IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

GLOBAL CO2 EMISSION PREDICTION USING MULTI-MODEL MACHINE LEARNING WITH INTERACTIVE VISUALIZATION

[1]Athiqa Zulequa, [2]M. Sai Vani, [3]P. Shirisha, [4]Jyotsna Tarigoppula, [5]Siva Yenikepalli, [6]Dr. K Raghu

[1,2,3] Student, [4,5] Assistant Professor, [6] Associate Professor

[1-6]Department of Computer Science and Engineering

[1-6]Geethanjali College of Engineering & Technology, Hyderabad, India

Abstract: The rapid increase in atmospheric carbon dioxide (CO2) emissions has emerged as a critical environmental challenge, significantly contributing to global climate change. Accurate prediction of CO2 emissions is essential for understanding emission patterns and supporting effective policy-making and sustainability planning. This work presents a machine learning-based system for predicting CO2 emissions using a globally representative dataset obtained from the Our World in Data (OWID) repository. The dataset includes key features such as emissions from coal, oil, gas, and cement production, along with population and gross domestic product (GDP) across multiple countries and years. The proposed system follows a structured pipeline consisting of data preprocessing, feature engineering, train-test splitting, feature scaling, model training, and performance evaluation. Multiple regression models are implemented and compared, including Linear Regression, Multi-Layer Perceptron (MLP) Regressor, XGBoost, LightGBM, and CatBoost. The models are evaluated using standard metrics such as Root Mean Squared Error (RMSE) and coefficient of determination (R^2 score). Experimental results indicate that the MLP Regressor achieves the best performance, with an R^2 score of approximately 0.994 and the lowest prediction error, demonstrating its ability to capture complex non-linear relationships in the data. In addition, a Streamlit-based interactive dashboard is developed to visualize model performance, analyze predictions, and enable real-time CO2 emission forecasting through user-defined inputs. The proposed system provides a scalable, accurate, and user-friendly solution for global CO2 emission prediction.

***Index Terms*—CO2 Emission Prediction, Machine Learning, Regression Models, XGBoost, LightGBM, CatBoost, MLP Regressor, Streamlit Dashboard**

I. INTRODUCTION

The increasing concentration of atmospheric carbon dioxide (CO2) has become one of the primary drivers of global climate change, posing serious environmental, economic, and societal challenges. Rising emission levels due to industrialization, energy consumption, and population growth have intensified the need for accurate monitoring and prediction systems. Reliable prediction of CO2 emissions is essential for effective climate policy formulation, sustainable development planning, and environmental risk assessment.

Traditional approaches for emission prediction, such as statistical and econometric models, often rely on assumptions of linearity and stationarity. However, real-world emission patterns are influenced by complex and non-linear interactions between multiple factors, including fossil fuel consumption, economic growth, and demographic changes. These limitations reduce the effectiveness of conventional models when applied to large-scale and diverse global datasets.

With the advancement of data-driven technologies, machine learning has emerged as a powerful tool for environmental analysis and prediction. Machine learning models can automatically learn hidden patterns from large datasets and capture complex relationships between input variables and output predictions. In particular, regression-based models and ensemble techniques have shown strong performance in modeling environmental data.

This paper presents a multi-model machine learning framework for predicting global CO2 emissions using a comprehensive dataset obtained from the Our World in Data (OWID) repository. The proposed system integrates multiple regression algorithms, including Linear Regression, Multi-Layer Perceptron (MLP) Regressor, XGBoost, LightGBM, and CatBoost, to evaluate and compare their predictive performance. In addition, an interactive dashboard is developed using Streamlit to provide visualization of results, performance comparison, and real-time prediction capabilities.

The main contributions of this work are as follows:

- Development of a scalable machine learning pipeline for global CO2 emission prediction
- Comparative analysis of multiple regression models on a unified dataset

- Integration of an interactive dashboard for real-time prediction and visualization
- Identification of key factors influencing CO2 emissions through feature analysis

II. RELATED WORK

Accurate prediction of carbon dioxide (CO2) emissions has gained significant attention due to its critical role in climate change mitigation and environmental planning. Various studies have explored statistical, econometric, and machine learning-based approaches to model and predict emission trends.

Traditional approaches such as linear regression and econometric models have been widely used for emission prediction. However, these methods often rely on assumptions of linear relationships and fail to capture the complex and non-linear interactions among influencing factors. As a result, their predictive performance is limited when applied to large-scale and heterogeneous datasets.

With the advancement of machine learning techniques, several researchers have proposed data-driven models for CO2 emission prediction. Zhu et al. (2022) utilized ensemble learning methods such as Random Forest and Gradient Boosting to predict CO2 emissions. Their study demonstrated improved accuracy compared to traditional models; however, it was limited to region-specific datasets and lacked global applicability.

Kang et al. (2023) applied machine learning techniques including XGBoost and Support Vector Regression for country-level CO2 emission prediction. Although their model achieved high prediction accuracy, it did not support real-time prediction or interactive visualization, limiting its practical usability.

Recent studies have also explored deep learning and hybrid approaches. Ajala et al. (2025) conducted a comparative analysis of machine learning and deep learning models, including CNN-LSTM architectures. While these models showed strong performance, they required high computational resources and large-scale datasets, making them less efficient for real-time applications.

Similarly, Zhang et al. (2025) proposed ensemble learning-based approaches using stacking techniques to improve prediction accuracy. Although effective, these models were complex and often tailored to specific datasets, reducing their generalization capability.

Despite these advancements, existing approaches suffer from several limitations, including lack of scalability, absence of real-time prediction systems, and limited integration of visualization tools for user interaction.

Table 1 presents a comparative analysis of existing CO2 emission prediction methods along with their methodologies and limitations.

Table 1. Comparison of Existing CO2 Emission Prediction Methods

Year	Author(s)	Methodology	Limitations
2022	Zhu et al.	Random Forest, Gradient Boosting	Region-specific, lacks global generalization
2023	Kang et al.	XGBoost, Support Vector Regression	No real-time prediction system
2025	Ajala et al.	CNN-LSTM, ML Models	High computational cost
2025	Zhang et al.	Ensemble Learning (Stacking)	Complex model, limited scalability
2026	Proposed Work	Multi-model ML (LR, MLP, XGBoost, LightGBM, CatBoost)	Improved accuracy with real-time visualization

From Table 1, it is evident that most existing approaches lack scalability, real-time prediction capability, or generalization across diverse datasets. To overcome these limitations, the proposed work integrates multiple machine learning models with an interactive dashboard, enabling accurate, scalable, and real-time CO2 emission prediction along with intuitive visualization.

III. PROPOSED SYSTEM

The proposed system is a scalable and modular machine learning framework designed to predict global CO2 emissions using historical environmental and socio-economic data. The system integrates data preprocessing, feature engineering, multi-model training, evaluation, and visualization into a unified pipeline. The primary objective is to develop an accurate, efficient, and user-friendly prediction system capable of supporting environmental analysis and decision-making.

The overall workflow of the proposed system is illustrated in Fig. 1. The system is divided into multiple stages, each responsible for a specific task in the prediction pipeline.

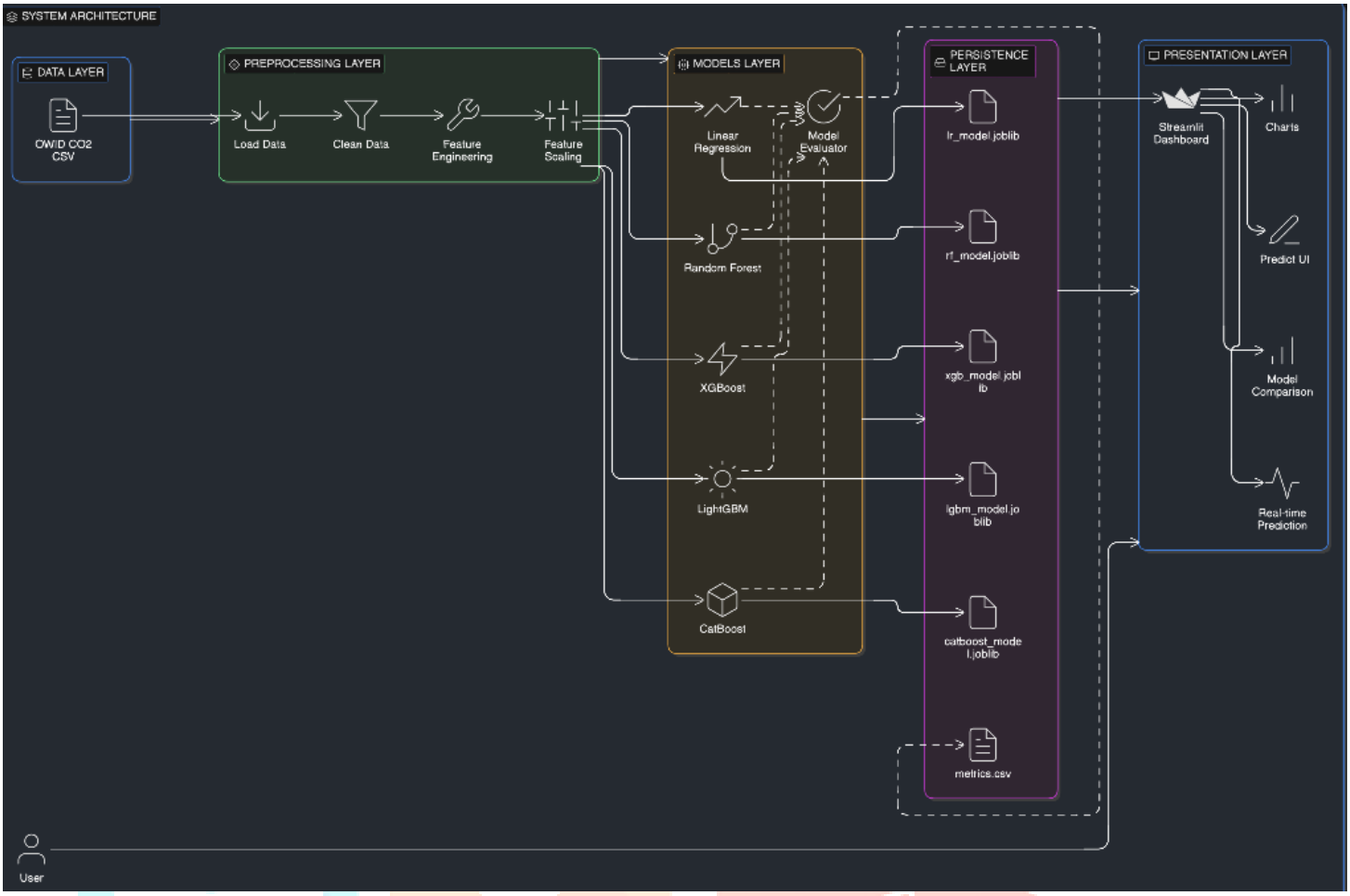

Fig. 1. System Architecture of CO2 Emission Prediction Framework

1. Data Collection and Preprocessing

The dataset used in this study is obtained from the Our World in Data (OWID) repository, which provides comprehensive global CO2 emission statistics along with relevant socio-economic features such as population, gross domestic product (GDP), and emissions from coal, oil, gas, and cement production.

Raw data often contains missing values, inconsistencies, and noise. Therefore, preprocessing is performed to ensure data quality and reliability before model training. The preprocessing steps include:

- Removal of records with missing or invalid CO2 emission values
- Handling of null and inconsistent entries
- Conversion of data types for numerical analysis
- Selection of relevant features for modeling.

These steps improve data consistency and enhance model performance.

2. Feature Engineering

Feature engineering is a critical step in enhancing the predictive capability and generalization of machine learning models. It involves selecting, transforming, and preparing input variables that effectively capture the underlying patterns influencing CO2 emissions.

The features used in this study include year, population, gross domestic product (GDP), and emission-related variables such as coal CO2, oil CO2, gas CO2, and cement CO2. These variables represent the primary drivers of carbon emissions.

Feature scaling is applied using standardization to ensure that all input variables are on a similar scale, preventing dominance of features with larger numerical ranges. This step is particularly important for models such as the Multi-Layer Perceptron (MLP), which are sensitive to input data distribution.

Feature importance is also analyzed using model-based techniques. The analysis indicates that emission-related variables such as coal, oil, and gas CO2 contribute most significantly to prediction accuracy, while GDP and population provide supporting contextual information.

3. Model Training

The model training phase involves implementing and comparing multiple regression algorithms to accurately predict CO2 emissions. The dataset is divided into training and testing sets using an 80:20 ratio to evaluate model generalization and ensure reliable performance on unseen data.

The following machine learning models are utilized in this study:

• Linear Regression: Serves as a baseline model by assuming a linear relationship between input features and CO2 emissions.

• Multi-Layer Perceptron (MLP) Regressor: A neural network-based model capable of capturing complex non-linear relationships within the data.
• XGBoost: An advanced gradient boosting algorithm that enhances prediction accuracy through sequential learning and error minimization.
• LightGBM: A fast and efficient gradient boosting framework optimized for handling large-scale datasets with improved training speed.
• CatBoost: A gradient boosting algorithm designed to handle categorical features effectively while reducing overfitting.
To optimize model performance, hyperparameters such as learning rate, number of estimators, depth of trees, and network architecture are carefully tuned. The trained models are then evaluated using appropriate performance metrics to determine their effectiveness in predicting CO2 emissions.

4. Model Evaluation

The performance of the trained models is evaluated using standard regression metrics such as Root Mean Squared Error (RMSE) and the coefficient of determination (R^2 score). These metrics provide a quantitative assessment of prediction accuracy and model reliability.

RMSE is defined as:

$$RMSE = \sqrt{[(1/n)\ \Sigma(yi - \hat{y}i)^2]}$$

where yi represents the actual CO2 emission and ŷi represents the predicted value. RMSE measures the average magnitude of prediction error, and lower values indicate better model performance.

The R^2 score is defined as:

$$R^2 = 1 - [\Sigma(yi - \hat{y}i)^2 / \Sigma(yi - \bar{y})^2]$$

where ȳ represents the mean of the actual values. The R^2 score indicates how well the model explains the variance in the data, with values closer to 1 representing better performance.
These evaluation metrics collectively provide a comprehensive understanding of model accuracy, consistency, and generalization capability.

5. Visualization and Dashboard Integration

The system includes a Streamlit-based interactive dashboard, as illustrated in Fig. 2, which enables real-time CO2 emission prediction and visualization.

Fig. 2. Interactive CO2 Emission Prediction Dashboard

The dashboard allows users to input parameters such as population, GDP, and emission-related variables to generate predictions dynamically. It also provides visualization tools, including actual versus predicted graphs, to facilitate better understanding of model behavior.
Key functionalities of the dashboard include:

• Real-time CO2 emission prediction
• Model selection and comparison
• Visualization of prediction results

The integration of machine learning models with an interactive dashboard enhances usability and makes the system accessible to both technical and non-technical users.
Overall, the proposed system provides a comprehensive, scalable, and user-friendly solution for accurate CO2 emission prediction.

IV. RESULTS AND DISCUSSION

This section presents the experimental evaluation of the proposed multi-model machine learning framework for CO2 emission prediction. The performance of the models is analyzed using standard regression metrics, along with a detailed visual interpretation of prediction results.
All visualization outputs, including prediction plots, error distribution, and feature importance analysis, are derived from the Linear Regression model to maintain consistency in analysis and interpretation.

A. Model Performance Comparison

Table 2 presents the performance comparison of the evaluated machine learning models based on Root Mean Squared Error (RMSE) and
the coefficient of determination (R^2 score). These metrics are used to assess the accuracy and effectiveness of each model in predicting CO2 emissions.

Table 2. Model Performance Comparison

Model	RMSE	R^2 Score
MLP Regressor	126.46	0.9943
XGBoost	192.66	0.9867
LightGBM	194.76	0.9864
CatBoost	207.95	0.9845
Linear Regression	432.63	0.9331

The results presented in Table 2 indicate that the MLP Regressor achieves the best performance, with the lowest RMSE and the highest R^2 score. This demonstrates its strong capability to model complex non-linear relationships in CO2 emission data.
The gradient boosting models, including XGBoost, LightGBM, and CatBoost, also demonstrate strong performance, with R^2 values exceeding 0.98. This indicates high predictive accuracy; however, they slightly underperform compared to the MLP model.
In contrast, Linear Regression exhibits comparatively lower performance due to its assumption of linear relationships, which limits its ability to capture complex emission patterns present in the dataset.

B. Prediction Analysis

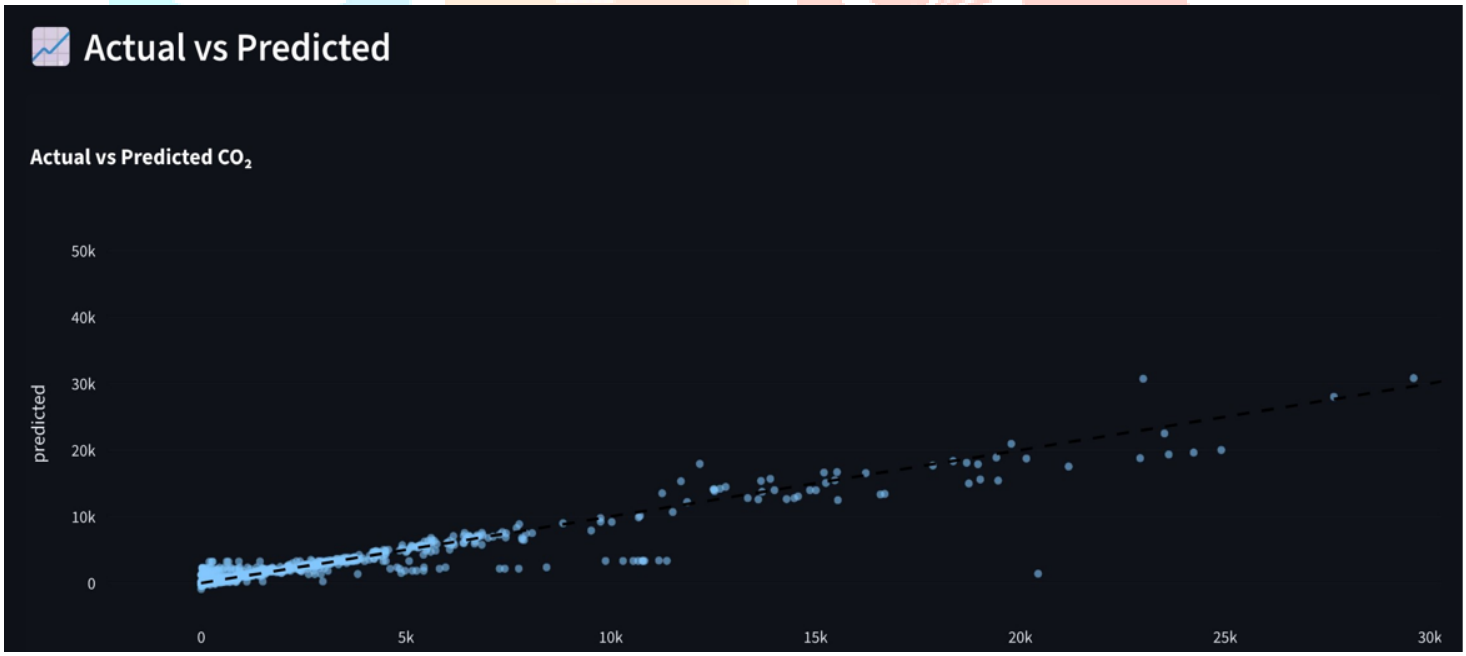

Fig. 3. Actual vs Predicted CO2 Emission Values

Fig. 3 illustrates the relationship between actual and predicted CO2 emission values. The data points are closely aligned along the diagonal line, indicating a high level of prediction accuracy.
The MLP model exhibits the most consistent alignment across the entire range of values, confirming its superior generalization capability. In contrast, Linear Regression shows noticeable deviations, particularly for higher emission values, indicating its limited ability to capture complex non-linear patterns in the data.

C. Error Distribution

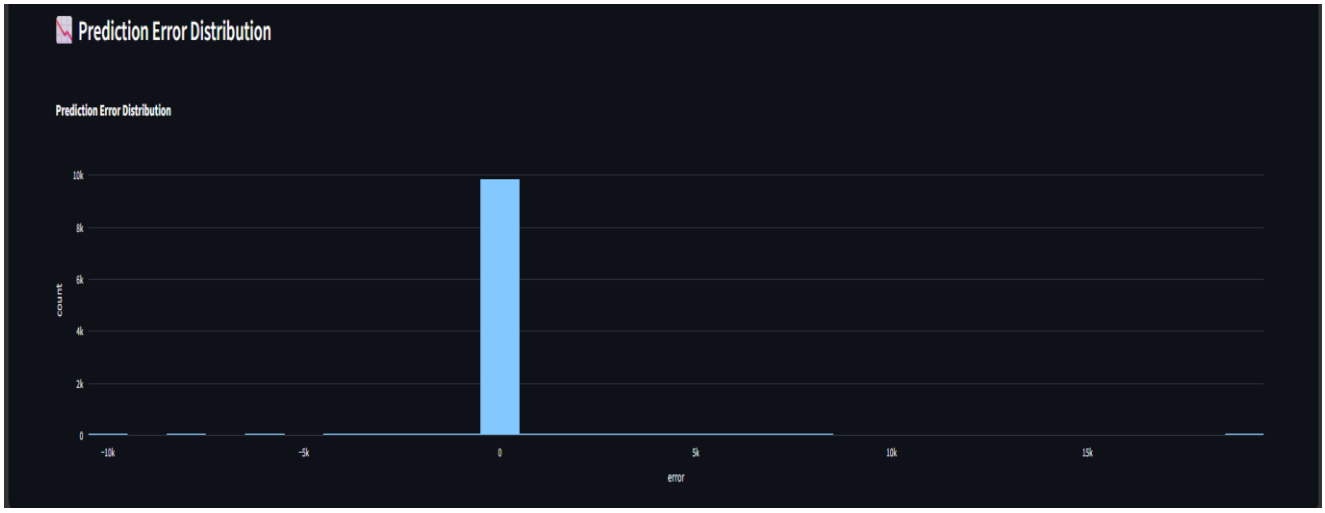

Fig. 4. Prediction Error Distribution

The error distribution shown in Fig. 4 indicates that most prediction errors are concentrated around zero, suggesting that the model produces stable and unbiased predictions.

The narrow spread of errors reflects low variance in the model's predictions, while a small number of outliers correspond to extreme emission values present in the dataset.

D. Feature Importance Analysis

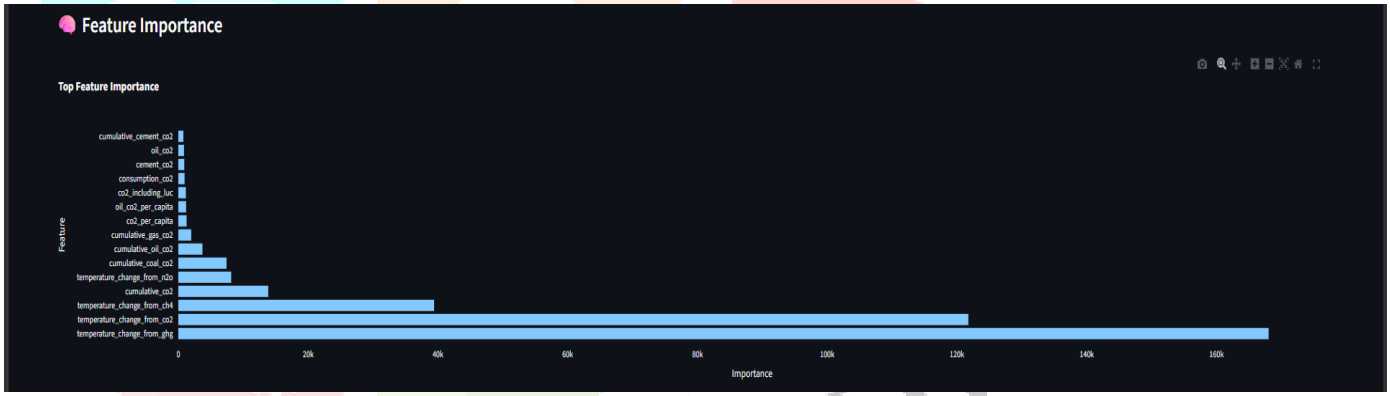

Fig. 5. Feature Importance Derived from Linear Regression Model

Fig. 5 presents the feature importance derived from the Linear Regression model. The results indicate that emission-related variables such

as coal, oil, and gas CO2 have the highest influence on total emissions.

Economic and demographic factors such as GDP and population also contribute to the prediction; however, their impact is relatively lower

compared to direct emission sources. These findings are consistent with real-world trends, where fossil fuel consumption remains the

dominant driver of CO2 emissions.

E. Discussion

The experimental results demonstrate that advanced machine learning models are highly effective for CO2 emission prediction. Among the

evaluated models, the MLP Regressor outperforms others due to its ability to capture complex non-linear relationships in the data.

Boosting algorithms also exhibit strong and stable performance owing to their ensemble learning mechanisms. However, they slightly lag

behind the neural network model in terms of prediction accuracy.

The combined analysis of performance metrics, prediction plots, and error distribution indicates strong predictive performance and consistent

model behavior on the test dataset.

The integration of visualization tools enhances usability, making the system suitable for applications such as environmental monitoring and

analytical studies.

Overall, the proposed framework provides a reliable, scalable, and efficient solution for predicting global CO2 emissions.

V. CONCLUSION AND FUTURE WORK

A. Conclusion

This paper presented a multi-model machine learning framework for predicting global CO2 emissions using environmental and socio-
economic features. The proposed system integrates data preprocessing, feature engineering, model training, evaluation, and visualization
into a unified pipeline.

Experimental results demonstrate that machine learning models are highly effective for CO2 emission prediction. Among the evaluated
models, the Multi-Layer Perceptron (MLP) Regressor achieved the best performance, indicating its strong capability to capture complex
non-linear relationships in emission data. Boosting models also showed competitive performance, while Linear Regression was limited due
to its assumption of linearity.

The analysis further highlights that emission-related features such as coal, oil, and gas CO2 are the most influential factors affecting total
emissions. The integration of an interactive dashboard enhances the practical usability of the system by enabling real-time prediction,
visualization, and model comparison.

Overall, the proposed framework provides a reliable, scalable, and user-friendly solution for CO2 emission prediction, with potential
applications in environmental monitoring and policy decision-making.

B. Future Work

Although the proposed system demonstrates strong performance and high prediction accuracy, several enhancements can be explored to
further improve its capability and real-world applicability:

1. Time-Series Forecasting: Future work can extend the current approach to time-series forecasting, enabling prediction of future emission trends. Advanced models such as Long Short-Term Memory (LSTM) and Transformer-based architectures can be used to capture temporal dependencies.
2. Advanced Hyperparameter Optimization: Techniques such as Grid Search, Random Search, and Optuna can be applied to fine-tune model parameters and further improve performance.
3. Integration of Additional Features: Incorporating variables such as renewable energy usage, industrial activity, and climate factors can provide deeper insights and improve prediction accuracy.
4. Geospatial Visualization: The dashboard can be enhanced with geographical visualizations, such as interactive maps showing emission levels across countries, making the system more intuitive and informative.
5. Sector-wise Emission Analysis: Extending the system to predict emissions across different sectors (e.g., energy, transport, and industry) can support more targeted policy analysis and decision-making.
6. Cloud Deployment and Scalability: Deploying the system on cloud platforms such as AWS or Streamlit Cloud can improve accessibility and enable large-scale usage without requiring local setup.
7. Explainable AI (XAI): Future work can incorporate explainable AI techniques to improve transparency and provide better understanding of model predictions.

REFERENCES

[1] M. Roser, H. Ritchie, and P. Rosado, “CO2 and Greenhouse Gas Emissions Data Explorer,” Our World in Data, 2023. [Online]. Available: https://ourworldindata.org/explorers/co2

[2] P. Friedlingstein et al., “Global Carbon Budget 2024,” Global Carbon Project, 2024. [Online]. Available: https://globalcarbonbudget.org/gcb2024/

[3] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.

[4] G. Ke et al., “LightGBM: A Highly Efficient Gradient Boosting Decision Tree,” in Advances in Neural Information Processing Systems, vol. 30, 2017.

[5] L. Prokhorenkova et al., “CatBoost: Unbiased Boosting with Categorical Features,” in Advances in Neural Information Processing Systems, vol. 31, 2018.

[6] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[7] X. Zhu et al., “Predicting China’s Provincial CO2 Emissions Using Ensemble Machine Learning Models,” Energy Policy, vol. 168, 2022.

[8] J. Kang et al., “Country-Level CO2 Emission Prediction Using Machine Learning Techniques,” Environmental Science and Pollution Research, vol. 30, 2023.

[9] M. W. Jones et al., “National Contributions to Climate Change Due to Historical Emissions,” Scientific Data, vol. 10, 2023.

[10] Our World in Data, “CO2 Dataset,” 2023. [Online]. Available: https://ourworldindata.org/co2-and-greenhouse-gas-emissions.