SAKARYA UNIVERSITY JOURNAL OF COMPUTER AND INFORMATION SCIENCES
http://saucis.sakarya.edu.tr/
e-ISSN: 2636-8129
Publisher: Sakarya University
Vol. 8, No. 1, 1-11, 2025
DOI: 10.35377/saucis...1564937
SAUCIS
SAKARYA UNIVERSITY
RESEARCH ARTICLE

Comparative Analysis of Machine Learning Models for CO Emission Prediction in Engine Performance

Beytullah Eren[1], **İdris Cesur**[2]

[1]Sakarya University, Faculty of Engineering, Department of Environmental Engineering, Sakarya, Türkiye, ror.org/04ttnw109
[2] Sakarya University of Applied Sciences, Faculty of Technology, Department of Mechanical Engineering, Sakarya, Türkiye, ror.org/01shwhq58

Corresponding author:
Beytullah Eren, Sakarya University,
Faculty of Engineering,
Department of Environmental Engineering
beren@sakarya.edu.tr

Article History:
Received: 10.10.2024
Revised: 17.12.2024
Accepted: 15.01.2025
Published Online: 27.03.2025

ABSTRACT
This study presents a comparative analysis of machine learning models for predicting carbon monoxide (CO) emissions in automotive engines. Four models—Linear Regression, Decision Tree, Random Forest, and Support Vector Regression—were evaluated using a dataset of engine performance parameters and emission measurements. Among these, the Random Forest model demonstrated the highest predictive accuracy, achieving an R^2 score of 0.8965. Feature importance analysis identified nitrogen oxides (NO_X), engine speed (RPM), and hydrocarbons (HC) as the most significant predictors of carbon monoxide emissions. Learning curve analysis provided insights into model generalization and highlighted potential limitations. The study underscores the value of data-driven approaches in optimizing engine design and controlling emissions. The findings contribute to the development of cleaner, more efficient vehicles, supporting sustainability efforts in the automotive industry. This research bridges data science and automotive engineering, offering a framework for advanced emission prediction and control that can be applied to other pollutants and engine types.

Keywords: Carbon monoxide emissions, Machine learning, Random Forest, Engine performance optimization, Emission control, Sustainability, Automotive engineering

1. Introduction

The automotive industry is at a critical crossroads, tasked with enhancing engine performance while significantly reducing harmful emissions. Among these emissions, carbon monoxide (CO) poses a considerable threat to both human health and the environment [1]. As a byproduct of incomplete combustion, CO can lead to severe respiratory issues and, at high concentrations, may even be life-threatening [2]. Moreover, CO contributes to ground-level ozone formation, a major component of smog that exacerbates air quality concerns [3]. To combat these challenges, stringent global regulations, such as the European Union's Euro 6 standards and the United States' Tier 3 regulations, have been implemented, driving the need for innovative emission reduction strategies [4, 5].

Traditional approaches to emission control in internal combustion engines, such as optimizing engine design and using after-treatment systems, often involve trade-offs with engine performance and fuel efficiency [6, 7]. These methods also struggle to address the complex, non-linear interactions between engine parameters and emission outputs, highlighting the limitations of conventional techniques.

Recent advancements in data analytics and machine learning (ML) have introduced new opportunities to tackle these challenges. ML techniques are particularly adept at modeling complex, non-linear relationships between variables, enabling more accurate emission predictions and optimized control strategies [8, 9]. Studies have applied machine learning algorithms to predict pollutants like NO_X and CO_2, demonstrating promising results. Artificial Neural Networks (ANNs) have been effective in predicting NO_X emissions, capturing intricate relationships between operating conditions and outputs [10]. Similarly, Support Vector Machines (SVMs) and Random Forests have shown success in estimating particulate matter and CO_2 emissions, respectively, due to their ability to manage complex feature interactions and robust performance across diverse datasets [11, 12]. Deep learning techniques, particularly hybrid models like Convolutional Neural Networks (CNN)

Cite as: B. Eren and İ. Cesur, "Comparative analysis of machine learning models for CO emission prediction in engine performance," *Sakarya University Journal of Computer and Information Sciences,* vol. 8, no. 1, pp. 1-11, 2025. doi: 10.35377/saucis...1564937

combined with Long Short-Term Memory (LSTM) networks, have further enhanced the modeling of temporal emission patterns, particularly under transient operating conditions [13].

Despite these advancements, limited research has focused on comparing multiple ML models specifically for CO emission prediction. CO emissions are uniquely sensitive to various engine parameters and operating conditions, making their prediction particularly challenging [14]. Moreover, studies rarely address feature importance, which is critical for understanding the key engine parameters driving CO emissions. The lack of comprehensive comparative analyses leaves gaps in identifying the most effective ML techniques and their practical implications for emission reduction.

This study addresses these gaps by evaluating the performance of four widely used machine learning models—Linear Regression, Decision Tree, Random Forest, and Support Vector Regression (SVR)—in predicting CO emissions based on engine performance parameters. The study aims to: (1) assess the predictive accuracy of these models across varying engine operating conditions, (2) identify the most influential engine parameters through feature importance analysis, (3) evaluate the generalization capabilities and limitations of the models using learning curves, and (4) provide actionable insights for integrating ML techniques into emission control strategies and engine design. By offering a comparative analysis of machine learning models and their practical applications, this research contributes to the growing body of knowledge on data-driven emission reduction approaches and provides a foundation for future studies and industrial applications in optimizing cleaner and more efficient automotive technologies.

2. Materials and Method

2.1. Dataset Description

The dataset used in this study was derived from a series of controlled engine performance tests conducted on a two-cylinder, electronic injection, naturally aspirated, water-cooled, four-stroke, spark ignition engine. This engine type was selected due to its widespread use in small- to medium-sized passenger vehicles, making it highly representative of real-world applications. Additionally, its relatively simple design and operation allow for a more focused analysis of the relationship between engine parameters and CO emissions without interference from complex subsystems, such as turbocharging or advanced after-treatment devices.

The experiments were performed under full-load conditions in a controlled laboratory environment to ensure consistency and reproducibility. Key engine parameters were varied systematically, including fuel composition (gasoline-ethanol blends of E10, E20, and E30) and engine speed (RPM) across a broad operational range. These conditions were chosen to represent typical and extreme scenarios encountered during real-world engine operation, providing a comprehensive dataset for modeling.

Emission values, including nitrogen oxides (NOx), hydrocarbons (HC), and carbon monoxide (CO), were measured using a calibrated MRU Delta 1600 L exhaust gas analyzer. To minimize variability, five repeated measurements were taken for each test point, and the average values were used for analysis. In total, 90 data points were collected, representing a diverse range of operating conditions and configurations relevant to CO emission prediction.

While the dataset provides a representative sample of typical spark ignition engine conditions, it is important to note its limitations. The controlled laboratory environment excludes external variables such as temperature, humidity, and altitude, which can influence real-world emission outputs. Furthermore, the focus on a single engine type does not account for variations in design and operation seen in turbocharged or diesel engines. Future research should expand the dataset to include these factors for broader applicability.

Feature scaling was applied using the StandardScaler function from the scikit-learn library to normalize input variables, ensuring consistency across features and facilitating model training [15].

2.2. Machine Learning Models

In this study, four machine learning models were implemented to predict carbon monoxide (CO) emissions based on the input features, each with unique characteristics and strengths for handling different data relationships. Below is a detailed explanation of each model:

Linear Regression: Linear Regression is a fundamental statistical model that serves as a baseline for comparison. It assumes a linear relationship between the independent variables (engine parameters) and the dependent variable (CO emissions). The model calculates coefficients for each feature by minimizing the sum of squared residuals, resulting in a straightforward predictive framework. While efficient and interpretable, this method struggles with non-linear relationships or feature interactions [16].

Decision Tree: The Decision Tree model is a non-linear algorithm that partitions the dataset into subsets based on feature values. The model splits the data at decision nodes to minimize an impurity measure, such as the Gini index or entropy, in

classification or Mean Squared Error (MSE) in regression. This model is particularly effective at capturing non-linear patterns and interactions between features. However, a single tree can be prone to overfitting, especially in datasets with noise [17].

Random Forest: Random Forest is an ensemble learning method that combines the predictions of multiple decision trees. Each tree is built on a random subset of data and features, introducing diversity and reducing overfitting. The final prediction is derived by averaging the predictions (for regression tasks) from all the individual trees. This robustness to overfitting and ability to handle complex, non-linear relationships makes Random Forest particularly suited for high-dimensional and noisy datasets [18].

Support Vector Regression (SVR): SVR is a versatile algorithm that maps input features into a high-dimensional space using a kernel function (e.g., linear, polynomial, or radial basis function) to capture linear and non-linear relationships. The model aims to fit a regression hyperplane within a tolerance margin, minimizing prediction error while controlling model complexity. This makes SVR particularly effective for datasets with intricate, non-linear patterns. However, its performance can be sensitive to parameter tuning, such as the choice of the kernel, regularization parameter (C), and epsilon (ε), which defines the margin of tolerance [19].

All models were implemented using the Python programming language and the scikit-learn library, a well-documented toolkit for machine learning applications [15]. The best-performing model also facilitated feature importance analysis, providing valuable insights into the relative contributions of input features to CO emission predictions.

2.3. Experimental Procedure

The experiments were conducted using a two-cylinder spark plug test engine on an engine test stand to determine engine performance parameters and exhaust emissions. An electric dynamometer was employed to provide full load conditions, while a mass-scale fuel measurement system was used to measure fuel consumption at varying engine speeds. Exhaust emissions were measured using an MRU Delta 1600L emission analyzer. The experiments were performed under full load conditions at engine speeds ranging from 1400 to 3400 rpm, in 400 rpm intervals. Initially, gasoline was used as the fuel, followed by ethanol-blended fuels. The experimental results obtained were subsequently used as input data for model development.

The experimental procedure included key steps to ensure a rigorous evaluation of machine learning models for predicting carbon monoxide (CO) emissions. The dataset consisted of 90 observations, incorporating engine performance parameters such as NO_X, RPM, HC, and fuel composition, along with the corresponding CO emission values. The data was split into training (80%) and testing (20%) subsets, allowing the models to learn patterns from the training data and evaluate their predictive performance on unseen testing data. The overall experimental workflow is illustrated in Figure 1.

The models' predictive performance was assessed using two primary metrics: the coefficient of determination (R^2) and the Mean Squared Error (MSE). R^2 quantified the proportion of variance in CO emissions explained by the model, reflecting its predictive accuracy, while MSE measured the average squared difference between predicted and actual CO values, highlighting the error magnitude. Together, these metrics provided a comprehensive evaluation of the models' performance.

To further enhance the interpretability of the best-performing model, a feature importance analysis was conducted to identify the relative contributions of input features to CO emission predictions [20]. Additionally, learning curves were generated to examine how model performance varied with the size of the training dataset, offering insights into the bias-variance trade-off for each model [21].

Finally, scatter plots and residual analyses were prepared for each model to visually compare actual and predicted values, and to identify any systematic patterns or biases in the predictions. These methodological steps ensured a robust evaluation process, facilitating reliable insights into the models' applicability for engine performance optimization and emission control strategies.

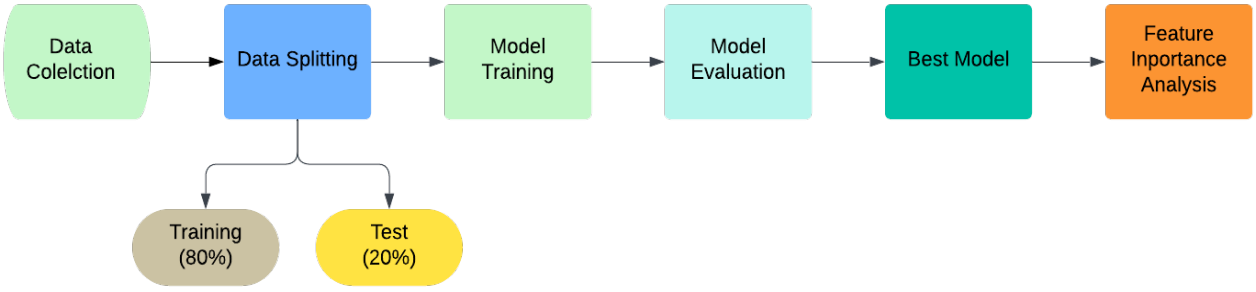

Figure 1. Workflow of the Study for CO Emission Prediction

3. Results and Discussion

In this study, we compared the performance of four different models - Linear Regression, Decision Tree, Random Forest, and Support Vector Regression (SVR) - in predicting CO emissions based on engine performance parameters. The results of our analysis are presented and discussed below.

3.1. Model Comparison and Evaluation

Figure 2 presents a comparative analysis of the performance of all four models. The graph shows each model's Mean Squared Error (MSE) and R2 scores. This analysis shows that the Random Forest model consistently outperformed the other models, achieving the lowest MSE and highest R^2 score. The Linear Regression model, on the other hand, showed the poorest performance, indicating that the relationship between the input features and CO emissions is likely non-linear. Linear Regression's inability to capture complex patterns stems from its strict assumption of linearity, which does not reflect real-world combustion dynamics. Similarly, SVR demonstrated moderate performance due to its sensitivity to parameter tuning and reliance on kernel functions. The limited dataset size may have also constrained SVR's ability to generalize effectively, highlighting the need for robust models like Random Forest in capturing non-linear relationships

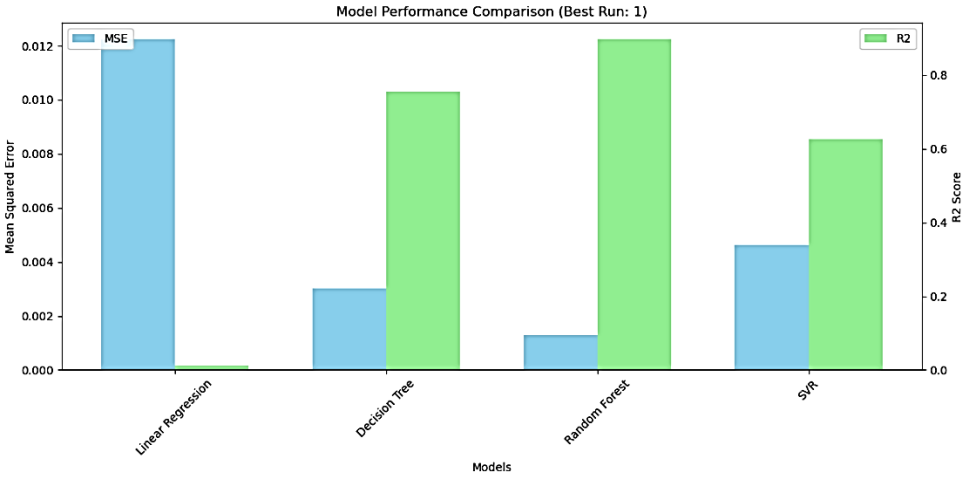

Figure 2. Comparative Performance of Machine Learning Models

To assess the consistency of our models across multiple runs, we analyzed the distribution of R^2 scores, as shown in Figure 3. This box plot reveals that the Random Forest model achieved the highest median R^2 score and demonstrated the least variability across runs. This suggests that Random Forest is not only the most accurate but also the most reliable model for this prediction task.

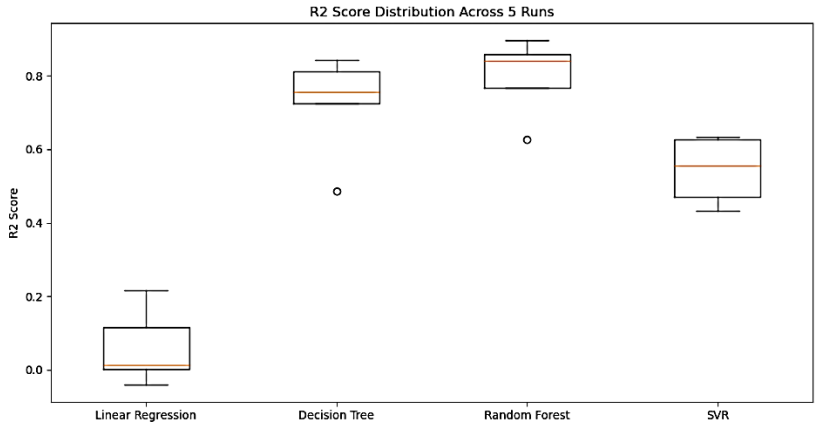

Figure 3. Distribution of R^2 Scores Across Multiple Runs

Figure 4 illustrates the importance of different features in predicting CO emissions, as determined by the Random Forest model. The analysis reveals that NO_X (nitrogen oxides), RPM (engine speed), and HC (hydrocarbons) are the most significant predictors of CO emissions. NO_X and RPM show almost equal importance, with the highest scores, followed closely by HC.

Interestingly, while the fuel composition (Gasoline and Ethanol) does impact CO emissions, their importance is considerably lower compared to the engine operation parameters and other emission components. This suggests that the engine's operating conditions and the formation of other pollutants play a more crucial role in determining CO emissions than the fuel mixture alone. These insights are particularly valuable for engine designers and environmental engineers focusing on emission reduction strategies. The high importance of NO_X and HC in predicting CO emissions indicates a complex interplay between different pollutants in the combustion process. As a top predictor, engine speed (RPM) suggests that optimizing engine operation across different speed ranges could be a key factor in controlling CO emissions.

The relatively lower importance of fuel composition (Gasoline and Ethanol) is noteworthy. While altering fuel mixtures is often considered a strategy for emission control, this analysis suggests that more significant gains might be achieved by focusing on engine operating parameters and technologies that simultaneously reduce NO_X, HC, and CO emissions. This feature importance analysis provides a clear direction for prioritizing efforts in emission control: focusing on technologies and strategies that address NO_X and HC emissions while optimizing engine speed could potentially yield the most significant reductions in CO emissions. Additionally, it highlights the importance of a holistic approach to emission control, considering the interdependencies between different pollutants and engine operating conditions.

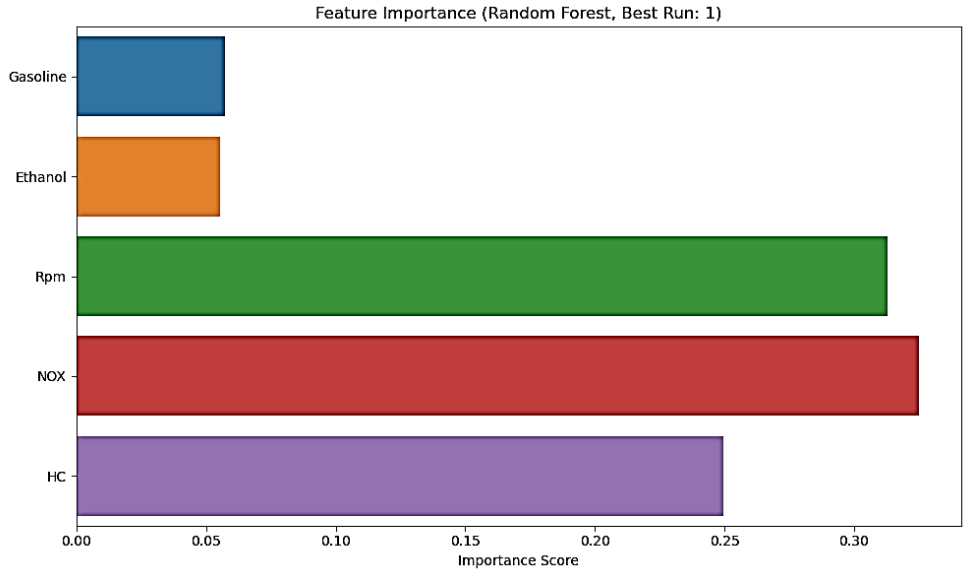

Figure 4. Feature Importance in CO Prediction

The correlation heatmap presented in Figure 5 reveals several significant relationships between input features, providing crucial insights into engine behavior and emission patterns. As expected, Gasoline and Ethanol show a perfect negative correlation (-1.0), reflecting their complementary nature in the fuel mixture. Engine speed (RPM) demonstrates strong correlations with both NO_X (0.7) and HC (-0.9), indicating that higher speeds tend to increase NO_X production while reducing HC emissions, likely due to changes in combustion conditions. The moderate negative correlation (-0.39) between NO_X and HC emissions highlights the typical trade-off in emission control strategies. Interestingly, Gasoline content correlates positively with both NO_X (0.59) and HC (0.32) emissions, while Ethanol shows inverse correlations of equal magnitude, suggesting that increasing ethanol content might help reduce these emissions. The absence of a significant correlation between RPM and fuel composition implies that engine speed-based optimization strategies could be effective across various fuel mixtures. These relationships underscore the complex relationship between engine operations, fuel composition, and emissions, emphasizing the need for a holistic approach in developing effective CO emission prediction models and reduction strategies.

3.2. Best Model Analysis

Figure 6 presents a comparative analysis of actual vs predicted CO values for four models, revealing significant variations in their predictive capabilities. The Random Forest model emerges as the best performer with an impressive R^2 score of 0.8965, demonstrating a strong alignment between predicted and actual values across the CO emission range. In contrast, the Linear Regression model (R^2: 0.0111) shows poor performance, indicating the highly non-linear nature of the CO emission prediction problem. The Decision Tree model (R^2: 0.7542) and SVR (R^2: 0.6236) fall between these extremes, with the

Decision Tree showing better consistency than SVR, particularly at lower CO values. Notably, even the top-performing Random Forest model exhibits slight tendencies to underpredict at higher CO values and overpredict at lower ones, suggesting potential areas for further refinement. These results underscore the complexity of CO emission processes and the superiority of ensemble methods like Random Forest in capturing the complex, non-linear relationships between engine parameters and emissions. The significant performance gap observed across models highlights the importance of selecting appropriate machine learning techniques for accurate CO emission predictions in engine performance analysis.

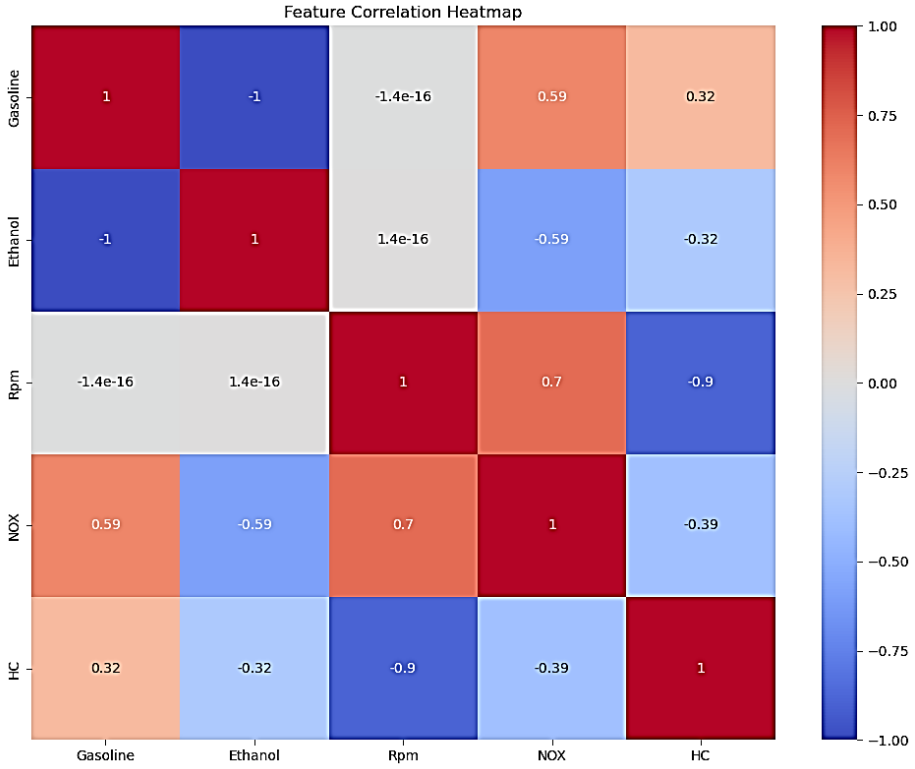

Figure 5. Correlation Analysis of Input Features

The residual plot for the best performing Random Forest model (Figure 7) displays a random scatter of points around the zero line, with residuals primarily falling within a -0.06 to 0.06 range. This pattern indicates that the model performs consistently across different CO emission levels without significant systematic bias. The even distribution of residuals above and below the zero line suggests that the model has captured most of the underlying patterns in the data. However, a slightly wider spread of residuals in the mid-range predictions (1.30 to 1.40) and the presence of a few outliers' hint at some complexity in CO emission behavior that the model doesn't fully capture. Notably, the residuals appear more concentrated around zero for lower and higher predicted values, potentially indicating better model performance at these extremes. These observations confirm the Random Forest model's strong overall predictive capability while highlighting areas for potential improvement, particularly in handling mid-range predictions and addressing outlier cases. The residual analysis thus supports the model's reliability while pointing to opportunities for further refinement in CO emission prediction accuracy.

Figure 8 presents the learning curve for the best performing Random Forest model. The graph shows a converging trend between the training and cross-validation scores as the number of training examples increases. Initially, there's a significant gap between the two scores, with the training score starting high and the cross-validation score starting low. As more data is introduced, this gap narrows considerably, stabilizing both scores at higher levels. The training score remains consistently high, slightly decreasing as more data is added, while the cross-validation score shows a steep increase before leveling off. This pattern indicates that the model has good generalization ability, effectively learning from the data without overfitting. The convergence of scores suggests that the model has reached a point of reducing returns in terms of performance gain from additional data. However, the slight upward trend in the cross-validation score at the highest number of training examples hints that there might still be small potential for improvement with even more data. This learning curve demonstrates that the Random Forest model is well-tuned for the current dataset, balancing complexity with generalization, and is likely to perform consistently on unseen data within the same domain.

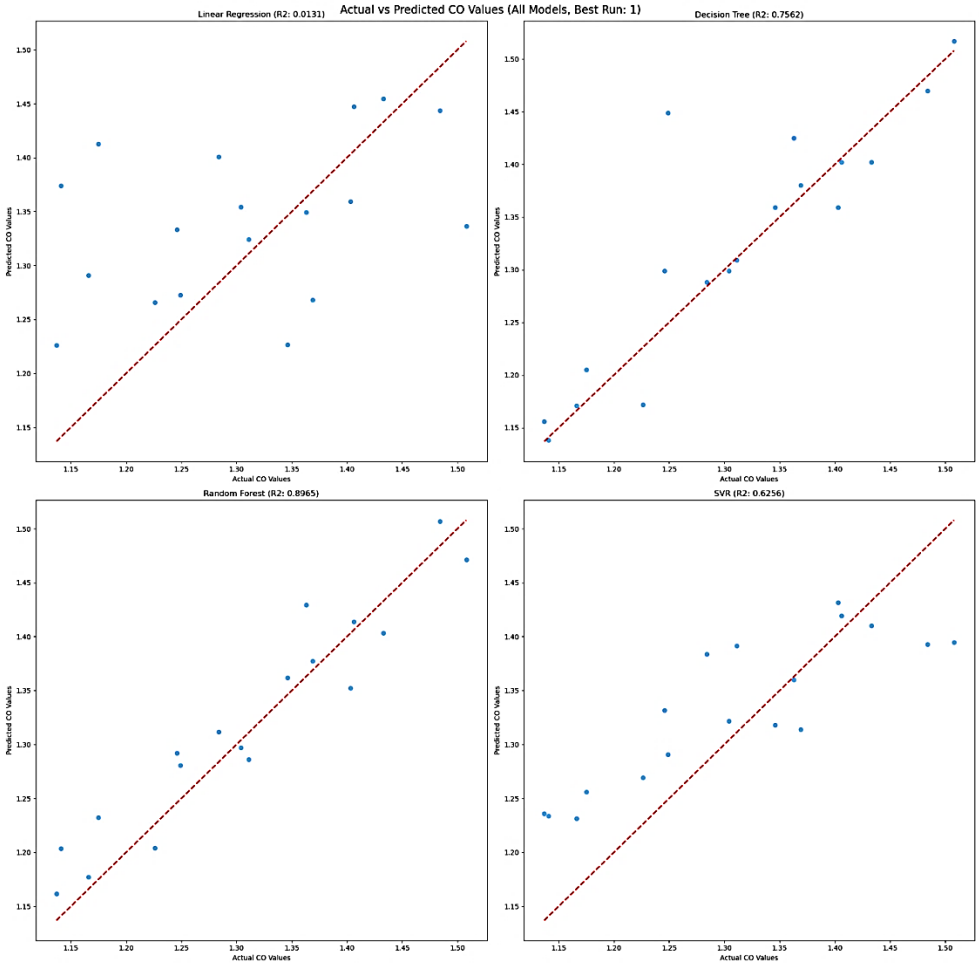

Figure 6. Actual vs Predicted CO Values for the Best Performing Model

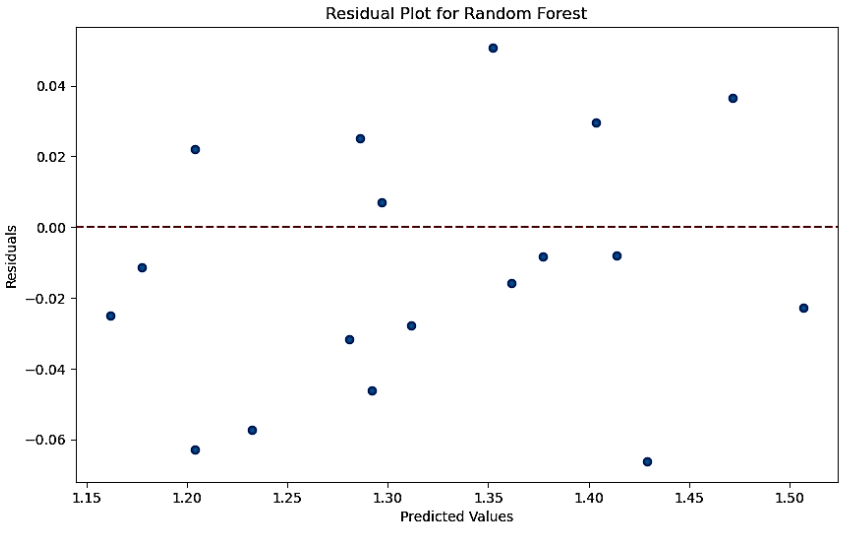

Figure 7. Residual Analysis of the Best Performing Model

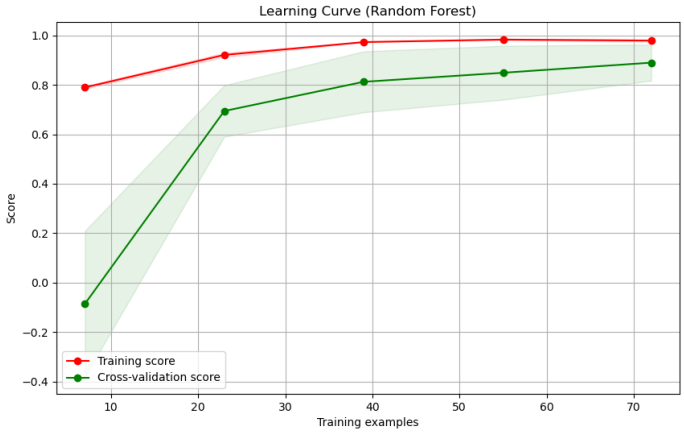

Figure 8. Learning Curve Analysis of the Best Performing Model

Finally, Figure 9 compares actual and predicted CO values for the best performing Random Forest model (R^2: 0.8965). This analysis reveals that our model captures the overall trends and significant fluctuations in CO emissions remarkably well across the sample range. The predicted values closely follow the actual values' pattern, accurately reflecting gradual changes and sharp peaks in CO emissions. The model demonstrates a strong ability to capture the trends and fluctuations in the actual CO values across the sample range. The Random Forest model demonstrates strong predictive capability, effectively tracking CO emissions' complex, non-linear behavior over the sample range. This performance underscores the model's potential for reliable CO emission forecasting in engine performance analysis.

The Random Forest model demonstrated superior performance, achieving an R^2 score of 0.8965, reflecting a strong alignment between actual and predicted CO values. This result underscores the model's ability to capture the complex, non-linear relationships inherent in the dataset. However, its applicability beyond the specific dataset warrants further investigation. While the dataset includes a range of engine speeds and fuel compositions, it is limited to a single engine type and controls laboratory conditions. Real-world scenarios introduce additional variables that may significantly influence combustion processes and emissions, such as variations in temperature, humidity, and altitude; long-term engine wear and maintenance inconsistencies; and diverse driving patterns, including stop-and-go traffic or highway cruising. These factors, absent from the current dataset, highlight the need for broader validation.

Future research should address these limitations by validating the Random Forest model on datasets encompassing diverse engine configurations, such as turbocharged and diesel engines, and incorporating external variables like environmental conditions. Such efforts would enhance the model's robustness across a wider range of operating scenarios. Additionally, while the model showed excellent performance, its tendency to underpredict higher CO emission values suggest opportunities for refinement. Incorporating advanced ensemble techniques, such as Gradient Boosting or hybrid models, could improve generalization across diverse datasets. Furthermore, integrating time-series data and transient operating conditions (e.g., rapid acceleration or deceleration) could enhance the model's adaptability to real-world applications. By addressing these gaps, the Random Forest model could become a more versatile tool for real-time CO emission monitoring and control in the automotive industry.

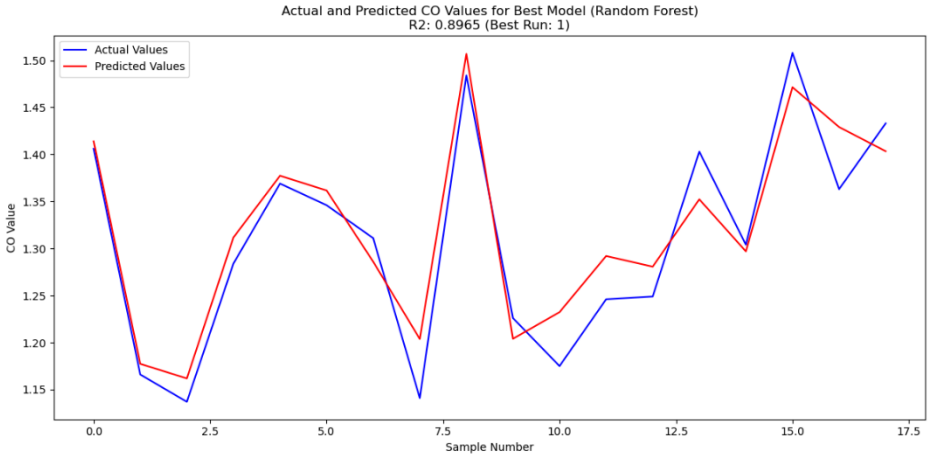

Figure 9. Comparison of Actual and Predicted CO Values

Our analysis demonstrates that machine learning models, particularly Random Forest, can effectively predict CO emissions based on engine performance parameters. The Random Forest model achieved the highest R^2 score of 0.8965, significantly outperforming other models such as Linear Regression, Decision Tree, and SVR. This superior performance can be attributed to Random Forest's ability to capture complex, non-linear relationships and its robustness to outliers. The model shows strong predictive capability across a range of CO emission values, as evidenced by the residual analysis and comparison. These findings have significant implications for engine design and emission control strategies. The feature importance analysis revealed that NO_X, RPM, and HC are the most important predictors of CO emissions, providing valuable insights for targeted emission reduction efforts. Manufacturers can optimize engine designs by accurately predicting CO emissions based on engine parameters to minimize emissions without compromising performance. For instance, they can focus on optimizing engine speed ranges and developing technologies that simultaneously address NO_X, HC, and CO emissions. Furthermore, the model's ability to capture trends in CO emissions can aid in real-time monitoring and control systems. This could lead to more adaptive and efficient emission control strategies, potentially improving overall engine performance while meeting stringent environmental regulations.

3.3. Analysis of Model Performance

The performance of the four machine learning models (Linear Regression, Decision Tree, Random Forest, and Support Vector Regression (SVR)) revealed distinct patterns in their ability to predict carbon monoxide (CO) emissions. Linear Regression and SVR exhibited relatively lower performance than Random Forest and Decision Tree, as shown in Figure 2. Below, we discuss the potential reasons for this underperformance.

Linear Regression, a baseline model, assumes a strict linear relationship between the independent variables (engine parameters) and the dependent variable (CO emissions). However, the combustion process in engines is inherently complex and involves non-linear interactions among various parameters such as NO_X, RPM, and HC. The model's inability to capture these non-linear relationships results in poor predictive performance, with an R^2 score of 0.0111. Additionally, Linear Regression lacks the flexibility to handle feature interactions, making it unsuitable for datasets where input variables exhibit strong dependencies, as observed in the correlation analysis in Figure 5.

Support Vector Regression demonstrated moderate performance (R^2: 0.6236) but underperformed compared to the ensemble-based Random Forest model. While SVR can handle non-linear relationships using kernel functions, its effectiveness depends on appropriate parameter tuning. In this study, an RBF (Radial Basis Function) kernel was used, which, while generally effective, may not have fully captured the complexity of CO emission patterns due to the dataset's limited size and diversity. Furthermore, SVR struggles with outliers, as the epsilon-insensitive loss function can inadvertently exclude essential data points, reducing predictive accuracy in highly variable scenarios.

In contrast, Random Forest consistently outperformed Linear Regression and SVR, achieving an R^2 score of 0.8965. This model's superior performance can be attributed to its ability to capture non-linear relationships, handle complex feature interactions, and remain robust to noise and outliers. By leveraging multiple decision trees and averaging their predictions, Random Forest effectively generalizes across a wide range of input conditions.

The underperformance of Linear Regression and SVR underscores the importance of selecting models that align with the underlying data characteristics. Models like Random Forest, which are inherently flexible and robust, are better suited for CO emissions' complex, non-linear nature. These findings highlight the necessity of comparative analyses to identify the most suitable algorithms for specific applications.

4. Conclusion

This study has demonstrated the efficacy of machine learning techniques, notably the Random Forest algorithm, in predicting carbon monoxide (CO) emissions based on engine performance parameters. Among the models evaluated (Linear Regression, Decision Tree, Support Vector Regression (SVR), and Random Forest), the ensemble-based Random Forest model achieved the highest predictive accuracy, with an R^2 score of 0.8965. Feature importance analysis revealed that NO_X, RPM, and HC levels were the most significant predictors of CO emissions, providing valuable insights into targeted emission reduction strategies. The learning curve analysis highlighted the generalization capabilities of the models and identified areas for further refinement.

The findings have significant practical applications in the automotive industry. The machine learning models developed in this study can be integrated into the engine design process to optimize configurations and reduce emissions without compromising performance. Real-time emission monitoring systems based on these models could dynamically adjust engine parameters to minimize CO emissions under varying conditions, promoting more environmentally friendly driving practices. Additionally, these systems could guide drivers with feedback on optimal acceleration patterns and gear shifts, supporting sustainability goals in the transportation sector.

This study has certain limitations that should be acknowledged. The dataset used was specific to certain engine types, which may restrict the generalizability of the findings across other engine designs or operating conditions. Future work should

address these limitations by including diverse engine configurations, external factors such as environmental conditions, and real-world driving scenarios to enhance the robustness and applicability of the models. The models also exhibited limitations in predicting extreme values, suggesting opportunities for improvement in handling outliers. Furthermore, real-world factors such as environmental conditions, fuel quality variations, and long-term engine wear were not considered in this study, potentially limiting the applicability of the models in broader contexts.

Future studies should incorporate data from diverse engine types (e.g., turbocharged and diesel) and environmental conditions (e.g., temperature, humidity, altitude) to improve model generalizability. Developing advanced ensemble methods or hybrid models could enhance prediction accuracy, particularly for extreme values. Integrating real-time data and time-series analysis techniques could enable the creation of adaptive predictive models capable of responding dynamically to changing conditions. Expanding the study to include more diverse scenarios and pollutants could further improve the utility and relevance of the models in supporting cleaner and more efficient engine technologies.

This research represents a significant step forward in applying machine learning to emission prediction in automotive engines. By providing accurate, data-driven insights, the models developed here contribute to more informed decision-making in engine design and emission control. As the automotive industry faces the dual challenges of enhancing performance while reducing emissions, the approach outlined in this study offers a promising pathway toward achieving both objectives simultaneously.

References

[1] World Health Organization, “Air pollution,” WHO, 2021. [Online]. Available: https://www.who.int/health-topics/air-pollution

[2] J. A. Raub, M. Mathieu-Nolf, N. B. Hampson, and S. R. Thom, “Carbon monoxide poisoning—a public health perspective,” Toxicology, vol. 145, no. 1, pp. 1–14, 2000.

[3] Environmental Protection Agency, “Ground-level Ozone Basics,” EPA, 2021. [Online]. Available: https://www.epa.gov/ground-level-ozone-pollution/ground-level-ozone-basics

[4] W. J. Requia, M. Mohamed, C. D. Higgins, A. Arain, and M. Ferguson, “How clean are electric vehicles? Evidence-based review of the effects of electric mobility on air pollutants, greenhouse gas emissions and human health,” Atmospheric Environment, vol. 185, pp. 64–77, 2018.

[5] T. Johnson and A. Joshi, “Review of vehicle engine efficiency and emissions,” SAE Int. J. Engines, vol. 11, no. 6, 2018.

[6] J. Gao, H. Chen, G. Tian, C. Ma, and F. Zhu, “An analysis of energy flow in a turbocharged diesel engine of a heavy truck and potential for recovery of exhaust heat,” Energy Convers. Manage., vol. 185, pp. 1040–1051, 2019.

[7] R. D. Reitz et al., “IJER editorial: The future of the internal combustion engine,” Int. J. Engine Res., vol. 21, no. 1, pp. 3–10, 2020.

[8] V. M. Janakiraman, X. Nguyen, and D. Assanis, “Stochastic gradient based extreme learning machines for stable online learning of advanced combustion engines,” Neurocomputing, vol. 177, pp. 304–316, 2016.

[9] J. D. Wu and J. C. Liu, “Development of a predictive system for car fuel consumption using an artificial neural network,” Expert Syst. Appl., vol. 38, no. 5, pp. 4967–4971, 2014.

[10] S. Roy, R. Banerjee, and P. K. Bose, “Performance and exhaust emissions prediction of a CRDI assisted single cylinder diesel engine coupled with EGR using artificial neural network,” Appl. Energy, vol. 119, pp. 330–340, 2014.

[11] T. Wang, M. Jerrett, P. Sinsheimer, and Y. Zhu, “Estimating PM2.5 in Southern California using remote sensing data and light use efficiency modeling: Implications for policy,” Environ. Sci. Technol., vol. 50, no. 9, pp. 4724–4733, 2016.

[12] S. C. De Lima Nogueira et al., “Prediction of the NOx and CO2 emissions from an experimental dual fuel engine using optimized random forest combined with feature engineering,” Energy, vol. 280, 128066, 2023.

[13] Q. Shen et al., “Prediction Model for Transient NOx Emission of Diesel Engine Based on CNN-LSTM Network,” Energies, vol. 16, no. 14, 5347, 2023.

[14] J. B. Heywood, Internal combustion engine fundamentals, New York, NY, USA: McGraw-Hill Education, 2018.

[15] F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011.

[16] D. C. Montgomery, E. A. Peck, and G. G. Vining, Introduction to linear regression analysis, Hoboken, NJ, USA: John Wiley & Sons, Inc., 2012.

[17] Y. Liu, Y. Wang, and J. Zhang, “A novel hybrid model based on data preprocessing and optimized decision tree for diesel engine NOx emission prediction under transient conditions,” Energy, vol. 239, 122207, 2022.

[18] L. Breiman, “Random forests,” Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.

[19] A. J. Smola and B. Schölkopf, “A tutorial on support vector regression,” Stat. Comput., vol. 14, no. 3, pp. 199–222, 2004.

[20] C. Strobl, A. L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, “Conditional variable importance for random forests,” BMC Bioinformatics, vol. 9, no. 307, 2008.
[21] R. L. Figueroa, Q. Zeng-Treitler, S. Kandula, and L. H. Ngo, “Predicting sample size required for classification performance,” BMC Med. Inform. Decis. Mak., vol. 12, no. 8, 2012.

Authors Contributions
Authors are solely responsible for the design, execution, analysis, and writing of this study.

Funding Statement
Authors have not received any financial support for the research, authorship, or publication of this study.

Conflict of Interest Declaration
Authors declares that there is no conflict of interest regarding the publication of this paper.

Ethical Approval and Informed Consent
It is hereby declared that all scientific and ethical standards were adhered to during the preparation of this study. All sources utilized in the study are appropriately cited in the bibliography.

Availability of Data and Materials
All data and materials related to this study are available from the corresponding authors upon reasonable request.

Plagiarism Statement
This article has been screened for plagiarism using iThenticate™ software and has been confirmed to be original.