Applied Data Science and Analysis
Vol.2024, **pp**. 173–188
DOI: https://doi.org/10.58496/ADSA/2024/014; ISSN: 3005-317X
https://mc04.manuscriptcentral.com/adsa

Research Article

Global Analysis and Prediction of CO2 and Greenhouse Gas Emissions across Continents

Fadya A. Habeeb[1], , Mustafa Abdulfattah Habeeb [2, *], Yahya Layth Khaleel [2], Fatimah N. Ameen [3],

[1] *Department of Mathematics, College of Education for Women, Tikrit University, Tikrit 34001, Iraq*

[2] *Department of Computer Science, Computer Science and Mathematics College, Tikrit University, Tikrit 34001, Iraq*

[3] *Institute of Automation and Info-Communication, Faculty of Mechanical Engineering and Informatics, University of Miskolc, Miskolc, Hungary*

ARTICLEINFO

Article History

Received 21 Aug 2024
Revised 18 Sep 2024
Accepted 20 Oct 2024
Published 25 Nov 2024

Keywords

CO2 Emissions
Greenhouse Gases
Climate Change
Machine Learning
Environmental Analytics

ABSTRACT

Understanding the concentrations of Carbon Dioxide (CO2) and greenhouse gases is very important in solving the problem of climate change. These emissions are the major cause of global warming, which, in turn, has many effects on the environment, economy and society. For this reason, the prediction models for these emissions must be precise to aid policy makers in planning for the effects of the climate in the future. To evaluate the emission data of different continents, this paper seeks to identify related patterns and findings that can help reduce emissions worldwide. The dataset used contains emission data and geographic information from several countries and allows the comparison of several ML models. The models that have been reviewed in this study are linear regression (LR), decision tree regression (DT), random forest regression (RF), support vector regression (SVR), k-nearest neighbor regression (KNN), the XGB regressor, the gradient boosting regressor, Ridge and Lasso. Among the models, the gradient boosting regressor was found to have the best prediction capability, with an R-squared value of 0. The highest value of the mean absolute error (MAE) was 929, and the lowest mean squared error (MSE) was 2535.30. This model outperforms the other models because of its excellent ability to identify the complex interactions between the input variables and emissions. The conclusions stress the possibility of using ensembles, such as gradient boosting, for emission forecasting and present a contribution to studies of this issue for researchers and policymakers. This is a nominal attempt in the ongoing global endeavour to gain insight and curb the determinable levels of CO2 and greenhouse gas emissions for effective decision-maki

1. INTRODUCTION

Carbon Dioxide (CO2) and greenhouse gas emissions are some of the biggest problems facing society today with respect to the environment. These emissions are mainly a result of human activities, which include industrial energy production, the cutting down of trees and the use of fossil fuels. CO2 and other greenhouse gas (GHG) emissions have increased global temperatures and thus adversely affect natural systems and their associated services, human well-being, and worldwide economies [1]. To overcome this challenge, the best approach is to understand the origin, current levels, and possible future behavior of these emissions. This study has the following objectives: offering such insights by estimating and calculating CO2 and greenhouse gas emissions via various modeling strategies. Figure 1 depicts the global carbon cycle.

Corresponding author. Email: mustafa@tu.edu.iq

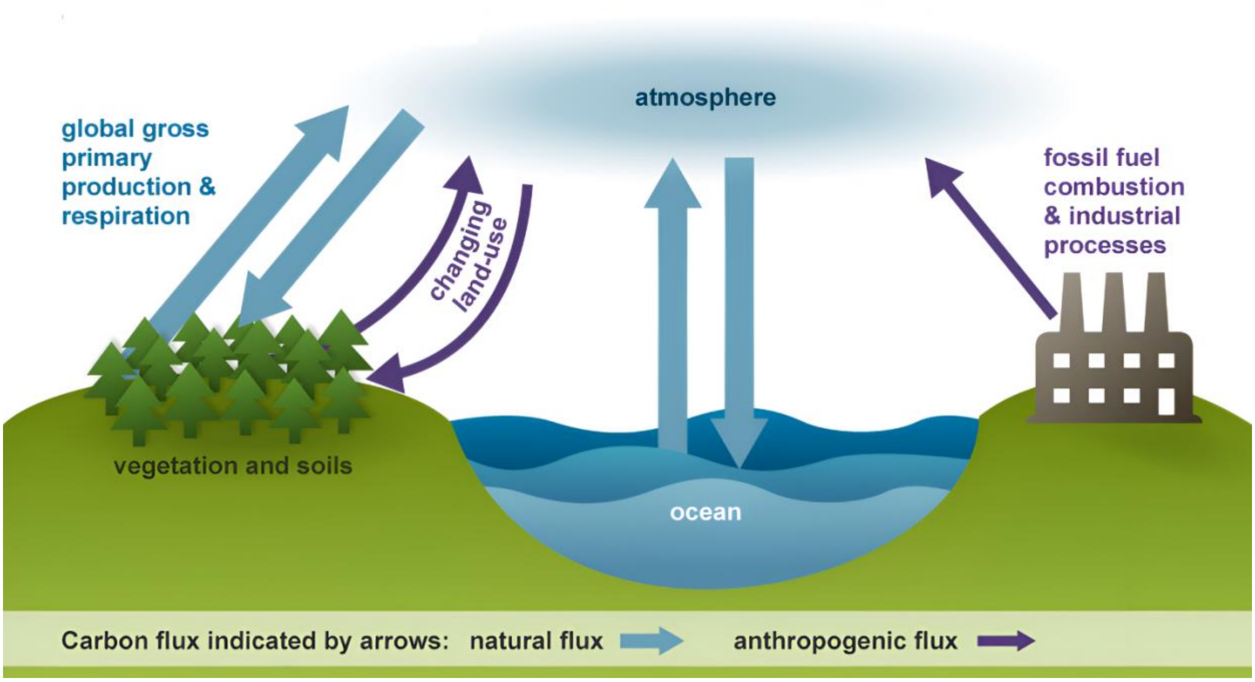
atmosphere
global gross primary production & respiration
changing land-use
fossil fuel combustion & industrial processes
vegetation and soils
ocean
Carbon flux indicated by arrows: natural flux
anthropogenic flux

Fig. 1. Global carbon cycle [2]

The major research questions answered in this study include the nature and extent of the geographical and sectoral dispersion of CO2 and greenhouse gases. Applying conventional methodologies to the results of a system to emissions makes it difficult for a sector to predict and prescribe policy measures to reduce emissions. Additionally, inadequate comprehensive models that can be used to project emissions for the next few years, depending on the trends that are anticipated and developed, hamper the development of effective measures for reducing emissions.

As stated earlier, the broad objective of this study is to formulate and test CO2 and greenhouse gas emission models that are capable of capturing the system dynamics of the emitting entities from one region to another and from one sector to another. Thus, the focus of this study is to compare several methods used in machine learning (ML) and artificial intelligence (AI) to determine the best methods for emissions forecasting and provide recommendations for policymakers and other stakeholders.

Artificial intelligence (AI) [3][4] and machine learning (ML) [5] are tools that may be helpful for improving the forecasting of CO2 and greenhouse emissions. When big datasets and efficient algorithms are involved, AI and ML have a better ability to find patterns that are not easily discerned otherwise. These techniques can also learn from new data and then perform better in the future than the previous methods do. In this context, AI and ML can be used to build not only forecaster models to predict probable future emissions but also policy impact models to determine how various interventions may impact emissions in the future, helping in formulating the right strategies to combat climate change.

2. BACKGROUND AND RELATED WORK

Global increases in greenhouse gas (GHG) emissions and carbon dioxide (CO2) in particular have been rapidly escalating, and are one of the most pressing global challenges to date [6], [7]. The emissions from which these results are primarily due to human activity, such as industrial processes, energy production, agriculture and transportation [8]–[10]. Rising global temperatures, extreme weather events, sea level rise, and disturbances in ecosystems have been caused by the build up in GHGs in the atmosphere [11]–[13]. Paris Agreement encompasses the need to take collective efforts to control the climate change and achieving carbon neutrality [14], [15]. Yet the dynamics of GHG emissions globally are difficult to understand, necessitating precisely collated data, analyzed, and predicted [16], [17].

Artificial Intelligence (AI) has fast become a game changer in dealing with environmental problems by providing sophisticated ways to work with large scale, disparate datasets [18], [19]. AI allows researchers to uncover patterns, predict future trends [20]–[22], inform and deliver actionable insights into CO2 and GHG emissions. For instance, machine learning algorithms are really good at uncovering correlations between socioeconomic factors and emissions level, and deep-learning methods are really good at running fancy atmospheric models [23], [24]. More often used nowadays are AI driven predictive models to predict emissions under a number of policy scenarios, helping policymakers to decide on an informed basis [25].

AI integration brings a special contribution to analysis with addressing regional emission disparities across continents. The variable industrialization levels, energy consumption patterns, and environmental policies result in diverse emission profiles for which this deserves comprehensive analysis [26]. The AI helps harmonize of datasets across multiple sources to make global estimates of emissions more precise and reliable [27]. In addition, real time monitoring of emission hotspots can be conducted using satellite imaging, natural language processing (NLP) and geospatial analytics for the development of targeted interventions with AI powered tools [28].

Even so, many challenges remain with the use of AI in environmental science. To enable any AI model to produce sustainable and impactful outcomes. So, the researchers must tackle issues including data availability, quality, and interpretability, not to mention the energy expense in building them. However, there is immense potential for the synergy between environmental research and AI to speed up progress towards global climate goals by allowing for the efficient and data informed solutions to reduce CO2 and GHG emissions.

ML models have been applied in recent years to forecast GHG emissions because of their ability to handle complex, variable data. A major study [29] undertook the use of classical regression, ML learning and deep learning (DL) algorithms to forecast the CO2 and nitrous oxide (N2O) fluxes of agricultural fields in the Quebec region. The applied algorithms yielded reasonably high performance, and the best result was given by the long short-term memory (LSTM) model, for which the R coefficient was 0. 87 for CO2 and 0.86 for N2O, with an RMSE of 30 and $p<0.05$ for both experiments at 86. 3 mg·m−2·hr−1 and 0. 19 mg·m−2·hr−1, respectively. On the other hand, classical models such as RF and SVM had comparatively low accuracies for peak N2O fluxes, and deeper ML models were comparatively less accurate overall and were also more sensitive to overfitting.

To predict total GHG emissions, including CO2, CH4, N2O and fluorinated gases, a predictive model of machine learning with mathematical programming as a mix of the two techniques was developed by [30] to measure Iranian energy sector emissions for the time series between 1990 and 2018. Among the nine selected ML algorithms, namely, ANN, AR, ARIMA, SARIMA, SARIMAX, RF, SVR, KNN, and LSTM, facet improvement with PSO and GWO provided enhanced results, with error rates of at least 31.7% and 12.8%, respectively. The forecast of the hybrid model was higher than 1096 Mt/year of Iran's GHG emissions by 2028 compared with the individually applied ML methods.

In Turkey, while the production capacity of energy has improved, the GHG emissions of the country are still high. In [31], the authors compared support vector machines (SVMs) and artificial neural networks (ANNs) for the prediction of CO2, CH4, N2O and fluorinated gases emitted by different sectors. Certain features of SVMs that generalize the details and importance of ANNs for classification and regression were illustrated from data obtained from the Turkish Statistical Institute for the period of 1990--2019. The analysis was performed via MATLAB 2019b, and the goal of the study was to increase the prediction efficiency and reliability.

In Study [32], OLS, SVM, and GBR were adopted for predicting CO2 emissions through transportation for 30 countries with high emissions. Hence, the gradient boosting regression model with combined features (GBR_ALL) seemed to provide an excellent performance model with the highest R^2 value of 0.9943, an rRMSE of 0.1165 and a mean absolute percentage error of 0.1408%. It was found that transportation features were somewhat significant for the whole set of cities; however, the transport features were much more significant. Nevertheless, the general transport standards of the city and socioeconomic features were shown to be important for Tier 1 and Tier 2 cities, with GDP and population being particularly important indicators.

[33] subsequently analyzed the performance of several ML methods in forecasting CO2 emissions from buildings up to 2050 via linear regression, ARIMA, and shallow and deep neural network univariate and multivariate methods. In particular, the study analyzed different methods of feature extraction, namely, lagged values and polynomial transformations, for various regions worldwide: Brazil, India, China, South Africa, the United States, Great Britain, the world average and the European Union. By means of this profound analysis, these key questions have been answered, offering an understanding of how to accomplish consistent projection of long-term CO2 emissions for the building sector.

3. PROPOSED FRAMEWORK

The proposed framework starts with a collection dataset on CO2 and greenhouse gas emissions. Next, we perform exploration data analysis, also known as EDA, to visualize the patterns, trends and relationships present in the dataset. Data cleaning and normalization as well as the handling of missing data are some of the preprocessing techniques that are used before modeling. There are many types of ML models that are used. Finally, the performance of the models is compared and tested, as shown in Figure 2, to determine how well they capture the actual emission trends.

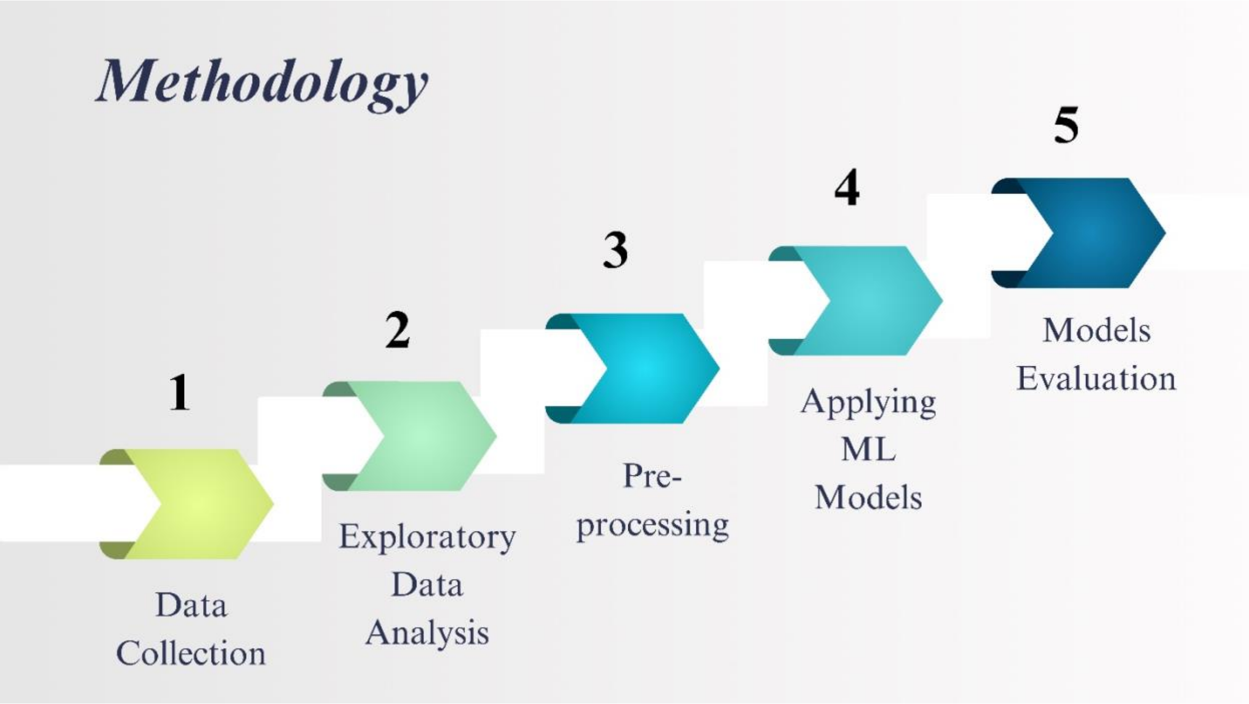

Fig. 2. Proposed Framework

3.1 Data collection

The dataset applied in this work is obtained from [34] and reports data on CO2 emissions and greenhouse gases from different countries and continents. This paper is organized in tabular form and contains a total of two hundred and four entries, where each row refers to a country/region and certain attributes.

Data fields:

a. Region: Region refers to the name of the country or the region where the material is located.
b. CO2 Emissions (Mt): Carbon dioxide emissions expressed in megatons that represent million metric tons.
c. Greenhouse gas emissions (Mt): This is the total, or the sum total, of the greenhouse gas emissions calculated in megatons. This column has some missing data, and information is available for 179 of the 204 fields.
d. Continent: The continent in which the country or the region is situated.

This dataset is useful for environmental science, social interaction planning and demonstrations for instruction and learning to understand worldwide CO2 and greenhouse gas emissions and their consequences for climate change. Table 1 presents CO2 and greenhouse gas emission data (in megatons) for various countries, along with their continent and total gas emissions. Missing values are indicated as "NaN."

TABLE I. CO2 AND GREENHOUSE GAS EMISSIONS

	Region	CO2 Emissions (Mt)	Greenhouse Gas Emissions (Mt)	Continent	Overall_gas_emission
0	*Afghanistan*	*5.68*	*98.9*	*Asia*	*104.58*
1	*Albania*	*4.49*	*10.1*	*Europe*	*14.59*
2	*Algeria*	*177.08*	*218.9*	*Africa*	*395.98*
3	*Angola*	*20.19*	*79.7*	*Africa*	*99.89*
4	*Anguilla*	*0.02*	*NaN*	*North America*	*NaN*
...	...	...	...	...	...
199	*Vietnam*	*327.91*	*376.5*	*Asia*	*704.41*
200	*Western Sahara*	*0.24*	*NaN*	*Africa*	*NaN*
201	*Yemen*	*12.26*	*21.8*	*Asia*	*34.06*
202	*Zambia*	*9.27*	*40.7*	*Africa*	*49.97*
203	*Zimbabwe*	*10.22*	*31.4*	*Africa*	*41.62*

3.2 Exploratory Data Analysis (EDA)

To enhance comprehension of the dataset's structure and uncover correlations that could enhance the predictive models for CO2 and greenhouse gas emissions, we conducted an initial analysis of its general characteristics, known as exploratory data analysis (EDA) [35]. This section presents an overview of the dataset along with the findings from this analysis that are directly relevant. Table 2 presents the cumulative greenhouse gas emissions (in megatons) across six continents. Asia is the highest emitter at 12,481.71 Mt, followed by Europe at 9,785.59 Mt and Africa at 6,069.65 Mt. North America and Oceania contribute 4,603.74 Mt and 4,436.94 Mt, respectively, while South America has the lowest emissions at 1,823.93 Mt. These data highlight regional disparities in greenhouse gas emissions globally.

TABLE II. CUMULATIVE GREENHOUSE GAS EMISSIONS ACROSS CONTINENTS

	Continent	Overall gas emission
1	*Asia*	*12481.71*
2	*Europe*	*9785.59*
0	*Africa*	*6069.65*
3	*North America*	*4603.74*
4	*Oceania*	*4436.94*
5	*South America*	*1823.93*

Figure 3 visualizes the distribution of CO2 gas emissions across different continents. Asia is the largest contributor, responsible for 40.1% of the emissions, followed by Europe at 32.2%. North America and Africa account for 9.1% and 10.7%, respectively. Oceania and South America had the lowest percentages, with 3.2% and 4.6%, respectively. This chart clearly shows emissions by region, with a focus on the greater influences of Asia and Europe on global CO2 and greenhouse gas emissions. These results imply the need for specific emission reduction measures in these high-emission zones.

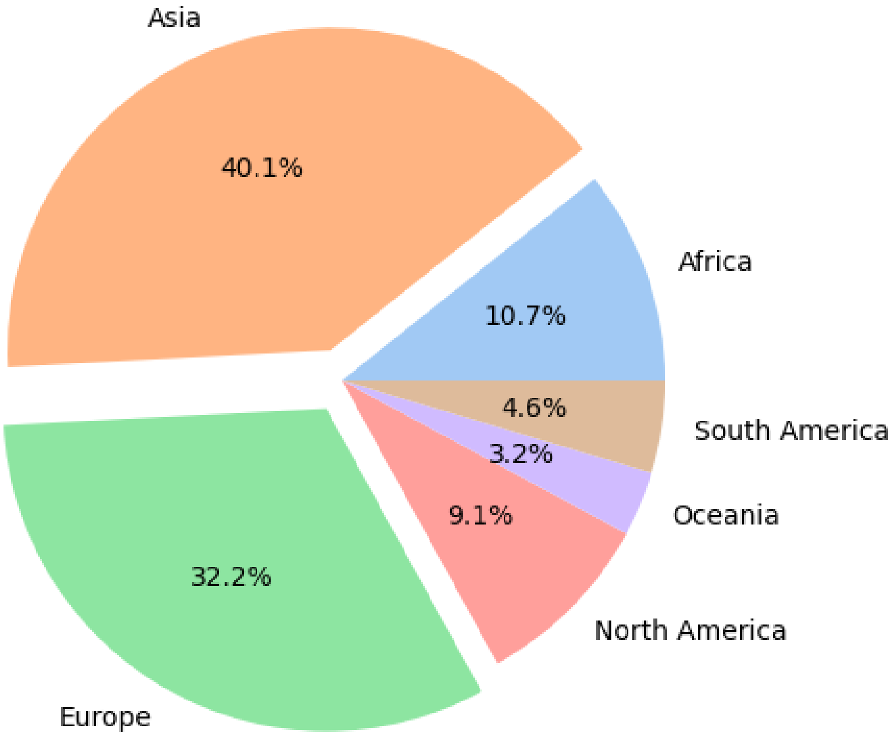

Fig. 3. Distribution of CO2 gas emissions across different continents

Figure 4 presents the distribution of greenhouse gas emissions across regions of the globe. Asia leads with 27.6%, which in itself contributes to the total emissions. Europe accounts for 21.2% of the total production, North America accounts for the smallest share, with only 13.1%, Africa and South America accounting for 17.9% and 4.7%, respectively. Oceania represents 15.5% of this distribution, indicating the unequal burden of climate change around the world and the need for a differentiated approach toward emissions around the world.

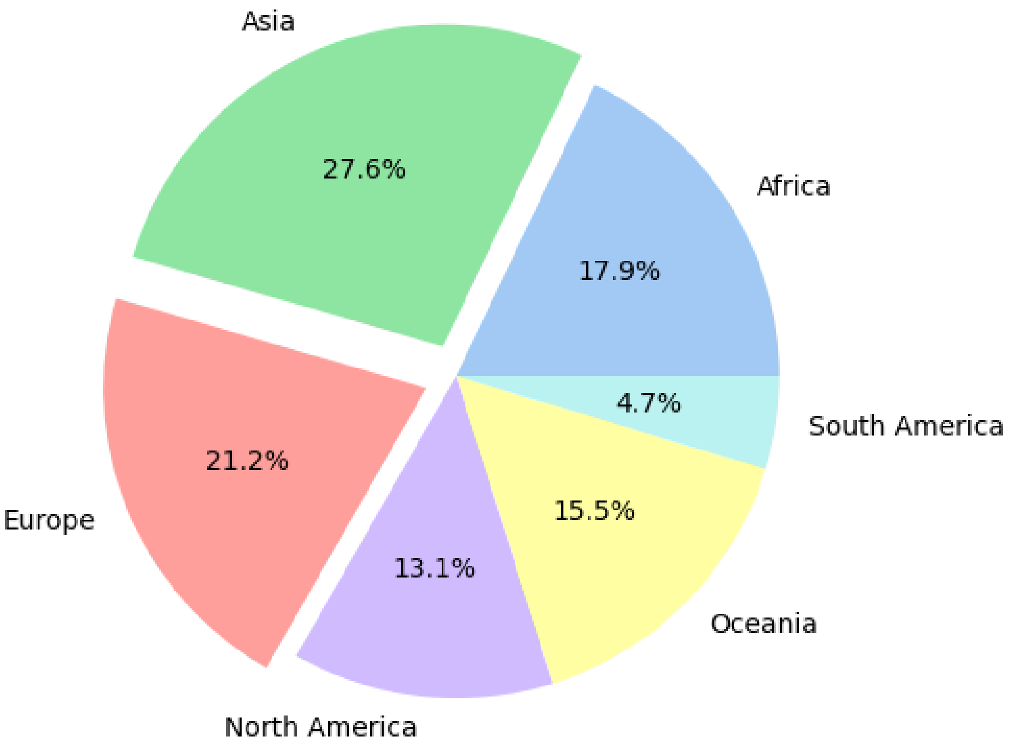

Fig. 4. Global greenhouse gas emissions by region.

The total gas emissions have hence been distributed across different continents, as shown in Figure 5. Asia remains the most dominant contributor of FDI, with 31.8 APEC of total emissions, whereas Europe accounts for 25.0%. Africa contributes 15.5%, highlighting its significant share. North America and Oceania contributed 11.7% and 11.3%, respectively. South America has the smallest share at 4.7%.

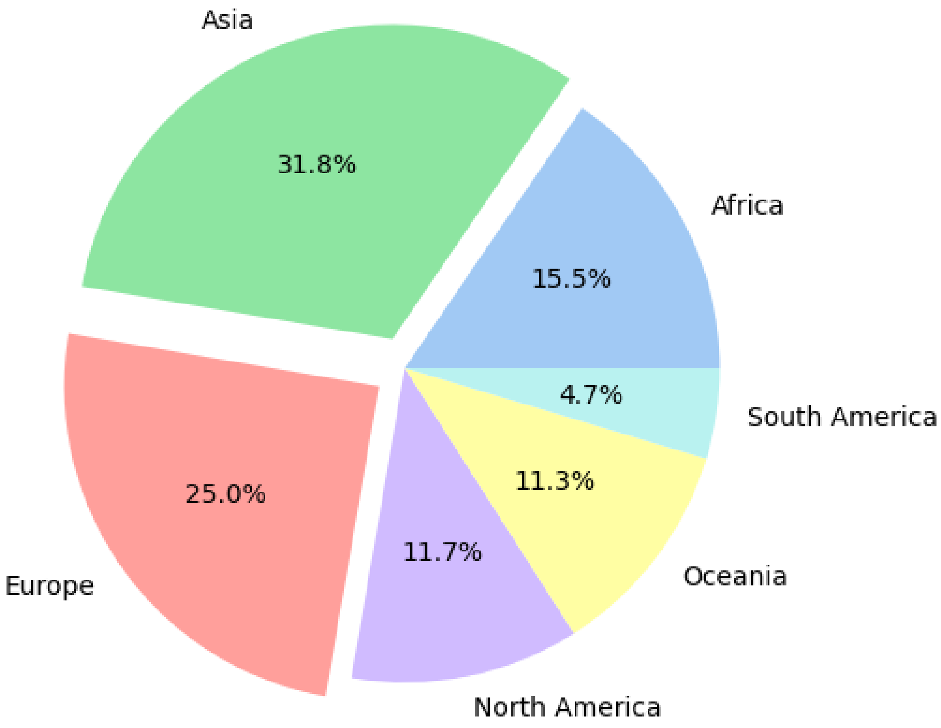

Fig. 5. The distribution of total gas emissions across different continents

Figure 6 compares CO_2 emissions and overall greenhouse gas emissions among the top ten countries in Asia. On the left, China has the highest CO_2 emission, at 12,670 Mt, which is significantly higher than that of India (2,690 Mt) and Japan (1,080 Mt). The other countries, including Indonesia and South Korea, contribute much smaller amounts. On the right, the greenhouse gas emissions chart shows a similar trend, with China again leading at 12,400,000 Mt. India and Japan remaining the next highest emitters, but the scale of emissions is much greater than that of CO_2 alone. This highlights China's critical role in Asia's environmental impact and underscores the urgency for emission reduction strategies in these leading countries.

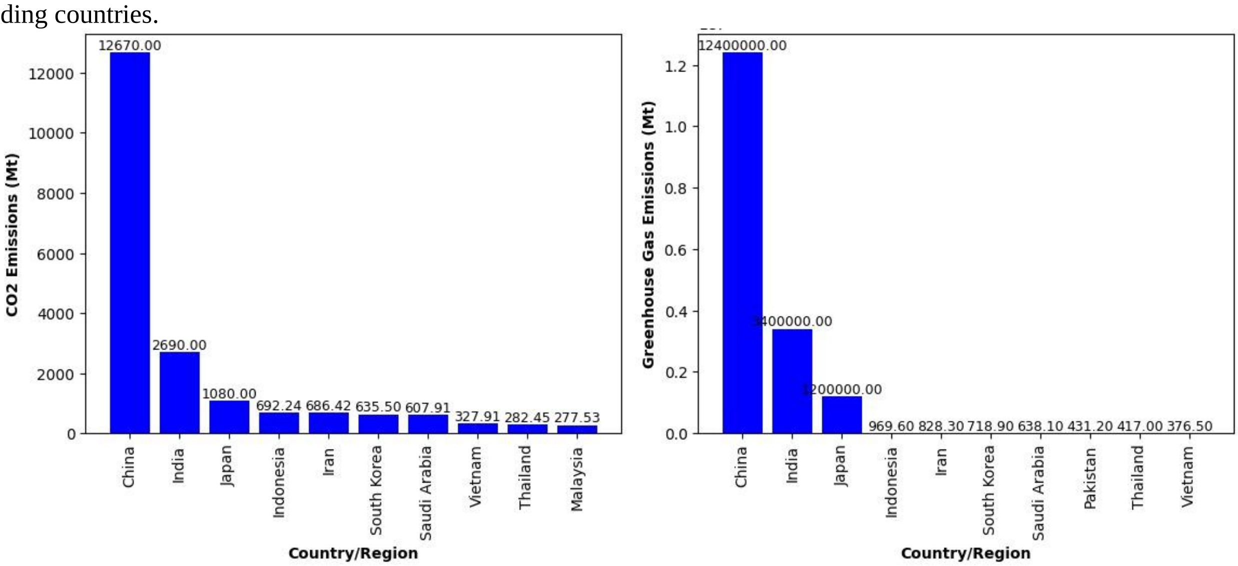

Fig. 6. Compares CO_2 emissions and overall greenhouse gas emissions among the top ten countries in Asia.

Figure 7 presents the CO_2 emissions and total greenhouse gas emissions for the top ten countries in Europe. In the CO_2 emissions chart, Russia stands out as the largest emitter at 1,910 Mt, followed by Germany (673.6 Mt) and Turkey (481.25 Mt). The remaining countries, including the UK and Italy, show significantly lower emissions, indicating a concentration of responsibility among a few nations. On the right, the chart of greenhouse gas emissions provides a similar picture, where Russia tops at 2,500,000 Mt, which is significantly higher than that of Germany and the UK. These data reflect the massive contribution of Russia to the overall emission level in Europe and therefore highlight the need to focus on effective climate policies that would affect high-emission countries only.

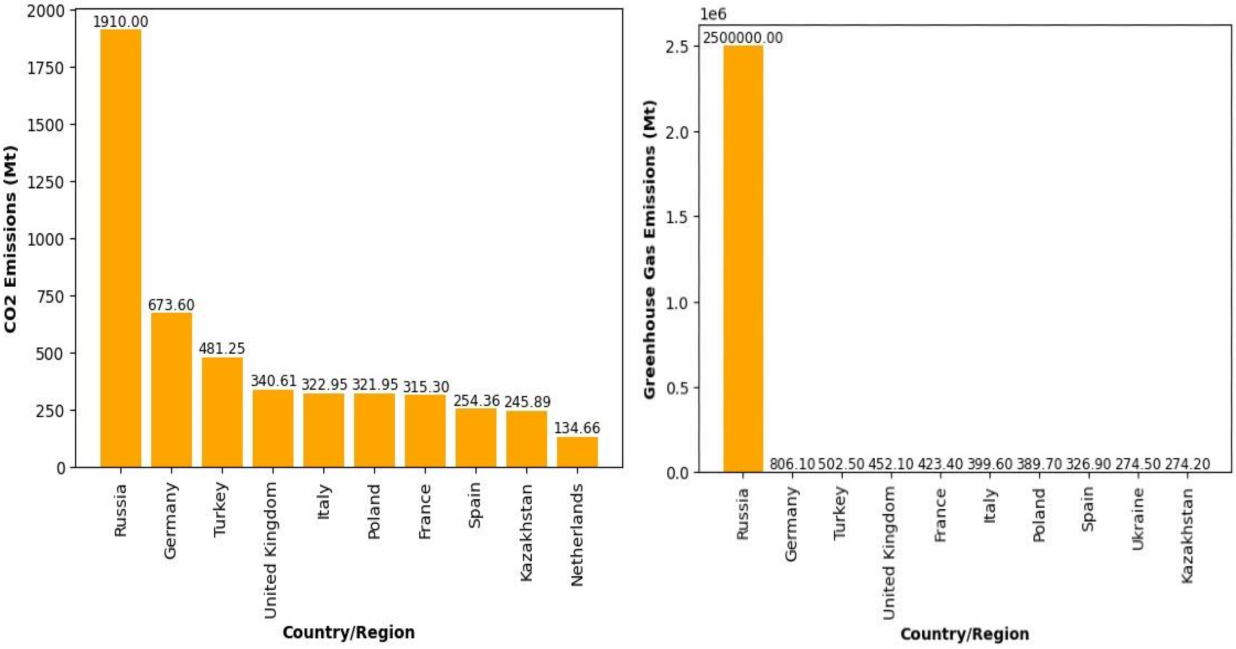

Fig. 7. CO_2 emissions and total greenhouse gas emissions for the top ten countries in Europe

The indicators of CO_2 emissions and the total greenhouse effect in the ten largest countries in South America are shown in Figure 8. Examining the chart of the CO_2 emissions, we find that Brazil is the top-ranked country and ranks 466.77, followed by Argentina, which contributes 184.04 Mt, and Venezuela, which contributes 96.92 Mt. Other countries, i.e., Chile and Peru, have lower total emissions, indicating that emissions are centralized. To the right, the emission of greenhouse gases chart also accentuates Brazil’s leadership, with the overall emission having scaled to 1,000,000 Mt. Argentina and Venezuela follow but at significantly lower levels, reinforcing Brazil's pivotal role in the region's environmental impact.

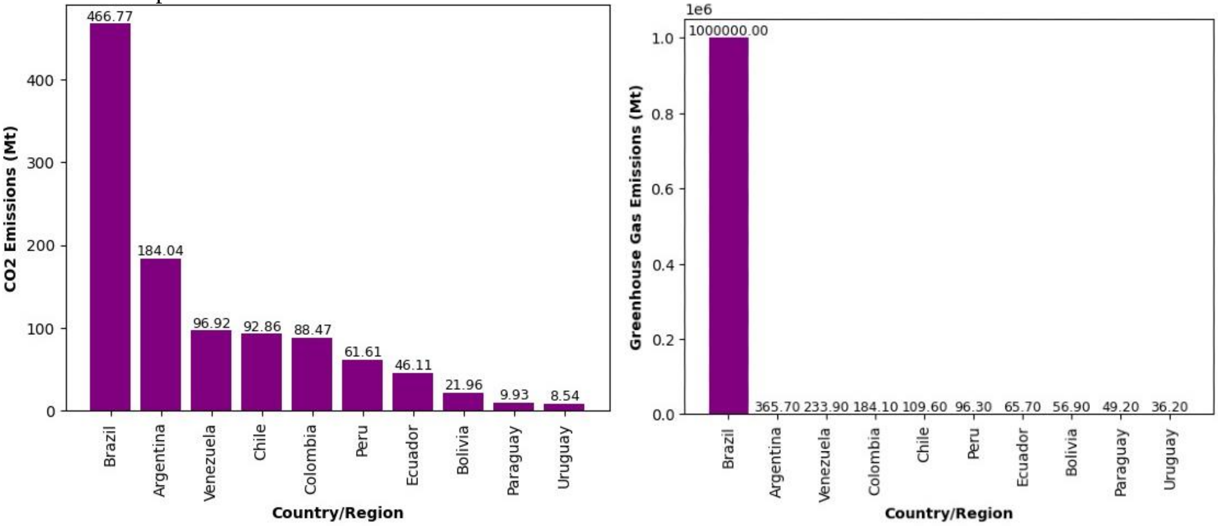

Fig. 8. CO_2 and total greenhouse gas emissions for the top ten countries in South America

Figure 9 depicts the CO_2 and greenhouse gas emissions from the top ten countries in North America. In the left chart, the United States is the predominant emitter, with a staggering 4,850 Mt, followed by Canada (582.07 Mt) and Mexico (487.77 Mt). Other nations, such as Trinidad and Tobago and Cuba, contribute much smaller amounts, highlighting the concentration of emissions among a few key players. On the right, the greenhouse gas emissions chart reinforces this trend, with the United States again leading at 6,000,000 Mt. Saint Lucia and Canada following, but their figures are significantly lower. These data emphasize the critical role of the United States in North America's emissions landscape and underscore the importance of focused climate policies in these leading countries to address environmental challenges effectively.

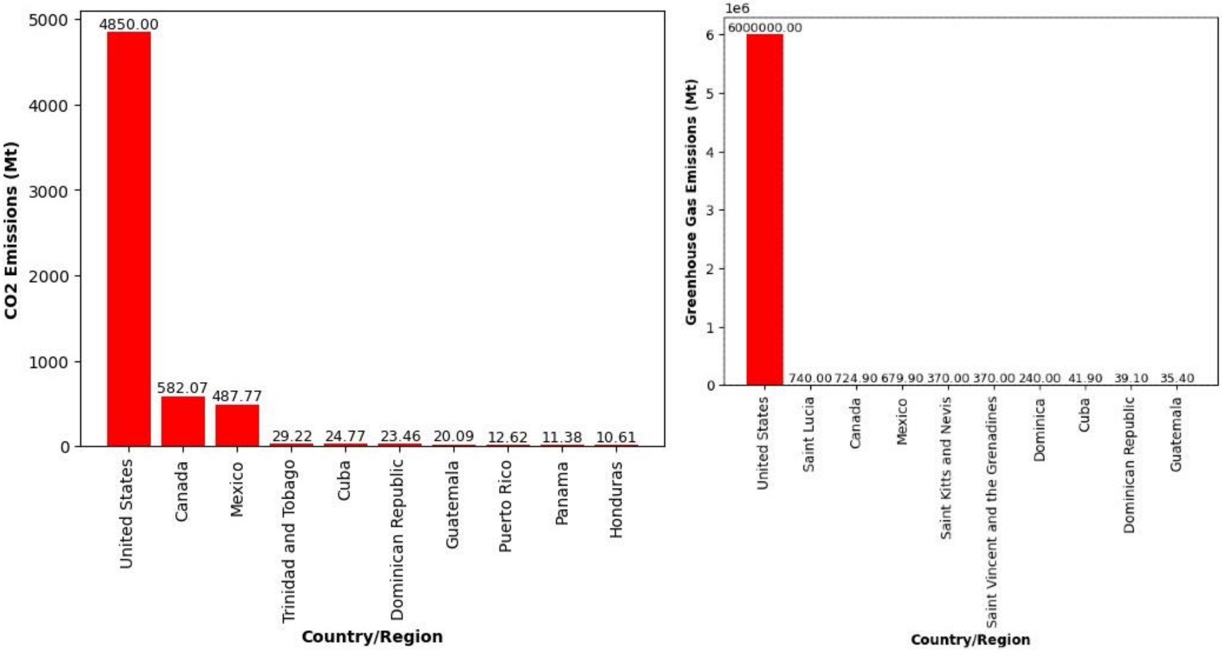

Fig. 9. CO_2 and greenhouse gas emissions from the top ten countries in North America

Figure 10 depicts the CO_2 and greenhouse gas emissions from the top ten countries in Africa. In the left chart, South Africa leads with 404.97 Mt, followed by Egypt (265.96 Mt) and Algeria (177.08 Mt). Other countries, such as Nigeria and Morocco, contribute smaller amounts, indicating a significant concentration of emissions among a few key nations. On the right, the greenhouse gas emissions chart shows Seychelles at the top with 780 Mt, followed by Comoros (590 Mt) and South Africa (513.40 Mt). Egypt and Nigeria also feature prominently.

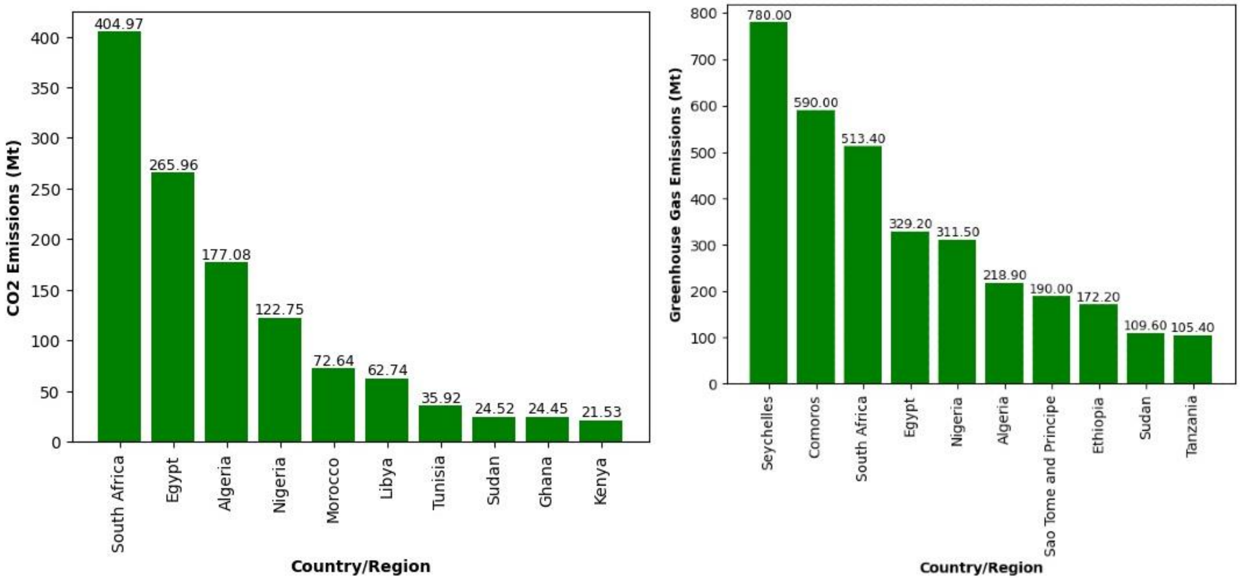

Fig. 10. CO_2 and greenhouse gas emissions from the top ten countries in Africa

Figure 11 presents the CO_2 and greenhouse gas emissions for the top ten countries in Oceania. In the CO_2 emissions chart, Australia stands out with 393.16 Mt, significantly outpacing its neighbors—New Zealand, for example, contributes only 32.37 Mt. Other countries, such as New Caledonia and Papua New Guinea, produce even less, reflecting Australia's major role in regional emissions. Conversely, the greenhouse gas emissions chart shows a different narrative. The Solomon Islands lead with 940 Mt, followed closely by Vanuatu at 870 Mt and Samoa at 690 Mt. Australia, while still being a significant emitter, which ranks lower in this context with 615.40 Mt.

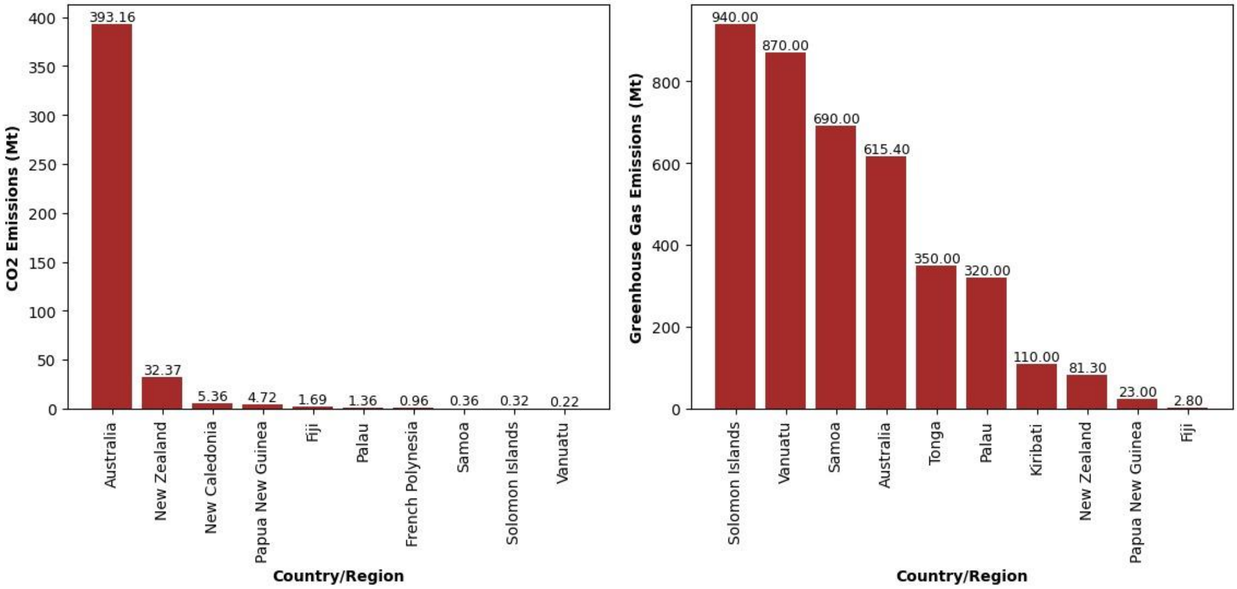

Fig. 11. CO_2 and greenhouse gas emissions for the top ten countries in Oceania

Finally, Figure 12 provides a comparative analysis of CO2 emissions and average greenhouse gas emissions across various regions, highlighting the top emitters. In the chart on the left, which focuses on total CO2 emissions, Indonesia and Iran lead with emissions just under 700 megatons (Mt), followed closely by Germany, South Korea, and Saudi Arabia, each surpassing 600 Mt. The remaining regions—Canada, Mexico, Turkey, Brazil, and South Africa—contribute between 500 and 400 Mt. On the right, the chart depicting average greenhouse gas emissions reveals that China is the overwhelming leader, with emissions significantly higher than those of other regions. The United States is second largest, followed by a significantly large difference from the third largest emitter, India. Some of the countries that emit a lower level of greenhouse gases than China does, and the U.S. includes Russia, Japan, Brazil and Indonesia.

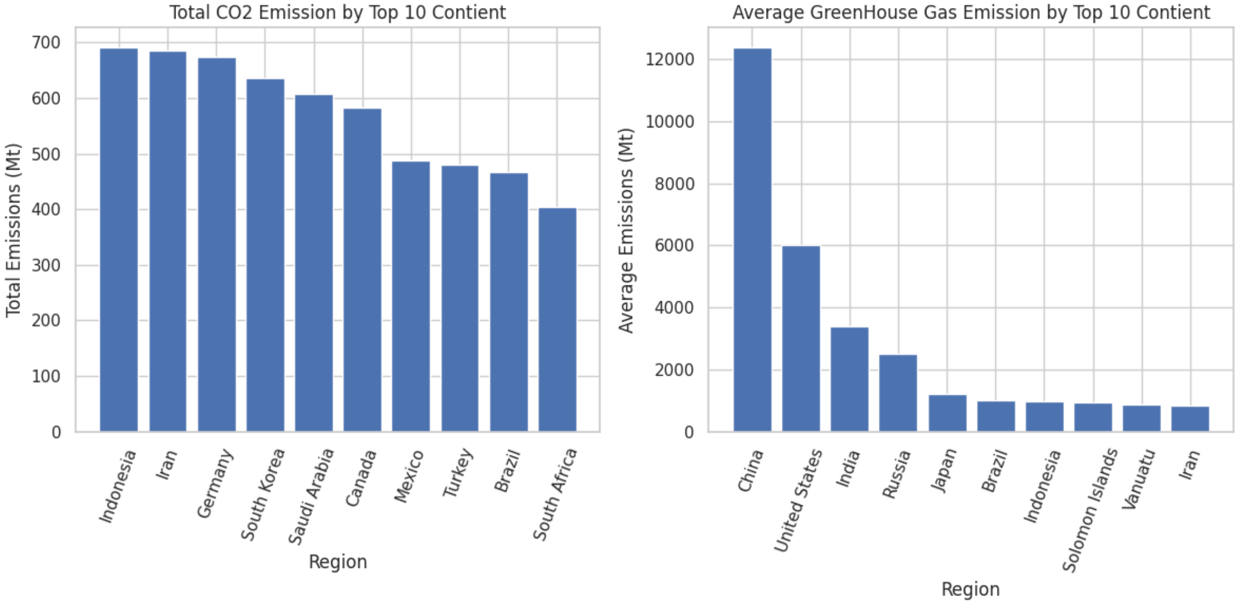

Fig. 12. Comparative analysis of CO2 emissions and average greenhouse gas emissions across the top emitters

This underscores the significant impact of specific countries on global emission levels, highlighting the need for targeted policies to address these disparities and mitigate climate change effectively.

3.3 Preprocessing

The first step of the preprocessing phase is data reading, which is performed with the help of the Pandas library [21]. This dataset provides data about CO2 emissions, greenhouse gas emissions and the continents to which they belong. These data were then transformed into a structure DataFrame so that it is in a better form to start working with.

[36] The management of missing data was important in this phase. Some cells had missing values, which were mostly for two variables: greenhouse gas emissions (Mt) and overall gas emissions. These missing cases were handled via proper techniques of imputation, such as the mean/median imputation, where gaps where data were missing were filled with data with equivalent mean or median values, respectively, or complete cases were dropped if there were more missing data that would affect the analysis.

Following the data cleaning process [37], the data were preprocessed by splitting them into training and test sets. The training set and test set were created with the training set comprising 80% of the entire dataset and the test set comprising 20% of the entire dataset. This division helped ensure that the majority of the data were given for training the model and that only a small portion of the data that the model had never encountered was given for testing the model. This process prepares the data for the next step, which involves training several of the ML models to forecast the emissions of CO2 and other greenhouse gases.

3.4 Applying ML models

This study employs a variety of ML models to analyze and predict CO2 and greenhouse gas emissions across different continents. Each model offers unique advantages, and the following models were used:

1. **Linear Regression (LR):** Linear regression models the relationship between a variable y and one or more independent variables X by fitting a linear equation [38]:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon \tag{1}$$

where β_0 is the intercept, $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients of the independent variables, and ϵ is the error term [39].

2. **Decision Tree Regression (DT):** Decision Tree Regression uses a tree-like model of decisions. The prediction is made by splitting the data into subsets at each node on the basis of the feature that results in the greatest reduction in variance:

$$\text{Variance Reduction } = \operatorname{Var}(y) - \left(\frac{n_{\text{left}}}{n} \operatorname{Var}(y_{\text{left}}) + \frac{n_{\text{right}}}{n} \operatorname{Var}(y_{\text{right}})\right) \tag{2}$$

where n is the total number of instances and where n_{left} and n_{right} are the instances in the left and right branches, respectively [40].

3. **Random forest regression (FR):** Random forest regression is an ensemble method that aggregates the predictions of multiple decision trees [41]:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} f_t(x) \tag{3}$$

where T is the number of trees, $f_t(x)$ is the prediction from the t-Th tree, and $\hat{y}$ is the final prediction [42].

4. **Support Vector Regression (SVR):** Support Vector Regression attempts to find a function $f(x)$ that deviates from the actual observed values y_i by at most ϵ for all training data while ensuring that this function is as flat as possible:

$$f(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x) + b \tag{4}$$

where $K(x_i, x)$ is a kernel function, α_i are the Lagrange multipliers, and b is the bias term [43].

5. **K-Nearest Neighbors Regression (KNN)**: K-Nearest Neighbor Regression predicts the value $\hat{y}$ of a new data point by averaging the values of its K-nearest neighbors:

$$\hat{y} = \frac{1}{K} \sum_{i=1}^{K} y_i \tag{5}$$

where y_i are the values of the K nearest neighbors [44].

6. **XGBRegressor:** This is a specific implementation of the gradient boosting framework, which optimizes the following objective function:

$$\text{Obj }(\theta) = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t)}\right) + \sum_{j=1}^{t} \Omega\left(f_j\right) \tag{6}$$

where l is a loss function, $\hat{y}_i^{(t)}$ is the prediction from the t-th tree, and $\Omega\left(f_j\right)$ is a regularization term for the tree structure [45].

7. **Gradient** boosting regressor: The gradient boosting regressor builds models sequentially, each one correcting the errors of its predecessor. The model is based on the minimization of the loss function:

$$L(\theta) = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(m)}\right) + \sum_{m=1}^{M} \gamma_m \tag{7}$$

where l is the loss function, $\hat{y}_i^{(m)}$ is the prediction at iteration$m$, and γ_m are the weights [46].

8. **Ridge Regression:** Ridge regression modifies the linear regression model by adding a regularization term to the loss function to penalize large coefficients:

$$\text{Minimize } \left(\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2\right) \tag{8}$$

where λ is the regularization parameter and where β_j are the coefficients [47].

9. **Lasso Regression**: LASSO regression adds a ℓ_1 regularization term to the linear regression model, which can shrink some coefficients to zero, effectively performing variable selection:

$$\text{Minimize } \left(\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \left|\beta_j\right|\right) \tag{9}$$

where λ is the regularization parameter and where β_j are the coefficients.

These models were selected to offer a comprehensive analysis of emission trends, each providing different insights and capabilities for predictive tasks [47].

3.5 Model evaluation

Evaluating the performance of the ML models was a critical part of the analysis. Two primary metrics were used for this evaluation: the mean squared error (MSE) and R-squared (R²).

The mean squared error (MSE) measures the average of the squares of the errors, that is, the average squared difference between the actual and predicted values. A lower MSE indicates a better fit of the model to the data. The MSE is calculated via the following equation [48]:

$$\text{MSE} \ = \frac{1}{m}\sum_{i=1}^{m} \ (X_i - Y_i)^2 \qquad (10)$$

$$(\text{best value} \ = 0; \ \text{worst value} \ = +\infty)$$

- ^{m} is the number of observations,
- X represents the actual values,
- Y represents the predicted value.

R-squared (R^2): a statistical measure that represents the proportion of the variance for the dependent variable that is explained by the independent variables in the model. It provides an indication of how well the model's predictions match the actual data. An R^2 value closer to 1 indicates a better fit. The R-squared value is calculated via the following equation [48]:

$$\text{R}^2 \ = 1 - \frac{\sum_{i=1}^{m} (X_i - Y_i)^2}{\sum_{i=1}^{m} (\acute{Y} - Y_i)^2} \qquad (11)$$

$$(\text{worst value} \ = -\infty; \text{best value} \ = +1 \ )$$

- X represents the actual values,
- Y represents the predicted value,
- Ý is the mean of the actual values.

In this study, both the MSE and R^2 were calculated for each of the ML models used, providing a comprehensive assessment of model performance. These metrics help determine which model is most effective in predicting CO2 and greenhouse gas emissions, with a focus on minimizing the error (MSE) and maximizing the explained variance (R^2).

4. RESULTS AND DISCUSSION

The results of applying various ML models to predict CO2 and greenhouse gas emissions reveal significant differences in model performance, as presented in Table 3.

TABLE III. THE RESULTS

ML Method	MSE	R^2
LR	4718.99	0.8678
DT	3684.03	0.8968
FR	3043.89	0.9147
SVR	36771.12	-0.0295
KNN	11757.80	0.6708
XGBRegressor	20658.38	0.4216
Gradient Boosting Regressor	2535.30	0.9290
Ridge	4718.99	0.8678
Lasso	4719.25	0.8678

LR demonstrated a good fit, achieving an R-squared value of 0.868 and an MSE of 4718.99. This finding indicates that while the model explains a substantial portion of the variance in the data, there is room for improvement, particularly in reducing the error. DT regression improved upon LR, with an MSE of 3684.03 and an R-squared of 0.897, showing its ability to capture nonlinear relationships within the data. RF regression further enhanced the predictive accuracy, achieving an MSE of 3043.90 and an R-squared value of 0.915, reflecting its effectiveness in handling the complexity and variability of the data.

The gradient boosting regressor emerged as the best-performing model, with the lowest MSE of 2535.30 and the highest R-squared value of 0.929. This suggests that gradient boosting is particularly well suited for this predictive task, as it effectively balances bias and variance. However, SVR performed poorly, with an MSE of 36771.13 and an R-squared of -0.030, indicating that it was unsuitable for this dataset. Similarly, KNN regression also struggled, with an MSE of 11757.81 and an R-squared of 0.671.

The XGB Regressor and Lasso models showed moderate performance, with the XGB Regressor achieving an MSE of 20658.39 and an R-squared of 0.422, whereas Lasso had an MSE of 4719.25 and an R-squared of 0.868. Ridge regression had similar results to linear regression, indicating that regularization did not significantly improve the model's performance for this dataset.

Figure 13 illustrates the relationship between the actual and predicted values across various regression models, with each point representing a prediction. The dashed diagonal line serves as a reference for perfect predictions; points closer to this line indicate better model performance. This analysis underscores the importance of selecting the appropriate regression model on the basis of the dataset's characteristics and prediction objectives.

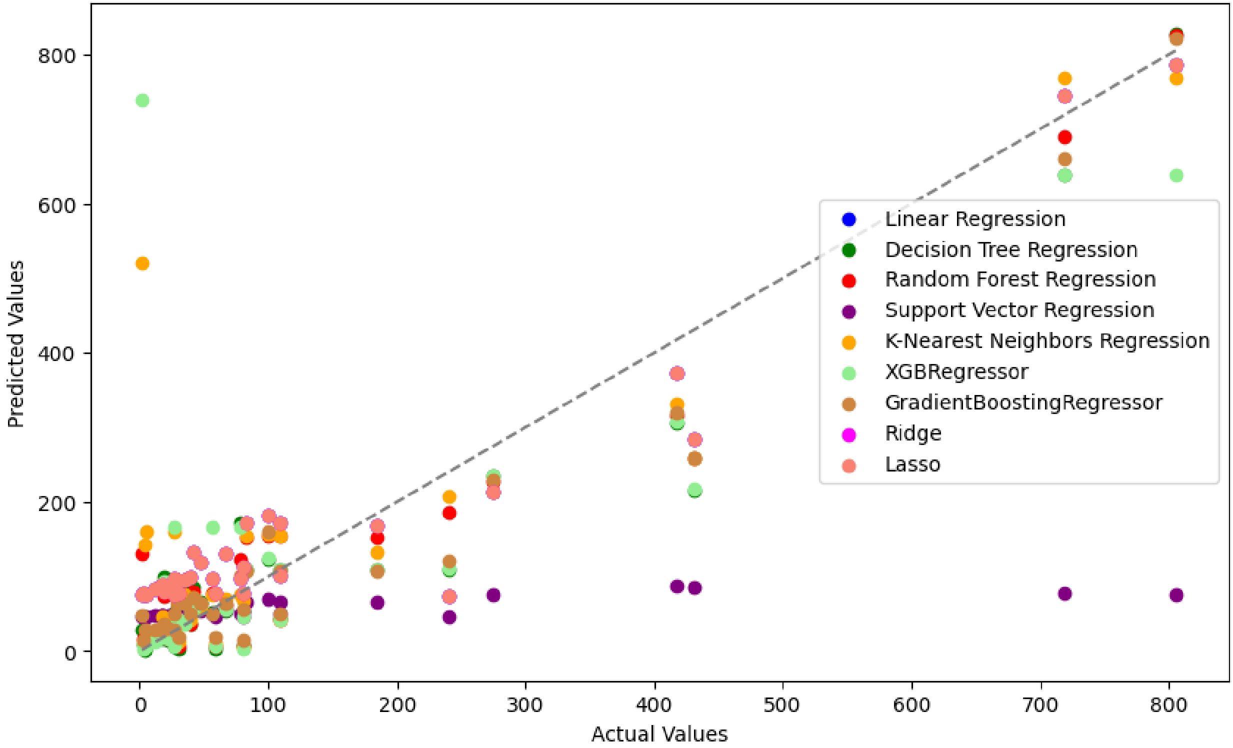

Fig. 13. Relationships between actual and predicted values across various regression models

These results highlight the effectiveness of ensemble methods, particularly gradient boosting, in handling complex environmental datasets. However, the limitations of some models, such as support vector regression and k-nearest neighbors, underscore the importance of model selection on the basis of the specific characteristics of the data.

5. CONCLUSION AND FUTURE WORK

In this work, emissions of CO2 and other greenhouse gases across various continents were assessed, and emission forecasts from regression models were used to determine the impacts of these gases on the world environment. Thus, the gradient boosting regressor is one of the most effective models, providing the highest accuracy, with an R squared of 0.929, and the lowest mean squared error (MSE), the minimum of which is recorded as 2535.30. The better performance of this model can be attributed to the increased handling capability for the emission data’s underlying pattern. There are other such models that are also helpful to some extent, but the accuracy and predictive ability of these models are not consistent and therefore make important the choice of the right model. Therefore, the findings of this study offer important information to political decision-makers and environmental scholars. Emission predictions are vital when measures are taken in the fight against global warming. The model’s capacity to highlight which configurations have the highest accuracy in various situations may help in determining the best direction for future studies to take and in improving existing measures to encourage environmental protection. As such, subsequent studies should examine other factors and more sophisticated methods for increasing the expected level of prognostic precision. Furthermore, it can consider actual time data and analyze regional characteristics to obtain more comprehensible emission trends. If analyses are repeated and improved upon, then the problems presented by global greenhouse gas emissions can be more suitably handled, and lasting solutions for earth can be provided.

Conflicts of Interest

The authors declare no conflicts of interest.

Funding

This research received no external funding.

Acknowledgment

The author extends appreciation to the institution for their unwavering support and encouragement during the course of this research.

References

[1] M. Kabir *et al.*, “Climate change due to increasing concentration of carbon dioxide and its impacts on environment in 21st century; a mini review,” *J. King Saud Univ. - Sci.*, vol. 35, no. 5, p. 102693, Jul. 2023, doi: 10.1016/j.jksus.2023.102693.

[2] “Energy and the environment explained,” 2024.

[3] A. S. Albahri *et al.*, “A systematic review of trustworthy artificial intelligence applications in natural disasters,” *Comput. Electr. Eng.*, vol. 118, p. 109409, 2024, doi: 10.1016/j.compeleceng.2024.109409.

[4] A. S. Albahri, Y. L. Khaleel, and M. A. Habeeb, “The Considerations of Trustworthy AI Components in Generative AI; A Letter to Editor,” *Appl. Data Sci. Anal.*, vol. 2023, pp. 108–109, 2023, doi: 10.58496/adsa/2023/009.

[5] L. A. E. Al-saeedi *et al.*, “Artificial Intelligence and Cybersecurity in Face Sale Contracts: Legal Issues and Frameworks ,” *Mesopotamian J. CyberSecurity*, vol. 4, no. 2 SE-Articles, pp. 129–142, Aug. 2024, doi: 10.58496/MJCS/2024/0012.

[6] K. O. Yoro and M. O. Daramola, “Chapter 1 - CO2 emission sources, greenhouse gases, and the global warming effect,” in *Advances in Carbon Capture*, M. R. Rahimpour, M. Farsi, and M. A. Makarem, Eds., Woodhead Publishing, 2020, pp. 3–28. doi: https://doi.org/10.1016/B978-0-12-819657-1.00001-3.

[7] R. Prasad *et al.*, “Role of Microalgae in Global CO2 Sequestration: Physiological Mechanism, Recent Development, Challenges, and Future Prospective,” *Sustainability*, vol. 13, no. 23, 2021, doi: 10.3390/su132313061.

[8] U. A. Bhatti *et al.*, “Global production patterns: Understanding the relationship between greenhouse gas emissions, agriculture greening and climate variability,” *Environ. Res.*, vol. 245, p. 118049, 2024, doi: https://doi.org/10.1016/j.envres.2023.118049.

[9] M. A. Salam and T. Noguchi, “Impact of Human Activities on Carbon Dioxide (CO2) Emissions: A Statistical Analysis,” *Environmentalist*, vol. 25, no. 1, pp. 19–30, 2005, doi: 10.1007/s10669-005-3093-4.

[10] W. F. Lamb *et al.*, “A review of trends and drivers of greenhouse gas emissions by sector from 1990 to 2018,” *Environ. Res. Lett.*, vol. 16, no. 7, p. 73005, 2021.

[11] K. R. Shivanna, “Climate change and its impact on biodiversity and human welfare,” *Proc. Indian Natl. Sci. Acad.*, vol. 88, no. 2, pp. 160–171, 2022, doi: 10.1007/s43538-022-00073-6.

[12] M. K. Jha and M. Dev, “Impacts of Climate Change,” in *Smart Internet of Things for Environment and Healthcare*, M. Azrour, J. Mabrouki, A. Alabdulatif, A. Guezzaz, and F. Amounas, Eds., Cham: Springer Nature Switzerland, 2024, pp. 139–159. doi: 10.1007/978-3-031-70102-3_10.

[13] A. Kumar, S. Nagar, and S. Anand, “1 - Climate change and existential threats,” in *Global Climate Change*, S. Singh, P. Singh, S. Rangabhashiyam, and K. K. Srivastava, Eds., Elsevier, 2021, pp. 1–31. doi: https://doi.org/10.1016/B978-0-12-822928-6.00005-8.

[14] L. Chen *et al.*, “Strategies to achieve a carbon neutral society: a review,” *Environ. Chem. Lett.*, vol. 20, no. 4, pp. 2277–2310, 2022, doi: 10.1007/s10311-022-01435-8.

[15] M.-T. Huang and P.-M. Zhai, “Achieving Paris Agreement temperature goals requires carbon neutrality by middle century with far-reaching transitions in the whole society,” *Adv. Clim. Chang. Res.*, vol. 12, no. 2, pp. 281–286, 2021, doi: https://doi.org/10.1016/j.accre.2021.03.004.

[16] H. Li, X. Jin, R. Zhao, B. Han, Y. Zhou, and P. Tittonell, “Assessing uncertainties and discrepancies in agricultural greenhouse gas emissions estimation in China: A comprehensive review,” *Environ. Impact Assess. Rev.*, vol. 106, p. 107498, 2024, doi: https://doi.org/10.1016/j.eiar.2024.107498.

[17] J. Udoh, J. Lu, and Q. Xu, “Application of Machine Learning to Predict CO2 Emissions in Light-Duty Vehicles,” *Sensors*, vol. 24, no. 24, 2024, doi: 10.3390/s24248219.

[18] A. H. Alamoodi, M. S. Al-Samarraay, O. S. Albahri, M. Deveci, A. S. Albahri, and S. Yussof, “Evaluation of energy economic optimization models using multi-criteria decision-making approach,” *Expert Syst. Appl.*, vol. 255, p. 124842, 2024, doi: 10.1016/j.eswa.2024.124842.

[19] M. Talal, A. H. Alamoodi, O. S. Albahri, A. S. Albahri, and D. Pamucar, “Evaluation of remote sensing techniques-based water quality monitoring for sustainable hydrological applications: an integrated FWZIC-VIKOR modelling approach,” *Environ. Dev. Sustain.*, vol. 26, no. 8, pp. 19685–19729, 2024, doi: 10.1007/s10668-023-03432-5.

[20] D. David *et al.*, “Correction: Sign language mobile apps: a systematic review of current app evaluation progress

and solution framework," *Evol. Syst.*, vol. 15, no. 5, p. 1989, 2024, doi: 10.1007/s12530-024-09600-w.

[21] M. A. Habeeb, Y. L. Khaleel, R. D. Ismail, Z. T. Al-Qaysi, and A. F. N., "Deep Learning Approaches for Gender Classification from Facial Images," *Mesopotamian J. Big Data*, vol. 2024, pp. 185–198, 2024, doi: 10.58496/MJBD/2024/013.

[22] T. J. Mohammed *et al.*, "A Systematic Review of Artificial Intelligence in Orthopaedic Disease Detection: A Taxonomy for Analysis and Trustworthiness Evaluation," *Int. J. Comput. Intell. Syst.*, vol. 17, no. 1, p. 303, 2024, doi: 10.1007/s44196-024-00718-y.

[23] A. J. Sietsma, "The First Generation of Machine Learning Applications for Tracking Climate Change Adaptation." University of Leeds, 2023.

[24] S. Arya, "Harnessing Big Data and Machine Learning for Climate Change Predictions and Mitigation," in *Maintaining a Sustainable World in the Nexus of Environmental Science and AI*, IGI Global, 2024, pp. 383–408.

[25] T. Li, A. Li, and Y. Song, "Development and Utilization of Renewable Energy Based on Carbon Emission Reduction—Evaluation of Multiple MCDM Methods," *Sustainability*, vol. 13, no. 17, 2021, doi: 10.3390/su13179822.

[26] A. Baklanov *et al.*, "Integrated urban services: Experience from four cities on different continents," *Urban Clim.*, vol. 32, p. 100610, 2020.

[27] M. SaberiKamarposhti, N. K. Why, M. Yadollahi, H. Kamyab, J. Cheng, and M. Khorami, "Cultivating a sustainable future in the artificial intelligence era: A comprehensive assessment of greenhouse gas emissions and removals in agriculture," *Environ. Res.*, p. 118528, 2024.

[28] S. E. Bibri, J. Krogstie, A. Kaboli, and A. Alahi, "Smarter eco-cities and their leading-edge artificial intelligence of things solutions for environmental sustainability: A comprehensive systematic review," *Environ. Sci. Ecotechnology*, vol. 19, p. 100330, 2024.

[29] A. Hamrani, A. Akbarzadeh, and C. A. Madramootoo, "Machine learning for predicting greenhouse gas emissions from agricultural soils," *Sci. Total Environ.*, vol. 741, p. 140338, Nov. 2020, doi: 10.1016/j.scitotenv.2020.140338.

[30] M. Emami Javanmard and S. F. Ghaderi, "A Hybrid Model with Applying Machine Learning Algorithms and Optimization Model to Forecast Greenhouse Gas Emissions with Energy Market Data," *Sustain. Cities Soc.*, vol. 82, p. 103886, Jul. 2022, doi: 10.1016/j.scs.2022.103886.

[31] I. Ulku and E. E. Ulku, "Forecasting Greenhouse Gas Emissions Based on Different Machine Learning Algorithms," 2022, pp. 109–116. doi: 10.1007/978-3-031-09176-6_13.

[32] X. Li, A. Ren, and Q. Li, "Exploring Patterns of Transportation-Related CO2 Emissions Using Machine Learning Methods," *Sustainability*, vol. 14, no. 8, p. 4588, Apr. 2022, doi: 10.3390/su14084588.

[33] S. Giannelos, F. Bellizio, G. Strbac, and T. Zhang, "Machine learning approaches for predictions of CO2 emissions in the building sector," *Electr. Power Syst. Res.*, vol. 235, p. 110735, Oct. 2024, doi: 10.1016/j.epsr.2024.110735.

[34] "CO2 and Greenhouse Gas Emissions by Region." https://www.kaggle.com/datasets/shahriarkabir/co2-and-greenhouse-gas-emissions-by-region/ (accessed Dec. 27, 2024).

[35] F. K. H. Mihna, M. A. Habeeb, Y. L. Khaleel, Y. H. Ali, and L. A. E. Al-saeedi, "Using Information Technology for Comprehensive Analysis and Prediction in Forensic Evidence," *Mesopotamian J. CyberSecurity*, vol. 2024, pp. 4–16, Mar. 2024, doi: 10.58496/MJCS/2024/002.

[36] M. A. Habeeb, Y. L. Khaleel, and A. S. Albahri, "Toward Smart Bicycle Safety: Leveraging Machine Learning Models and Optimal Lighting Solutions," in *Proceedings of the Third International Conference on Innovations in Computing Research (ICR'24)*, K. Daimi and A. Al Sadoon, Eds., Cham: Springer Nature Switzerland, 2024, pp. 120–131.

[37] Y. L. Khaleel, M. A. Habeeb, and G. G. Shayea, "Integrating Image Data Fusion and ResNet Method for Accurate Fish Freshness Classification," *Iraqi J. Comput. Sci. Math.*, vol. 5, no. 4, p. 21, 2024.

[38] R. D. Ismail, Q. A. Hameed, M. A. Habeeb, Y. L. Khaleel, and F. N. Ameen, "Deep Learning Model for Hand Movement Rehabilitation," *Mesopotamian J. Comput. Sci.*, vol. 2024, no. SE-Articles, pp. 134–149, Oct. 2024, doi: 10.58496/MJCSC/2024/011.

[39] F. M. Ottaviani and A. De Marco, "Multiple Linear Regression Model for Improved Project Cost Forecasting," *Procedia Comput. Sci.*, vol. 196, pp. 808–815, 2022, doi: 10.1016/j.procs.2021.12.079.

[40] A. A. Mahamat *et al.*, "Decision Tree Regression vs. Gradient Boosting Regressor Models for the Prediction of Hygroscopic Properties of Borassus Fruit Fiber," *Appl. Sci.*, vol. 14, no. 17, p. 7540, Aug. 2024, doi: 10.3390/app14177540.

[41] H. M. Abdulfattah, K. Y. Layth, and A. A. Raheem, "Enhancing Security and Performance in Vehicular Adhoc Networks: A Machine Learning Approach to Combat Adversarial Attacks," *Mesopotamian J. Comput. Sci.*, vol. 2024, pp. 122–133, 2024, doi: 10.58496/MJCSC/2024/010.

[42] Y. O. Ouma *et al.*, "Land-Use Change Prediction in Dam Catchment Using Logistic Regression-CA, ANN-CA and

Random Forest Regression and Implications for Sustainable Land–Water Nexus," *Sustainability*, vol. 16, no. 4, p. 1699, Feb. 2024, doi: 10.3390/su16041699.

[43] M. S. Ahmad, S. M. Adnan, S. Zaidi, and P. Bhargava, "A novel support vector regression (SVR) model for the prediction of splice strength of the unconfined beam specimens," *Constr. Build. Mater.*, vol. 248, p. 118475, Jul. 2020, doi: 10.1016/j.conbuildmat.2020.118475.

[44] G. Lin, A. Lin, and D. Gu, "Using support vector regression and K-nearest neighbors for short-term traffic flow prediction based on maximal information coefficient," *Inf. Sci. (Ny).*, vol. 608, pp. 517–531, Aug. 2022, doi: 10.1016/j.ins.2022.06.090.

[45] Zhagparov, Z. Buribayev, S. Joldasbayev, A. Yerkosova, and M. Zhassuzak, "Building a System for Predicting the Yield of Grain Crops Based On Machine Learning Using the XGBRegressor Algorithm," in *2021 IEEE International Conference on Smart Information Systems and Technologies (SIST)*, IEEE, Apr. 2021, pp. 1–5. doi: 10.1109/SIST50301.2021.9465938.

[46] D. A. Otchere, T. O. A. Ganat, J. O. Ojero, B. N. Tackie-Otoo, and M. Y. Taki, "Application of gradient boosting regression model for the evaluation of feature selection techniques in improving reservoir characterisation predictions," *J. Pet. Sci. Eng.*, vol. 208, p. 109244, Jan. 2022, doi: 10.1016/j.petrol.2021.109244.

[47] L. A. Geraldo-Campos, J. J. Soria, and T. Pando-Ezcurra, "Machine Learning for Credit Risk in the Reactive Peru Program: A Comparison of the Lasso and Ridge Regression Models," *Economies*, vol. 10, no. 8, p. 188, Jul. 2022, doi: 10.3390/economies10080188.

[48] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Comput. Sci.*, vol. 7, p. e623, Jul. 2021, doi: 10.7717/peerj-cs.623.