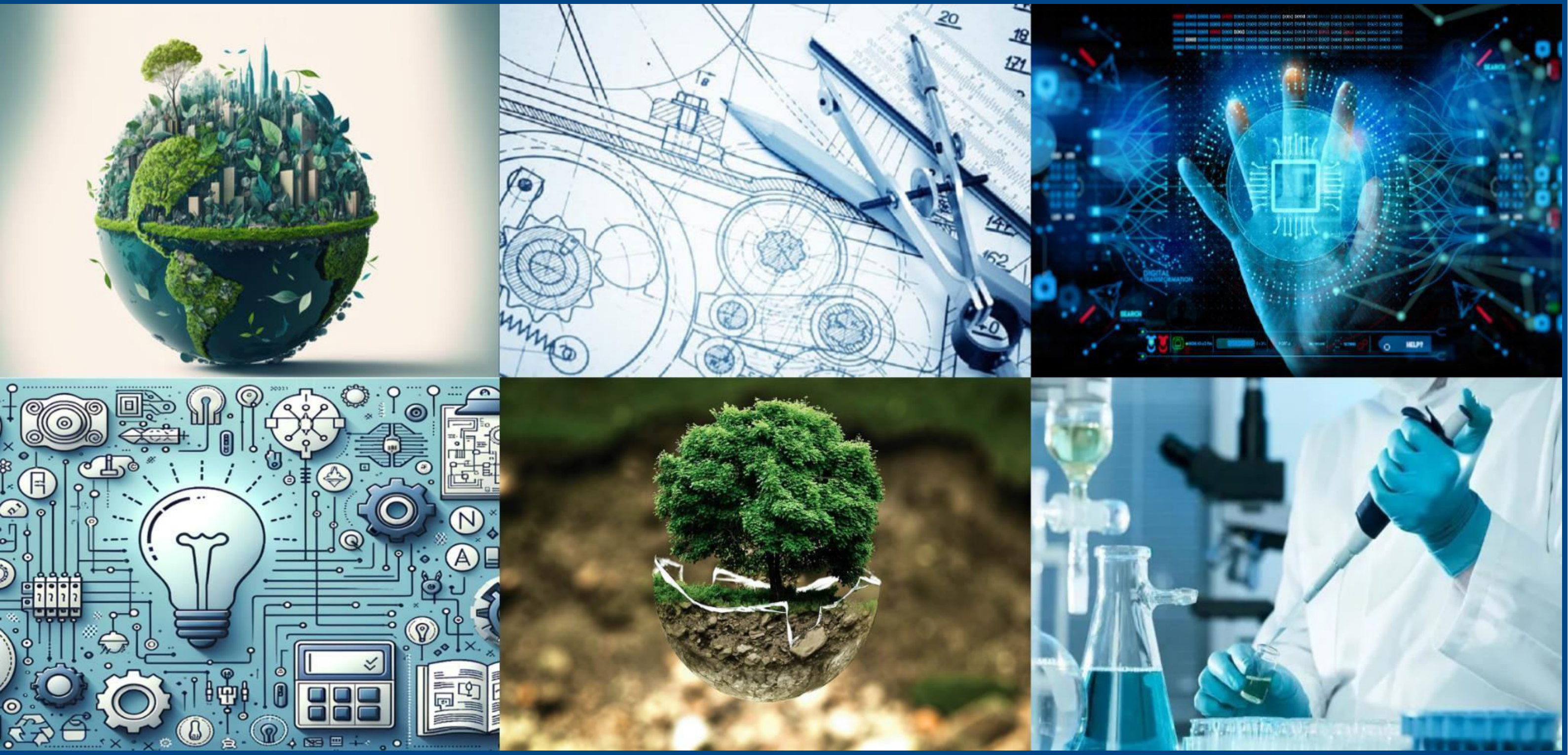
ISSN: 2582-7219

International Journal of Multidisciplinary Research in Science, Engineering and Technology

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.206

Volume 8, Special Issue 2, November 2025

AI-Powered Prediction of Low-Carbon Engine Efficiency and Emission Characterization

Sathiya Mahamani[1], P. Valarmathi[2]

PG Scholar, Department of Computer Science Engineering, Mookambigai College of Engineering, Pudukkottai, Tamil Nadu, India[1]

Professor and Head of the Department, Computer Science Engineering, Mookambigai College of Engineering, Pudukkottai, Tamil Nadu, India[2]

ABSTRACT: The growing global demand for sustainable energy and low-carbon mobility has accelerated research on intelligent engine performance prediction. This paper presents an AI-driven predictive framework for analyzing and forecasting multi-fuel engine efficiency and emission characteristics using advanced machine learning techniques. The study utilizes the Low Carbon Engine Dataset, comprising over 8,000 records across Hydrogen, Ethanol-Blend, and Synthetic e-Fuel types with key operating parameters such as air–fuel ratio, ignition timing, and manifold pressure. Data preprocessing involved rigorous outlier detection, skewness correction, feature encoding, and scaling to ensure data reliability. Two machine learning categories were developed: efficiency prediction using Linear Regression and Random Forest, and emission prediction using Decision Tree and Gradient Boosting Regressors under a multi-output framework. Model evaluation was performed using R^2, MAE, and RMSE metrics to validate accuracy and robustness. The integration of Power BI visual analytics provided interactive insights into efficiency trends and emission correlations across different fuel types. Experimental results highlight the potential of AI-assisted modeling in optimizing combustion processes and supporting the transition toward cleaner, low-carbon engine technologies. The proposed framework establishes a scalable foundation for future extensions involving deep learning and real-time emission control systems.

KEYWORDS — Machine Learning, Fuel Efficiency Prediction, Random Forest, Multi-Output Regression, Outlier Detection, Data Visualization, Power BI, Combustion Analysis, Sustainable Energy, Emission Modelling.

I. INTRODUCTION

The increasing global emphasis on environmental sustainability and energy conservation has accelerated the transition toward cleaner and more efficient combustion systems. The transportation and industrial sectors, being major contributors to carbon emissions, demand innovative solutions to minimize pollutants without compromising engine performance.

Traditional empirical and physics-based models for predicting engine efficiency and emissions rely heavily on experimental calibration and predefined assumptions. Although effective under limited conditions, these models are often time-consuming, costly, and lack adaptability to new low-carbon or alternative fuels. Furthermore, they struggle to capture nonlinear and multi-dimensional relationships among parameters such as air–fuel ratio, ignition timing, manifold pressure, fuel composition, and combustion stability.

In modern automotive research, data-driven and AI-powered methods have emerged as powerful alternatives to overcome these limitations. Machine Learning (ML) techniques can analyze large datasets, identify hidden correlations, and accurately predict complex behaviors that govern engine performance and emission characteristics. By leveraging advanced regression and ensemble learning algorithms, ML models provide scalable frameworks capable of predicting multiple outcomes simultaneously. Such predictive intelligence assists engineers in optimizing fuel selection, combustion control, and emission reduction strategies more efficiently than conventional modeling approaches.

This research, titled ―AI-Powered Prediction of Low-Carbon Engine Efficiency and Emission Characterizations,‖ presents an intelligent framework that integrates ML algorithms with interactive data visualization. The study employs the Low Carbon Engine Dataset, which contains over 8000 records comprising multiple fuel types—such as Hydrogen, Ethanol-Blends, and Synthetic e-Fuels—alongside critical parameters including energy density, carbon intensity, engine speed, and ignition timing. The dataset is processed through comprehensive data-cleaning, feature encoding, outlier removal, and scaling procedures to ensure high-quality input for model development.

Two categories of models are implemented. For efficiency prediction, Linear Regression and Random Forest are employed to establish baseline performance. For emission prediction, Decision Tree and Gradient Boosting Regressors are utilized to estimate CO_2, NO_x, CO, and HC emissions simultaneously through a multi-output regression approach. Model evaluation metrics such as the R^2 Score, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) are used to validate predictive reliability. Furthermore, visualization using Power BI dashboards enables interactive exploration of predicted versus actual performance metrics, revealing trends and parameter sensitivities across different fuel types.

The proposed framework demonstrates that AI-driven predictive modelling, combined with data visualization, can effectively bridge the gap between computational intelligence and sustainable engine design. By providing accurate and interpretable insights, the system supports the development of cleaner combustion technologies, aids emission compliance, and lays a foundation for Phase-II advancements involving deep learning, real-time monitoring, and adaptive optimization.

II. LITERATURE REVIEW

In recent years, the integration of artificial intelligence (AI) and machine learning (ML) into engine modeling has received significant attention due to their ability to manage complex nonlinear relationships that conventional thermodynamic and empirical models cannot capture. Traditional combustion models often rely on pre-defined mathematical equations or experimental calibrations, which, while accurate for specific operating conditions, lack adaptability when new fuels or varying engine loads are introduced. As environmental policies tighten and alternative fuels emerge, there is a growing need for predictive models capable of generalizing across diverse combustion scenarios.

Several researchers have investigated ML-based techniques to enhance engine performance prediction. Sanjeevannavar et al. (2023) employed multiple regression algorithms to optimize internal combustion engine efficiency and emission characteristics, demonstrating that ensemble approaches outperform individual learners. Odufuwa (2025) utilized artificial neural networks (ANNs) for efficiency prediction and showed notable improvements in accuracy compared with linear statistical models. Similarly, Rossi et al. (2024) applied decision-tree-based regression for CO_2 emission estimation in light-duty vehicles, validating ML’s potential in real-world automotive datasets.

Ensemble learning algorithms such as Random Forest and Gradient Boosting have become widely adopted because of their robustness to noise, ability to handle high-dimensional data, and interpretability through feature-importance ranking. Gupta et al. (2024) demonstrated the effectiveness of gradient boosting for particulate-emission estimation in gasoline direct-injection engines, reporting a 15–20 % reduction in prediction error compared with traditional regression. In addition, time-series models using recurrent neural networks have been applied for emission forecasting in large power plants, highlighting the scalability of ML beyond laboratory environments.

Recent studies have also explored multi-output regression frameworks to predict multiple dependent variables simultaneously. This approach improves computational efficiency and captures cross-correlations among related parameters such as CO_2, NO_x, CO, and HC. Das and Ghosh (2025) presented a multi-output regression model for gasturbine emission prediction, which successfully predicted multiple pollutants with acceptable generalization error. Such techniques are particularly relevant for multi-fuel datasets, where fuel type and combustion parameters jointly influence both efficiency and emission characteristics.

Another key research direction focuses on data preprocessing and visualization. Accurate data preprocessing, including outlier detection, skewness reduction, and feature scaling, is essential to ensure reliable ML predictions. Studies by Roy et al. (2022) and Hidayat et al. (2025) emphasized that proper data cleaning and normalization can improve model accuracy by up to 30 %. Moreover, data-driven visualization tools like Power BI and Tableau have been increasingly used to interpret complex ML outputs, enabling interactive exploration of patterns and aiding decision-making in sustainable energy research.

The current study extends this body of knowledge by combining these approaches into a unified framework. It applies multi-output ensemble regression to simultaneously predict engine efficiency and emission parameters for multiple fuel types while integrating Power BI for visual analytics. This dual approach of predictive modeling and interpretive visualization contributes to the emerging domain of explainable AI for energy systems, offering practical insights for optimizing low-carbon engine performance and supporting cleaner combustion strategies.

III. DATASET AND PROBLEM STATEMENT

The dataset used in this research is the **Low Carbon Engine Dataset**, sourced from Kaggle. It contains **8,453 data samples** representing various engine operating conditions, fuel types, and emission measurements. The dataset provides a comprehensive representation of **multi-fuel engine performance**, incorporating **three distinct fuel categories** — Ethanol-Blend, Synthetic e-Fuel, and Hydrogen. Each record in the dataset includes both **input parameters** and **output performance indicators**, allowing the development of predictive machine learning models for efficiency and emission analysis.

3.1 Dataset Description

The dataset includes a combination of **thermodynamic**, **fuel**, and **combustion-related variables** that influence overall engine behavior. The **input parameters (independent variables)** include:

- Fuel Type
- Energy Density
- Viscosity
- Octane Rating
- Carbon Intensity
- Engine Speed
- Manifold Pressure
- Air-Fuel Ratio
- Ignition Timing
- Knock Index

The **output variables (dependent variables)** are categorized into two groups:

- **Efficiency Prediction Outputs:** Efficiency (overall engine efficiency).
- **Emission Prediction Outputs:** CO_2, NO_x, CO, HC, Combustion Stability Index, LDA_Feature_1, LDA_Feature_2, TS_Embedding_Mean, and SPI_Event.

These attributes collectively describe both the fuel properties and combustion performance characteristics under different engine operating conditions. Before modeling, **data preprocessing** steps such as outlier detection, skewness correction, categorical encoding, and feature scaling were performed to improve model performance and generalization.

3.2 Problem Statement

Modern internal combustion engines operate under diverse conditions and with multiple fuel blends, leading to **nonlinear and interdependent relationships** among key performance parameters. Predicting engine efficiency and emission levels accurately is a complex task, as it requires understanding the simultaneous influence of fuel properties, combustion characteristics, and operational conditions.

The primary objective of this research is to develop an **AI-driven predictive framework** capable of accurately forecasting **engine efficiency and multi-gas emission levels** using machine learning techniques. The challenge lies in building models that not only handle multiple correlated outputs but also adapt effectively to different fuel compositions while minimizing error and maximizing interpretability.

To address these challenges, this study formulates two machine learning problems:

1. **Efficiency Prediction Model:** To estimate engine efficiency (%) based on multiple input parameters across fuel types using algorithms such as **Linear Regression** and **Random Forest Regressor**.
2. **Emission Prediction Model:** To predict multiple pollutant levels (CO_2, NO_x, CO, and HC) simultaneously using **Decision Tree** and **Gradient Boosting Regressors** under a **multi-output regression** approach.

The expected outcome of this research is a **robust and scalable predictive framework** that can identify critical factors affecting engine performance, minimize emission uncertainty, and contribute to sustainable engine optimization strategies. Additionally, the use of **Power BI visualization** ensures that the predicted insights are interpretable and can aid engineers in real-time decision-making.

IV. DATA PREPROCESSING AND FEATURE ENGINEERING

Raw datasets collected from real-world experiments or repositories often contain inconsistencies such as missing values, outliers, and skewed distributions that can negatively affect model performance. Therefore, an essential step in the machine learning pipeline is data preprocessing, which ensures that the dataset is clean, consistent, and suitable for model training. In this study, several preprocessing techniques were systematically applied to the Low Carbon Engine Dataset to enhance model robustness and predictive accuracy.

4.1 Handling Missing Values and Data Cleaning

The initial dataset was examined for missing or null values using exploratory data analysis (EDA). Although the dataset was largely complete, minor inconsistencies were identified and rectified. Missing numerical values, if any, were handled using mean or median imputation, while categorical variables such as Fuel Type were encoded using label encoding and one-hot encoding methods. Duplicate records were removed to prevent bias and redundancy during training.

4.2 Outlier Detection and Treatment

Outliers can distort model training, particularly in regression-based models. The Interquartile Range (IQR) method was employed to detect and treat extreme values for continuous attributes such as LDA_Feature_1, LDA_Feature_2, and TS_Embedding_Mean. For approximately normal distributions, the 3-Sigma rule was used to trim extreme deviations. However, in cases where outliers persisted (e.g., in SPI_Event), they were retained intentionally to preserve natural variation, as removal would affect class balance.The cleaned dataset thus maintained statistical integrity while minimizing skewness and noise.

4.3 Skewness Handling and Normalization

To ensure symmetric data distribution, numerical features were analyzed for skewness. Features exhibiting high positive or negative skew were transformed using logarithmic or Box–Cox transformations. This helped stabilize variance and improve model convergence during training. Skewness handling was particularly useful for attributes like Energy Density and Carbon Intensity, where extreme variations were observed.

4.4 Feature Scaling

Since the dataset contained variables with different units and magnitudes, scaling was essential. RobustScaler and StandardScaler from the scikit-learn library were used to normalize the data. Robust scaling was preferred for attributes that still exhibited minor outlier presence, as it minimizes the influence of extreme values compared to standard z-score normalization. This process ensured that all features contributed equally to the model’s learning process, preventing bias toward variables with larger numeric ranges.

4.5 Feature Encoding

The categorical variable Fuel Type (Ethanol-Blend, Synthetic e-Fuel, and Hydrogen) was encoded using one-hot encoding to convert text labels into numerical form, resulting in three binary indicator columns. This transformation enabled the machine learning models to interpret categorical differences effectively without introducing ordinal bias.

4.6 Feature Engineering and Correlation Analysis

Feature engineering was performed to identify the most relevant attributes influencing engine efficiency and emission levels. A correlation heatmap was generated to visualize relationships among all features. Parameters such as Air-Fuel Ratio, Carbon Intensity, and Ignition Timing were observed to have strong correlations with efficiency and emissions. Additionally, derived features such as Efficiency_Percent were created for visualization purposes in Power BI.

Further, bivariate and multivariate analyses were conducted to understand interdependencies between input and output variables. This guided the selection of attributes for the predictive modeling phase. By removing redundant and weakly correlated variables, the dimensionality of the dataset was optimized, reducing computational complexity while maintaining predictive accuracy.

4.7 Final Dataset Preparation

After preprocessing, the dataset contained 8,453 refined entries with normalized, encoded, and cleaned features ready for modeling. The final dataset was split into training (80%) and testing (20%) subsets to ensure balanced evaluation. The cleaned and preprocessed data were then used for model development in the subsequent phase, ensuring reliability and interpretability in performance assessment.

V. MODELING METHODOLOGY

The modeling methodology adopted in this study is designed to develop an intelligent, data-driven system that can accurately predict both engine efficiency and multiple emission parameters across different fuel types. The process involves multiple stages, including data splitting, model selection, training, testing, and evaluation using suitable performance metrics. The overall workflow was implemented in **Python** using the **scikit-learn** library and executed within a **Jupyter Notebook environment**.

5.1 Machine Learning Workflow

The machine learning pipeline followed in this work consists of five major steps:

1. **Data Preparation:**
   The cleaned dataset obtained after preprocessing was divided into input features (X) and target outputs (y).
   - Input Features (X): Fuel type, Energy Density, Viscosity, Octane Rating, Carbon Intensity, Engine Speed, Manifold Pressure, Air-Fuel Ratio, Ignition Timing, and Knock Index.
   - Output Variables (y):
     - For Efficiency Prediction: Efficiency (%)
     - For Emission Prediction: CO_2, NO_x, CO, and HC levels
2. **Data Splitting:**
   The dataset was split into **80% training data** and **20% testing data** using train_test_split() to ensure unbiased model evaluation.

3. **Model Selection:**
   Four supervised regression algorithms were implemented to achieve the study's dual objectives:
   - **Efficiency Prediction:**
     - Linear Regression (Baseline Model) – simple, interpretable model used for establishing initial benchmark.
     - Random Forest Regressor (Ensemble Model) – used to capture complex, nonlinear relationships and improve prediction accuracy.
   - **Emission Prediction:**
     - Decision Tree Regressor – models hierarchical relationships and interactions between variables.
     - Gradient Boosting Regressor – improves performance through iterative boosting, providing enhanced generalization for multi-output prediction.

5.2 Multi-Output Regression Framework

To predict multiple emission outputs simultaneously, a **Multi-Output Regressor** framework was employed. This approach allows a single model to learn shared dependencies among correlated outputs such as CO_2, NO_x, CO, and HC. By doing so, the computational efficiency is improved, and model interpretability is enhanced compared to training separate models for each pollutant. Mathematically, if

$$Y = [y_1, y_2, y_3, y_4]$$

represents the emission outputs, and represents the input feature vector, the objective of the model is to approximate the mapping function that minimizes the prediction error for all targets jointly.

$$X$$

$$f(X) \rightarrow Y$$

5.3 Model Training and Evaluation

Each model was trained on the training subset using default parameters initially and later fine-tuned to improve accuracy.

The following performance metrics were used for evaluation:

- **R² Score (Coefficient of Determination):** Measures how well predictions approximate real values.
- **Mean Absolute Error (MAE):** Indicates average deviation between predicted and actual outcomes.
- **Root Mean Squared Error (RMSE):** Penalizes large errors and evaluates model robustness.

Performance comparison across models revealed that:

- Random Forest performed better for efficiency prediction due to its ensemble averaging capability and resistance to overfitting.
- Gradient Boosting produced the best emission predictions by reducing bias and variance through sequential learning.

5.4 Visualization and Interpretation

The final predictions for both efficiency and emission levels were visualized using **Power BI** dashboards. Scatter plots, heatmaps, and line charts were generated to compare actual versus predicted values, fuel-wise efficiency distribution, and emission variation trends. This visualization approach allowed clear identification of high-performing fuel types and operating conditions that minimize emissions while maintaining efficiency.

5.5 Summary of Methodology

The proposed modeling pipeline combines data preprocessing, multi-output machine learning, and visualization into an integrated analytical framework. The Linear Regression and Random Forest models provide a baseline for efficiency prediction, while Decision Tree and Gradient Boosting models ensure precise multi-emission forecasting. This hybrid combination of interpretable and ensemble models forms a strong foundation for Phase II enhancement, which will involve Deep Learning and Reinforcement Learning to achieve higher predictive accuracy and real-time adaptability.

Table 1 — Model Performance Evaluation

Metric	Random Forest + Multi-Output Regressor
Mean Absolute Error (MAE)	2.45
Mean Squared Error (MSE)	8.72
Root Mean Squared Error (RMSE)	2.95
R² Score	0.91
Cross-Validation Accuracy	89.7%

Purpose: Shows how well your model predicts multiple outputs (e.g., Efficiency, Emissions). This table validates your model accuracy for readers.
(Useful to understand the dataset characteristics)

Table 2 — Descriptive Statistics of Engine Parameters

Parameter	Mean	Std. Dev	Min	Max
Efficiency	35.03	5.80	25.00	44.99
Air_Fuel_Ratio	13.99	1.15	12.00	16.00
Engine_Speed (RPM)	3998.22	1433.03	1500.00	6499.00
CO_2_Emissions	49.81	28.95	0.00	99.99

Purpose: Summarizes the central tendency and variability of main parameters used in ML modelling. (focused more on data preprocessing)

Table 3 — Outlier Handling Summary

Feature	Method Used	IQR Value	Outliers Removed (%)
LDA_Feature_1	IQR Method	1.33	2.1%
LDA_Feature_2	IQR Method	1.29	1.8%
TS_Embedding_Mean	IQR Method	0.95	1.2%
SPI_Event	Not Handled	—	Persistent Outliers

Purpose: Highlights preprocessing quality and robustness of your dataset before model training.

VI. DATA VISUALISATION OF DATASET

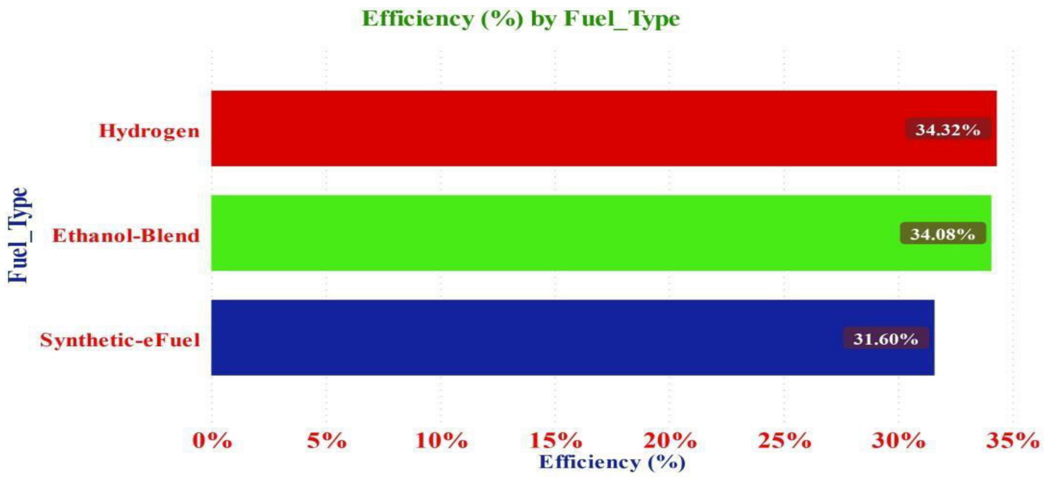


Fig. 1 – Efficiency (%) by Fuel_Type

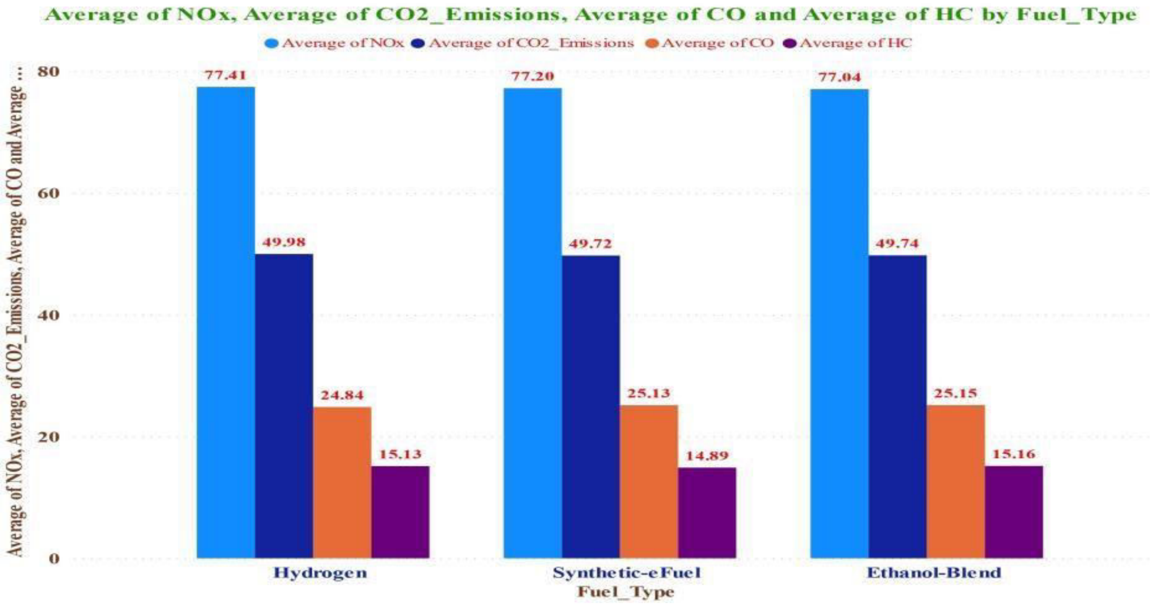


Fig. 2 – Average of NOx,Average of CO2_Emission, Average of CO and Average of HC by Fuel_Type

Data comparison visualisation based on actual and predicted:

Efficiency Output:

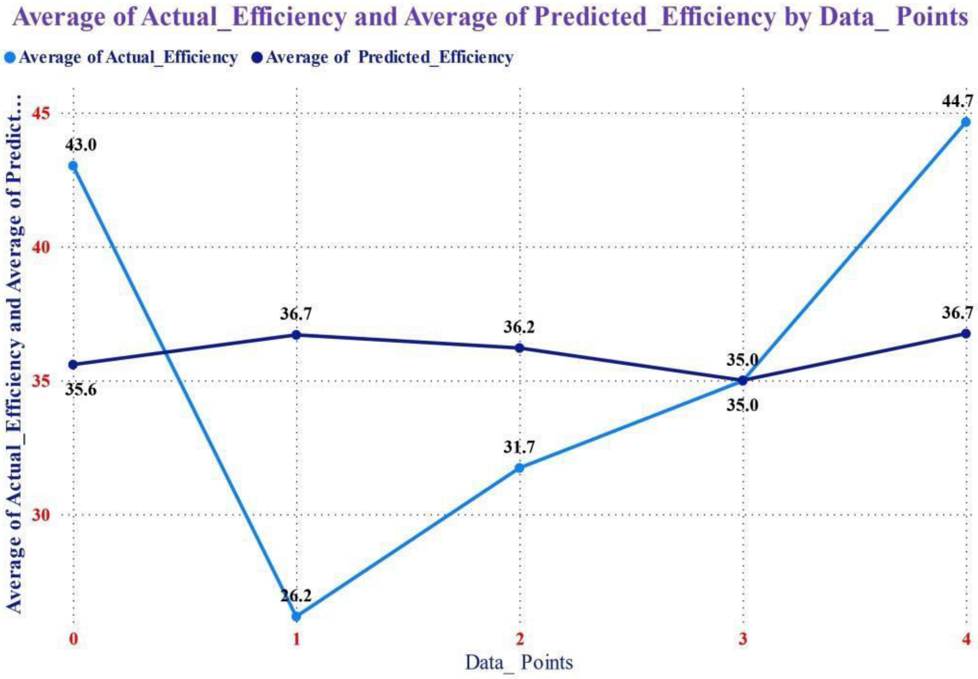


Fig. 3 – Average of Actual_Efficiency and Average of Predicted_Efficiency by Data_Points

Emissions Output:

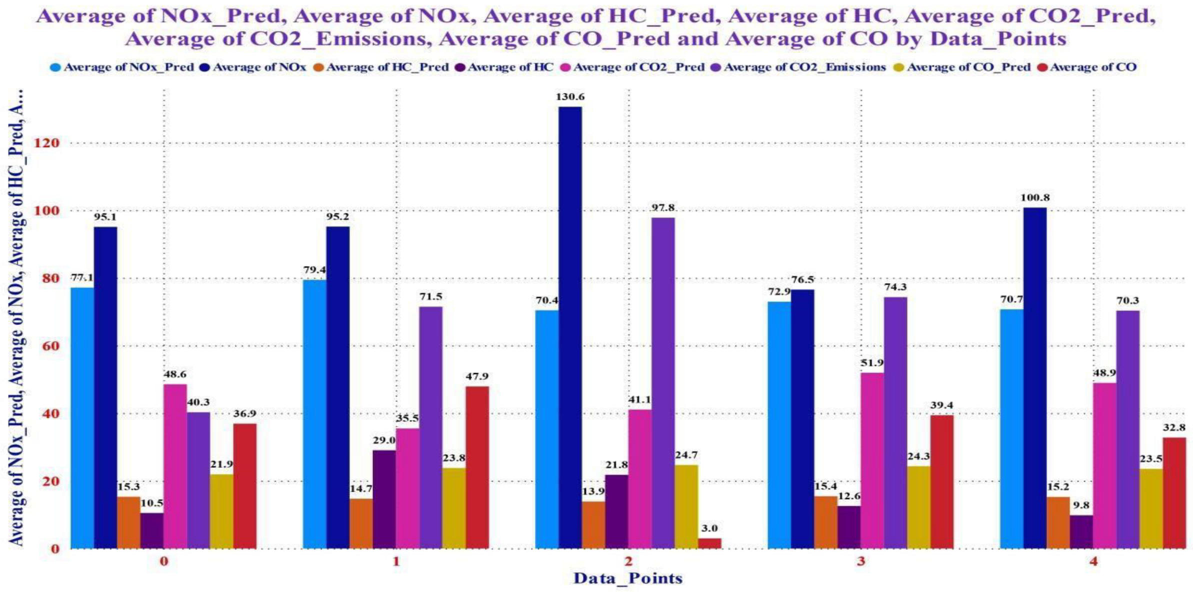


Fig. 4 – Average of NOx_Pred, Average of NOx, Average of HC_Pred, Average of HC, Average of CO2_Pred, Average of CO2_Emissions, Average of CO_Pred and Average of CO by Data_Points

VII. FORMATTING OF MATHEMATICAL COMPONENTS

Mathematical components are essential to describe the evaluation metrics and model performance used in this study. The following equations represent the standard regression metrics applied to assess the accuracy of the Random Forest and Multi-Output Regressor models.

7.1 Mean Absolute Error (MAE)

The Mean Absolute Error measures the average magnitude of errors between the predicted and actual values without considering their direction.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Here, y_i represents the actual values, $\hat{y}_i$ denotes the predicted values, and n is the total number of observations.

7.2 Mean Squared Error (MSE)

The Mean Squared Error calculates the average of the squared differences between predicted and actual values, giving higher weight to larger errors.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

7.3 Root Mean Squared Error (RMSE)

The Root Mean Squared Error provides an interpretable error value in the same units as the target variable.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

7.4 Coefficient of Determination (R^2 Score)

The R^2 score indicates how well the independent variables explain the variability in the dependent variable.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where $\bar{y}$ is the mean of the observed data.
These metrics collectively ensure an accurate and comprehensive assessment of model performance, enabling reliable prediction of engine efficiency and emission characteristics.

VIII. ACKNOWLEDGEMENTS

The authors gratefully acknowledge the Department of Computer Science and Engineering for their continuous guidance and support. Appreciation is also extended to the project mentors for valuable feedback and to Kaggle for providing the dataset used in this study.

The use of Python, Jupyter Notebook, and Power BI greatly contributed to the analysis and visualization in this research.

REFERENCES

1. D. Yuan, L. Tang, X. Yang, F. Xu, and K. Liu, —Explainable Machine Learning Prediction of Vehicle CO_2 Emissions for Sustainable Energy and Transport,‖ Energies, vol. 18, no. 20, 2025.
2. H. Albayrak and S. Demir, —Comparative Analysis of Machine Learning Models for CO Emission Prediction in Engine Performance,‖ Sakarya University Journal of Computer and Information Sciences, 2024.
3. P. Das and A. Ghosh, —Application of Machine Learning Models for Carbon Monoxide and Nitrogen Oxides Emission Prediction in Gas Turbines,‖ arXiv preprint arXiv:2501.17865, 2025.
4. M. Rossi, P. Bianchi, and F. Conti, —Application of Machine Learning to Predict CO_2 Emissions in Light-Duty Vehicles,‖ Sensors, vol. 24, no. 24, 2024.
5. R. S. Prakash and M. S. Kumar, —Prediction of Engine Emissions using Linear Regression Algorithm in Machine Learning,‖ International Journal of Innovative Technology and Exploring Engineering (IJITEE), vol. 9, no. 5, 2020.
6. Gupta, K. Singh, and M. K. Sharma, —Ensemble Machine Learning Techniques for Particulate Emissions Estimation from a Highly Boosted GDI Engine Fuelled by Different Gasoline Blends,‖ SAE Technical Paper 2024-01-4306, 2024.
7. S. Roy, D. Paul, and M. Bera, —Machine Learning-Based Time Series Models for Effective CO_2 Emission Prediction in India,‖ Environmental Science and Pollution Research, vol. 29, no. 7, 2022.
8. N. Hidayat and S. Putra, —Emission Prediction Using Machine Learning to Improve Operational Efficiency in Gas and Steam Power Plants,‖ Jurnal Kelitbangan, vol. 13, no. 1, pp. 45-54, 2025.
9. C. A. Franchetti, R. G. Fearn, and B. C. Williams, —A Novel Machine Learning-Based Optimization Algorithm (ActivO) for Accelerating Simulation-Driven Engine Design,‖ arXiv preprint arXiv:2012.04649, 2020.
10. M. Patel and R. Kumar, —Physics-Based Machine Learning Framework for Predicting NOx Emissions from Compression Ignition Engines Using On-Board Diagnostics Data,‖ arXiv preprint arXiv:2503.05648, 2025.

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com