot Before and After Scaling

Data Splitting – Train & Test

- **Purpose:** To evaluate model performance on unseen data and prevent overfitting.
- **Method:** Dataset divided into two parts.
- **Split Ratio:** 80% for Training, 20% for Testing.
- **Tool:** train_test_split (from scikit-learn).

6. Machine Learning Models Implemented

i. **Linear Regression** – for simple linear relationships

ii. **Random Forest** – reduces overfitting, good accuracy

iii. **SVR (SVM)** – handles non-linear, high-dimensional data

iv. **XGBoost** – fast, accurate boosting algorithm

v. **Ensemble Methods** – combines models (Averaging, Voting)

vi. **Parameter Tuning** – via Grid Search, Cross-Validation

6. Deep Learning Models Implemented

ANN – captures complex non-linear patterns

- Architecture Overview:
  - Model: Sequential
  - Input Layer: Dense(64), ReLU
  - Hidden Layer: Dense(32), ReLU
  - Output Layer: Dense(1), Linear

6. Deep Learning Models Implemented

- Training Configuration:
  - Optimizer: Adam
  - Loss: Mean Squared Error (MSE)
  - Epochs: 50
  - Batch Size: 32
  - Validation Split: 20%

7. Hyper parameter Tuning & Cross-Validation

Hyperparameter tuning was done using **GridSearchCV** with **K-Fold Cross-Validation** for **SVR** and **XGBoost**.

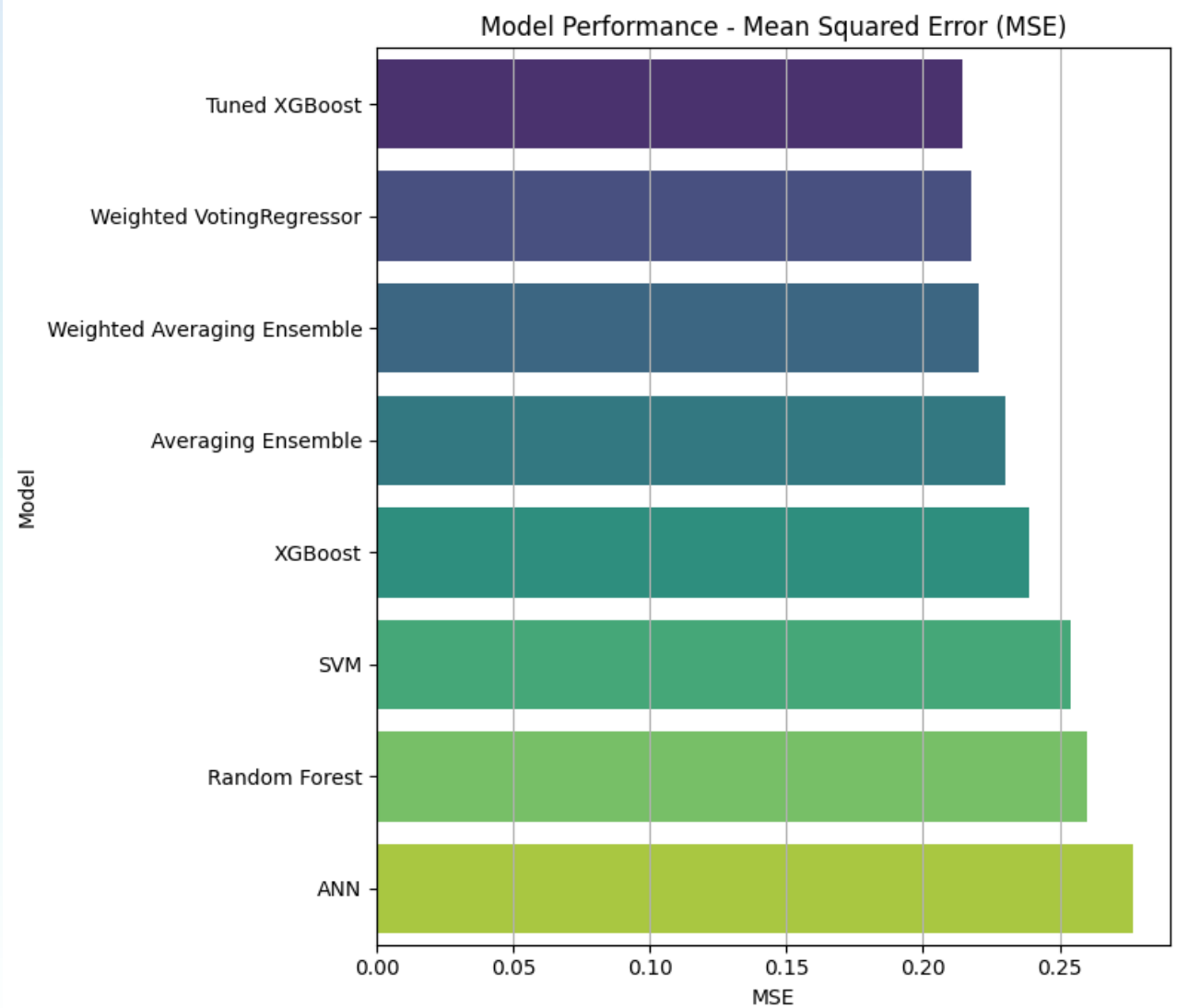
Performance was evaluated using Mean Squared Error (MSE) and R-squared (R^2) to ensure accuracy and generalization.

7. Hyper parameter Tuning & Cross-Validation

Model	MSE	R-squared
Best Cross-Validated SVM	0.6180	N/A
Best Tuned XGBoost	20.1%	90%

Models Comparison

Model	Accuracies
Best Tuned XGBoost	**90.08%**
Ensemble (Weighted Averaging)	89.74%
Ensemble (Averaging)	89.25%
XGBoost Regressor	88.96%
SVM Regressor	88.27%
Random Forest Regressor	87.97%
ANN	86.79%



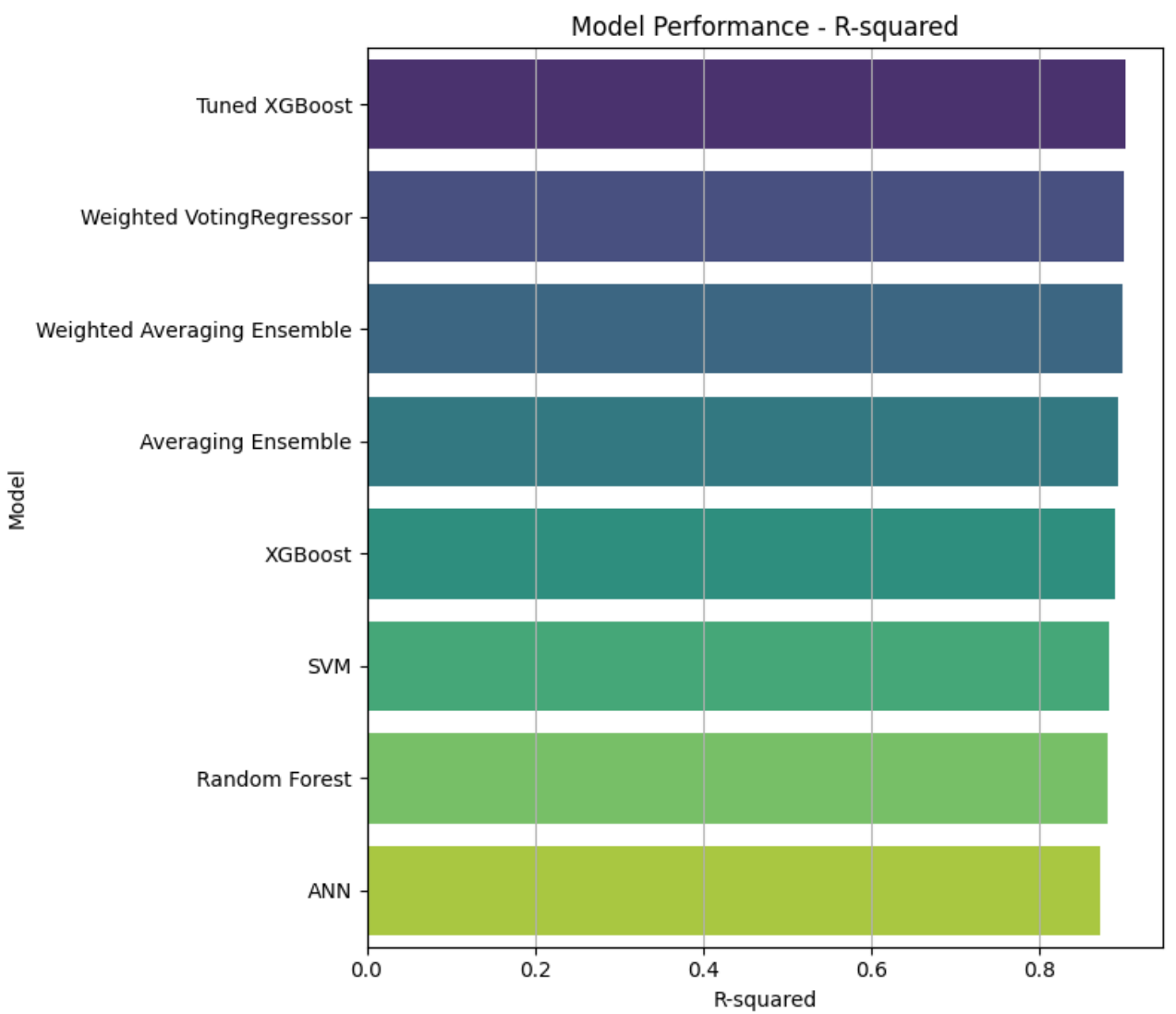


Fig: Model comparison

8. Results & Discussion

Our models were evaluated using Mean Squared Error (MSE) and R-squared (R2), a metric indicating the proportion of variance in the dependent variable that is predictable from the independent variables.

Model	MSE	R-squared
XGBoost Regression	0.2387	0.8896
SVM Regression	0.2536	0.8827
Best Cross-Validated SVM	0.6180	N/A*
ANN	0.2856	0.8679
Random Forest Regression	0.2601	0.8797
Best Tuned XGBoost	**0.2145**	**0.9008**
Ensemble (Averaging)	0.2325	0.8925
Ensemble (Weighted Averaging)	0.2219	0.8974

8. Results

The **Best Tuned XGBoost Regressor** emerged as the top-performing individual model, achieving an impressive R-squared of **0.9008 (90.08%)**. This indicates it explains over 90% of the variance in Carbon Monoxide concentrations.

Model	MSE	R-squared
XGBoost Regression	0.2387	0.8896
SVM Regression	0.2536	0.8827
Best Cross-Validated SVM	0.6180	N/A*
ANN	0.2856	0.8679
Random Forest Regression	0.2601	0.8797
Best Tuned XGBoost	**0.2145**	**0.9008**
Ensemble (Averaging)	0.2325	0.8925
Ensemble (Weighted Averaging)	0.2219	0.8974

Reference

[Air Quality,UCI]

Thank You

Q&A