SSRG International Journal of Civil Engineering
ISSN: 2348-8352/ https://doi.org/10.14445/23488352/IJCE-V12I8P116

Volume 12 Issue 8, 176-183, August 2025

Original Article

Role of Machine Learning to Analyze the Impact of Construction Industry on the Carbon Footprint

Anupama Chadha[1], Anjali Gupta[2]

[1]*Department of Computer Applications, MRIIRS, Haryana, India.*
[2]*Department of Civil Engineering, MRIIRS, Haryana, India.*

[1]*Corresponding Author : anupma.sca@mriu.edu.in*

Received: 10 June 2025 Revised: 15 July 2025 Accepted: 14 August 2025 Published: 29 August 2025

Abstract - *The construction sector is one of the top contributors to the world's carbon footprint, owing to its energy requirements and resource consumption. This study aims to understand the impact of Machine Learning (ML) on estimating and assessing the carbon footprint of construction activities using multi-sector carbon datasets. Key components such as energy usage, emissions at the sector level, GDP, population, and the proportion of renewable energy were used to fine-tune and assess multiple regression-based ML algorithms. Six baseline models were created: Linear Regression, Ridge, Lasso, Support Vector Regression (SVR), Decision Tree, and Random Forest, as well as advanced ensemble methods XGBoost and LightGBM. Additional feature engineering was utilized to develop normalized emission ratios, such as per capita and per GDP. A broad range of evaluation indicators was used, including R^2 score, RMSE, MAE, MAPE, MSLE, median absolute error, and explained variance. The outcome indicated that traditional linear models were more predictable ($R^2 \approx -0.007$, RMSE ≈ 85.97) while tree-based Random Forest models struggled ($R^2 \approx -0.034$, RMSE ≈ 87.12), which means none of the parallel models outperformed the emission variance. XGBoost and LightGBM achieved similar yields; xGBoost earned $R^2 = -0.235$, RMSE = 95.20, illustrating that a model based on complex, high-dimensional environmental data is difficult to construct. In a hypothetical situation where the use of renewable energy sources was increased by 20%, most models still forecasted only slight emission reductions (for example, Random Forest: Change from 152.92 to 152.87 metric tons). SHapley Additive exPlanations (SHAP) explainability pointed to energy demand, industrial CO_2, and proportion of renewables as the main contributors to emissions. Further cluster analysis revealed distinct emission profiles by region, which can inform focused environmental policy. This research analyzes the possibility of applying machine learning to discover structural features in carbon emissions and assesses the impact of renewable energy policies in the construction industry. The study also stresses the role of explainability and feature engineering on environmental simulations.*

Keywords - *Carbon footprint, Construction industry, Environment, Machine Learning, Renewable energy policies.*

1. Introduction

The construction sector has an impact on the environment due to the extensive use of cement, steel, and construction activities, as well as the substantial demand for electricity and transportation. The construction and buildings sector has an even greater impact, accounting for almost 39% of global CO_2 emissions and joining other industries as one of the leading sectors that are in dire need of a global sustainability agenda. To mitigate construction emissions, there is a need to provide focus on Machine Learning (ML) approaches due to their predictive abilities. These methodologies stand in stark contrast to traditional Life Cycle Assessment (LCAs) and Input Output (IO) modeling methodologies, which, although rigorous, lack granularity and forecasting abilities. ML facilitates modeling complex non-linear relationships in construction activity involving emissions such as economic indicators, energy type, and demographic statistics.

This area has recently begun to attract attention [1]. studied urban datasets, focusing on construction CO_2 emissions, and applied ensemble learning techniques. The study concluded that boosting methods achieved higher predictive accuracy compared to linear baselines [2]. focused on carbon 'regional' emission models built using SVR and random forests, and noted energy consumption and intensity of industrial activities as significant factors [3]. augmented SHAP explainers with Machine Learning (ML) models to emissions data, clarifying where policies require transparent information, and enhancing the interpretability of the data. In the same context, [4] forecasted emissions at the building level using XGBoost and emphasized the usefulness of feature selection in optimizing performance.

Regardless of these efforts, the literature regarding comprehensive studies that integrate several machine learning techniques and conduct policy impact evaluations is limited. There is a lack of models that can assess and explain emissions policies and normalize custom emission features. This research fills these gaps by performing a multi-dimensional machine learning analysis on industrial, agricultural, domestic, and transport CO_2 emissions from several regions over a span of years.

The main contributions of the study are as follows:

1. A collection of machine learning algorithms is evaluated based on Linear Regression, Ridge Regression, Lasso Regression, Support Vector Regression (SVR), Decision Tree, Random Forest, XGBoost, and LightGBM. Such algorithms receive vast attention, especially regarding their evaluation through RMSE, R^2, MAPE, MSLE, and Explained Variance Score.
2. Average CO_2 emissions across all regions are predicted, and it is discovered that there is a small change post the intervention. For example, Random Classifier goes from 152.92 to 152.87 metric tons.
3. The analysis of features for predictive models incorporates SHAP analysis for diverse models with the aim of explaining energy consumption. Industrial emissions and the proportion of renewable energy, alongside the level of consumption, have always ranked among the most regarded.
4. Emissions on a per capita basis alongside GDP are engineered to achieve normalization for the population and the economy. It also reveals regional groupings and behavior based on emissions, which K-Means clustering offers insights into intervention based on robust policy.

In the subsequent sections, the paper is structured as follows: In Section 2, a review of related work in carbon emissions with machine learning is presented. Section 3 highlights the discussion around the dataset, methods of feature engineering, and other procedures taken in cleaning the data. Section 4 describes the machine learning models utilized alongside the evaluation metrics, assessment, or measurement standards. Section 5 describes the results of the model benchmarking in conjunction with the SHAP analysis and policy simulation. The last section presents the overall discussion and the future scope of the work.

2. Related Work

The carbon footprint of construction remains an area of concern in environmental research. There is an increasing body of work that aims to use Machine Learning (ML) to analyze, anticipate, and alleviate emissions associated with construction activities. These methodologies have the capacity to model intricate dynamics regarding emissions drivers and environmental impacts, oftentimes surpassing traditional statistical techniques in precision and versatility. The study by [5] focuses on predicting CO_2 emissions during the demolition phase of buildings. They devised an optimal machine learning framework that highlighted the GBM, Decision Tree, and Random Forest algorithms. The authors had access to a dataset that included 186 demolition projects within South Korea. Out of the models tested, GBM proved to be the most accurate with an R2 of 0.984 on validation data, displaying exceptional generalization ability. The analysis also revealed equipment type and floor area as the major emission predictors.

[6] focused on embedded carbon emissions and used machine learning algorithms to create predictive models tied to the building design phase. While specific metrics were not disclosed, the authors noted material selection and spatial parameters as important features in emissions forecasting. Their models were trained on datasets rich in building component attributes, indicating that emissions control is possible in the design stage. [7] Implemented a hybrid machine learning method to create a carbon emission forecasting model for cities in China. The authors have used Random Forest, SVR and XGBoost machine learning models. The data was processed and substituted into the machine learning models. The prediction results show that random forest is better than SVR and XGBoost in terms of accuracy. [8] developed a hybrid model that integrates BIM alongside Machine Learning to estimate the carbon emissions of buildings throughout the construction process. By leveraging real-time data provided by BIM, these environments permitted ML models to forecast emissions relative to both material inventory and timelines scheduled for execution. Although precise metrics were not disclosed, the authors claimed significant improvements in accuracy for estimates made during the early stages of predictions and enhanced support for decision-making processes.

[9] Used Neural Networks together with Multiple Linear Regression (MLR) to estimate CO_2 emissions during the construction phase of buildings. A variety of tests were conducted to evaluate the predictive performance of the selected ML techniques. [10] developed a first-of-its-kind machine learning program that estimates Product Carbon Footprints (PCFs) with very few parameters as inputs. Although they did not provide conventional measures of accuracy such as R^2 or RMSE, their approach focused on usability and scalability for PCFs in small and medium construction enterprises. The authors in [11] examined the leading causes of carbon emissions in China using panel data from 254 cities during the years 2011-2020 and focused on six machine learning models in comparison to traditional econometric techniques. The results indicated that the machine learning models greatly outperformed in interpretability and predictive capability. This study further confirmed that energy consumption remains the primary driver of growth in carbon emissions. [12] Conducted an assessment of building emissions forecasting using Random Forest, Support Vector Regression and Gradient Boosting. Based on datasets of building projects, the Random Forest method outperformed other techniques in accuracy. The above-discussed work is summarized in Table 1.

Table 1. Comparison of various machine learning methodologies

Paper Name/Year	Methodology	Dataset	Performance Metrics (Numerical)
[5]/2025	GBM, Decision Tree, Random Forest	186 demolition projects (Korea)	GBM: R^2 = 0.997 (train), 0.983 (test), 0.984 (val)
[6]/2024	12 ML models to develop 72 alternative models	Building design parameters & materials	The gradient boosting model gave superior performance with R2 and MAPE values of 0.917 and 0.038
[7]/2025	Random Forest, XGBoost, SVR	Chinese cities	Random Forest: Highest R^2, Lowest MSE
[8]/2024	BIM + ML ensemble models	BIM-integrated construction data	XGBoost demonstrates a relatively higher degree of accuracy and minimal errors, with the RMSE of 206.62 and *R*2 of 0.88
[9]/2024	Linear Regression, Neural Networks	The building sector across the world	Mean Absolute Percentage Error (MAPE)
[10]/2024	Light ML framework for PCF	Construction product/process data	Lightweight model; no R^2 or RMSE reported
[11]/2024	Comparison of 6 ML models vs. econometric	Panel data from 254 Chinese cities	ML models outperformed econometric models. SE, MAPE and R^2, to evaluate the predictive performance of traditional OLS and six machine learning algorithms. The Extra-trees is superior by providing the smallest MSE and MAPE values and the largest R^2
[12]/2021	RF, SVR, GB	Building construction data	RF model with the highest value of R^2=99.88%

A few more systematic reviews also contributed to this field. [13] examined the use of Artificial Intelligence (AI) techniques such as ANN, CNN, RF, and SVR in the calculations and predictions of carbon emissions associated with buildings. They highlighted the usefulness of AI in smart construction for real-time tracking and optimization of operations. In a similar manner, [14] performed a science-mapping review of AI applications for net-zero emissions in sustainable buildings, analyzing 154 papers, which revealed LCA, energy, and decision-making tools as key dominating subject areas. The authors called attention to the need for more interpretable AI that connects models and crucial decision-making outputs. [15] analyzed different AI applications for carbon footprinting in construction. Their assessment of the capabilities of current ML-based platforms revealed a distinct trend: while adoption of such technologies is increasing, most lack adequate system-wide interoperability and interpretability frameworks that impact environmental policy. Analyzing the latest literature on the application of machine learning for analyzing the carbon footprint in construction indicates important areas of research. First, the majority of the studies, like [5, 9], evaluate single algorithms or a max of two, and do not test numerous models in the linear, tree-based, and ensemble categories. More broad studies like [7] also seem to have this issue, as they focus on urban-level emission domains. Second, not many studies include normalized indicators such as emissions per capita or emissions over GDP, which are critical for ideal cross-regional comparisons. This does not increase the generalizability of findings for a diverse dataset. Third, in the rising interest in predictive modeling, there seems to be a huge gap in policy scenario simulations in the literature. Some studies not only consider model outputs, but even fewer explore the impact of increasing renewable energy, a focus for international climate policy. Fourth, very few except [3] make use of “explainability” techniques like SHAP, which has decreased the transparency in predicting outputs of the models. Lastly, emission profiling behavior across regions using clustering and other unsupervised learning techniques is limited, which reduces opportunities for strategic policy creation. To address these gaps, the current research performs a detailed benchmarking of eight machine learning models on a practical multi-sector emissions dataset: Linear Regression, Ridge, Lasso, SVR, Decision Tree, Random Forest, XGBoost, and LightGBM. It adds designed features such as emissions and GDP per capita, models a 20% increase in renewable energy supply to assess policy responsiveness, uses SHAP for explainability, and implements KMeans clustering to reveal regional emission profiles. This approach improves accuracy while offering clear, actionable information critical for designing effective policies to mitigate carbon emissions in the construction industry.

3. Dataset Description and Preprocessing

3.1. Dataset Overview

The research dataset, "Carbon Emissions Dataset" (from Kaggle repository), used in the study, comprised 4,385 records and 16 original features. These records encompass several years, along with a variety of countries and regions, to capture factors that influence carbon emissions. The variable that is predicted is Co2EmissionsMetricTons, which signifies carbon emissions in metric tons for each record. The dataset has a range of variables like energy consumption, population, gdp, rate of urbanization, emissions from the industrial sector and the proportion of renewable energy, which allows for comprehensive emissions analysis. Figure 1 represents the heatmap of the pairwise Pearson correlation coefficients of the dataset's numeric variables. Perfect correlation, represented with on-diagonal values of 1.0, indicates strong self-relations. Most features have weak direct correlations with Co2EmissionsMetricTons, which tends to indicate that indirect relations - maybe complex and non-linear - exist, thus justifying machine learning approaches instead of simple regression. Emissions values are most strongly correlated with features such as population, energy consumption, and industrial output.

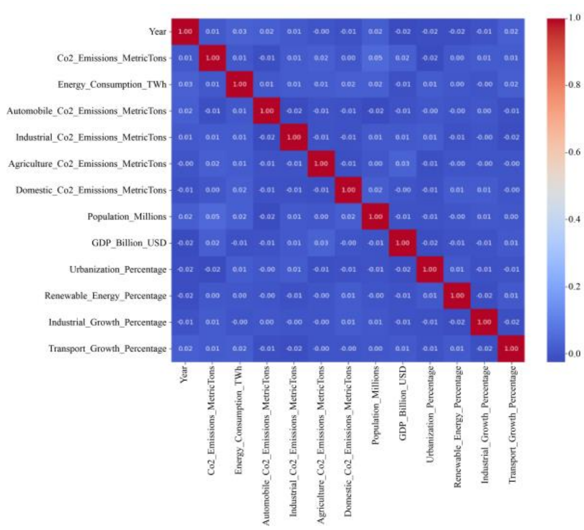

Fig. 1 Correlation heatmap between features and CO_2 emission

3.2. Feature Selection and Engineering

As a first step, non-numeric and non-predictive fields like Country and Region were removed to concentrate on the numerical and encoded categorical data. Two engineered features, along with original ones, were added:

- Emissions per Capita: the ratio of Co2EmissionsMetricTons to population millions, allows for emission figures to be normalized.
- Emissions per GDP: The ratio of Co2EmissionsMetricTons to GDP billion USD allows for emission context in terms of economic activity.

These metrics allow regions with different populations and economies to be compared more effectively. For categorical variables like IndustryType, one-hot encoding was utilized with the pd.getdummies() command and the first category was dropped to eliminate multicollinearity.

3.3. Data Splitting and Scaling

The dataset was split into training and testing subsets with an 80/20 train-test split. Standardization was done on the dataset using StandardScaler to avoid features with richer ranges in numeric dominating. The scaler was fitted on the training data and later applied to both sets to avoid data leakage. Figure 2 represents Energy Consumption TWh and Industrial Co2Emissions_MetricTons, which are two selected features out of four, as a set. Both features' raw distributions demonstrate high variability as well as non-standard ranges. Unscaled, the variance in the values of the features can introduce bias for distance-based or gradient-based algorithms where variables measured on larger scales dominate the outcome. Figure 3 is the same component after standard scaling is applied to it, showing the two features: Energy Consumption TWh and Industrial CO_2 Emissions_MetricTons. Both have means and standard deviations adjusted to around zero and one, respectively. This transformation will ensure that all features work without discrimination during the training phase, enhancing convergence, particularly on algorithms sensitive to feature scale like Support Vector Regression and Gradient Boosting.

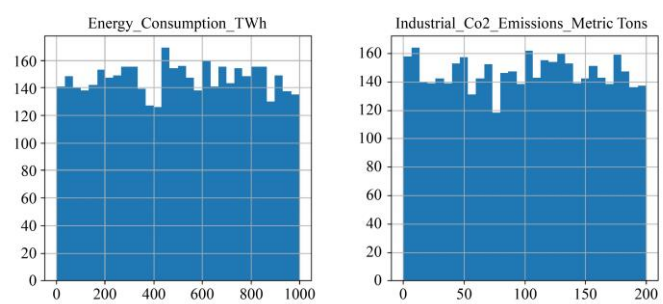

Fig. 2 Feature distributions before scaling

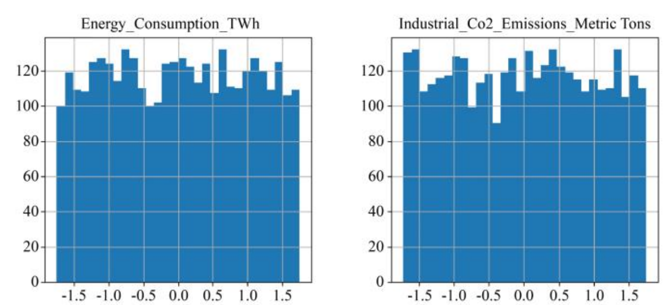

Fig. 3 Feature distributions after scaling

4. Machine Learning Models and Evaluation Metrics

4.1. Machine Learning Models

This research uses eight supervised regression models to predict carbon dioxide emissions in relation to construction features:

4.1.1. Linear Regression (LR)

It is a simple model where a linear relationship between the input features and the target variable is posited.

4.1.2. Ridge and Lasso Regression

These are linear regression models with L2 and L1 regularization, respectively, which help reduce overfitting and improve generalizability.

4.1.3. Support Vector Regression (SVR)

This is a kernel method model that locates a hyperplane that minimizes the error within a designated margin.

4.1.4. Decision Tree Regressor

A model that describes a set of observations using an integrated tree structure. This model is non-parametric and cuts the feature space to achieve a minimum prediction error.

4.1.5. Random Forest Regressor

A collection of decision trees constructed using bootstrapped samples. It introduces randomness within a controlled set of features to increase accuracy and robustness.

4.1.6. XGBoost and LightGBM

Efficient and fast-acting gradient boosting frameworks mostly used for structured table data.

All models were trained using scaled features with the 80/20 train-test split, applying the same held hyperparameters and features to ensure unbiased evaluations of each model's performance.

4.2. Evaluation Metrics

Model performance was assessed using standard regression metrics:

4.2.1. Coefficient of Determination (R^2 Score)

Indicates the proportion of variance in the dependent variable explained by the model.

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

4.2.2. Root Mean Squared Error (RMSE)

Measures the square root of the average squared prediction error.

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

4.2.3. Mean Absolute Error (MAE)

Represents the average absolute difference between predicted and true values.

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

4.2.4. Mean Absolute Percentage Error (MAPE)

Evaluates prediction accuracy as a percentage useful for interpretability.

$$MAPE = \frac{100\backslash\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

4.2.5. Mean Squared Logarithmic Error (MSLE)

Useful when target values vary across several orders of magnitude.

$$MSLE = \sum_{i=1}^{n}(\log(1 + y_i) - \log(1 + \hat{y}_i))^2$$

These metrics provide a comprehensive view of model performance, balancing error magnitude, interpretability, and sensitivity to outliers.

5. Results and Analysis

To determine the efficiency of different approaches of machine learning in forecasting carbon emissions, eight models were implemented and evaluated using the processed dataset: Linear Regression, Ridge, Lasso, SVR, Decision Tree, Random Forest, XGBoost and LightGBM. Models were assessed based on multiple evaluation criteria: R^2, RMSE, MAE, MAPE, MSLE, MedAE, and Explained Variance Score given by Table 2. Of all the models, SVR performed the most adequately, achieving R^2 of -0.0023, RMSE of 85.77, MAPE of 2.61% and MSLE of 0.96.

Ensemble models are supposed to be more powerful, showing larger variation: The Random Forest model obtained an RMSE of 87.12; however, a negative R^2 of -0.034 suggests poor generalization.

XGBoost was the worst performer in RMSE and R^2, recording 95.20 and -0.235, respectively, while LightGBM also did not generalize as well, reporting R^2 of -0.077 and RMSE of 88.90.

The Decision Tree model performed the worst out of all models, overfitting with an R^2 of -0.964 and RMSE of 120.07, as shown in Table 2.

Table 2. Evaluation criteria of various machine learning models

Model	R^2 Score	RMSE	MAE	MAPE (%)	MSLE	MedAE	Explained Variance
Support Vector Regression	-0.00	85.77	74.08	2.61	0.96	72.04	-0.00
Lasso Regression	-0.00	85.8	74.21	2.57	0.95	72.81	-0.00
Ridge Regression	-0.01	85.97	74.26	2.58	0.95	72.49	-0.01
Linear Regression	-0.01	85.97	74.26	2.58	0.95	72.5	-0.01
Random Forest	-0.03	87.12	74.84	2.64	0.96	72.51	-0.03
LightGBM	-0.08	88.9	-	2.65	0.97	73.06	-0.08
XGBoost	-0.24	95.2	-	2.61	1.02	75.66	-0.24
Decision Tree	-0.96	120.07	98.37	2.94	1.63	85.82	-0.96

These quantitative results now have accompanying visual analytics. The heatmap in Figure 4 contains a summary of all model results in relation to the evaluation metrics on a single consolidated image. In this Figure, dark and light color shades indicate lower normalized error values and higher performance scores, respectively.

SVR appears to be the most favorable model across nearly all metrics, and Decision Tree dominates the high-error zones, confirming it as a weak performer. Also, the radar chart in Figure 5 displays the individual and relative attribute proportions of the models, allowing an outline of each model's strengths and weaknesses.

The ideal models are positioned towards the center for relative error metrics and stretch towards the outer circle for performance metrics. SVR, Lasso and Ridge models display balanced shapes suggesting stability, while the Decision Tree model shows obfuscation with extremely unbalanced contour shapes.

As depicted in the correlation heatmap in Figure 1, the strong linear relationships between the features and the target variable generally influence an algorithm's domain generalization prowess. Thus, these results suggest that simpler models, especially SVR, may demonstrate better generalization within this domain.

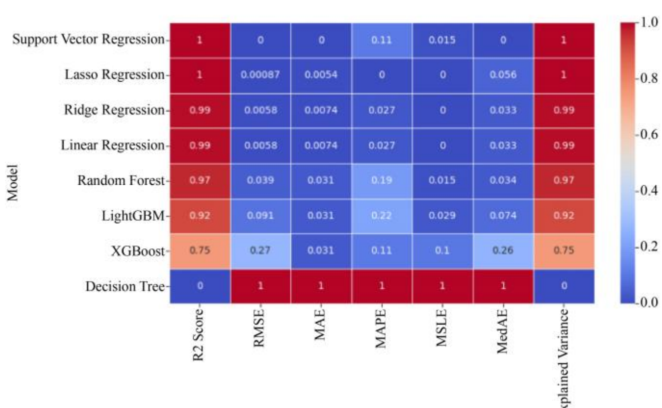

Fig. 4 Normalized heatmap of model metrics

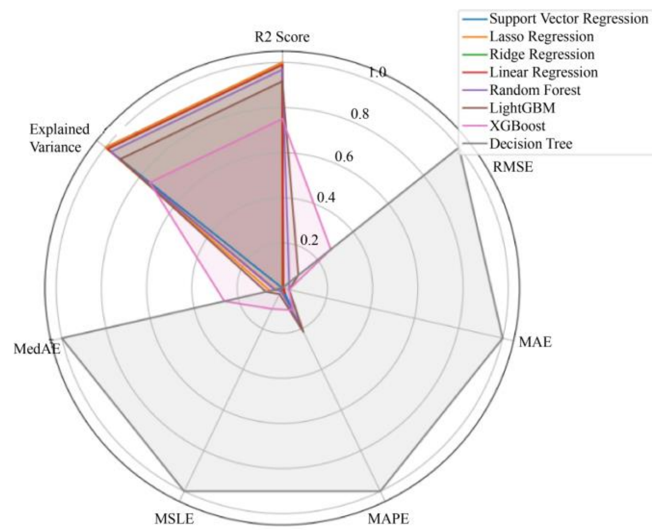

Fig. 5 Spider chart of model metrics

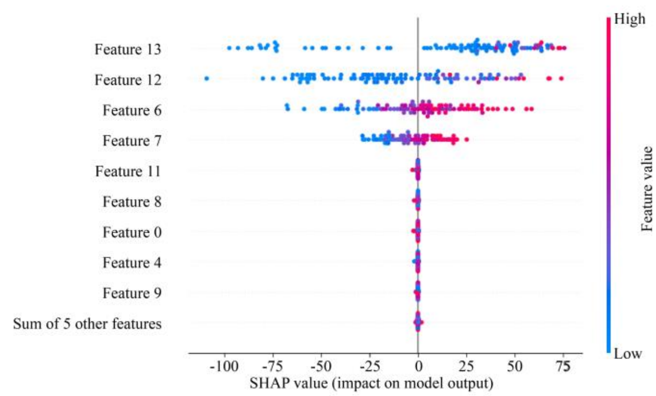

Fig. 6 SHAP values vs Feature Vallues

To comprehend the internal decision processes of the models more thoroughly, SHAP values were calculated for all models that support SHAP integration, such as Random Forest and XGBoost, as shown in Figure 6. SHAP analysis showed that the features Energy_Consumption_TWh, Industrial_Co2 _Emissions_MetricTons, Renewable_Energy_ Percentage, and Population_Millions were the most impactful on CO_2 emissions predictions across models. These features also dominated the rankings in the SHAP beeswarm plots, thus confirming their strong influence on prediction outcomes. This assessment strengthens the need to strategically direct policies and investments towards industrial decarbonization and renewable energy infrastructure to reduce emissions in construction and energy systems.

A policy simulation was performed further to assess the effect of renewable energy on emission reduction. In this experiment, a simulation with an increase of 20% in the Renewable_Energy_Percentage parameter for every piece of test data is conducted. Subsequently, the models are retrained and generate fresh predictions to determine CO_2 emission alterations.

The Random Forest model estimated a minor decrease in average emissions from 152.92 tons to 152.88 tons, while the Decision Tree model anticipated a slightly larger reduction from 152.95 to 152.71 metric tons. Emissions decreased from 151.91 to 151.89 metric tons in Lasso Regression and other linear models exhibiting minimal variation. Regardless of the small absolute changes, the consistent emissions reduction proposed by every model indicates the strong effect of additional renewable energy usage on the emission-reduction policy framework for the decline strategy climate policy.

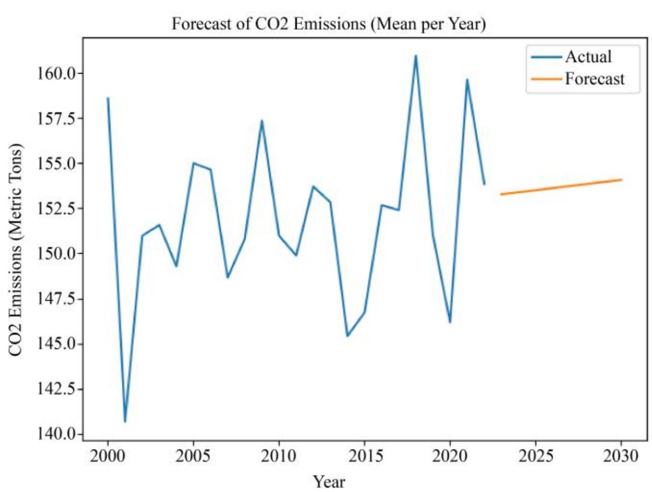

Fig. 7 Forecast of CO_2 emissions (mean per year) from 2023 to 2030

Figure 7 displays historical CO_2 emissions (in metric tons) from the year 2000 to 2022 along with projected emissions from 2023 to 2030. The blue line shows the actual yearly average emissions, while the orange line shows the predicted trend from a time-series model.

The forecast demonstrates emissions will gradually increase in the future if current industrial and energy consumption habits continue. The visual assists in understanding where carbon emissions could escalate and encourages anticipatory policies aimed at counteracting the upward trend.

6. Conclusion and Future Scope

The study investigates the application of machine learning techniques for forecasting and analyzing carbon emissions within the construction sector, demonstrating promising potential in this area. The research offers deep insights into various algorithms' prediction and generalization strengths by applying and benchmarking eight different regression models on a comprehensive dataset spanning multiple sectors. Among them, the simpler Support Vector Regression (SVR) model consistently outperformed XGBoost Ensemble and Decision Tree ensembles in robustness and accuracy, especially considering the weak linear correlations among the feature set and target variable. In addition, the constructed features, including per capita emissions and emissions per GDP, enhanced interpretability and cross-regional comparability while aligning regional predictions closer to actual policy metrics.

The most important contribution of the work is the integration of SHAP-based explainability, which highlighted dominant predictors such as energy consumption, industrial emissions, population, and the share of renewable energy. These findings fundamentally support known emission drivers and also help formulate targeted policy development aimed at emission reduction in energy-intensive industries. Moreover, the scenario simulation with a 20 percent increase in the share of renewables demonstrated that even with modest policy changes, emissions are consistently predicted to decrease across all models, indicating the potential for substantial emission reduction over time as a result of energy policy decisions made today. This is further corroborated by forecasting trends indicating the steady rise of emissions, underscoring the need for proactive carbon mitigation strategies.

In future, this work can be extended in many ways. For example, understanding local and seasonal emission patterns could be greatly improved with predictive time-series forecasting models if more granular temporal and spatial data were integrated. Second, including values for embodied carbon from material sourcing and transportation from other constituents would further enrich the scope of emission attribution, expanding the dataset's scope of value. Third, high-heterogeneous, high-dimensional datasets may improve predictive outcomes with the integration of hybrid models such as deep learning frameworks with attention mechanisms or graph-based ML.

Finally, enhanced policy simulation modules stand to gain from incorporating counterfactual analysis, economic cost modeling, and multi-objective optimization for structured carbon-neutral infrastructures. All together, these approaches are fundamental towards developing advanced intelligent construction ecosystems, underpinned by data and AI, which are responsive to the sustainability challenge.

References

[1] Gang Du, "Nexus between Green Finance, Renewable Energy, and Carbon Intensity in Selected Asian Countries," *Journal of Cleaner Production*, vol. 405, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[2] Zhonghua Cheng, and Shiyu Kong, "The Effect of Environmental Regulation on Green Total-Factor Productivity in China's Industry," *Environmental Impact Assessment Review*, vol. 94, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[3] Aditya Bhattacharya, *Applied Machine Learning Explainability Techniques Make ML Models Explainable and Trustworthy for Practical Applications Using LIME, SHAP, and More*, Packt Publishing, pp. 1-304, 2022. [Google Scholar] [Publisher Link]

[4] Yan Wang, and Xuelei Sherry Ni, "A XGBoost Risk Model via Feature Selection and Bayesian Hyper-parameter Optimization," *Arxiv Preprint*, pp. 1-17, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[5] Gi-Wook Cha, and Choon-Wook Park, "Development of an Optimal Machine Learning Model to Predict CO_2 Emissions at the Building Demolition Stage," *Buildings*, vol. 15, no. 4, pp. 1-22, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[6] Xiaocun Zhang et al., "Predictive Models of Embodied Carbon Emissions in Building Design Phases: Machine Learning Approaches Based on Residential Buildings in China," *Building and Environment*, vol. 258, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[7] Jian Sun et al., "Construction and Analysis of China's Carbon Emission Model Based on Machine Learning," *Scientific Reports*, vol. 15, no. 1, pp. 1-13, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[8] Haining Wang et al., "Integrating BIM and Machine Learning to Predict Carbon Emissions Under Foundation Materialization Stage: Case Study of China's 35 Public Buildings," *Frontiers of Architectural Research*, vol. 13, no. 4, pp. 876-894, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[9] Spyros Giannelos et al., "Machine Learning Approaches for Predictions of CO_2 Emissions in the Building Sector," *Electric Power Systems Research*, vol. 235, pp. 1-6, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[10] Silvio Lang et al., "A Simplified Machine Learning Product Carbon Footprint Evaluation Tool," *Cleaner Environmental Systems*, vol. 13, pp. 1-9, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[11] Wenmei Yu, Lina Xia, and Qiang Cao, "A Machine Learning Algorithm to Explore the Drivers of Carbon Emissions in Chinese Cities," *Scientific Reports*, vol. 14, pp. 1-15, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[12] Emmanuel B. Boateng et al., *Predicting Building-Related Carbon Emissions: A Test of Machine Learning Models*, Enabling AI Applications in Data Science, Springer, Cham, pp. 247-266, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[13] Jianmin Hua et al., "Artificial Intelligence for Calculating and Predicting Building Carbon Emissions: A Review," *Environmental Chemistry Letters*, vol. 23, pp. 783-816, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[14] Yanxue Li et al., "Artificial Intelligence in Net-Zero Carbon Emissions for Sustainable Building Projects: A Systematic Literature and Science Mapping Review," *Buildings*, vol. 14, no. 9, pp. 1-26, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[15] Peter Mésároš et al., "The Potential of Using Artificial Intelligence (AI) to Analyse the Impact of Construction Industry on the Carbon Footprint," *Mobile Networks and Applications*, vol. 29, pp. 1038-1052, 2024. [CrossRef] [Google Scholar] [Publisher Link]