Predicting Flue Gas Emissions and Turbine Energy Yield
A Data Science Approach

Rishu Bhadani
Department of Mechanical Engineering
IIT Bombay
Mumbai, India
22b2130@iitb.ac.in

***Abstract*—This study applies a data science approach to analyze gas turbine sensor data and predict key operational metrics, including flue gas emissions (CO and NOx) and turbine energy yield (TEY). Using advanced exploratory data analysis (EDA), feature engineering, and machine learning models, the research identifies critical variables and achieves high predictive accuracy. Random Forest emerged as the top model for emissions and TEY predictions, balancing interpretability and performance. The findings provide actionable insights to optimize turbine efficiency and minimize emissions, underscoring the potential of data-driven strategies in energy systems.**

***Index Terms*—Gas turbine performance, flue gas emissions, machine learning, feature engineering, exploratory data analysis, dimensionality reduction, predictive modeling.**

I. Introduction

Gas turbines are pivotal in modern energy systems, serving as primary power generation units in various industrial and commercial settings. Their performance is critically evaluated based on two key metrics: emissions and efficiency. The environmental impact of carbon monoxide (CO) and nitrogen oxides (NOx) emissions, coupled with the economic implications of turbine energy yield (TEY), underscores the necessity for comprehensive predictive analytics. With increasing global emphasis on sustainability and operational efficiency, accurate modeling of these metrics has become essential.

This study leverages advanced data science methodologies to analyze gas turbine sensor data collected in Turkey between 2011 and 2015. The dataset encompasses 36,733 hourly records, capturing a range of operational parameters such as ambient temperature (AT), pressure (AP), and humidity (AH), as well as turbine-specific variables like exhaust pressure (GTEP) and inlet temperature (TIT). These parameters offer a rich foundation for exploring the intricate relationships governing turbine emissions and energy yield.

The analysis begins with exploratory data analysis (EDA) to identify patterns, correlations, and potential outliers in the dataset. Key findings, such as the oscillatory behavior of NOx emissions and the linear dependence of TEY on compressor discharge pressure (CDP), guide the subsequent feature engineering process. Techniques like log transformations, standardization, and dimensionality reduction using PCA and UMAP are employed to enhance model readiness.

Predictive modeling forms the core of this research. A suite of machine learning models, including Random Forest, LightGBM, CatBoost, and Neural Networks, are evaluated for their efficacy in predicting CO and NOx emissions as well as TEY. Notably, the Random Forest model emerges as the top performer, achieving high accuracy with robust interpretability. Additionally, a multi-output regression approach using Neural Networks demonstrates the potential for joint prediction of emissions, leveraging shared patterns between CO and NOx.

Beyond predictive accuracy, this study emphasizes the importance of feature importance analysis for actionable insights. For instance, parameters such as AT, TIT, and AFDP are consistently identified as critical drivers of emissions, highlighting their role in gas turbine performance optimization. These findings not only validate the models but also provide practical recommendations for operational enhancements.

In summary, this research illustrates the transformative power of data science in tackling complex challenges in energy systems. By integrating state-of-the-art machine learning techniques with domain-specific knowledge, the study paves the way for more efficient and environmentally sustainable gas turbine operations. Future work aims to expand on these findings by incorporating additional data features and exploring real-time predictive systems for broader applicability.

II. Dataset Overview

The dataset used in this study contains **36,733 hourly records** of gas turbine sensor data collected in Turkey from **2011 to 2015**. It includes the following variables:

- **Operational Parameters:**
  1) Ambient Temperature (AT), °C
  2) Ambient Pressure (AP), mbar
  3) Ambient Humidity (AH), %
  4) Air Filter Difference Pressure (AFDP), mbar
  5) Gas Turbine Exhaust Pressure (GTEP), mbar
  6) Turbine Inlet Temperature (TIT), °C
  7) Turbine After Temperature (TAT), °C
  8) Compressor Discharge Pressure (CDP), mbar

9) Turbine Energy Yield (TEY), MWh
10) Carbon Monoxide (CO), mg/m^3 (target variable)
11) Nitrogen Oxides (NOx), mg/m^3 (target variable)

The dataset spans four years, capturing a wide range of operational and environmental conditions. The following key insights were derived from an initial exploratory data analysis (EDA):

- **Outliers:** Approximately 6% of the data points are outliers. These were retained to ensure the models can generalize well to real-world variations.
- **Correlations:** Strong positive correlations were observed between Turbine Energy Yield (TEY) and variables like Compressor Discharge Pressure (CDP) and Turbine Inlet Temperature (TIT).
- **Distributions:** Variables such as Ambient Temperature (AT) and Ambient Pressure (AP) follow normal distributions, while others like Carbon Monoxide (CO) and Nitrogen Oxides (NOx) exhibit significant skewness, addressed via log transformations.
- **Feature Importance:** Parameters like AT, TIT, and AFDP were identified as critical drivers of emissions and energy yield, offering actionable insights for optimization.

This rich dataset provides a robust foundation for predictive modeling of turbine performance metrics, enabling the study of complex interactions and development of accurate machine learning models.

III. Methodology

A. Exploratory Data Analysis (EDA)

Preliminary analysis included visualizing relationships among variables:

- Scatter plots for CO and NOx emissions, as shown in Fig. 1, reveal distinct patterns: CO exhibits high variability without a clear periodic trend, while NOx demonstrates oscillatory behavior, resembling sine or cosine functions. The Plotly-based visualization in Fig. 2 provides a more detailed and interactive view, enabling better identification of trends and patterns in the emissions data.
- Figure 3 illustrates the relationships among several numerical variables using a pairplot. The relationship between **TEY (Turbine Energy Yield)** and **CDP (Compressor Discharge Pressure)** is highly linear, indicating a strong positive correlation. Similarly, variables such as **TEY**, **GTEP (Gas Turbine Exhaust Pressure)**, **CDP**, and **TIT (Turbine Inlet Temperature)** exhibit strong linear relationships, suggesting significant interdependence among them. Certain pairs of variables display cluster-like patterns that may indicate the presence of specific operational regimes or conditions under which these variables are tightly grouped. Although many relationships are linear, some, such as the relationship between **TEY** and **TIT**, exhibit slightly curved trends in specific ranges, hinting at possible nonlinear interactions or diminishing returns. The diagonal histograms provide insight into the distributions of individual variables; for example, **TEY** and **TIT** have relatively uniform or unimodal distributions, whereas **CDP** exhibits a more concentrated range. Finally, strong correlations among variables such as **TEY**, **CDP**, **GTEP**, and **TIT** suggest potential multicollinearity, which could affect the performance and interpretability of predictive models. These observations provide valuable guidance for feature selection and engineering decisions in predictive modeling tasks.
- The correlation matrix in Figure 4 reinforces the observations explained in Figure 3. Provides a quantitative representation of the relationships among the variables, confirming the strong correlations identified in the pairplot.
- The box plot shown in Figure 5 highlights the presence of some outliers in the data. However, I have decided to retain these outliers in the dataset. This decision is based on the following considerations:
  First, the proportion of outliers is approximately 6%, which is relatively small and can be effectively handled through data processing steps such as log transformation. Removing such a small percentage of data is unlikely to significantly impact the modeling process but may lead to a loss of valuable information.
  Second, the models considered for fitting the data—such as **Random Forest**, **LightGBM**, **CatBoost**, **Neural Networks**, and **Support Vector Machines**—are robust enough to handle this small percentage of outliers. These models are designed to generalize well even in the presence of noise or irregularities in the data.
  Third, retaining the outliers can contribute to reducing overfitting in the models. Outliers introduce diversity into the data, encouraging the model to learn broader patterns rather than focusing on specific trends or noise in the majority of the data. This can enhance the generalization capability of the model, particularly when applied to unseen data.
  Finally, the presence of outliers also reflects real-world variations in the data that should not be ignored. Addressing these variations through robust modeling approaches ensures the results are more realistic and applicable to practical scenarios.
  Considering these factors, I concluded that it is more beneficial to retain the outliers and handle them during the data preprocessing and modeling stages rather than removing them outright.
- The histogram in Figure 6 illustrates the distribution of each feature, including the target variable. Analyzing these distributions is crucial before fitting any machine learning model to ensure that the data is not biased towards any particular feature and to address potential issues related to skewness or non-normality.
  From the histograms, the following insights can be drawn:
  - Several features, such as **AT (Ambient Temperature)** and **AP (Ambient Pressure)**, appear to have approx-

imately normal distributions. These features are well-behaved and do not require significant preprocessing. - Features like **GTEP (Gas Turbine Exhaust Pressure)** and **CDP (Compressor Discharge Pressure)** exhibit a multimodal distribution, which might indicate the presence of multiple operational regimes or clusters within the data. Further investigation or segmentation could provide more clarity. - The target variables **CO (Carbon Monoxide)** and **NOX (Nitrogen Oxides)** show significant right-skewness, with a majority of the values concentrated at lower ranges. Log transformation or other scaling techniques might be helpful to reduce skewness and stabilize variance. - **TAT (Turbine After Temperature)** shows an extremely narrow range with values concentrated around 550, suggesting a potential constant or near-constant relationship, which might limit its predictive importance. - The variable **TEY (Turbine Energy Yield)** shows a multimodal distribution, indicating the possibility of capturing different behaviors or operating conditions in the dataset.

Ensuring that features are as close to uniform or normal distributions as possible helps prevent bias in model fitting. This step is particularly important for models sensitive to feature distributions, such as linear regression or neural networks, though tree-based models like Random Forest or CatBoost are more robust in this regard.

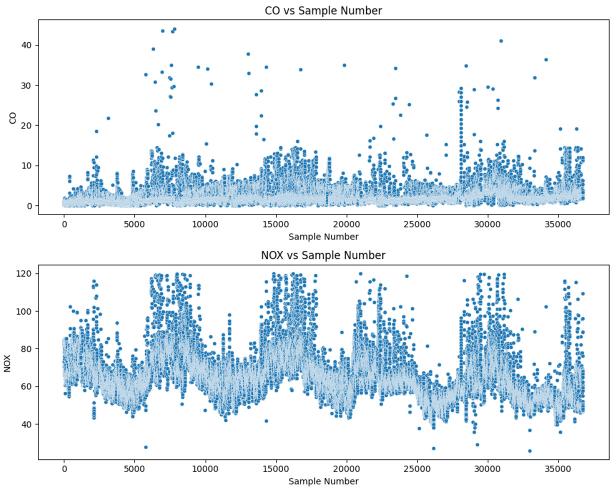

Fig. 1. Scatter plots for CO and NOx emissions.

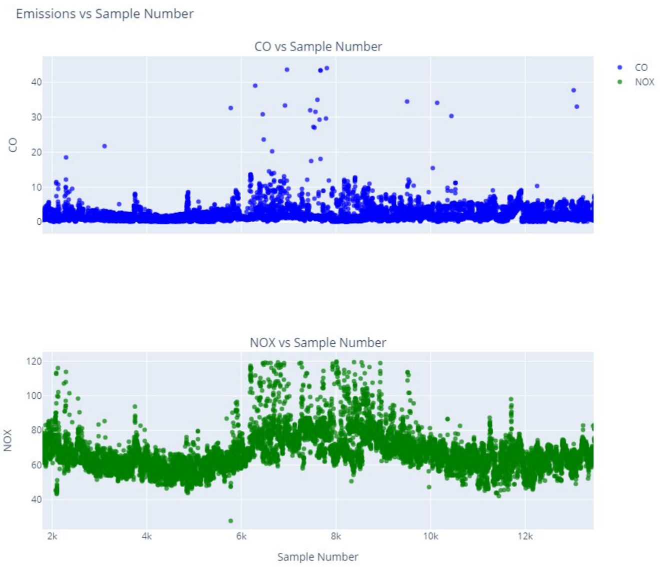

Fig. 2. Scatter plots for CO and NOx emissions.

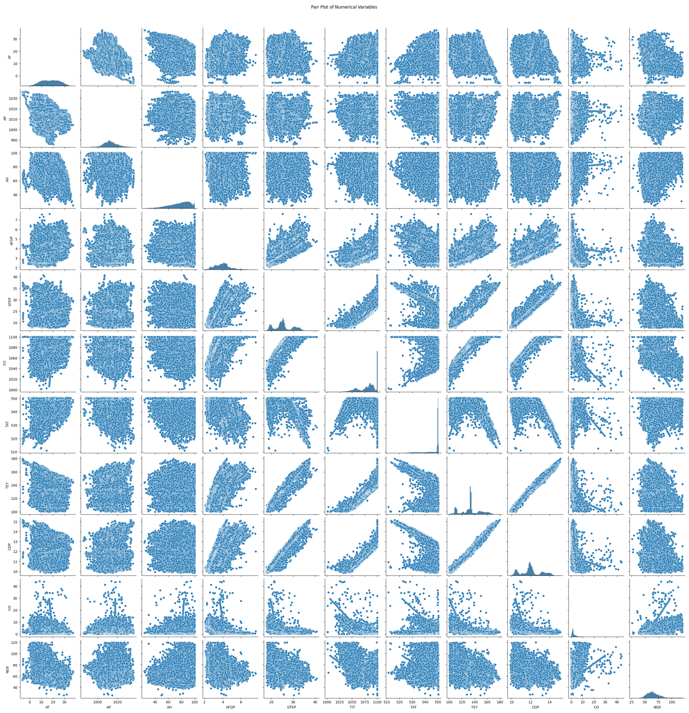

Fig. 3. Pairplot of Numerical Variables

B. Data Preprocessing

- To address the issues of skewness and distribution discussed earlier in Figure 6, I applied the **log1p transformation** to selected features. This transformation not only helps in reducing skewness but also handles zero values effectively by applying the transformation as $\log(1+x)$. The impact of this transformation is evident in Figure 7, where we observe the following changes: - The distribution of **CO (Carbon Monoxide)** has shifted significantly towards a near-normal distribution. This improvement reduces skewness and ensures that the feature is better suited for models that assume normality in data. - Similarly, the distribution of **AFDP (Air Flow Discharge Pressure)** has become more symmetrical, enhancing its suitability for predictive modeling. - Features like **AP (Ambient Pressure)** and **NOX (Nitrogen Oxides)** have also been positively impacted by the transformation, with their distributions showing improved alignment with

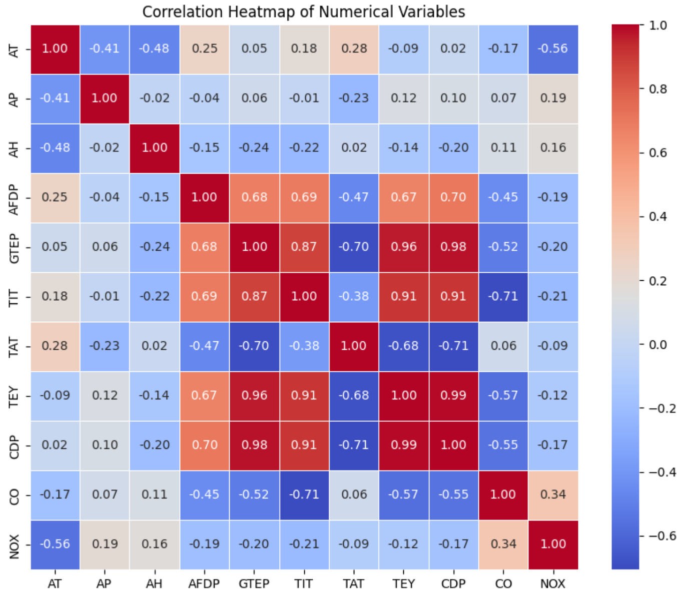

Fig. 4. Correlation Matrix of Numerical Variables

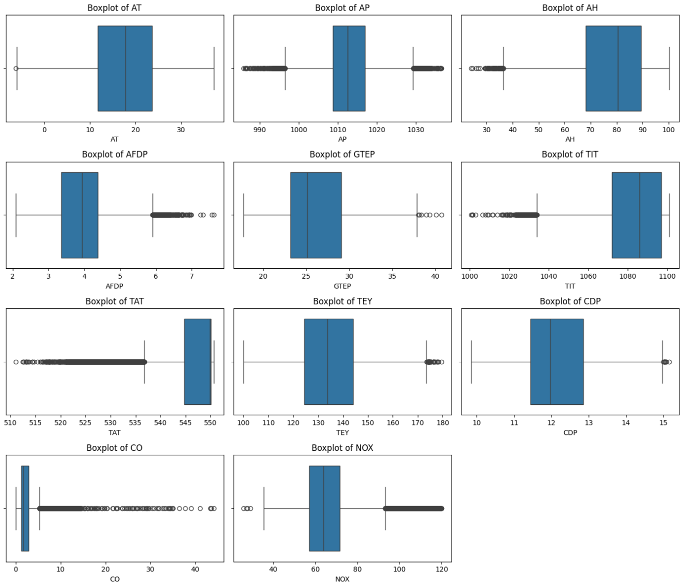

Fig. 5. Box Plot Highlighting Outliers in the Data

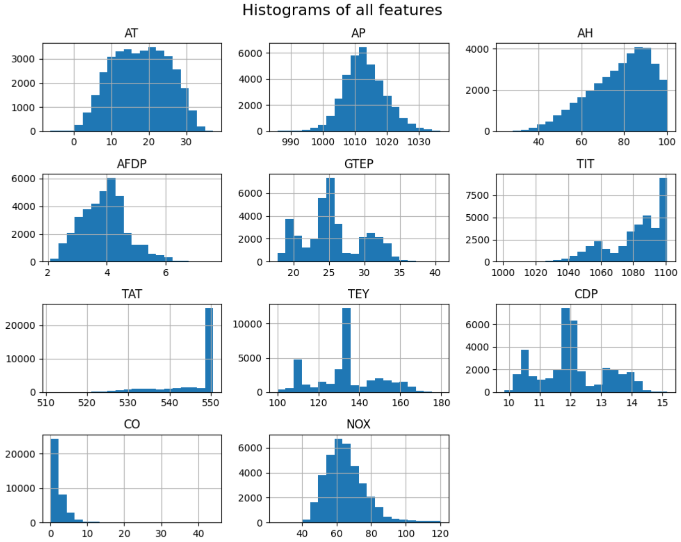

Fig. 6. Histograms Showing the Distribution of Each Feature

normality. - Other features that already exhibited relatively normal distributions, such as **AT (Ambient Temperature)**, remain unaffected by the transformation, as expected.

By transforming these features, the data is now better prepared for model fitting, reducing potential biases introduced by highly skewed distributions. Moreover, this step ensures a more balanced contribution of all features to the models, improving overall robustness and predictive performance.

- To standardize the feature values and ensure uniformity across different scales, I applied standardization using the **StandardScaler**. This transformation scales the features to have a mean of 0 and a standard deviation of 1. Standardization is essential for algorithms sensitive to feature magnitudes, such as Support Vector Machines or Neural Networks. After scaling, the standardized values replaced the original feature values in the dataset. This step enhances the model's performance by improving convergence and ensuring balanced contributions from all features.

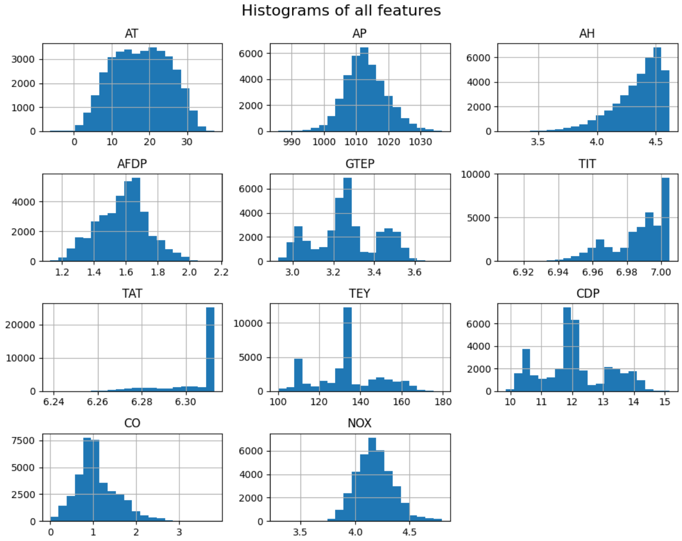

Fig. 7. Histograms Showing the Distribution of Features After log1p Transformation

C. Feature Engineering

As part of the feature engineering process, dimensionality reduction techniques like **Principal Component Analysis (PCA)** [1] and **UMAP (Uniform Manifold Approximation**

and Projection) [2] were applied to optimize the dataset for modeling. PCA was used to reduce the dimensionality of the data while preserving as much variance as possible. Mathematically, PCA works by finding the orthogonal directions (principal components) that maximize the variance in the dataset, which is achieved by solving the eigenvalue decomposition of the covariance matrix. Five principal components were extracted, explaining a significant proportion of the variance in the data, as observed through the explained variance ratio. This reduction not only helps in improving computational efficiency but also minimizes noise and redundancy in the dataset.

In addition to PCA, UMAP, a nonlinear dimensionality reduction technique, was employed to capture complex structures in the data. UMAP constructs a weighted graph of the high-dimensional data and optimizes a low-dimensional representation that preserves the global and local structure. By embedding the data into a 5-dimensional space, UMAP enhances feature representation and aids models in better capturing nonlinear relationships within the data. Both techniques effectively prepare the data by retaining essential patterns while mitigating issues associated with high-dimensional spaces, such as overfitting or increased computational cost. These transformations are particularly valuable when dealing with complex datasets where interpretability and scalability are crucial.

As part of the feature engineering process, the outputs of PCA and UMAP were scaled using **StandardScaler** to ensure uniformity in their ranges. The transformed PCA and UMAP features were then converted into meaningful DataFrames with columns labeled as **PC1, PC2, ...** for PCA and **UMAP1, UMAP2, ...** for UMAP. These newly generated features were concatenated with the original dataset to create a comprehensive feature set, enhancing the model's ability to capture both linear and nonlinear relationships in the data.

D. Modeling

Multiple machine learning models were employed to predict the target variables (**CO** and **NOX**), leveraging two distinct approaches. The first approach utilized raw features from the dataset, while the second combined these raw features with dimensionality-reduced features obtained from **PCA** and **UMAP**. For both approaches, the dataset was split into features (**X**) and targets (**y**), with an 80:20 train-test ratio. During evaluation, predictions underwent an inverse `log1p` transformation to interpret results in their original scale. A variety of models were implemented in the first approach, including **Linear Regression**, **Random Forest Regressor**, **Support Vector Regressor (SVR)**, **LightGBM**, **CatBoost Regressor**, and **Neural Networks**. The Random Forest (RF) model [3], [4], known for its ensemble learning capabilities, builds multiple decision trees using different feature subsets and combines their outputs for robust predictions. LightGBM (LGBM) [4], [5], a gradient boosting framework, is memory-efficient, accurate, and includes regularization techniques to mitigate overfitting. CatBoost [4], [6] introduces ordered boosting and efficient handling of categorical variables without inflating feature space through one-hot encoding. Neural Networks, although highly effective for complex, non-linear relationships, were excluded from the second approach due to their ability to generate feature representations, rendering dimensionality-reduced features redundant. LightGBM and CatBoost were also excluded from the second approach, as simpler models like RF provided comparable performance with better computational efficiency and interpretability when using combined datasets. In the second approach, **Linear Regression**, **Random Forest Regressor**, and **SVR** were used to evaluate the impact of dimensionality reduction on prediction accuracy.

To evaluate the performance of these regression models, three key metrics were calculated: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2). The R^2 metric, a well-known statistical measure for regression and forecasting, is defined as [4], [7], [8]:

$$R_i^2 = 1 - \frac{\sum_{k=1}^{m} \left(y_{i,k} - \hat{y}_{i,k}\right)^2}{\sum_{k=1}^{m} \left(\bar{y}_i - y_{i,k}\right)^2},$$

where i represents the output variable ($i = 1$ for **CO** and $i = 2$ for **NOX**), m is the sample size, $y_{i,k}$ is the k-th observed (experimental) output, $\hat{y}_{i,k}$ is the k-th predicted output, and $\bar{y}_i$ is the mean of the observed output. R^2 ranges from 0 to 1, with a higher value indicating better model accuracy [4].

The RMSE is given by:

$$RMSE_i = \sqrt{\frac{\sum_{k=1}^{m} \left(y_{i,k} - \hat{y}_{i,k}\right)^2}{m}},$$

and the MAE is defined as:

$$MAE_i = \frac{\sum_{k=1}^{m} |y_{i,k} - \hat{y}_{i,k}|}{m}.$$

These metrics provided insights into the accuracy, precision, and robustness of each model. The results highlighted the trade-offs between leveraging raw feature information for simplicity and integrating dimensionality-reduced features for capturing additional patterns and reducing noise. Ensemble and gradient boosting methods excelled with raw features, while dimensionality reduction enhanced the performance of simpler models, demonstrating the importance of aligning model complexity with dataset characteristics and objectives.

E. Joint Prediction of CO and NOX Using Advanced Models

For the joint prediction of **CO** and **NOX** emissions, a multi-output regression model was implemented using a **Neural Network (MLPRegressor)**. This approach enables simultaneous prediction of both targets, leveraging potential correlations and shared patterns between them. The network architecture consisted of two hidden layers with 80 and 30 neurons, respectively, and was trained for a maximum of 1000 iterations to ensure convergence. The predictions for each target were evaluated using standard metrics such as MAE, RMSE, and R². For **CO** emissions, the model achieved an MAE of 0.5339, an RMSE of 1.1437, and an R^2 of 0.7594. For **NOX** emissions, the model yielded an MAE of 3.5940, an RMSE of 5.2309, and an R^2 of 0.7935. These results demonstrate the model's capacity to effectively predict both **CO** and **NOX** emissions, with relatively better performance in predicting NOX, as indicated by the higher R^2 and lower error metrics.

This joint prediction method underscores the capability of neural networks to handle multi-output regression tasks efficiently, offering a holistic approach to emissions modeling.

IV. Results and Discussion

A. *Flue Gas Emissions Predictions*

The performance metrics for predicting **CO** emissions using raw features are summarized in Table I. As shown, different models exhibit varying levels of prediction accuracy. The Random Forest model achieved the lowest MAE of 0.4776 and the highest R^2 of 0.7571, indicating the best performance among the models tested. The Neural Network model also performed well with an MAE of 0.5096 and an R^2 of 0.7457, showcasing its capability to capture complex relationships in the data.

Model	MAE	RMSE	R^2
Linear Regression	0.741477	1.470335	0.602422
Random Forest	0.477630	1.149306	0.757081
Support Vector Machine	0.548161	1.215324	0.728372
LightGBM	0.534814	1.213269	0.729290
CatBoost	0.506799	1.138645	0.761567
Neural Network	0.509648	1.175819	0.745744

TABLE I
Performance metrics for **CO** predictions using raw features.

The results for predicting **NOX** emissions using raw features are presented in Table II. Random Forest again outperforms other models, achieving the lowest MAE of 2.5958 and the highest R^2 of 0.8748. The Neural Network model, with an MAE of 3.6492 and an R^2 of 0.7908, also demonstrates strong predictive ability but lags behind Random Forest and CatBoost in terms of accuracy.

Model	MAE	RMSE	R^2
Linear Regression	5.776058	8.211201	0.491202
Random Forest	2.595846	4.073922	0.874756
Support Vector Machine	3.756686	5.115796	0.802504
LightGBM	3.203057	4.693861	0.833738
CatBoost	2.871623	4.242472	0.864178
Neural Network	3.649236	5.265005	0.790815

TABLE II
Performance metrics for **NOX** predictions using raw features.

In Table III, the performance metrics for **CO** predictions using combined data show that Random Forest still delivers the best results with an MAE of 0.4707 and an R^2 of 0.7566. The Neural Network model in this case slightly improves over its raw data performance, though it remains close to other models in terms of MAE and R^2 values.

Model	MAE	RMSE	R^2
Linear Regression	0.716515	1.462004	0.606914
Random Forest	0.470727	1.150498	0.756577
Support Vector Machine	0.534021	1.216171	0.727994

TABLE III
Performance metrics for **CO** predictions using combined data.

The results for **NOX** predictions using combined data, shown in Table IV, indicate that Random Forest once again provides the best predictive performance with an MAE of 2.5301 and an R^2 of 0.8789. The Neural Network model, with an MAE of 3.6823 and an R^2 of 0.8094, maintains a strong performance but falls short of Random Forest and CatBoost.

Model	MAE	RMSE	R^2
Linear Regression	4.959301	6.969707	0.633426
Random Forest	2.530141	4.005472	0.878929
Support Vector Machine	3.682298	5.025128	0.809442

TABLE IV
Performance metrics for **NOX** predictions using combined data.

Finally, Table V summarizes the results for joint predictions of **CO** and **NOX** emissions using a multi-output Neural Network model. For **CO**, the model achieved an MAE of 0.5339, an RMSE of 1.1437, and an R^2 of 0.7594. For **NOX**, the MAE was 3.5940, RMSE was 5.2309, and R^2 was 0.7935. These results highlight the ability of the multi-output regression approach to effectively predict both **CO** and **NOX** emissions, with the model performing slightly better for **NOX** than for **CO**, as indicated by the higher R^2 and lower error metrics for NOX.

Emission	MAE	RMSE	R^2
CO	0.5339	1.1437	0.7594
NOX	3.5940	5.2309	0.7935

TABLE V
Performance metrics for joint prediction of **CO** and **NOX** emissions using a multi-output Neural Network model.

B. *Model Selection Justification*

Upon comparing all the models, Random Forest stands out as the best-performing model for both **CO** and **NOX** emissions prediction. This model consistently achieves the lowest MAE and the highest R^2 values, particularly for the **NOX** predictions where it achieves an MAE of 2.5958 and an R^2 of 0.8748, the highest across all models.

While the multi-output Neural Network model performed well for joint predictions, its performance did not surpass that of Random Forest. The Neural Network achieved an R^2 of 0.7594 for **CO** and 0.7935 for **NOX** in joint prediction, which, though respectable, is lower than Random Forest's R^2 of 0.8789 for **NOX** when using combined data.

C. *Feature Importance Analysis*

To gain further insights into the predictive models, feature importance was analyzed for both **CO** and **NOX** emissions predictions using the Random Forest and LightGBM models.

Feature importance highlights the contribution of individual features to the predictions and helps identify the most influential parameters.

Figures 8 and 9 illustrate the feature importance rankings for LightGBM models predicting **CO** and **NOX** emissions, respectively. Similarly, Figures 10 and 11 present the feature importance rankings for the Random Forest models.

For **CO** emissions, LightGBM identified `AFDP` (Air Flow Discharge Pressure) as the most important feature, followed by `AT` (Ambient Temperature) and `AP` (Ambient Pressure). On the other hand, the Random Forest model highlighted `TIT` (Turbine Inlet Temperature) as the most critical feature, followed by `TAT` (Turbine After Temperature) and `AFDP`. These differences indicate that LightGBM and Random Forest prioritize different aspects of the data for predicting **CO** emissions.

For **NOX** emissions, LightGBM again ranked `AFDP` and `AT` as the top contributors, emphasizing their relevance. The Random Forest model, however, identified `AT` as the most influential feature, followed by `TIT` and `GTEP` (Gross Turbine Energy Production). These variations further underline the differences in feature prioritization across models.

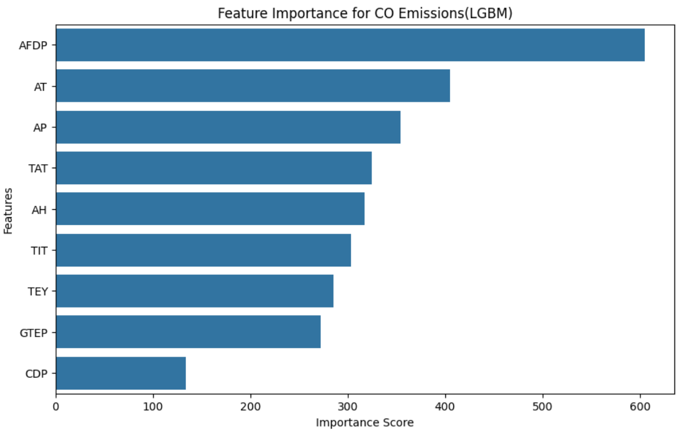

Fig. 8. Feature importance for **CO** emissions predictions using LightGBM.

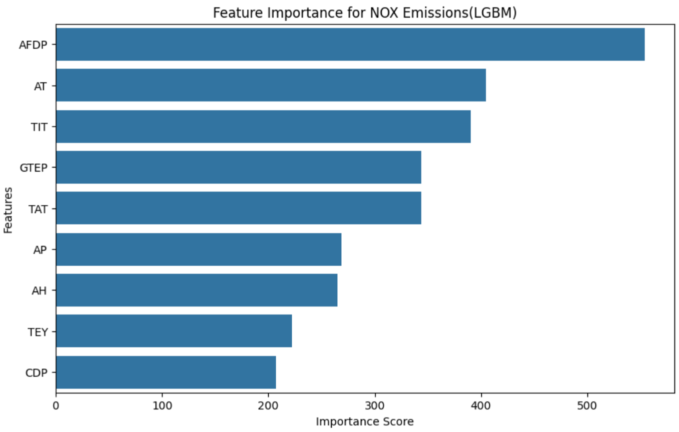

Fig. 9. Feature importance for **NOX** emissions predictions using LightGBM.

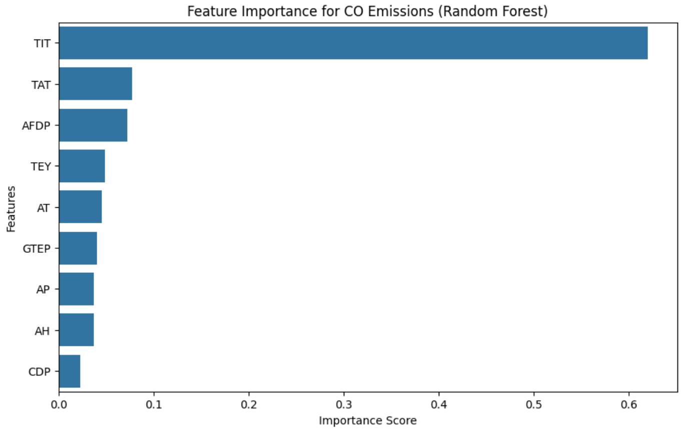

Fig. 10. Feature importance for **CO** emissions predictions using Random Forest.

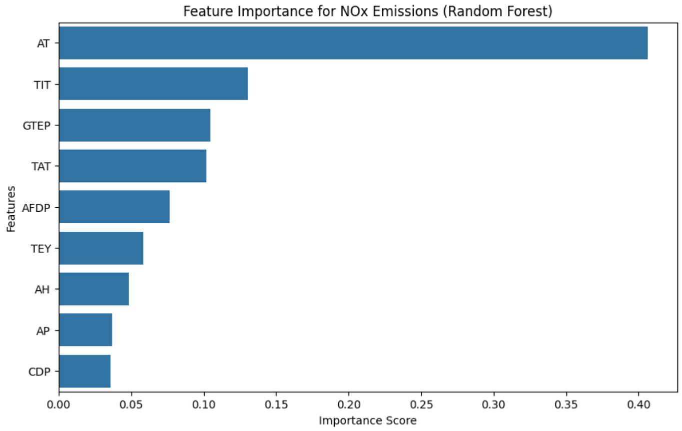

Fig. 11. Feature importance for **NOX** emissions predictions using Random Forest.

The feature importance analysis provides valuable insights into the driving factors behind **CO** and **NOX** emissions. While models such as LightGBM and Random Forest leverage different aspects of the data, the consistently high ranking of features like `AFDP`, `AT`, and `TIT` across models underscores their critical role in emissions predictions.

V. Turbine Energy Yield Predictions

This section focuses on the additional interest of predicting **Turbine Energy Yield (TEY)** using various machine learning models and analyzing the corresponding feature importance.

A. Scatter Plots of TEY vs Features

To understand the relationships between TEY and key emissions parameters (**CO** and **NOX**), scatter plots were generated. Figures 12 and 13 illustrate the trends.

B. Data Preprocessing and Scaling

To improve model performance, logarithmic transformations were applied to certain features (`AP`, `AH`) to handle non-linear relationships and scaling was performed using StandardScaler.

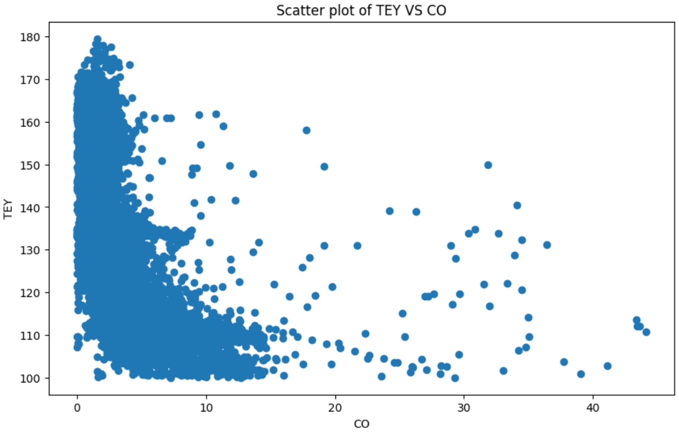

Fig. 12. Scatter plot of TEY vs CO emissions.

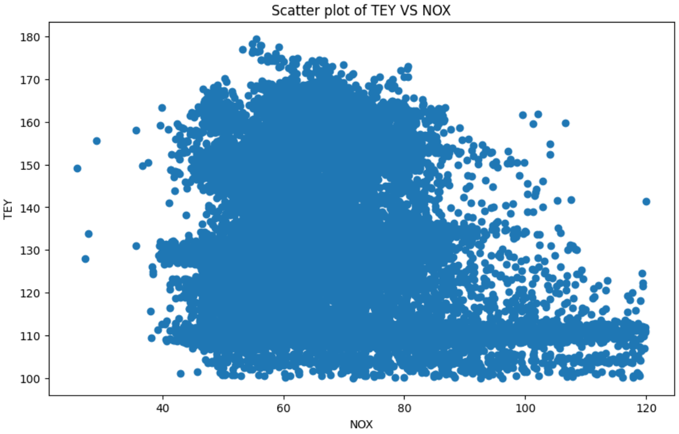

Fig. 13. Scatter plot of TEY vs NOX emissions.

This ensured that all features (AP, AT, AH) had a mean of 0 and a standard deviation of 1.

C. Model Training and Results

Six machine learning models, including **Linear Regression**, **Random Forest**, **Support Vector Machine**, **LightGBM**, **CatBoost**, and a **Neural Network**, were trained on the scaled features (AP, AT, AH) to predict TEY.

The models were evaluated using the following metrics:

- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)
- R^2 (Coefficient of Determination)

The summarized results for each model are presented in Table VI.

D. Feature Importance Analysis

The Random Forest model was chosen for feature importance analysis. The top contributing features were identified as AT, AH, and AP, as shown in Figure 14.

TABLE VI
MODEL PERFORMANCE METRICS FOR TEY PREDICTIONS.

Model	**Target**	**MAE**	**RMSE**	R^2
Linear Regression	TEY	5.776	8.211	0.491
Random Forest	TEY	2.596	4.074	0.875
Support Vector Machine	TEY	3.757	5.116	0.803
LightGBM	TEY	3.203	4.694	0.834
CatBoost	TEY	2.872	4.242	0.864
Neural Network	TEY	3.649	5.265	0.791

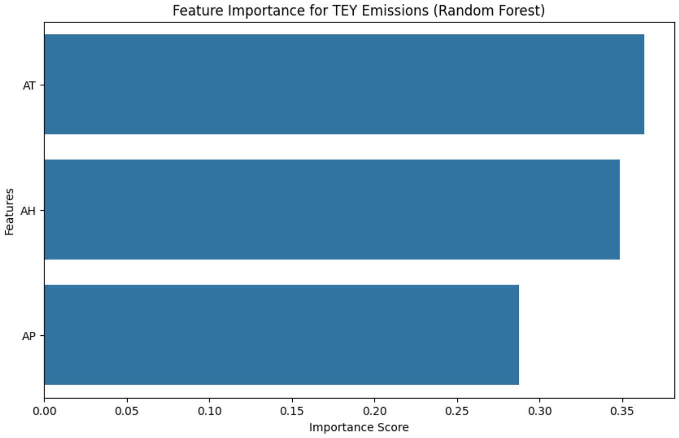

Fig. 14. Feature importance for TEY predictions using Random Forest.

VI. Conclusion

This report presented a comprehensive analysis of gas turbine operational and emission parameters, with a particular focus on predicting turbine energy yield (TEY). The following key conclusions were drawn from the study:

A. Operational and Emission Insights

A detailed exploratory data analysis highlighted the relationships between critical operational parameters such as ambient temperature (AT), ambient pressure (AP), and ambient humidity (AH), and emissions metrics like carbon monoxide (CO) and nitrogen oxides (NOX). Scatter plots demonstrated nonlinear trends between these variables and turbine energy yield. Logarithmic transformations and feature scaling were applied to address non-linearity and improve the performance of machine learning models.

B. Machine Learning for Emission Predictions

Multiple machine learning models, including Linear Regression, Random Forest, Support Vector Machine, LightGBM, CatBoost, and Neural Networks, were implemented to predict both CO and NOX emissions. Random Forest and CatBoost demonstrated superior predictive performance, with Random Forest achieving the lowest error metrics (MAE, RMSE) and the highest R^2 scores for both CO and NOX predictions. This highlighted their robustness in capturing complex, nonlinear relationships in the data.

C. Turbine Energy Yield Predictions

An additional focus on predicting TEY revealed that Random Forest and CatBoost models provided the best performance, with Random Forest achieving an R^2 score of 0.875. Feature importance analysis for TEY predictions identified **AT**, **AH**, and **AP** as the most influential factors, emphasizing the role of environmental conditions in determining turbine energy output.

D. Feature Importance and Interpretability

- Feature importance visualizations using Random Forest provided valuable interpretability for both emission and TEY predictions, enabling a better understanding of the underlying drivers for these targets.
- Ambient temperature (**AT**) consistently emerged as a critical factor across all models, reinforcing its central role in gas turbine performance and emissions.

E. Overall Contributions

This study demonstrates the power of machine learning in analyzing and predicting gas turbine operational metrics and emissions. By integrating data preprocessing, model evaluation, and interpretability techniques, the findings provide actionable insights for optimizing turbine performance and reducing emissions. The results further underscore the importance of adopting data-driven approaches for efficient and sustainable gas turbine operations.

F. Future Work

While the models developed in this study achieved high accuracy, further exploration can focus on:

- Expanding the dataset with additional features, such as fuel composition and turbine load variations.
- Leveraging advanced deep learning architectures to capture even more complex relationships.
- Implementing real-time prediction systems for monitoring and optimizing turbine performance in operational environments.

In conclusion, this report provides a solid foundation for future research and practical implementation in gas turbine operations, contributing to the overarching goals of efficiency and sustainability in energy systems.

The combination of insights from scatter plots, preprocessing, and advanced machine learning models provides a comprehensive approach to predict and analyze the turbine energy yield.

References

[1] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[2] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[3] R. Espinosa, J. Palma, F. Jiménez, J. Kamińska, G. Sciavicco, and E. Lucena-Sánchez, "A time series forecasting based multi-criteria methodology for air quality prediction," *Applied Soft Computing*, vol. 113, p. 107850, 2021.

[4] L. dos Santos Coelho, H. V. H. Ayala, and V. C. Mariani, "Co and nox emissions prediction in gas turbine using a novel modeling pipeline based on the combination of deep forest regressor and feature engineering," *Fuel*, vol. 355, p. 129366, 2024.

[5] M. Mahdaviara, M. Sharifi, S. Bakhshian, and N. Shokri, "Prediction of spontaneous imbibition in porous media using deep and ensemble learning techniques," *Fuel*, vol. 329, p. 125349, 2022.

[6] Z. Huang, R. Li, and Z. Chen, "Integration of data-driven models for dynamic prediction of the sagd production performance with field data," *Fuel*, vol. 332, p. 126171, 2023.

[7] R. G. da Silva, S. R. Moreno, M. H. D. M. Ribeiro, J. H. K. Larcher, V. C. Mariani, and L. dos Santos Coelho, "Multi-step short-term wind speed forecasting based on multi-stage decomposition coupled with stacking-ensemble learning approach," *International Journal of Electrical Power & Energy Systems*, vol. 143, p. 108504, 2022.

[8] X. Wang, X. Wang, B. Ma, Q. Li, C. Wang, and Y. Shi, "High-performance reversible data hiding based on ridge regression prediction algorithm," *Signal Processing*, vol. 204, p. 108818, 2023.