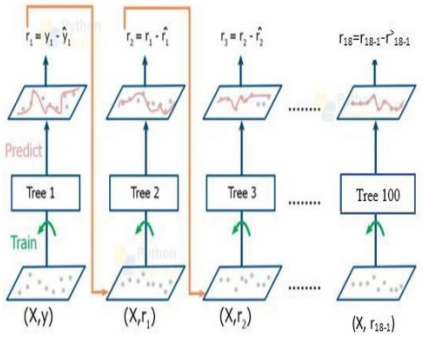
A comparative analysis and prediction of carbon emission in India using Machine learning models

Venkatesa Palanichamy Narasimma Bharathi[1], Kalpana Muthuswamy[2*], Balakrishnan Natarajan[3*], Shantha Sheela M[4], Gitanjali Jothiprakash[5], Kavitha Shanmugam[6], Karthiba Loganathan[7], Balamurugan Vasudevan[8], Suresh Appavu[8], Rajavel Marimuthu[9], Dhivya Rajaram[10] and Syndhiya Ranjan[11]

[1]Dean (Agriculture), Tamil Nadu Agricultural University, Coimbatore – 641 003,
[2*]Professor (Computer Science), Office of the Dean (Agriculture), Tamil Nadu Agricultural University, Coimbatore – 641 003
[3*]Teaching Assistant, Office of the Dean (Agriculture), Tamil Nadu Agricultural University, Coimbatore – 641 003.
[4*]Professor (Agrl Extension), Directorate of Agribusiness Development, Tamil Nadu Agricultural University, Coimbatore – 641 003
[5]Teaching Assistant, Centre for Post Harvest Technology, AEC & RI, Tamil Nadu Agricultural University, Coimbatore – 641 003.
[6]Associate Professor (Seed Science), Department of Seed Science and Technology, Tamil Nadu Agricultural University, Coimbatore – 641 003
[7]Assistant Professor, Department of Pulses, Tamil Nadu Agricultural University, Coimbatore – 641 003
[8]Research Scholar, Office of the Dean (Agriculture), Tamil Nadu Agricultural University, Coimbatore – 641 003
[9]Associate Professor (Crop Physiology), Office of the Public Relations, Tamil Nadu Agricultural University, Coimbatore – 641 003
[10]Assistant Professor, Department of Business Administration, PSGR Krishnammal College for Women, Coimbatore – 641004
[11]Research Associate, Centre for Agricultural Nanotechnology, Tamil Nadu Agricultural University, Coimbatore – 641 003
*Corresponding author: blkrshnnnatarajan@gmail.com

Graphical Abstract



Abstract

Among the most significant issues that will impact and pose a major threat to our future are global warming and climate change. This study focuses on forecasting India's carbon emissions, specifically examining the role of agriculture and agri-related industries, using time-series data from 1990 to 2020. To achieve this, a machine learning-based approach was proposed, utilizing five advanced models AdaBoost, XGB, Gradient Boosting, Random Forest, and Linear Regression. These models were selected for their ability to process large datasets, identify complex patterns, and provide precise predictions, offering a comprehensive framework for estimating carbon emissions. The approach aims to identify the most accurate predictive model for estimating carbon emissions, facilitating informed policy interventions and strategic planning. The Carbon Budget 2023 research highlights that India's carbon emissions increased by 8.2% in 2023, raising significant concerns for environmental sustainability. Using the proposed machine learning models, the study facilitates robust data prediction. Among the five models, XGB demonstrates superior performance with evaluation metrics such as a Mean Absolute Error (MAE) of 19843, Root Mean Square Error (RMSE) of 24022, Mean Squared Error (MSE) of 5770998, R^2 value of 0.9, and test accuracy of 99%. These findings establish XGB as the most reliable model for forecasting carbon emissions. Consequently, XGB emerges as a powerful tool for accurately estimating emissions and addressing the challenges posed by climate change.

Keywords: Carbon emission, India, Machine learning, Prediction

Introduction

These days, countries all across the world are concerned about global warming. More than 95% of scientists who collaborate with the Intergovernmental Panel on Climate Change (IPCC) concur with the main causes of global warming are other human (anthropogenic) actions and growing greenhouse gas concentrations [1]. Increase in the amount of corrosive gases which are frequently referred to as greenhouse gas (GHGs) emissions and include perfluorocarbons (PFC), hydrofluorocarbons (HFC), nitrous oxide (N_2O), carbon dioxide (CO_2), and methane (CH_4)-affects the equilibrium between the Earth and atmosphere. Specifically, one of the main contributors to global warming is CO_2 [2]. It's well recognized that the greenhouse effect causes CO_2 emissions to play a major role in global warming. The question of whether or not CO_2 causes climate change is up for debate [3], but it is generally accepted [4,5] that CO_2 emissions are the main factor contributing to global warming. Thirteen

percent of greenhouse gas emissions and, by extension, climate variation are attributable to agriculture. Climate variability can impact agriculture through a variety of means, such as altered precipitation patterns, an increased frequency of floods, fluctuations in the average temperature, and rising sea levels [6]. Global agricultural output is significantly impacted by climate change. The United Nations Department of Economic and Social Affairs (UN-DESA) has projected that, there will be 9.7 billion people on Earth by 2050. Producing enough food to meet the demands of the expanding population is imperative. As a result, farmers employ a range of strategies, such as fertilizer application, crop residue management, soil management, land management, etc. These methods have an adverse effect on the environment and often cause an increase in the emissions of greenhouse gases (GHGs) in order to meet needs such as food consumption, crop yield, etc. Reducing greenhouse gas emissions is one way to mitigate climate variation. As a result, it's essential to forecast and assess the greenhouse gas emissions produced by the agricultural sector. Carbon dioxide (CO_2) is the primary component that contributes to the release of greenhouse gasses, which accounts for 81% of all gas emissions although making up only 1% of atmospheric gases [7]. The main cause of climate change is the atmospheric release of carbon dioxide, which endangers many aspects of human existence. The primary cause of climate change is CO_2 emissions, which also make global warming an increasing issue. [8]. Countries all across the world are tackling this issue head-on as the state of the global climate declines. India, the world's third-largest emitter, reported CO_2 emissions in 2022 of 2830 $MtCo_2$, or 7.6% of global CO_2 emissions. It is projected that India, which contributes 8% of global emissions, will experience an 8.2% (range 6.7% to 9.7%) increase in emissions in 2023 compared to 2022. The country's emissions are expected to increase from coal (+9.5%), natural gas (+5.6%), oil (+5.3%) and cement (+8.8%). One of the main factors contributing to coal's rise is the fast rise in electricity consumption, which new renewable energy sources are unable to meet. Consolidated data indicates that since 2022, India's emissions have exceeded those of the European Union [9].

Related works

Several studies have been written with the intention of utilizing various prediction models to forecast carbon emissions. The prediction of emissions frequently used traditional statistical models like, ARIMA, GM, SARIMAX, and numerous other models. In contrast, the volume, instability, and nonlinearity of time series data are increasing, and this makes standard statistical prediction techniques obsolete [10]. Researchers have thus focused on machine learning methods, notably deep learning techniques, to address the challenges encountered with

time series data that is not linear. These prediction-based machine learning models include, among others RF, ANN, LSTM and SVM. The increasing prevalence of these models can be attributed to their capacity to manage nonlinearity and detect complex patterns within the data. Recent advancements in high-dimensional models and supervised machine learning (ML), such as the use of ensemble and hybrid models with different inputs and linear and non-linear forecasting techniques [11]. A recurrent neural network model with LSTM was utilized to forecast the carbon emissions originating from Rwandan agriculture. High levels of prediction accuracy (97.64%) and loss accuracy (2.36%) were shown in their findings [12]. B. Khoshnevisan *et al.* have employed ANN to forecast greenhouse gas emissions related to strawberry cultivation. Their data indicated that the primary cause of greenhouse gas emissions was chemical fertilizer. The 11-6-10-2 structure of the ANN model works best when forecasting the generated electricity and emissions of greenhouse gases, resulting in lower coefficients for RMSE and MAE measures [13]. Kalra *et al*. employed DT, ANN, ANN, and linear regression to replicate the 65-year association between atmospheric concentrations of CO_2, N_2O, and CH_4. According to the authors, ANN outperformed the other methods on the basis of the MSE [14]. Magazzino and Mele investigated the connection between the nation's GDP, energy use, and emissions of carbon dioxide using yearly statistics ranging from 1970 to 2017. Additionally, they developed a brand-new machine learning method known as causal direction through dependency. The researchers' findings [15] provided evidence in favor of the three elements' causal relationships. As a result, Ahmed *et al.* [16] further utilized three state-of-the-art machine learning techniques-ANN, SVM, RF and LSTM to evaluate how energy consumption affects the release of greenhouse gases in the top emitters of the world. The findings demonstrated that although greenhouse gas emissions were decreasing in Russia and the USA, they were increasing in China and India in the future [17]. In another study that also employed the LSTM model, the factors that significantly contributing to greenhouse gas emissions in China and India were population increase, GDP expansion, energy consumption, and the development of the financial sector. Bakay and Abulut [18] used Deep Learning techniques, SVM and ANN models to anticipate greenhouse gas emissions from Turkey's power production industry using a dataset that covered the years 1990 to 2018. The models' R2 values, which varied from 0.861% to 0.998%, led the authors to the conclusion that each model could accurately predict greenhouse gas emissions. From 2000 to 2018, Li *et al*. [19] looked at the connections between China's CO_2 emissions and its industrial trends, urbanization, R&D spending, foreign investment, and energy consumption. Artificial Neural Networks (ANN),

ensemble models, and k-nearest neighbors (KNN) were employed. The KNN model yielded the highest accurate forecasts compared to the other models used in the investigation.

The literature review reveals several key insights into the prediction of carbon emissions. The increasing complexity of predicting carbon emissions due to the growing volume and nonlinearity of data. Traditional statistical models like ARIMA and SARIMAX are found to be less effective as they cannot handle the instability and nonlinearity present in large datasets. In contrast, machine learning models, particularly deep learning techniques such as LSTM, ANN, and ensemble models, are proving more adept at managing these complexities. Previous studies demonstrated the effectiveness of models such as ANN and LSTM in capturing complex patterns, inspiring the use of advanced algorithms like RF and XGB in this research. The literature underscored the importance of accurate predictive tools for policy interventions, aligning with the study objective of identifying the most reliable model for forecasting carbon emissions. This knowledge base helped refine the research methodology and validate the choice of algorithms for better accuracy and performance.

Materials and methods

Data Collection and preparation

This study utilized carbon emissions data spanning 30 years, from 1990 to 2020, sourced from FAOSTAT, a globally recognized database that consolidates data across all industries and greenhouse gases [20]. FAOSTAT was selected for its comprehensiveness and reliability in providing high-quality datasets for climate research. The data highlights that approximately 60% of global emissions are attributed to the top 10 emitters, as shown in Figure 1. The collected data was initially stored in Excel format for ease of accessibility. However, as the analytical tools used in the study (Python) did not support direct processing of Excel files in this context, the data was converted to the Comma-Separated Values (CSV) format. This format was chosen for its compatibility with both software, allowing seamless integration into the machine learning workflow.

To ensure accuracy and model reliability, pre-processing steps were meticulously performed on the dataset. These steps included cleaning missing or inconsistent data, normalizing values, and ensuring uniformity across all records. Following this, the dataset was split into training and testing samples (80:20 ratio) using an appropriate ratio to enable effective model training and evaluation. The training data was used to build and fine-tune machine

learning models, while the test data was reserved for assessing the predictive performance and accuracy of these models.

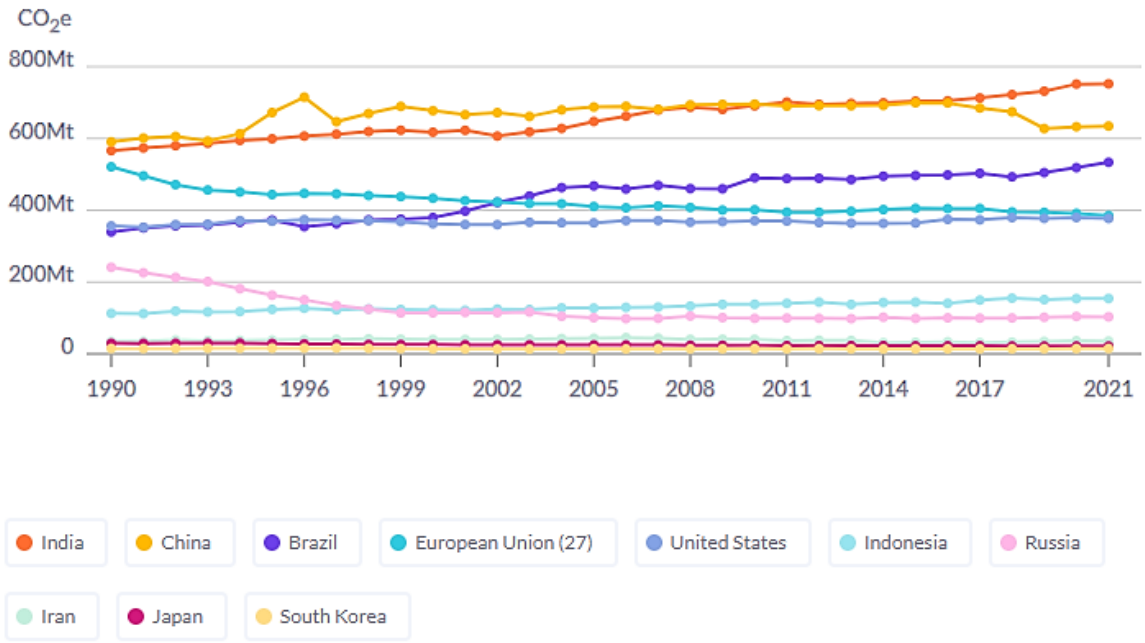


Figure 1. illustrates annual carbon dioxide equivalent emissions (in **$MtCO_2e$**, metric tons of CO_2 equivalent) from 1990 to 2020 for the top global emitters, including China, the United States, the European Union (27 countries), India, Russia, Indonesia, Japan, Brazil, Iran, and South Korea. Each line represents the emission trends for a specific country or region, with CO_2 emissions on the y-axis and years on the x-axis. These trends are important to understand the global contribution to climate change, and they provide context for the study's focus on India's rising carbon emissions, particularly from the agricultural and agri-related sectors. (Data Source: Climate Watch. 2024).

Machine Learning Models (MLMs) proposed

As was covered in the preceding part, India's carbon emissions are rapidly rising and posing a major threat to the ecosystem and, eventually, all life. Therefore, lowering these emissions should be the top priority for both the government and business. An accurate estimation of emissions might be beneficial for developing and implementing policies. Because of this, humans possess certain traits and data on carbon emissions for the last 30 years, from 1990 to 2020. In order to handle missing values, pre-processing and cleaning of a univariate dataset were required. Based on these historical statistics, we have used machine learning algorithms to project India's carbon emissions for the next ten years. Given the dataset's

stationary nature and rising trend, we employed five machine learning models for it: XGB, adaboost, Gradient Boosting, Random Forest, and Linear Regression. The proposed technique differs from existing models by incorporating multiple advanced machine learning algorithms to enhance the accuracy and reliability of carbon emissions forecasting. Unlike traditional statistical methods, which may rely on linear assumptions and fixed parameters, the machine learning models used in this study, such as XGB, AdaBoost and Random Forest, are capable of learning from complex, non-linear relationships in large datasets. These algorithms adapt to changes in the data and can detect intricate patterns without the need for predefined equations or assumptions about the data structure. Furthermore, the use of multiple models allows for comparison and validation, ensuring that the most accurate model is chosen for forecasting. This approach provides a more dynamic, flexible, and data-driven framework, distinguishing it from conventional models that may not fully capture the complexities of the factors influencing carbon emissions over time.

The main advantage of using machine learning algorithms for the predictive model is their ability to analyze large, complex datasets and identify intricate patterns that traditional statistical methods might overlook. These algorithms can handle non-linear relationships, interactions, and high-dimensional data, which enables more accurate and robust predictions, especially when forecasting complex phenomena like carbon emissions. MLMs can continuously improve their accuracy as more data becomes available, making them highly adaptable and effective for long-term forecasting.

Linear Regression

The Linear Regression technique was used because of the magnitude of the dataset and the fact that the prediction was quantitative rather than qualitative. Using linear regression, this approach looks into the distribution of a response variable, Y, that changes depending on how the intervening variable, X, is calculated. To offer a prediction, the value of the response variable must be eliminated by an accurate computation of the explanatory variable. Conversely, regression is the process of using the best straight line to ascertain the relationship between eighteen independent variables, one dependent variable. Below is an illustration of the regression equation.

$$Y = a + (b*X_1, b*X_2 \ldots\ldots b*X_{18}) + d$$

Total emission – Dependent variable (Y)

Crop Residues Year Rice cultivation, depleted organic soils (CO_2), production of pesticides, food consumption in households, disposal of waste from agrifood systems, production of fertilizers, application of manure to soils, the management of manure and energy use on farms Male and female populations, as well as those in rural and urban areas, Minimum Temp, Maximum Temp and Mean Rainfall) - Independent variable (X_1, X_2…… X_{18})

a) The intercept; b) The slope and d) The residual (error)

Random Forest Regressor

A supervised technique to perform classification regression is the Random Forest (RF) Regressor. In order to increase accuracy, the regression method uses ensemble learning, a bagging technique that combines individual decision trees. Time series forecasting can also be done with this method, albeit the outcomes might not be predictable. Before fitting, the data for this model must be correctly prepared. The data has been divided according to carbon emissions. The random forest algorithm's ability to tolerate numbers that are not present while maintaining accuracy is one of its advantages [21-24]. In this study, 100 trees are created, and predictions are made from each tree. The average of all the forecasts is also determined for carbon emission, as Figure 3 illustrates.

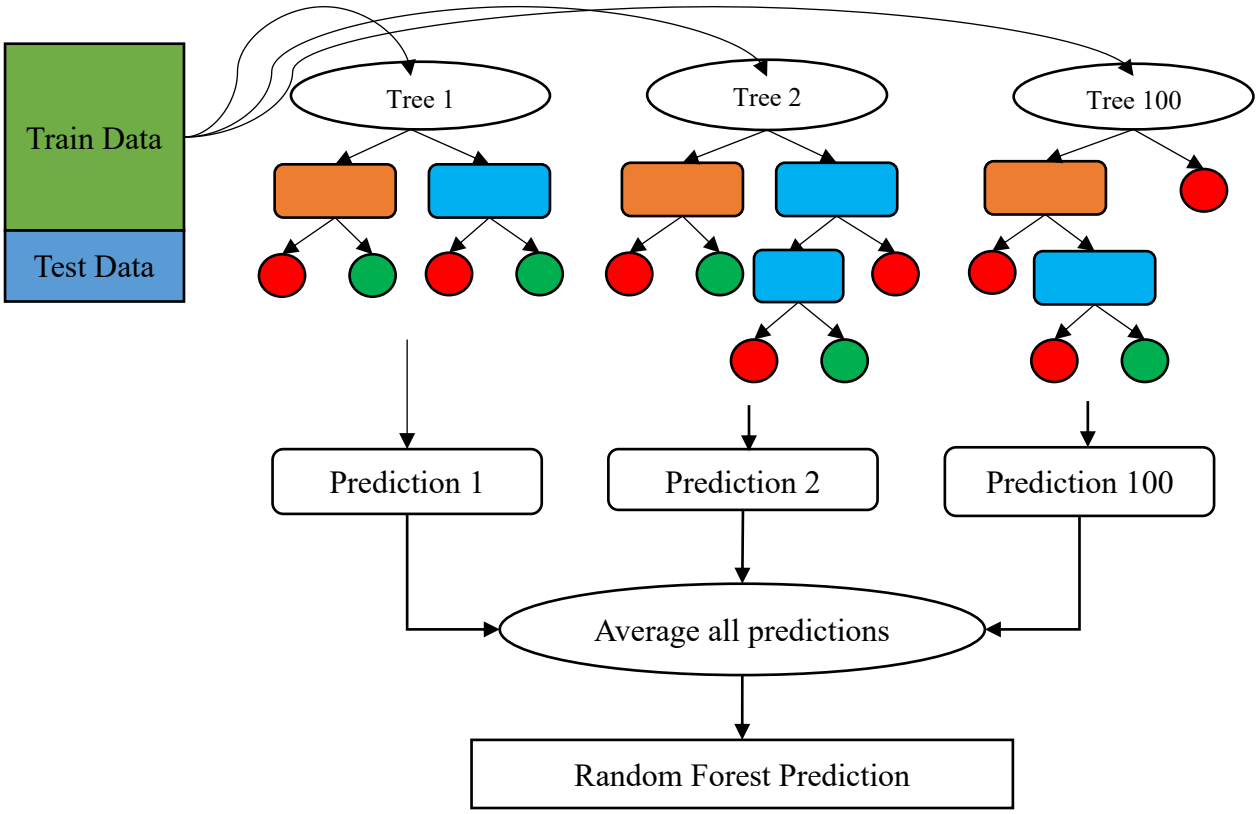


Figure 2. Random Forest Regressor

Gradient Boosting Regressor (GBR)

Initially, machine learning researchers developed boosting strategies to tackle classification problems. Several rudimentary models, referred to as "weak learners," are frequently combined in this technique to produce a "strong learner" with a higher prediction accuracy. Essentially, the GBR method is a numerical optimization process designed to find an addition model that minimizes the loss function. Consequently, in order to minimize the loss function, the GBR technique iteratively builds a new decision tree at each step. More specifically, in regression, the algorithm starts with an initial guess for the model, typically a decision tree which decreases the loss function of the regression. The new decision tree is subsequently integrated to the old model to update it, with the current residual being applied to it at each step. The process keeps repeating until the user-specified maximum number of repeats is achieved [25-28]. Because this technique is stage-wise, the decision trees that were introduced to the framework in earlier steps are not updated at each subsequent phase.

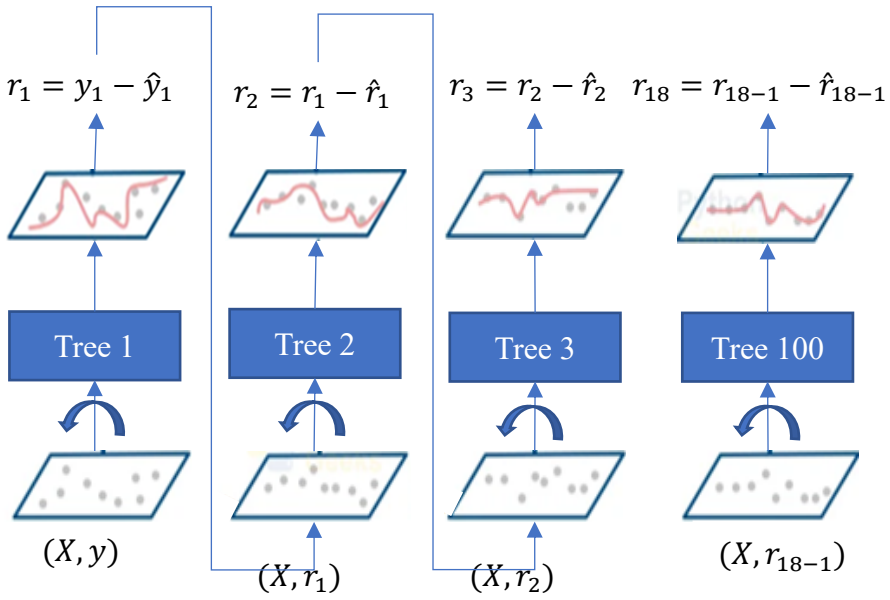


Figure 3. working model of Gradient boosting regressor and shows how decision trees are fitted to the residuals to enhance the model in areas where it performs poorly.

Adaboost regressor

A collection of numerous weak learner decision trees called adaptive boosting all outperform random guessing by an inadequate a margin. The AdaBoost technique that adapts improves the error of succeeding trees by utilizing the gradient of earlier trees. As a result of the subsequent tree learning at each stage, a powerful learner emerges. The final forecast consists of the weighted average of each tree's forecasts. Moreover, the method feeds future trees with knowledge from older trees so they can concentrate on challenging training samples [29-32]. AdaBoost employs 50 Decision Tree Classifiers in this investigation.

Algorithm,

1. Initialize the weights $w_{1,1}...w_{18,1}$ to (1/18) and also the overall count of weak students (h) for the training set (x_i, y_i).

2.Regarding g in 1 through 18.

I. Use the square loss function, E = L ($f(x_i)$, y_i), to calculate the error of each leaner.

II. Choose the less proficient learner that reduces error.

III. Concatenate it with the learning rate, A, in the tree-building algorithm

Fg(x) = Fg-1(x) + A*hgi.

IV. Revision of weights $w_{1,1}... w_{18,1}$.

3. The ultimate forecast is FG(x).

XGBoost regressor

Regression using an extreme gradient boost is called XGBoost regression. It performs exceptionally when measured against other learning algorithms. The structure of XGBoost, which consists of target functions and base learners, indicates that it is among the most effective supervised learning algorithms. The regularization term is used to show how much the actual value deviates from the anticipated value, the deviation between the actual and anticipated values is displayed by the decrease in the activity, which happens to be a component of the end result value [33-36]. XGBoost's ensemble learning uses numerous models, known as base learners, to forecast a single result. Since not every base learner is expected to produce a negative prediction, the poor predictions cancel out the good ones when put together. A regressor is a device that fits a model with predetermined features in order to anticipate an unknown output value. In Figure 5, the XGBoost regressor is displayed.

The XGBoost model is

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F$$

where K =100(number of decision-tree)

In the 100th decision-tree, $f_k(x_i)$ represents the input function, ˆ y_i represents the value that was predicted, and F represents the set of potential CARTs.

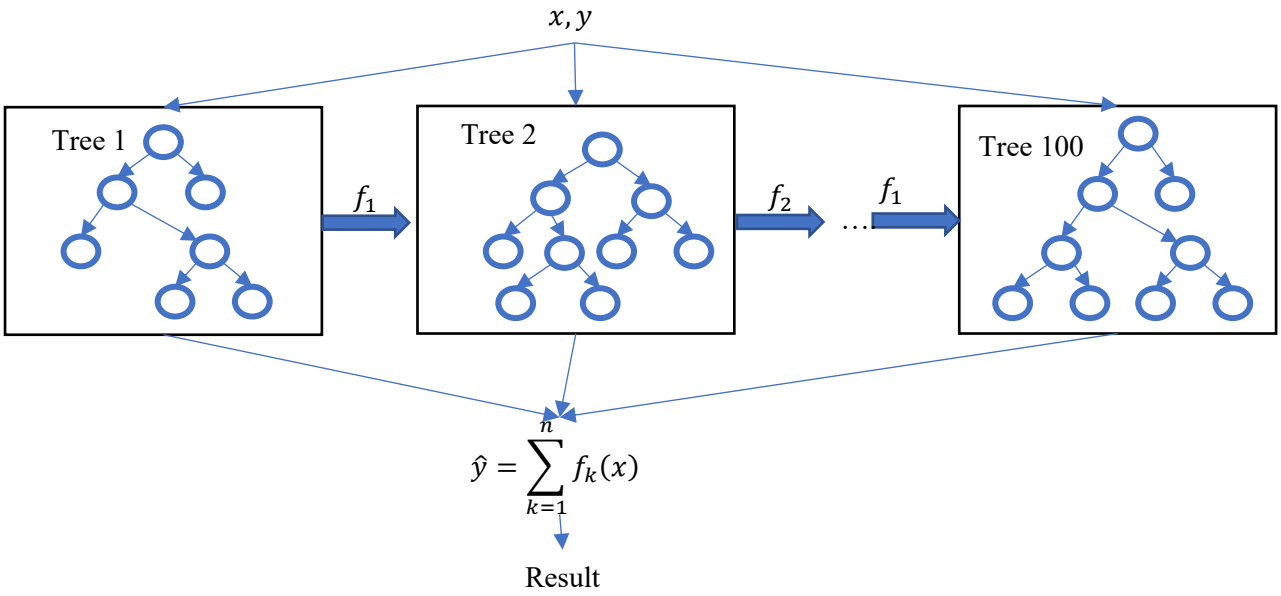


Figure 4. XGBoost regressor

The proposed model's framework is presented in the flowchart showed in Figure 5. It demonstrates how the machine learning models we used before have been implemented.

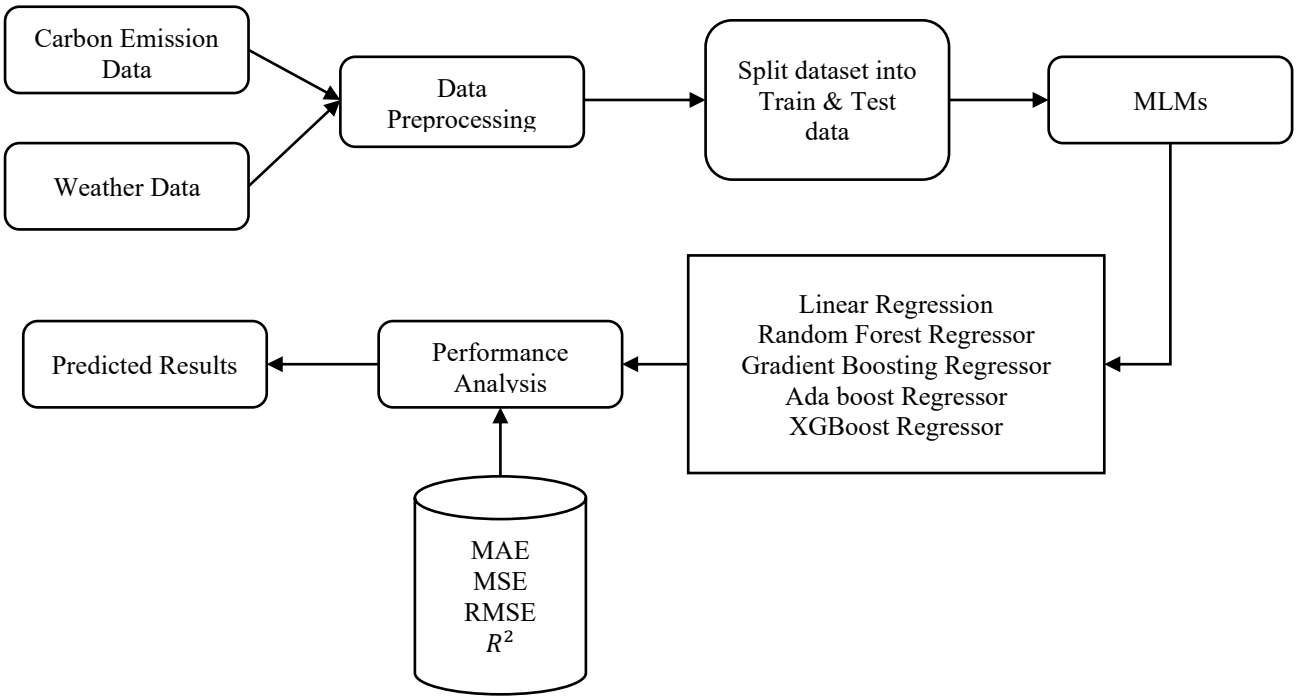


Figure 5. Proposed Framework for carbon emission forecasting

Four performance indicators were employed in order to assess the implemented models' correctness. For the ensuing ten years, from 2020 to 2030, carbon emissions are projected using the most effective models. Before the models were applied, the data underwent pre-processing to account for the missing variables. Training and testing sets of the pre-processed data were created. The models' performance on test data indicates that they are capable of accurately predicting carbon emissions for the ensuing ten years, up to 2030.

Evaluation metrics

Performance measures were covered in this part as a means to determine the models' efficiency. Regression is used by these models to predict carbon emissions. We used nine distinct evaluation factors to assess the models' efficacy. Before employing these measurements, we need to determine the residual error ($y-\hat{y}$). The values y and $\hat{y}$ represent the expected and actual values. The performance metric listed below was used to assess the models.

The sum of the absolute residual errors is known as the Mean Absolute Error (MAE). This suggests that whether it's positive or negative doesn't matter. The mean absolute error formula is given in Equation 1.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}| \quad (1)$$

The calculation of Mean Squared Error (MSE) involves first squaring the residual error for each data point, as previously mentioned, and then calculating the average. The mean squared error formula is given in Equation 2.

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

Where, respectively, y_i and $\hat{y}_i$ stand for the actual and predicted values. 'n' denotes the quantity of data points...

Root Mean Square Error (RMSE) for short, is a measurement of prediction errors. The dispersion of these mistakes is calculated using residuals, which quantify the deviation of data points from the regression line. In other words, RMSE quantifies how closely the data points cluster around the line of best fit.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2} \quad (3)$$

The degree to which the values fit together in respect to the initial values is indicated by the R-squared. Values ranging from 0 to 1 represent the percentages. The greater the value, the better the model

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2} \quad (4)$$

where in above three equations, y^ is the production's actual value and y is its expected value.

Tools used:

Python is the chosen programming language for this study, covering all stages from data extraction to model evaluation. The suggested design's performance is assessed through the creation of an application using the Python 3.6 programming environment. This choice is based on the fact that Python has strong library support in the fields of AI and machine learning, making it perfect for handling problems in the real world. Because Python is portable, it does

not require a specific operating system, which increases the project's accessibility and flexibility.

4. Results and Discussion

The implementation of the model covered in the previous section is shown in this section. Before using the model, we perform a thorough study of the data to get knowledge about the dataset. Because of this, the recommended model's performance is assessed once the statistical data are presented in this section.

Table 1 shows the evaluation criteria values that were determined by each model used in this study.

Evaluation metrics	LR	RF	GDR	Ada Boost	XGB
MAE	89052.05	33181.41	22889.01	22889.03	19843.28
MSE	14755187	95942629	7976705	7960569	5770998
RMSE	121470.90	50933.90	28243.06	28645.07	24022.09
R^2	0.75	0.956	0.987	0.986	0.990
Test accuracy (%)	75	95	98	87	99
Train accuracy(%)	99	99	99	99	99

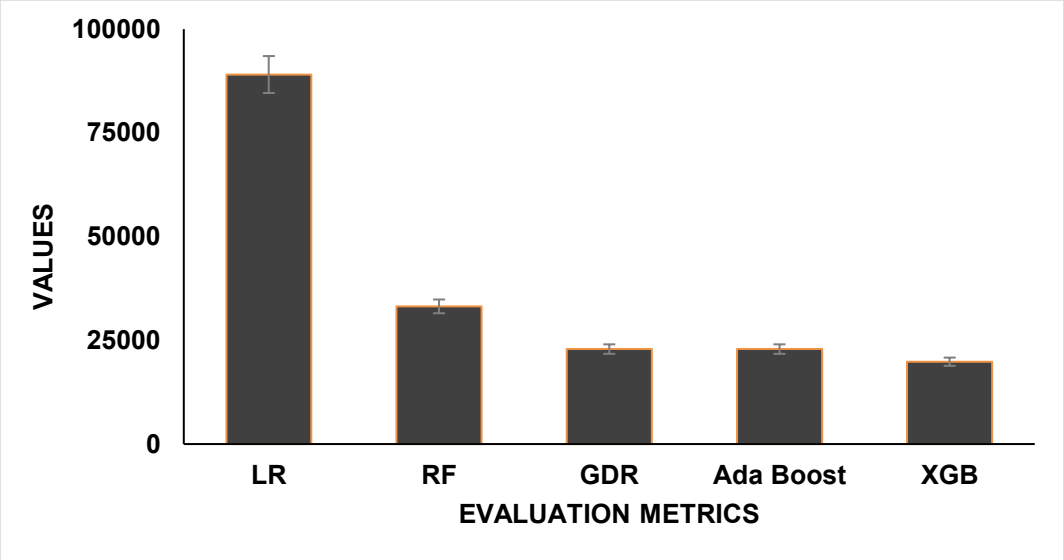


Figure 6. Performance analysis of MAE

A lower Mean Absolute Error (MAE) number indicates superior model performance, as can be seen by comparing the MAE values of the various models. With the lowest MAE value of 19843 and exceptional 0.99% test accuracy among the evaluated devices, the XGB model is clearly the best performance. With test accuracy of 98% and MAE values of 22889.01, the

GDR model trails closely behind. With a test accuracy of 95% and a higher MAE value of 33181, the RF model is still rather respectable. Lastly, the linear regression model has a lower test accuracy but the greatest MAE value (89052.05). The XGB model is the ideal one since it performs better than the others in terms of test accuracy and MAE values (Figure-6).

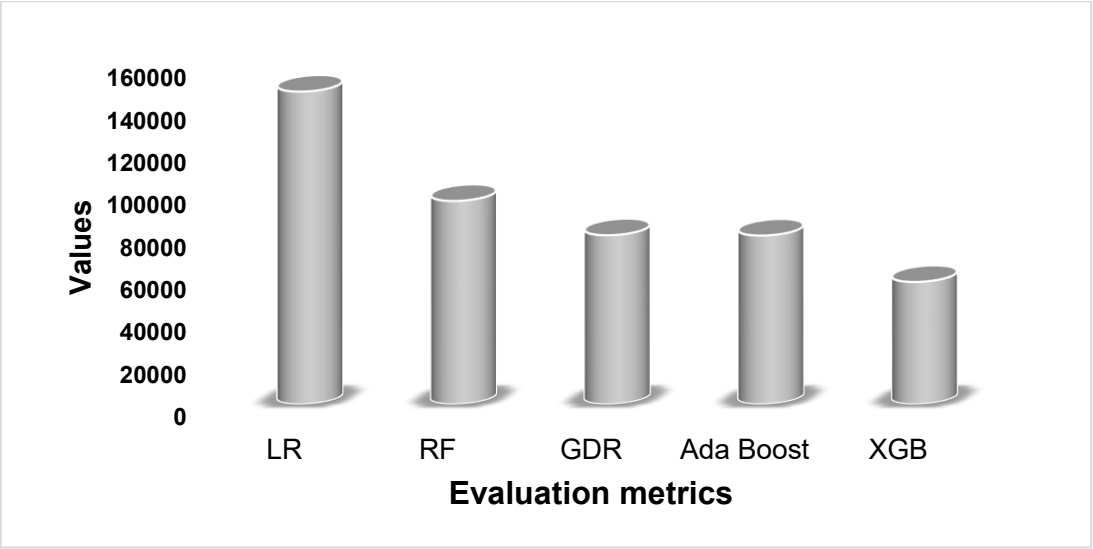


Figure 7. Performance analysis of MSE

As per the MSE definition, a lower score indicates a more robust prediction skill. In this experiment, the MSE values varied from 14755187 to 5770998. The results demonstrate that, in terms of MSE values, machine learning models perform better than statistical methods. At 5770998, the XGB model has the lowest MSE value. It is followed by the RF model at 9594262 and the Ada boost regressor model at 7976705. At last, the greatest value of 14755187 was obtained by linear regression (Figure-7).

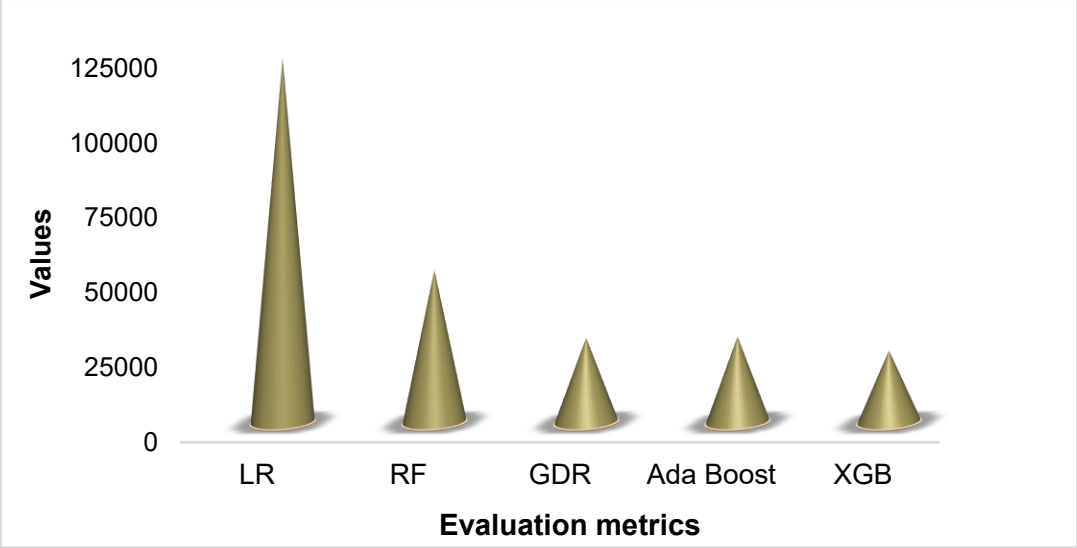


Figure 8. Performance analysis of RMSE

When assessing predictive ability, a reduced Root Mean Square Error (RMSE) indicates greater performance. With an RMSE score of 24022.09, The most effective model among those that was examined at was the XGB model. The AdaBoost Regressor model produced a value of 28645.07, the GDR model produced a value of 24243.06, and the Linear Regression model produced the highest value of 121470.9. The RMSE values varied from 121470.9 to 24022.09 (Figure-8). The statistics clearly shows the extent to which the XGB model performs in comparison to the other types that were looked at in terms of prediction accuracy.

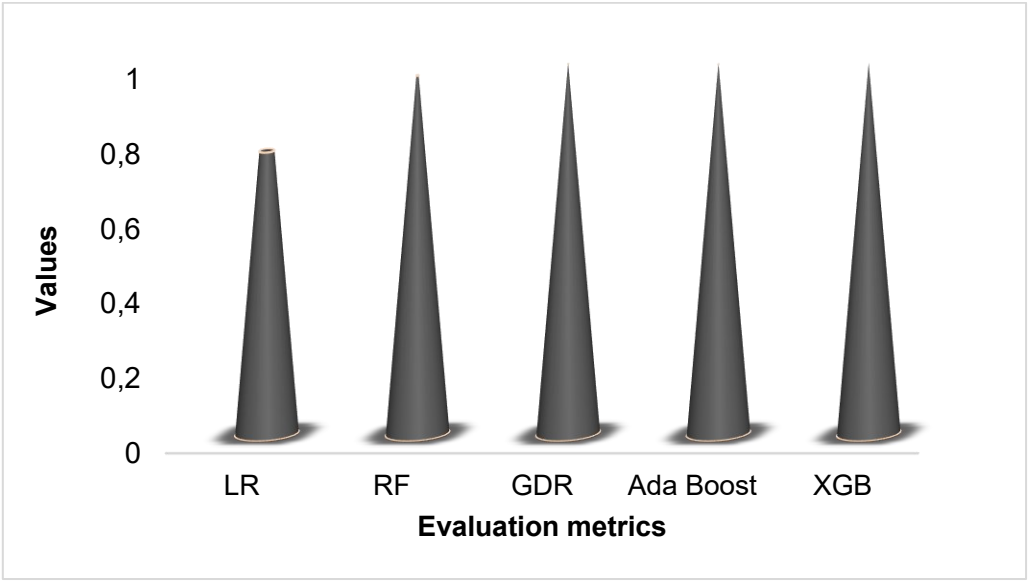


Figure 9. Performance analysis of R^2

An indicator of a model's predictive power is its R-squared value. R-squared values for the five models under investigation range from 0.75 to 0.99. The models based on statistics all display

lower values in comparison to the average values obtained from the data, suggesting that their forecasts are less dependable. The LSTM model obviously outperforms other models, with a coefficient R-squared of 0.99, a value that is extremely close to 1. This indicates that it has more predictive power.

The five machine learning models consistently outperformed all evaluation criteria out of all the models that were evaluated. With remarkable MSE, RMSE, MAE, R^2, and other measures, the XGB model proved to be the most effective at fitting daily data that was collected in close to real-time. In terms of performance, the AdaBoost Regressor and GDR models trail closely behind. Conversely, the LR model had a restricted fit for forecasting long daily time series data, as seen by its poor performance in MSE, RMSE, R^2, and test accuracy. Consequently, among all the models assessed for this specific task, the LR model is the least chosen one.

The XGB model outperformed others due to its ability to handle complex datasets, integrate decision trees effectively through gradient boosting, and minimize loss with regularization. Its evaluation metrics were superior, with the lowest MAE (19843.28), MSE (5770998), RMSE (24022.09), and the highest R^2 value (0.99). These attributes enable XGB to learn intricate patterns and interactions in the data, ensuring high prediction accuracy and robust results compared to other models. Other models like Random Forest, Gradient Boosting, and AdaBoost lack the XGB models optimized regularization features, which reduce overfitting and improve generalization. Linear Regression, being a simpler model, could not effectively capture the non-linear patterns in the data. While Gradient Boosting and AdaBoost were competitive, their tuning capabilities and feature handling were less advanced, resulting in slightly inferior prediction metrics.

On the test dataset, we concurrently displayed the actual and anticipated values of carbon missions, as seen in Figure. 10. The orange curve shows the values that the model was expected, whereas the curve in the blue colour shows the values that were actually observed.

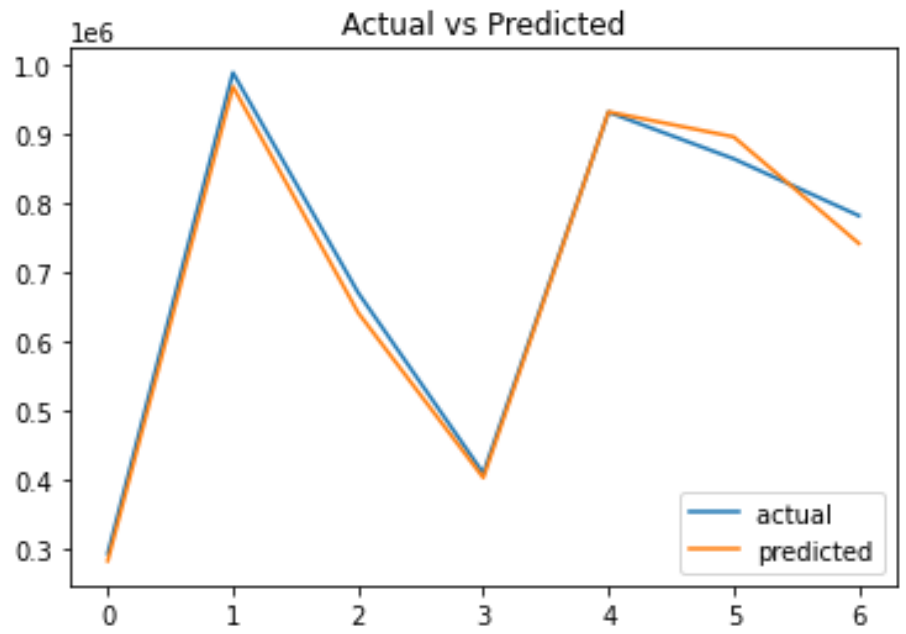


Figure 10. shows a comparison between the actual and predicted values of a particular dataset. The blue line represents the actual values, while the orange line indicates the predicted values. Both lines follow a similar trend, suggesting that the prediction model closely matches the actual data. Using machine learning models, we provided the carbon forecast findings for the test set, as Figure 10 illustrates. It is apparent that machine learning models' prediction values outperform previous, yielding precise and consistent results. These models' exceptional predictive accuracy amply illustrates their significant potential. In addition, an analysis of the prerequisite values reveals how the XGB model performed better than comparable models in the computation of daily carbon dioxide emissions, producing extremely satisfactory outcomes.

Table 2. Carbon emission prediction for next 10 years (2020 to 2030)

Year	Total Emission
2020	975465.78
2021	638020.9
2022	714591.75
2023	861775.4
2024	343444.28
2025	715862.25
2026	395854.72
2027	686273.44
2028	893278.56
2029	794890.75
2030	281816.28

We analyze data from 1990 to 2019 and predict the value for 2020 in order to test the accuracy of the anticipated values. In 2020, 97,975,465.78, was the actual value; the expected value was 93,458,754.15. There is very little variation between the actual and projected figures for 2020. The next step is to project carbon emissions for the ensuing ten years. The positive findings of this experiment demonstrate the efficacy of our predictive modelling.

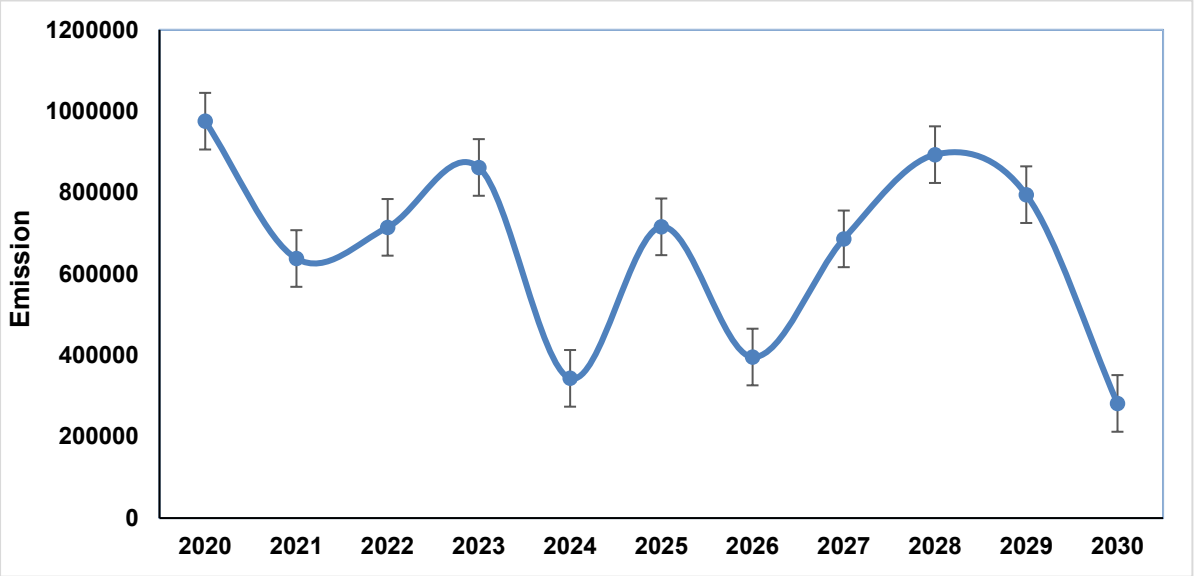


Figure 11. Carbon emission predicted for next 10 years (2021 to 2030)

This analysis provides important insights into improving existing systems for monitoring and mitigating carbon emissions, particularly in the agricultural and agri-related sectors. Using advanced MLMs, the study addresses the limitations of traditional statistical methods, such as their inability to handle nonlinear and complex data patterns effectively. The accurate predictions made by models like XGB can help policymakers and stakeholders identify key emission sources and trends with high precision. The study findings also facilitate the creation of early-warning systems to monitor potential emission spikes, helping governments and industries take proactive measures. The adoption of machine learning models can encourage innovation in carbon tracking systems, integrating diverse datasets for comprehensive and dynamic analysis. This alignment between advanced technology and policy frameworks ensures a more sustainable approach to addressing climate change challenges.

Conclusion

The government's attempts to prevent and regulate climate change in India depend heavily on the ability to anticipate carbon emissions in near real time with accuracy. This study attempts to predict carbon emissions in India by assessing the forecasting capacity of five

probabilistic machine learning models with contemporaneous information. The study, which focuses on agriculture and other industries connected to agriculture in India, makes use of a dataset that spans from 1990 to 2022. The study examines and discusses the five assessment criteria, namely MSE, RMSE, MAE, and R^2, to assess the predictive abilities of the models and identify the most successful model for future predictions. The five machine learning models consistently outperformed all evaluation criteria out of all the models that were evaluated.

The XGB model was the best at fitting daily data that was gathered in almost real-time, with impressive MSE, RMSE, MAE, R^2, and high-test accuracy values (Zhu *et.al*). Precise estimates of carbon emissions will additionally enable prompt policy and managerial adaptations during the current energy transition. Early identification of changes in carbon emissions allows policy adjustments to be made more quickly, perhaps saving 1-2 years, when compared to calculating yearly emissions. Government-applicable legislation should be implemented by all industries in order to successfully reduce carbon emissions. This makes a better future for the environment and is in line with the goals of sustainable development. One of the analysis's drawbacks is that it only uses linear historical data to predict carbon emissions, ignoring other relevant variables like GDP, energy consumption, and possible future governmental actions. Future research should examine the major impact that exogenous variables can have on emission patterns.

The proposed technique can be improved in the future by integrating more advanced machine learning and deep learning approaches that can further enhance the model's prediction accuracy and adaptability. The potential improvement is the use of ensemble learning techniques, which combine multiple models to reduce overfitting and improve prediction robustness. For instance, combining XGB with Random Forest or Gradient Boosting could yield even better results by capturing a wider range of patterns in the data. Deep learning models like LSTM or Recurrent Neural Networks (RNNs) can be explored for their ability to handle sequential data and model long-term dependencies, which might improve accuracy for forecasting emissions over extended periods. Incorporating additional variables such as technological advancements in carbon capture, shifts in agricultural practices, or the impact of climate policies, the model could offer more detailed and accurate predictions.

Reference

1. IPCC 2014 Climate Change 2014 Synthesis Report: Headline statements from the Summary for Policymakers.
2. IPCC 2007 Synthesis Report. Contribution of Working Groups I, II and III to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change
3. Zhong, W.; Haigh, J.D. The greenhouse effect and carbon dioxide. Weather 2013, 68, 100–105.
4. Cook, J.; Oreskes, N.; Doran, P.T.; Anderegg, W.R.; Verheggen, B.; Maibach, E.W.; Rice, K. Consensus on consensus: A synthesis of consensus estimates on human-caused global warming. Environ. Res. Lett. 2016, 11, 048002.
5. Myers, K.F.; Doran, P.T.; Cook, J.; Kotcher, J.E.; Myers, T.A. Consensus revisited: Quantifying scientific agreement on climate change and climate expertise among Earth scientists 10 years later. Environ. Res. Lett. 2016, 16, 104030.
6. Safwan Mohammed, Karam Alsafadi, István Takács & Endre Harsányi, (2019) Contemporary changes of greenhouse gases emission from the agricultural sector in the EU-27," Geology, Ecology, and Landscapes. DOI:10.1080/24749508.2019.1694129.
7. Amarpuri, N. Yadav, G. Kumar, S. Agrawal, Prediction of CO2 emissions using deep learning hybrid approach: a case study in indian context, In: 2019 twelfth international conference on contemporary computing (IC3) IEEE, (2019) 1–6.
8. K. Dong, X. Dong, C. Dong, Determinants of the global and regional CO2 emissions: what causes what and where? Appl. Econ., 51 (2019) 5031–5044.
9. https://globalcarbonbudget.org/ [GCB,2023]
10. W.H. Zhou, B. Zeng, X.Z. Liu, (2021). Forecasting Chinese carbon emissions using a novel grey rolling prediction model, Chaos Solitons Fract, 147.
11. Masini RP, Medeiros MC, Mendes EF (2021) Machine learning advances for time series forecasting. J Econ Surv. https://doi.org/ 10.1111/joes.12429.
12. Leopord Uwamahoro, Dr. Papias Niyigena, (2019). "Deep Learning in Greenhouse Gases Emissions from Agriculture Activities in Rwanda using Long Short-Term Memory Recurrent Neural Network," International Research Journal of Engineering and Technology (IRJET), Volume: 06 Issue: 10, e-ISSN: 2395-0056, p-ISSN: 2395-0072.
13. B. Khoshnevisan, S. Rafiee, M. Omid, H. Mousazadeh (2013). Developing an Artificial Neural Networks Model for Predicting Output Energy and GHG Emission of Strawberry Production," International Journal of Applied Operational Research, Vol. 3, No. 4, pp. 43 54.

14. Kalra, S.; Lamba, R.; Sharma, M. Machine learning based analysis for relation between global temperature and concentrations of greenhouse gases. J. Inf. Optim. Sci. 2020, 41, 73–84.
15. Magazzino, C.; Mele, M. A new machine learning algorithm to explore the CO2 emissions-energy use-economic growth trilemma. Ann. Oper. Res. 2022, 1–19.
16. Ahmed,M.;Shuai, C. Analysis of energy consumption and greenhouse gas emissions trend in China, India, the USA, and Russia. Int. J. Environ. Sci. Technol. 2022, 1–16.
17. Ahmed, M.; Shuai, C.; Ahmed, M. Influencing factors of carbon emissions and their trends in China and India: A machine learning method. Environ. Sci. Pollut. Res. 2022, 29, 48424–48437.
18. Bakay, M.S.; Agbulut (2020). Electricity production-based forecasting of greenhouse gas emissions in Turkey with deep learning, support vector machine and artificial neural network algorithms. J. Clean. Prod. 285, 125324.
19. Li, S.; Siu, Y.W.; Zhao, G. Driving factors of CO2 emissions: Further study based on machine learning. Front. Environ. Sci. 2021, 9, 721517.
20. Food and Agriculture Organization of the United Nations (FAO). (2023). *FAOSTAT domain emissions totals: Methodological note* (Release October 2023). Retrieved from https://www.fao.org/faostat/en/#data
21. Climate Watch data: Climate Watch. 2024. GHG Emissions. Washington, DC: World Resources Institute. Available at: https://www.climatewatchdata.org/ghg-emissions.
22. Kalaivani, K., Kshirsagarr, P.R., Sirisha Devi, J., Bandela, S.R., Colak, I., Nageswara Rao, J. and Rajaram, A., 2023. Prediction of biomedical signals using deep learning techniques. Journal of Intelligent & Fuzzy Systems, 44(6), pp.9769-9782.
23. Pushpavalli, M., Dhanya, D., Kulkarni, M., Rajitha Jasmine, R., Umarani, B., RamprasadReddy, M., Garapati, D.P., Yadav, A.S. and Rajaram, A., 2024. Enhancing Electrical Power Demand Prediction Using LSTM-Based Deep Learning Models for Local Energy Communities. *Electric Power Components and Systems*, pp.1-18.
24. Fang, X., Liu, W., Ai, J., He, M., Wu, Y., Shi, Y., Shen, W., & Bao, C. (2020). Forecasting incidence of infectious diarrhea 362 using random forest in Jiangsu Province, China. BMC Infectious Diseases, 20(1), 1–8.
25. Raju, K. N., Sudha, V., Kshirsagar, P. R., Tirth, V., & Rajaram, A. (2024). Intelligent traffic prediction system using hybrid convolutional neural networks for smart cities. Multimedia Tools and Applications, 1-19.

26. Chandrika, V. S., Kumar, N. M. G., Kamesh, V. V., Shobanadevi, A., Maheswari, V., Sekar, K., ... & Rajaram, A. (2024). Advanced LSTM-Based Time Series Forecasting for Enhanced Energy Consumption Management in Electric Power Systems. *International Journal of Renewable Energy Research (IJRER)*, *14*(1), 127-139.
27. Touzani, S., Granderson, J., & Fernandes, S. (2018). Gradient boosting machine for modeling the energy consumption of commercial buildings. *Energy and Buildings*, *158*, 1533-1543.
28. Selvarajan, S., Manoharan, H., Iwendi, C., Alsowail, R. A., & Pandiaraj, S. (2023). A comparative recognition research on excretory organism in medical applications using artificial neural networks. *Frontiers in Bioengineering and Biotechnology*, *11*, 1211143.
29. Pradeep, J., Raja Ratna, S., Dhal, P. K., Daya Sagar, K. V., Ranjit, P. S., Rastogi, R., ... & Rajaram, A. (2024). DeepFore: A Deep Reinforcement Learning Approach for Power Forecasting in Renewable Energy Systems. *Electric Power Components and Systems*, 1-17.
30. Patil, S., Patil, A., & Phalle, V. M. (2018, December). Life prediction of bearing by using adaboost regressor. In *Proceedings of TRIBOINDIA-2018 An International Conference on Tribology*.
31. Rajaram, A., Padmavathi, K., Ch, S. K., Karthik, A., & Sivasankari, K. (2024). Enhancing Energy Forecasting in Combined Cycle Power Plants using a Hybrid ConvLSTM and FC Neural Network Model. *International Journal of Renewable Energy Research (IJRER)*, *14*(1), 111-126.
32. Singh, S., Subburaj, V., Sivakumar, K., Anil Kumar, R., Muthuramam, M. S., Rastogi, R., ... & Rajaram, A. (2024). Optimum Power Forecasting Technique for Hybrid Renewable Energy Systems Using Deep Learning. *Electric Power Components and Systems*, 1-18.
33. Ma, J., Yu, Z., Qu, Y., Xu, J., & Cao, Y. (2020). Application of the XGBoost machine learning method in PM2. 5 predictions: A case study of Shanghai. *Aerosol and Air Quality Research*, *20*(1), 128-138.
34. Saravanan, A., Farook, S., Kathir, I., Pushpa, S., Padmashini, R. K., Logeswaran, T., & Rajaram, A. (2024). Adaptive Solar Power Generation Forecasting using Enhanced Neural Network with Weather Modulation. *International Journal of Renewable Energy Research (IJRER)*, *14*(2), 275-292.

35. Shinde, S. K., Tirlangi, S., Devaraj, V., DVSSSV, P., Jithesh, K., Sathyamurthy, R., & Rajaram, A. (2024). Enhancing Wind Power Generation Forecasting with Advanced Deep Learning Technique using Wavelet-Enhanced Recurrent Neural Network and Gated Linear Units. *International Journal of Renewable Energy Research (IJRER)*, *14*(2), 324-338.

36. Zhu, X., Chu, J., Wang, K., Wu, S., Yan, W., & Chiam, K. (2021). Prediction of rockhead using a hybrid N-XGBoost machine learning framework. *Journal of Rock Mechanics and Geotechnical Engineering*, *13*(6), 1231-1245.